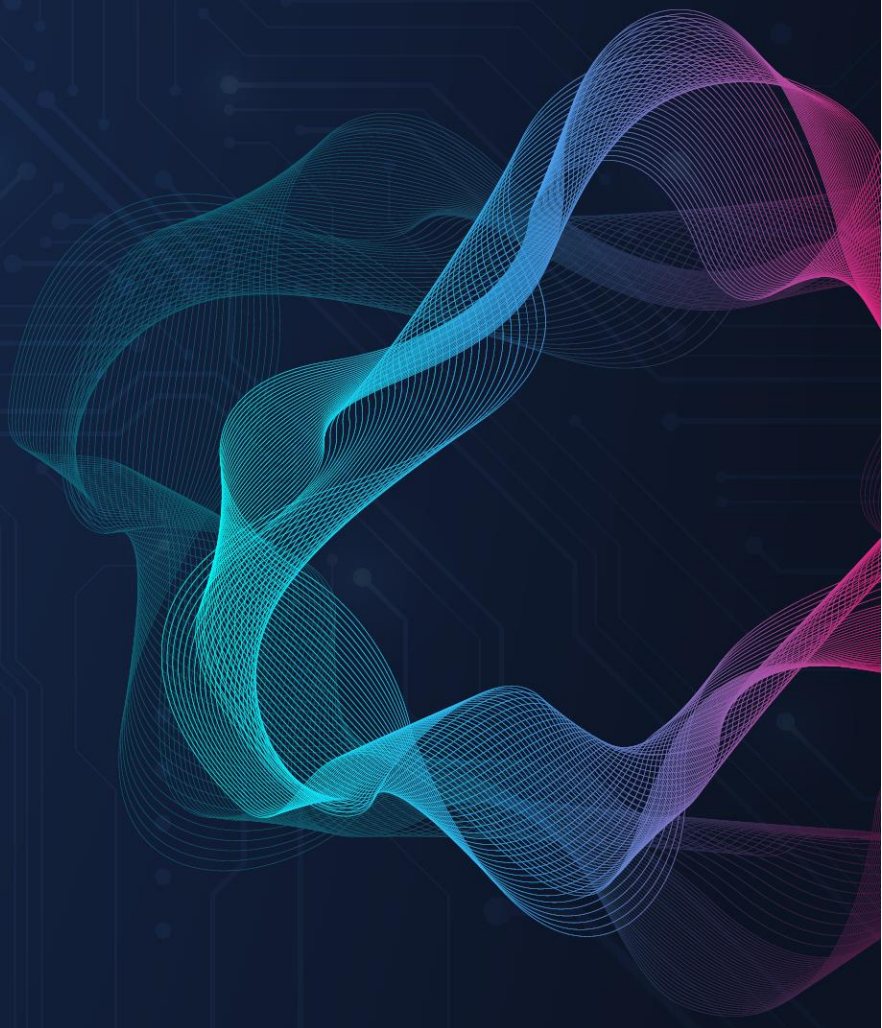
PD Dr. Reinhard Kreissl, Roger von Laufenberg, PhD.
Wiener Zentrum für Sozialwissenschaftliche Sicherheitsforschung (VICESSE)

Risiken und Gefahren der „Künstlichen“ „Intelligenz“

Bericht aus dem Forschungsprojekt
„Künstliche Intelligenz, Mensch und Gesellschaft“
Juli 2022



GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

KI.Me.Ge.

ZUSAMMENFASSUNG

Die Debatte über das Für und Wider von KI wird auf der einen Seite mit dem Argument der Optimierung menschlichen Handelns und Wirkens geführt, auf der anderen Seite dienen Horrorszenarien einer alles einnehmende Technologie als Gegenbeispiel. Dabei mangeln beide Argumentationsstränge häufig einer realistischen Einschätzung, Beobachtung und Analyse der Möglichkeiten und Grenzen von KI, inklusive der damit einhergehenden realen Risiken und Gefahren. Die Erwartungen an das „Können der KI“ sind häufig eher illusorischer Natur und unter- bzw. überschätzen dadurch auch die Risiken, denn bei genauerer Analyse wird ersichtlich, dass der Begriff KI in vielen Fällen irreführend ist – weder künstlich noch intelligent – in welcher die Fehleinschätzung über das Können der KI begründet ist. In diesem Beitrag gehen wir dabei auf diese verzweigte Risiko-Debatte ein, analysieren die Aspekte der Künstlichkeit und Intelligenz der KI, bevor wir auf die unterschiedlichen Stränge der KI-Risiko Debatten eingehen.

KEYWORDS

Algorithmus, Datafizierung, maschinelle Datenverarbeitung, maschinelles Lernen, menschliche Intelligenz, Mensch-KI-Interaktion, Regulierung, Technikoptimismus, Technikpessimismus

INHALT

1. WAS IST EIGENTLICH KI?

„KI ist weder *künstlich* noch *intelligent* / AI is neither *artificial* nor *intelligent*" (Crawford 2021: 8)

KI findet mittlerweile in allen Bereichen der Gesellschaft Anwendung, sei es im Bereich der Finanzen und Versicherungen, im Konsum, in der Medizin und in der Pflege, in der Bildung, Polizei und Militär setzen Hoffnungen auf KI im Bereich der inneren und äußeren Sicherheit. Somit ist KI auch in der aktuellen politischen und gesellschaftlichen Diskussion ein stark besetzter und häufig genutzter Begriff. Bei genauerer Betrachtung der Diskussionen rund um KI finden sich allerdings sowohl verschiedene Definitionen des Begriffs als auch gegensätzliche Prognosen hinsichtlich der Möglichkeiten und Probleme der Technologie (Christen et al. 2020). Unter den – teils inflationär verwendeten – Begriff KI, wird eine Vielzahl an Technologien, Systemen und Modellen subsumiert, die mal mehr oder mal weniger wirklich als Künstliche Intelligenz gelten (Kitchin 2016). Dies Spektrum reicht von einfachen, datenbasierten und (teil-)automatisierten Berechnungen bis hin zu komplexen Modellen des maschinellen Lernens mittels künstlicher neuronaler Netze.

