ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

PROJEKTBERICHT

WIEN, FEBRUAR/2022
ITA-2022-01
WWW.OEAW.AC.AT/ITA

KÜNSTLICHE INTELLIGENZ

VERSTEHBARKEIT UND TRANSPARENZ

ITA

KÜNSTLICHE INTELLIGENZ

VERSTEHBARKEIT UND TRANSPARENZ

ENDBERICHT

Institut für Technikfolgen-Abschätzung
der Österreichischen Akademie der Wissenschaften
Projektleitung: Walter Peissl
Autor*innen: Titus Udrea
Daniela Fuchs
Walter Peissl

Studie in Kooperation mit der Bundesarbeitskammer

Wien, Februar/2022

IMPRESSUM

Medieninhaber:
Österreichische Akademie der Wissenschaften
Juristische Person öffentlichen Rechts (BGBl 569/1921 idF BGBl I 31/2018)
Dr. Ignaz Seipel-Platz 2, A-1010 Wien

Herausgeber:
Institut für Technikfolgen-Abschätzung (ITA)
Apostelgasse 23, A-1030 Wien
www.oeaw.ac.at/ita

Die ITA-Projektberichte erscheinen unregelmäßig und dienen der Veröffentlichung der Forschungsergebnisse des Instituts für Technikfolgen-Abschätzung.
Die Berichte erscheinen in geringer Auflage im Druck und werden über das Internetportal „epub.oeaw" der Öffentlichkeit zur Verfügung gestellt:
epub.oeaw.ac.at/ita/ita-projektberichte

ITA-Projektbericht Nr.: ITA-2022-01 (Wien, Februar/2022)
ISSN: 1819-1320
ISSN-online: 1818-6556
epub.oeaw.ac.at/ita/ita-projektberichte/ITA-2022-01.pdf

INHALT

ABBILDUNGSVERZEICHNIS

ZUSAMMENFASSUNG

Der Begriff der „Künstliche Intelligenz“ (KI) ist in der öffentlichen Debatte allgegenwärtig. Flächendeckende Anwendungen von KI, basierend auf umfassenderen und genaueren Daten-Analysen, wecken Hoffnungen auf eine effizientere, genauere und objektivere Gestaltung von wirtschaftlichen, arbeitsbezogenen oder im weiteren Sinn gesellschaftlichen Prozessen. Gleichzeitig bleiben Konsequenzen und „Nebeneffekte“ eines breiten Einsatzes dieser Anwendungen weitestgehend ungeklärt. Um KI zu regulieren, legte die Europäische Kommission im April 2021 einen Entwurf des „Artificial Intelligence Act“[1] vor, um KI-Innovation unter größtmöglichem Schutz zu ermöglichen. Eine zentrale Forderung des AI Acts ist die Forderung nach Nachvollziehbarkeit bzw. Transparenz. Während der AI Act aber hauptsächlich auf Transparenz für Entwickler*innen, Produzent*innen und den kommerziellen Vertrieb abstellt, bleiben andere, potenziell (Mit-)Betroffene wie Konsument*innen, Verbraucher*innen oder Endnutzer*innen weitgehend unbeachtet.

Künstliche Intelligenz ist allgegenwärtig und von Erwartungen und Befürchtungen begleitet

In diesem Spannungsfeld zwischen technischer Machbarkeit und gesellschaftlicher Auswirkungen lassen sich einige grundlegende Punkte zusammenfassend festhalten, die auch für gegenwärtige und zukünftige Regulierungsversuche relevant sind:

Die Definition von KI bleibt umstritten: Einerseits vereint der Begriff der KI verschiedene gesellschaftsrelevante Ansprüche („schwache“ und „starke“ bzw. „generelle“ und „spezifische“ KI), deren Umsetzung von Akteuren sehr unterschiedlich eingeschätzt wird. Andererseits dient er als Sammelbecken für unterschiedliche (z. T. zusammenhängende) technische Ansätze, wie Algorithmen, maschinelles Lernen oder algorithmische Entscheidungssysteme. Durch ein Verständnis von KI als soziotechnischem System, wie es in diesem Bericht vorgeschlagen wird, werden neben technischen Spezifizierungen auch soziale Kontexte verstärkt in den Blick genommen. Das rückt nicht nur die technische Gestaltbarkeit bestimmter Anwendungen in den Mittelpunkt des Interesses, sondern betont die vielfältigen Möglichkeiten, die sich aus der Einbindung von KI in soziale Kontexte, das Zusammenwirken von sozialen und technischen Aspekten und die Umgestaltung von sozialen Praktiken durch KI ergeben. Besonders der Begriff der algorithmischen Entscheidungssysteme[2] ist hier von Bedeutung, weil die unmittelbaren Auswirkungen auf Konsument*innen, Endnutzer*innen und Betroffene hier einen hohen Stellenwert einnehmen.

KI als soziotechnisches System

Ähnlich vielgestaltig präsentiert sich der Begriff der „Transparenz“. Bei der Frage von Vertrauen in KI-Anwendungen kommt dem Begriff der Transparenz, von dem dieser Bericht seinen Ausgangspunkt nimmt, eine zentrale Rolle zu. Transparenz ist allerdings vielschichtig (bezogen auf den Algorithmus, den Prozess, den Kontext) und in manchen Bereichen auch ambivalent (z. B. Geschäfts-

Transparenz ist zentral, aber vielgestaltig

[1] Vorschlag für eine „Verordnung des europäischen Parlaments und des Rates Zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union“, *https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_1&format=PDF*.

[2] Algorithmic Decision Makingsystem (ADM).

geheimnisse). Sie kann einerseits technisch (als die detaillierte Offenlegung von Codes), andererseits prozedural (als zielgerichtete Kommunikation um mehr Verständnis bei der intendierten Zielgruppe zu erzeugen) verstanden werden. Durch diese inhärente Ambivalenz ist eine reine Forderung nach Transparenz nicht ausreichend, um KI verantwortungsvoll zu entwickeln, in Umlauf zu bringen und zu regulieren, auch, weil sie in einer rein technischen Interpretation auch zur Verschleierung von Verantwortung beitragen kann.

technische Ansätze zur Erhöhung von Transparenz

Technische Ansätze wie die erklärbare KI (XAI) versuchen, Transparenz durch technische externe Lesehilfen herzustellen und außerhalb des Systems zu kommunizieren. Darüber hinaus konzentriert sich das Feld auf Möglichkeiten, die Erklärungen der Modellentscheidungen in die algorithmische Entwicklung einzubeziehen. Interpretierbarkeit stellt darauf ab, technisch nachvollziehbare Ergebnisse zu generieren. Nachvollziehbarkeit wiederum stellt auf ein Verständnis von verschiedenen Akteuren ab (so beispielsweise auch Konsument*innen oder Betroffene).

Auf technischer Ebene stellt sich die Frage, ob die in diesem Bericht angesprochenen Ansätze der „erklärbaren KI“ (XAI) ausreichen um (technische) Transparenz – und in weiterer Folge soziale Nachvollziehbarkeit und Verantwortung – herzustellen. Aktuelle Forschungsansätze liefern interessante Ergebnisse, befinden sich aber noch in der Entwicklung. Da sie jedoch soziale Kontexte der Anwendungen und einen bewussten Umgang mit Wertentscheidungen nicht berücksichtigen, können Ansätze des XAIs auch nur eingeschränkt Wirkung in Richtung Verantwortung entfalten.

Transparenz allein reicht nicht, zusätzlich notwendig: Veränderbarkeit und Eingreifbarkeit

In Erweiterung des Transparenzbegriffs betont ein Fokus auf *Nachvollziehbarkeit (*statt *Transparenz)* die soziale Dimension des Verstehens von KI-Anwendungen und damit der informierten (Endnutzer*innen- und Anwender*innen-)Entscheidung bzw. des verantwortungsvollen Umgangs mit KI. Abgestufte Transparenzregeln (je nach Adressatengruppe) erscheinen sinnvoll, wobei der institutionelle Rahmen entsprechend gestaltet sein muss. Die Bereitstellung notwendiger Ressourcen und die Erwerbung solcher Kompetenzen sind durch verantwortliche (öffentliche) Stellen zu garantieren. Unabhängige Institutionen müssen eingerichtet und mit entsprechenden Kompetenzen besetzt werden, um Transparenzforderungen, Beschwerden oder Klagen zu überprüfen und gegebenenfalls adäquat handeln zu können – gesellschaftliche Akzeptanz wird neben der notwendigen Transparenz nur durch entsprechende Regulierung und Institutionen herzustellen sein.