Ein prominentes Beispiel für eine fahrlässige Verwendung des Etiketts KI ist der Algorithmus des Arbeitsmarktservices (AMS) in Österreich, welcher mittels algorithmischem Profiling Arbeitssuchende klassifiziert, um die Effizienz des Beratungsprozesses und die Wirksamkeit der aktiven Arbeitsmarktprogramme zu erhöhen (siehe Allhutter et al. 2020). Die Aussichten von Arbeitssuchenden auf dem Arbeitsmarkt werden dabei über eine gewisse Anzahl an Variablen statistisch ausgewertet (z.B. Geschlecht, Herkunft, Bildungsabschluss, Arbeitsbereich etc.), woraufhin die Kund:innen in drei Kategorien eingeteilt werden können: (1) mit hohen Chancen, innerhalb eines halben Jahres einen Arbeitsplatz zu finden, (2) mit mittelmäßigen Aussichten auf dem Arbeitsmarkt und (3) mit schlechten Aussichten auf eine Beschäftigung in den nächsten zwei Jahren. Während diese Art der Auswertung mittels recht simpler statistischer Modelle und Algorithmen – als eine Art Anleitung – automatisiert werden kann, wird sie in politischen und gesellschaftlichen Debatten dennoch häufig unter dem Begriff der KI subsumiert.

Dabei unterscheiden sich solche (semi-)automatisierten Anwendungen und Auswertungen zum Teil erheblich von dem, was im technischen Sinne unter KI verstanden wird. Denn als KI werden Systeme definiert, die nicht nur in der Lage sind, Daten nach fest vorgegebenen algorithmischen Regeln zu verarbeiten, sondern darüber hinaus neue Zusammenhänge und Strukturen in den verarbeiteten Daten identifizieren können. Aufgrund dieser Fähigkeit Regelmäßigkeiten, Muster oder Zusammenhänge in maschinenlesbaren Datensätzen zu entdecken. werden diese Systeme als ‚intelligent' bezeichnet (Marwala 2014). Dies wird als Maschinelles Lernen (ML) bezeichnet. ML erfordert sehr große Mengen an Daten, wobei deren Qualität von entscheidender Bedeutung für das Ergebnis des maschinellen Lernprozesses ist. In der KI-Entwicklung wird dies als GIGO Problem bezeichnet: Garbage in – Garbage out.

Ein bekanntes Beispiel für ML sind sogenannte Bilderkennungssysteme, die z.B. mit einer großen Menge von unterschiedlichen Tierbildern gefüttert werden. Bilder auf denen Katzen abgebildet sind, werden für das System erkennbar als ‚Katze' gekennzeichnet alle Bilder von anderen Tieren als ‚nicht Katze'. Durch die statistische Auswertung der visuellen Muster in den Bildern ‚erlernt' das System, Katzen von nicht-Katzen zu unterscheiden. Komplexe KI-Modelle verwenden dazu künstliche neuronale Netzwerke (KNN), ein Programmierverfahren, das von der Funktionsweise des menschlichen Gehirns inspiriert ist. Lässt man viele untereinander verbundene KNN miteinander interagieren, um einen Output zu generieren, entstehen komplexe, selbst für die Entwickler:innen nicht mehr durchschaubare Strukturen der Verarbeitung von Daten, die es ermöglichen, hochgradig abstrahierte Strukturen und Zusammenhänge zu modellieren (Arai/Kapoor 2020). Solche KI-Modelle finden sich zum Beispiel bei selbstfahrenden Autos im Einsatz, da hier das System eine Menge an unterschiedlichen ‚Fähigkeiten' ausführen muss: die Umwelt wahrnehmen, sich selbst in dieser verorten, Vorhersagen über die Umwelt treffen und eigene Entscheidungen auf Basis dieser Inputs treffen.

2. RISIKEN UND GEFAHREN – EINE VERKÜRZTE DEBATTE

In den politischen und gesellschaftlichen Diskursen wird bei der Debatte über KI häufig nicht zwischen algorithmen- oder datenbasierten Auswertungen und hochkomplexen Anwendungen wie KNNs unterschieden. Dies führt in der Auseinandersetzung über das Für und Wider von KI häufig zu sehr unrealistischen Einschätzungen (Markus/Davis 2019). Auf der einen Seite wird diese Debatte mit dem Argument der Optimierung menschlichen Handelns und Wirkens geführt (siehe z.B. Moorstedt 2022; Smith 2021). Auf der anderen Seite dienen Horrorszenarien einer omnipotenten – übermenschlichen und unkontrollierbaren – Technologie als Gegenbeispiel (O'Connell 2017). Da diese Erwartungen an das ‚Können der KI' allerdings eher illusorischer Natur sind – entsprungen aus einer Vermischung unterschiedlicher (verwandter) Methoden – unter- bzw. überschätzen diese dadurch auch die Risiken. Dies zeigt sich auch an den kulturindustriellen Dramatisierungen und ihren dystopischen und utopischen Zukunftsprojektionen, die auch in der wissenschaftlichen Diskussion immer wieder Niederschlag finden (siehe auch Jansen 2022).

Bei näherer Betrachtung der Debatten über KI lassen sich mehrere Stränge typisieren, die die Möglichkeiten und Risiken von KI zwar diskutieren, in ihrer Ursachenanalyse allerdings verkürzt sind:

(1) Überaus optimistische KI-Perspektiven befeuern dystopische wie utopische Szenarien gleichermaßen. Sie verlaufen dabei parallel mit entsprechenden kulturindustriellen Popularisierungen. Als KI-optimistisch lassen sich jene Ansätze und Positionen bezeichnen, die eine zunehmende Leistungsfähigkeit der KI bei der Lösung von komplexen Problemen prognostizieren und an deren End-Punkt nicht selten eine Art künstliche Superintelligenz steht. Stichworte wie Transhumanismus und Singularität befeuern Utopien wie Dystopien gleichermaßen (O'Connell 2017).

(2) Pessimistische KI-Perspektiven, die auf die nach wie vor begrenzte Leistungsfähigkeit von KI verweisen, sehen Risiken und Gefahren weniger in den wachsenden technologischen Kapazitäten als in dem unbedachten Einsatz von Systemen, die vorgeben ‚intelligente' Entscheidungen zu treffen, aber immer das Risiko von folgenreichen Fehlurteilen beinhalten (z.B. Lazer et al. 2014).

Eine Gemeinsamkeit dabei ist, dass die in der Diskussion über KI unterstellten Vorstellungen von ‚Künstlichkeit' und ‚Intelligenz' selten explizit und eher unklar sind. Das führt zu nicht immer sehr produktiven Vergleichen zwischen menschlicher Intelligenz und Prozessen maschineller Datenverarbeitung (Crawford 2021).

Sinnvoll und notwendig ist daher eine Debatte, die eine Vorstellung von Intelligenz als praktischer Fähigkeit zur Problemlösung entwickelt. Verschiedene Szenarien oder Diskurse über Risiken von KI können besser verglichen und bewertet werden, wenn man ein theoretisch fundiertes Konzept von den Kernbegriffen der KI – ‚Künstlich' und ‚Intelligenz' – als Maßstab und Orientierung zugrunde legt. Damit lassen sich zugleich praktische Handlungs- bzw. Anwendungskontexte differenzieren, in denen KI-Systeme eingebettet sind. Für die Auseinandersetzung mit Risiken und Gefahren der KI ergibt sich daraus eine vielschichtige Ausgangslage.

3. KI – EINE „KÜNSTLICHE“ „INTELLIGENZ“

Ein Verständnis über das Begriffspaar ‚Künstliche‘ ‚Intelligenz‘ ist dementsprechend notwendig, um auch die realen Risiken und Gefahren einer solchen besser einschätzen zu können. Angefangen mit dem Begriff ‚künstlich‘: in der Debatte um KI wird dieser gerne mit einer maschinellen oder digitalisierten Darstellung der Realität gleichgesetzt bzw. als ein maschinelles Pendant der menschlichen Intelligenz angesehen. Dabei werden allerdings grundlegende Aspekte der praktischen, datenbasierten KI übersehen. Um einem KI-System durch maschinelles Lernen etwas ‚beizubringen‘, benötigt dieses System eine Fülle an Daten über den Ausschnitt der Welt, über den etwas ‚erlernt‘ werden soll. Gängige Datensätze für ML-Systeme zur Gesichtserkennung bestehen aus mehreren Tausenden von Bildern und die Menge der verfügbaren Daten nimmt im Zeitalter von Big Data kontinuierlich zu. Trotz dieser Fülle an Daten, bleiben solche ML-Systeme an der Oberfläche. Sie reduzieren menschliche Gesichter auf vergleichbare visuelle Muster und blenden die breite Variabilität kultureller, individueller und gesellschaftlicher Unterschiede und Bedeutungen aus. Jeder KI-Datensatz wird immer nur einen gewissen, engen Ausschnitt der Welt abbilden oder verarbeiten können und ein KI-System das mit diesen Daten gelernt hat, kennt nur diesen Ausschnitt der Realität (van Dijck 2014; Symons/Alvarado 2016).

Hinzu kommt, dass diese Datengenerierung, -sammlung, und -verarbeitung nur unter erheblichem Einsatz menschlicher Dienstleistungen zustande kommt, somit also auch nur bedingt wirklich ‚künstlich‘ ist. In allen Schritten entlang des KI-Prozesses stehen menschliche Entscheidungen, Handlungen und Tätigkeiten, die die Funktionsweise des KI-Systems maßgeblich beeinflussen: Datensätze müssen manuell erstellt werden; durch sogenanntes Mikro-Tasking beschriften und kategorisieren Arbeiter:innen Daten; Entwickler:innen entscheiden über Variablen, Modelle, Gewichtungen und Feineinstellungen des KI-Systems, also darüber, welche Faktoren für die KI-Auswertung möglicherweise relevant sind und welche nicht, in welchem Verhältnis, etc. Da diese menschlichen und damit sozialen Handlungen und Entscheidungen allerdings innerhalb der KI im Verborgenen bleiben, entsteht ein solcher Trugschluss der ‚künstlichen‘ und ‚maschinellen‘ Intelligenz (Crawford 2021).

Folglich ist auch der Begriff der ‚Intelligenz‘ in der KI irreführend, da dieser einen direkten Zusammenhang zwischen der menschlichen Intelligenz und der maschinellen ‚Intelligenz‘ herstellt – während letztere eher als eine maschinelle Datenverarbeitung verstanden werden sollte. Dabei sind die Differenzen zwischen menschlicher Intelligenz und maschineller Datenverarbeitung erheblich. Am besten darstellen lässt sich dies durch die Dekonstruktion menschlicher Intelligenz als auch der maschinellen Datenverarbeitung. Bei der menschlichen Intelligenz stehen dabei die neurobiologischen Prozesse sowie die realweltliche, soziale Situierung des Individuums im Zentrum (Dreyfus 1992). Menschliche Informationsverarbeitung ist ein kontinuierlicher und verkörperter Vorgang, d.h. der menschliche Organismus verarbeitet fortlaufend eine Vielzahl von über seine verschiedenen Sinnesorgane eingehenden Umweltreizen, die sich kontinuierlich mit der Änderung der Position des menschlichen Körpers in seiner Umwelt ändern (Schütz 1971). Bei der Lösung alltäglicher Probleme nutzen menschliche Akteure in realweltlichen Situationen eine Vielzahl von unterschiedlichen Informationen, die in ihrer konkreten Situierung in einer strukturierten dreidimensionalen Umwelt verfügbar sind (Hutchins 1995). Wichtig ist dabei auch die enge und kontinuierliche Kopplung von interner Informationsverarbeitung mit körperlicher Bewegung (Verhalten, Handeln). Durch eine Veränderung der Blickrichtung werden neue Informationen zugänglich, durch das Berühren eines Objekts erhält der menschliche Organismus durch den Tastsinn zusätzliche Informationen über dessen Beschaffenheit.