In Bezug auf Governance sind die Ansprüche uneinheitlich: ob Transparenz (im Sinne der Offenlegung von Codes) oder Interpretierbarkeit ausreichen oder aufgrund sozialer Gegebenheiten (z. B. Geheimhaltungen) durchführbar sind, oder ob es wesentlicher ist, Entscheidungen für Betroffene nachvollziehbar zu gestalten, ist nach wie vor Gegenstand der wissenschaftlichen und politischen Debatte in der Gestaltung von verantwortungsvoller KI.

soziale Auswirkungen von KI als zentraler Faktor in der Regulierung

Während zur Gestaltung von KI eine Auseinandersetzung mit dem technischen Status-quo unerlässlich ist, können aufgrund rein technischer Spezifikationen keine Aussagen über soziale Auswirkungen von KI-Systemen getroffen werden. Ansätze des (regulatorischen) Umgangs mit KI können sich daher nicht ausschließlich über technische Aspekte definieren, sondern müssen Auswirkungen von KI auf den Menschen in den Mittelpunkt des Interesses rücken. Eine Engführung einer Definition von KI tendiert dazu, Gefahren zu ignorieren, die durch den Einsatz etablierter IT-gestützter Ansätze bereits bestehen (z. B. Diskri-

minierung durch statistische Verfahren). Daher plädiert der Bericht für die Beibehaltung einer breiten Definition des KI-Begriffs, wie im AI Act vom im April 2021 ursprünglich vorgesehen. Während ein risikobasierter Ansatz beibehalten werden sollte, kann ein zu starker Fokus auf Hochrisiko-Technologien allein ebenfalls dazu führen, dass Anwendungen mit weitreichenden Konsequenzen *per definitionem* unreguliert (oder unterreguliert) bleiben. Um den risikobasierten Ansatz weiter auszubuchstabieren, gibt es Ansätze zu abgestuften Transparenzforderungen wie etwa Krafft/Zweig (2019).

öffentliche Einbindung notwendig, um Sozialverträglichkeit zu gewährleisten

Die Verantwortung gegenüber Betroffenen von KI-Anwendungen durch regulierende Stellen, aber auch Entwickler*innen, Produzent*innen und Verwender*innen im AI Act muss hervorgehoben und gestärkt werden. Die aktive Einbindung von Betroffenen in Gestaltungsprozesse von KI oder algorithmischen Entscheidungssystemen durch inklusive partizipative und deliberative Ansätzen ist unerlässlich, um einen grundlegenden Schritt in Richtung Sozialverträglichkeit von KI-Anwendungen zu machen. Hierbei geht es um eine breite gesellschaftliche Diskussion, welche Werte in KI-Anwendungen eingeschrieben werden (sollen) und gesellschaftlich akzeptabel erscheinen. Gleichzeitig ermöglichen partizipative Ansätze, möglichst vielfältig potenzielle Auswirkungen von KI-Anwendungen aufzuzeigen.

Verantwortungsvakuum durch menschliche Aufsicht vermeiden

In KI-Anwendungen intendierte autonome Lernprozesse können nicht immer vollständig antizipiert werden. Entsprechend wird in der Literatur teilweise argumentiert, dass Entwickler*innen (auch bei vollständiger Transparenz in Bezug auf Codes) nicht vollständig für Konsequenzen des Einsatzes von KI verantwortlich gemacht werden können. Damit ergibt sich die Gefahr eines Verantwortungsvakuums. Aus Sicht von Betroffenen könnte es hier schwierig sein, individuelle Rechte einzufordern. Daher ist menschliche Aufsicht („human oversight“) als eine Grundbedingung zu sehen, um KI-Anwendungen auf den Markt zu bringen. Sollte dies derzeit nicht möglich sein, erscheint ein Moratorium für bestimmte Methoden und Anwendungen sinnvoll, bis eine solche Aufsicht – im Sinne von Erklärbarkeit, Interpretierbarkeit und Nachvollziehbarkeit – garantiert werden kann.