Hinzu kommt, dass menschliche Intelligenz eng mit sozialer Interaktion und Kooperation verknüpft ist. Auch soziales Handeln lässt sich als eine Abfolge situierter kognitiver Prozesse analysieren. Ego und Alter bilden füreinander als Element ihrer jeweiligen Umwelt eine privilegierte Datenquelle. Die durch gegenseitige Beobachtung vermittelte Imitation von Bewegungen stellt eine Art Urform von bedeutungsvermitteltem

Verstehen dar. Die Fähigkeit, das Verhalten von Artgenoss:innen als absichtsvolle Gesten zu deuten, d.h über eine theory of mind zu verfügen, ist eine evolutionär wichtige Grundlage für die Entwicklung menschlicher Intelligenz durch kulturell situierte soziale Interaktion (Tomasello 2010). Dabei sind vor allem symbolische Beziehungen das Verbindungsglied von Intelligenz und Kultur. Sie repräsentieren praktische Handlungen und ermöglichen die Speicherung und Vermittlung von nützlichem Wissen über die Welt. Sie entstehen durch soziales Handeln und strukturieren die Interaktion zwischen zwei Akteuren, sind daher ein wichtiger Schlüssel zum Verständnis der Funktionsweise menschlicher Intelligenz. Sie ermöglichen die Strukturierung von Sinneswahrnehmungen im Prozess menschlicher Informationsverarbeitung zur Orientierung des Individuums in der Umwelt. Zugleich entlasten sie diese Prozesse, da zumeist ein Bruchteil der verfügbaren Umweltreize ausreicht, um eine zur Bearbeitung akuter Anforderungen brauchbare mentale Repräsentation zu erstellen. G.H. Mead (1934) hat diese soziale Dimension menschlicher Intelligenz grundlegend analysiert und dabei das Primat der Dyade gegenüber dem isolierten Individuum betont. Die soziologische Handlungstheorie in der Tradition des symbolischen Interaktionismus bietet eine Reihe ungenutzter Anschlussmöglichkeiten für einen erweiterten Begriff von menschlicher Intelligenz, der nicht das isolierte Individuum, sondern die Handlenden in der sozialen Situation als Ausgangspunkt nimmt.

Im Gegensatz dazu reduziert KI die hier skizzierten Prozesse auf ein Problem der Mustererkennung durch Vergleich von eingehenden mit gespeicherten Daten in einem monadisch gedachten System. Findet sich ein gespeichertes Datenmuster, das dem Input entspricht oder zeigt der Vergleich nacheinander verarbeiteter Inputdaten eine wiederkehrende Regelmäßigkeit, wirft das KI-System ein Ergebnis aus. Der Befund lautet dann entweder X (die Daten des Input) gleicht den gespeicherten Daten, die auf ein Exemplar der Kategorie Y verweisen oder in der großen Menge der als Input verarbeiteten Daten findet sich eine Reihe ähnlicher Verbindungen, die es erlauben, aus dieser Menge eine Ordnung von strukturierten Einheiten zu aggregieren. Dabei ergibt sich eine Reihe wesentlicher Unterschiede zur menschlichen Informationsverarbeitung. Die Bandbreite der in der menschlichen Informationsverarbeitung durch multimodale Sinneswahrnehmung anfallenden Daten ist wesentlich größer. Auch stehen für die Verarbeitung flexiblere und komplexere Verfahren zur Verfügung, die es erlauben, gespeichertes Wissen mit verfügbarem Input zu vergleichen. Zudem genügt in den meisten Fällen ein vorläufig erzielter Match zwischen Input und gespeichertem Wissen, um gezielt über weitere Schritte in einem strukturierten Handlungsvollzug zu entscheiden. Menschliche Intelligenz greift bei der Verarbeitung sensorischer Information auf gespeichertes Wissen über die materiellen Strukturen der Welt, symbolische Beziehungen und kulturelle Muster zur selektiven Steuerung der Aufmerksamkeit zurück, um daraus eine vorläufig passende (satisficing) Definition der aktuellen Situation zu entwickeln. Die spezifisch menschliche Fähigkeit, diese Definition ad-hoc, sozusagen im laufenden Betrieb zu ändern und dadurch die Aufmerksamkeit auf Neues zu lenken und Ziele oder Handlungspläne zu ändern, steht der maschinellen Datenverarbeitung auch in der komplexesten Ausführung nicht zur Verfügung.

4. RISIKEN UND GEFAHREN – EINE EINORDNUNG DER DEBATTEN

Die hier kurz skizzierte Konkretisierung der Begriffe ‚künstlich' und ‚Intelligenz' erleichtert einen sozialwissenschaftlich informierten Zugang in der Debatte über die Risiken und Gefahren der KI. Eine kritische Analyse konkreter Anwendungen von KI-basierten Systemen jenseits dystopischer und utopischer Versprechen und Befürchtungen erlaubt einen evidenzbasierten Blick auf Risiken und Gefahren. Jede konkrete Anwendung von KI produziert ein spezifisches Verhältnis zwischen menschlichem Nutzer und technischem KI-System, beide eingebettet in einen konkreten sozialen Kontext, der ihr Verhältnis prägt. Mit diesen drei Elementen

(Mensch, Maschine, Welt) lassen sich typisierte Szenarien entwickeln, die für eine differenziertere, kontextspezifische Analyse von Risken und Gefahren nützlich sein können.

(1) Das *forensische* Szenario: Die KI fungiert als Assistent, den ein menschlicher Nutzer gezielt in einem sozialen Kontext zur Erledigung kontextspezifischer Aufgaben einsetzt. So nutzt etwa ein Grenzpolizist, der bei Einreisenden Passkontrollen durchführt, ein KI-basiertes System zur Gesichtserkennung, das einen Input in der Form eines Passfotos durch Rückgriff auf eine große Datenbasis digitalisierter Fotografien mit den gespeicherten Bildern vergleicht. Wird ein passendes Bild gefunden, können gespeicherte Informationen über die der Fotografie zugeordnete Person abgerufen und das Ergebnis an den Nutzer zurückgemeldet werden.

(2) Das *strategische* Szenario: KI und menschlicher Akteur treten hier in einem durch explizite Regeln vollständig und eindeutig definierten geschlossenen Universum in Kontakt, in dem nur bestimmte Aktionen erlaubt sind. KI-basierte Anwendungen im Bereich der Molekularbiologie oder Schachprogramme zeigen die Unterlegenheit menschlicher Nutzer in einem solchen, regelbasierten Universum mit limitierten Freiheitsgraden und klar definierten Kriterien. KI-basierte Schachprogrammen verweisen auf die Bedeutung des Kontexts für die Interaktion von KI und Nutzer. Je enger und strukturierter die Welt ist, in der KI und Mensch agieren, desto besser schneidet die künstliche Intelligenz ab und desto weniger kann die Flexibilität menschlicher Intelligenz produktiv genutzt werden.

(3) Das *Überwachungsszenario:* Die KI sammelt unbemerkt Daten über einen menschlichen Nutzer in einem technisch-medial vermittelten sozialen Kontext. Jede Aktivität eines Nutzers in diesem Kontext techno-sozialen Alltagshandelns produziert personenbezogene Daten für KI-basierte Analysen zur Erstellung von Konsumentenprofilen, um Nutzerverhalten durch die Präsentation von gezielt ausgewählten Inhalten, Angeboten oder Informationen zu beeinflussen (Zuboff 2020).

(4) Das *Eliza* Szenario: KI und menschlicher Nutzer treten hier im Kontext einer face-to-face Situation in Kontakt, wobei der Nutzer auf der Basis der Annahme agiert, dass er es mit einem menschlichen Gegenüber zu tun hat. Avancierte KI-basierte Sprachprogramme sind im sprachlichem Konversationsmodus kaum mehr von menschlichen Gesprächspartnern zu unterscheiden. Auch hier spielt der Kontext der Nutzer und KI ohne Handlungsbezug nur über das Medium Sprache verbindet, eine wichtige Rolle sind.

Mit Hilfe solcher Szenarien lassen sich konkrete Anwendungen von KI differenziert betrachten, unterscheiden und vergleichen, wobei die hier beispielhaft beschriebenen Konstellationen nur zur Illustration dienen. Durch die Unterscheidung von Nutzer, KI-System und Kontext ergeben sich einerseits Anhaltspunkte für die Identifikation möglicherweise bisher unberücksichtigter Risiken und Gefahren. Andererseits macht diese Unterscheidung die Rolle des jeweiligen sozialen Kontexts deutlich, in dem KI-basierte Systeme zum Einsatz kommen. Viele medial großflächig verbreitete Meldungen über epochale Durchbrüche in der KI-Forschung, die dann auch die öffentliche Debatte über Fluch und Segen der KI befeuern, verdanken ihren Erfolg der gezielten Gestaltung des Kontexts ihrer Anwendung (Jansen 2022). Gleichzeitig können sich hoch riskante und gefährliche KI-Anwendungen, medial unbeachtet und für Nutzer nicht erkennbar im Kontext der Gesellschaft verbreiten.

Der hier vorgeschlagene differenzierte und empirisch informierte Zugang zu den diversen konkreten Anwendungsfällen von KI liefert Ansatzpunkte für eine sozialwissenschaftlich informierte Kritik eines (oft implizit bleibenden) technologischen Determinismus (Grosman/Reigeluth 2019), der die Einheit in der Differenz von utopisch-optimistischen und dystopisch-apokalyptischen Diskursen über KI garantiert. Dies lässt sich exemplarisch anhand einer Auswahl aktueller Debatten und Kontroversen demonstrieren.

(1) *Transhumanismus*. Diese Debatte basiert im Wesentlichen auf Prognosen über zukünftige Entwicklungen der KI-Forschung. Kritiker wie Anhänger des Transhumanismus skizzieren eine (mehr oder weniger nahe) Zukunft, in der sich domänenspezifische KI-Lösungen zu einer umfassenden bereichsunabhängigen Künstlichen Intelligenz transformieren. Diese zur materiellen Reproduktion fähige AGI (Artifical General Intelligence) wird – so die utopischen Optimisten – die Menschheit retten, KI und Mensch entwickeln sich in friedlicher Ko-Evolution und lösen globale Probleme. In der dystopischen Variante mutiert der Mensch zum Haustier der AGI oder verschwindet gänzlich vom Planeten, weil er entweder evolutionär überflüssig oder zum Opfer einer aus dem Ruder gelaufenen und nicht mehr kontrollierbaren AGI wird (O'Connell 2017). Beide Seiten nehmen schwere Hypotheken auf die Zukunft der technologischen Entwicklung auf und operieren mit einem reichlich unterkomplexen und naiven Verständnis von Gesellschaft. Der an sich sinnvolle Versuch, zukünftige technologische Entwicklungen zu skizzieren und sie auf mögliche Risiken und Gefahren abzuklopfen bleibt jedoch ohne Einbeziehung möglicher gesellschaftstheoretisch informierter Anwendungsszenarien unbefriedigend (siehe auch Bostrom/Yudkowsky 2014).

(2) *Biologische Hybridisierung*. Diese Debatte entzündet sich an den technologischen Möglichkeiten und Folgen von KI-basierten Systemen, die in den menschlichen Körper implantiert werden, um dessen Funktionsweise zu steuern. Sie liefert damit gleichsam Evidenz und Anschauungsmaterial für den Transhumanismus. Einschlägige Forschungsfelder wie Neuro-Enhancement ziehen prominente Investoren wie Elon Musk an, dessen Unternehmen Neuralink an der Entwicklung avancierter Schnittstellen von menschlichem Organismus und KI arbeitet. Die Befürworter treten mit geradezu biblischen Versprechen an: Blinde werden sehend, Lahme können wieder gehen und selbst Dumme werden gescheit – dank direktem Anschluss ihres Gehirns an Künstliche Intelligenz. Zudem gebe es bereits Implantate wie Herzschrittmacher, KI-basiertes Neuro-Enhancement stelle nur einen weiteren Schritt des medizinischen Fortschritts dar (Pisarchik et al. 2019; Neuralink 2022). Die Kritiker sehen in dieser Technologie einen Angriff auf die konstitutiven Bestandteile eines humanistischen Menschenbildes. Hinter den Versprechungen, bisher unbehandelbare körperliche Beeinträchtigungen zu lindern lauert die Gefahr des Verlusts menschlicher Autonomie. Der Mensch als Amalgam technischer und (neuro-)biologischer Prozesse verliert seine Freiheit und den freien Willen. Die Debatte über Risiken und Gefahren von KI zeigt hier eine gewisse Ähnlichkeit mit Kontroversen im Bereich der Medizin: mögliche Risiken werden neutralisiert bzw. in Kauf genommen, um Heilungserfolge jenseits erprobter Therapieansätze zu ermöglichen. Dabei geraten jedoch Risiken und Gefahren einer erwartbaren Kommerzialisierung und Vermarktung dieser Technologie jenseits des engen Feldes medizinischer Anwendungen nicht in den Blick.

(3) *Soziale Hybridisierung*. Im Gegensatz zu den kontroversen Debatten über Transhumanismus und biologische Hybridisierung, die sich auf mögliche zukünftige Entwicklungen von Technologie und Gesellschaft und KI-Anwendungen im experimentellen Stadium als vielversprechende emerging technology konzentrieren, handelt es sich bei sozialer Hybridisierung um ein sozio-kulturelles Massenphänomen digitalisierter Gegenwartsgesellschaften. Die Schnittstelle zwischen KI und Nutzer sind hier nicht Implantate, sondern alltäglich genutzte digitalisierte Endgeräte, mit denen Nutzer (auch untereinander) mit der virtuellen Welt des Internet in Kontakt treten. KI-Anwendungen befördern soziale Hybridisierung, da sie in einer Art kognitiven Arbeitsteilung mit und für den Nutzer bestimmte Schritte bei der Erledigung von unterschiedlichsten Aufgaben übernehmen können. Diese Form kognitiver Arbeitsteilung zwischen KI-basierten Systemen und menschlichen Akteuren unterscheidet sich von Interaktionsformaten wie etwa der Anfrage eines Nutzers an eine Suchmaschine. Suchmaschinen sind ein passives Medium, sie reagieren auf Anfrage und bieten Information zu einem vom Nutzer definierten Thema an, die möglicherweise Entscheidungen und Verhalten beeinflussen, aber sie übernehmen keine Aufgaben. Das unterscheidet sie von KI-basierten Formen der sozialen Hybridisierung durch GPS-Navigationssysteme, Fitness Tracker, online Lernplattformen oder niederschwellig zugängliche sprachgesteuerte Assistenten wie Alexa und Siri. Während die Verteidiger dieser Entwicklung sozialer Hybridisierung den Convenience Faktor betonen und auf die mit

der Verbreitung solcher Dienstleistungen einhergehende Entlastung im Alltag oder die Vorteile eines für- und vorsorglichen KI-basierten Gesundheitsmonitoring verweisen, sehen die Kritiker in der fortschreitenden sozialen Hybridisierung eine Entwicklung, die dem Verlust wichtiger evolutionärer und kultureller Fertigkeiten Vorschub leistet, den Horizont der Nutzer in vielfacher Hinsicht einschränkt oder manipuliert und zudem die letzten Barrieren zum (Daten-)Schutz der Privatsphäre beseitigt. Je leistungs-, lern- und anpassungsfähiger KI-basierte Systeme werden, desto invasiver wird die kognitive Arbeitsteilung zwischen KI und Nutzer (Doctorow 2022). Solche KI-basierten Monitoring- und Assistenzsysteme interagieren mit ihren Nutzern nicht nur in definierten und beschränkten Kontexten, sondern intervenieren und steuern realweltliches soziales Handeln (z.B. durch unterschiedliche Preisgestaltung basierend auf persönliche Profile; vgl. Mattioli 2012; Zuboff 2020).

(4) *Bias und Diskriminierung.* Unter dieser Überschrift lassen sich die Beiträge zu Risiken und Gefahren der KI zusammenfassen, die auf unbeabsichtigte Nebenfolgen beim Einsatz KI-basierter Entscheidungssysteme verweisen (Crawford 2021). Benachteiligungen entlang der traditionellen Dimensionen, race, class and gender konnten in einer Vielzahl von Studien über den Einsatz von KI-basierten Entscheidungssystemen nachgewiesen werden (u.a. O'Neil 2016; Edwards/Veale 2017; Aghostino et al. 2019). In den meisten Fällen resultieren solche Verzerrungen aus den für ML verwendeten Trainingsdaten. Ein KI-System reproduziert dabei zuverlässig Verzerrungen, Stereotypen und kulturelle Vorurteile, die es in den Daten, die es zur Entwicklung seiner ‚Intelligenz' verwendet, identifizieren kann. Die Diskussion über Risiken und Gefahren KI-induzierter Diskriminierung muss über eine Kritik an den Prozessen und der Leistungsfähigkeit der Datenverarbeitung hinausgehen und um eine kritische Auseinandersetzung mit der Struktur und Qualität der verwendeten Daten erweitert werden (Crawford 2021).

(5) *Mangelnde Flexibilität*. Diese Debatte verweist auf die Nebeneffekte die sich aus der begrenzten Leistungsfähigkeit von KI-basierten Systemen beim Einsatz in wenig strukturierten Alltagssituationen ergeben, in denen menschliche Akteure kooperativ und lösungsorientiert interagieren können. Die Entscheidungsrationalität von KI, die in einem solchen Setting an die Stelle eines menschlichen Gegenübers tritt, ist weder einsichtig noch im Kontakt mit einem menschlichen Akteur veränderbar. KI-Systeme reagieren nicht auf Gründe, sondern auf Daten. Aufgaben und Handlungsprobleme können nur entlang der dem KI-System verfügbarer Schritte bearbeitet werden. Situationsangemessene Flexibilität und Kreativität sind in der Interaktion mit algorithmischen Prozessen nicht anschlussfähig (Gillespie 2016).

(6) *Physikalisch-technische Vulnerabilität von datenbasierten Systemen*: Mit zunehmender Abhängigkeit von KI-unterstützten Prozessen steigt in einer Gesellschaft die Anfälligkeit für die Folgen von technischen Fehlfunktionen, gezielten Angriffen auf die Datenverarbeitung oder einer schlichten Unterbrechung der Stromversorgung (Blackout). Die Debatte über die Vulnerabilität moderner Gesellschaften reicht über mehrere Jahrzehnte zurück (z.B. Perrow 1999) und ist häufig von einem technikkritischen Unterton geprägt. Mit dem Vordringen und der Vernetzung von KI-unterstützten Prozessen in modernen Gesellschaften, entstehen neue Bedrohungen und Risiken, etwa durch den Einsatz von KI als Waffe im militärischen Bereich des sogenannten ‚Cyber Warfare'. In der Debatte über Vulnerabilität und den Umgang mit neuen technologischen Risiken treffen technikkritische und technikaffine Positionen aufeinander. Während die einen zur Vorsicht beim Einsatz von KI rät, fordern die anderen mehr und bessere technologie-basierte Lösungen zur Erhöhung gesellschaftlicher Resilienz und zum Schutz vor feindlichen Angriffen.

Betrachtet man die hier exemplarisch skizzierten unterschiedlichen Szenarien der Anwendung von KI-Systemen, so zeigt sich, dass die Möglichkeiten und Modalitäten der Interaktion von menschlichen Akteuren und KI erst durch die Berücksichtigung des jeweiligen sozialen Kontexts sinnvoll erschlossen und im Hinblick auf mögliche Risiken und Gefahren analysiert werden können. Durch eine solche Kontextualisierung lassen sich aus einer sozialwissenschaftlich informierten Perspektive die unterschiedlichen Positionen und Kontro-

versen, die sich entlang der diversen Diskursstränge entwickeln und die jeweils als Beleg vorgebrachten Anwendungsbeispiele einer kritischen Bewertung unterziehen. Dadurch können kontroverse Dramatisierungen in der Debatte über Risiken und Gefahren vermieden werden, deren diskursive Rahmung die öffentliche Aufmerksamkeit in eine bestimmte Richtung steuert, sodass Hinweise auf unterschätzte Gefahren oder Risikopotentiale unterhalb der Wahrnehmungsschwelle bleiben.

5. RISIKEN UND GEFAHREN – GESELLSCHAFTLICHE GESTALTUNGSMÖGLICHKEITEN

Diese Debatten sowohl um die realistische Einschätzung über das Können einer ‚künstlichen' ‚Intelligenz', die weder künstlich noch intelligent ist, als auch über die unterschiedlichen Diskursstränge der Risiken und Gefahren in der Interaktion zwischen KI und Mensch, sollen primär dazu dienen, Klarheit in dieser verworrenen Debatte zu schaffen. Es soll deutlich machen, dass eine Einordnung von KI jenseits extremer Utopie oder Dystopie an konkreten, beobachtbaren Szenarien ansetzen sollte. Erst dann ist eine realistische Einschätzung möglich, was KI ist, wie sie eingesetzt werden kann bzw. wird und was die Folgen daraus sein werden – sowohl in der Gegenwart als auch in der Zukunft. Die in diesem Beitrag skizzierten Beispiele über Risiken und Gefahren sind real existierend und haben reale Folgen für menschliche Akteure, die mit KI-Systemen in Interaktion treten, ob gewollt oder nicht. Diese müssen auf politischer und gesellschaftlicher Ebene angegangen werden.

Der europäische Entwurf einer KI-Regulierung, bei der auch die konkrete Anwendung und Interaktion zwischen menschlichen Akteuren und den KI-Systemen im Vordergrund stehen, ist ein erstes Beispiel für eine Einschätzung des Risikos von KI (low risk bis high risk), welche dadurch reguliert werden soll. Dabei sollen auch die Vorteile von KI-Systemen nicht ignoriert werden. Hierbei ist es ebenfalls von Vorteil, wenn Klarheit über die wirkliche Funktionalität des Systems herrscht. Über die rechtliche Komponente hinausgehend besteht also gesellschaftlicher Aufklärungsbedarf, um einen Weg heraus aus dem kommerziell getriebenen Hype zu finden. Wie bei jeder technologischen Entwicklung benötigt es individuelle und gesellschaftliche Kompetenz im Umgang mit solchen Technologien. Bestrebungen, digitale Grundkenntnisse an Schulen zu unterrichten, sind – spät aber doch – ein erster Schritt dort hin, wenn auch die Curricula den technologischen Entwicklungen auf dem Gebiet der Digitalisierung und KI meilenweit hinterherhinken. Ebenso benötigt es ein öffentliches und mediales Umdenken, dass nicht jeder Fortschritt in der KI-Entwicklung als Schritt zum Transhumanismus (v-)erklärt wird.

REFERENZEN

Agostinho, Daniela, Annie Ring, Kristin Veel, Catherine D'Ignazio, und Nanna Bonde Thylstrup. 2019. Uncertain Archives: Approaching the Unknowns, Errors, and Vulnerabilities of Big Data through Cultural Theories of the Archive. Surveillance & Society 17 (3/4): 422–441.

Allhutter, Doris, Florian Cech, Fabian Fischer, Gabriel Grill und Astrid Mager. 2020. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. Frontiers in Big Data 3: 326. doi: 10.3389/fdata.2020.00005.

Arai, Kohei und Kapoor, Supriya. 2020. Advances in computer vision. Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1. 1st ed. 2020. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing, 943).

Bostrom, Nick und Eliezer Yudkowsky. 2014. The ethics of artificial intelligence. In The Cambridge Handbook of Artificial Intelligence, ed. Keith Frankish and William M. Ramsey, 316–334: Cambridge University Press.

Christen, Markus, Clemens Mader und Johann Cas. 2020. Wenn Algorithmen für uns entscheiden. Chancen und Risiken der künstlichen Intelligenz. TA-Swiss.

Crawford, Kate. 2021. Atlas of AI. Yale University Press.

Doctorow, Cory. 2022. Revenge of the Chickenized Reverse-Centaurs. *OneZero*, April 17. https://onezero.medium.com/revenge-of-the-chickenized-reverse-centaurs-b2e8d5cda826. Zugriff 26 Juli 2022.

Dreyfus, Hubert. 1992. What Computers Still can't do. A Critique of Artificial Reason. MIT Press Cambridge/Mass.

Edwards, Lilian und Michael Veale. 2017. Slave to the Algorithm?: Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. 16 Duke Law & Technology Review 18. doi: 10.2139/ssrn.2972855.

Gillespie, Tarleton. 2016. Algorithm. In Digital keywords: A vocabulary of information society and culture, ed. Benjamin Peters, 18–30. Princeton studies in culture and technology. Princeton [New Jersey], Boston, Massachusetts: Princeton University Press; Credo Reference.

Grosman, Jérémy, und Tyler Reigeluth. 2019. Perspectives on algorithmic normativities: Engineers, objects, activities. Big Data & Society 6 (2): 205395171985874. doi: 10.1177/2053951719858742.

Hutchins, Edwin. 1995. Cognition in the Wild. MIT press Cambridge/Mass.

Jansen, Sue Curry. 2022. What Was Artificial Intelligence?: mediastudies.press.

Kitchin, Rob. 2016. Thinking critically about and researching algorithms. Information, Communication & Society 20 (1): 14–29. doi: 10.1080/1369118X.2016.1154087.

Lazer, David, Ryan Kennedy, Gary King und Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. Science 343: 1203–1205.

Marcus, Gary, und Ernest Davis. 2019. Rebooting AI. Building artificial intelligence we can trust. New York: Vintage Books, a division of Penguin Random House LLC.

Marwala, Tshilidzi. 2014. Artificial Intelligence Techniques for Rational Decision Making. Cham: Springer International Publishing.

Mattioli, Dana. 2012. On Orbitz, Mac Users Steered to Pricier Hotels. *The Wall Street Journal*, August 23. https://www.wsj.com/articles/SB10001424052702304458604577488822667325882, Zugriff 28 Juli 2022.

Mead, George Herbert. 1934. Mind, self and society (Vol. 111). Chicago: University of Chicago press.

Moorstedt, Michael. 2022. Sind Maschinen die besseren Menschen? *Süddeutsche Zeitung*, Juli 25. https://www.sueddeutsche.de/kultur/democratic-ai-gerechtigkeit-kuenstliche-intelligenz-netzkolumne-1.5626822. Zugriff 26 Juli 2022.

Neuralink. 2022. https://neuralink.com/. Zugriff 28 Juli 2022.

O'Connell, Mark. 2017. The Techno-Libertarians Praying for Dystopia. *Intelligencer*, April 30. https://nymag.com/intelligencer/2017/04/the-techno-libertarians-praying-for-dystopia.html. Zugriff 26 Juli 2022.

O'Neil, Cathy. 2016. Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy, 1st edn. New York: Crown.

Perrow, Charles. 1999. Normal accidents: Living with high-risk technologies. Princeton University Press.
Pisarchik, Alexander N., Vladimir A. Maksimenko, und Alexander E. Hramov. 2019. From Novel Technology to Novel Applications: Comment on "An Integrated Brain-Machine Interface Platform With Thousands of Channels" by Elon Musk and Neuralink. Journal of medical Internet research 21 (10): e16356. doi: 10.2196/16356.
Schütz, Alfred. 1971. Das Problem der sozialen Wirklichkeit (Bd.1 Gesammelte Aufsätze). Nijhoff Den Haag
Smith, Craig. 2021. A.I. Here, There, Everywhere. *The New York Times*, Februar 23. https://www.nytimes.com/2021/02/23/technology/ai-innovation-privacy-seniors-education.html. Zugriff 26 Juli 2022.
Symons, John und Ramón Alvarado. 2016. Can we trust Big Data?: Applying philosophy of science to software. Big Data & Society 3 (2): 1–17. doi: 10.1177/2053951716664747.
Tomasello, Michael. 2010. Origins of human communication. MIT press, Cambridge/Mass
van Dijck, Jose. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. Surveillance & Society 12 (2): 197–208.
Zuboff, Shoshana. 2020. *The age of surveillance capitalism. The fight for a human future at the new frontier of power.* New York, NY: PublicAffairs.