Eine realistische Einschätzung von KI (oder ADMs) basiert daher auf einer technischen und sozialen Komponente, die jeweils unterschiedlich zu bewerten sind. Schon gegenwärtig verändern sich Sozialsysteme unter dem Einsatz von IT- und KI-Systemen (Jobsuche, Zuteilung von Sozialleistungen, Credit Scoring). Gleichzeitig zieht der flächendeckende Einsatz von KI-Systemen weitere Veränderungen nach sich. Wichtiger als technisch-basierte Zugänge erscheinen regulatorische Ansätze, die gesellschaftliche Realitäten und Anwendungskontexte zumindest gleichwertig berücksichtigen, um eine aktive Gestaltung von KI-Systemen (und auch etablierter Methoden) zu ermöglichen.

Empfehlungen zur sozialverträglichen Gestaltung von KI

Um das Potenzial von KI ausschöpfen und eine verantwortungsvolle und sozialverträgliche Gestaltung voranzubringen erscheint es notwendig:

- die Berücksichtigung der Interessen von Konsument*innen und Betroffenen im KI-Diskurs und in der Entwicklung verstärkt einzubinden und einen möglichst ausdifferenzierten Kriterienkatalog zu erarbeiten,
- eine weit gefasste Definition von KI zu verwenden, die neben KI-Systemen im engeren Sinn auch etablierte IT-Systeme umfasst, deren Vorschläge und Entscheidungen Menschen in ihrer physischen, psychischen oder ökonomischen Existenz betreffen,

- Transparenz in ihren verschiedenen Dimensionen wie etwa dem Recht auf Information, als Nachvollziehbarkeit für Menschen und als institutionell verankerte Transparenz inklusive der notwenigen Verfahren (Rechtsbehelfe) umzusetzen,
- alle KI-Systeme die mit Menschen interagieren bzw. deren Entscheidungen sich auf das Leben und die Entfaltungsmöglichkeiten der Menschen auswirken, zu registrieren, einem grundlegenden Zertifizierungsprozess zu unterwerfen und im Verkehr kenntlich zu machen,
- einen flexiblen Regulierungsansatz zu verfolgen, der periodisch evaluiert wird,
- Forschung in jenen Bereichen zu fördern, die sich mit erklärbarer KI (XAI), Fairness, Gerechtigkeit, Rechenschaftspflicht und Verantwortlichkeit und mit gesellschaftlichen Auswirkungen der KI beschäftigen,
- KI-Systeme, die bestehende Grundrechte und Freiheiten sowie die Demokratie als solche schwer beschädigen oder aus ethischen Grundsätzen (Menschenwürde, Gleichheit, die Unantastbarkeit des Lebens usw.) zu verurteilen sind, zu verbieten und nicht zuletzt
- sollte die Letztverantwortung immer bei natürlichen oder juristischen Personen verbleiben. Solange Transparenz und Nachvollziehbarkeit nicht ausreichen, um menschliche Kontrolle zu gewährleisten, sollten Moratorien für bestimmte KI-Systeme überlegt werden.

1 EINLEITUNG

Der Begriff der „Künstliche Intelligenz“ (KI) ist in der öffentlichen Debatte zunehmend allgegenwärtig. Flächendeckende Anwendungen von KI, basierend auf umfassenderen und genaueren Datenanalysen, wecken Hoffnungen auf eine effizientere, genauere und objektivere Gestaltung von wirtschaftlichen, arbeitsbezogenen oder im weiteren Sinn gesellschaftlichen Prozessen. Gleichzeitig bleiben Konsequenzen, Folgen und unerwünschte Nebeneffekte eines breiten Einsatzes dieser Anwendungen weitestgehend ungeklärt. KI löst nicht nur positive Assoziationen aus, sondern zeigt auch vielfach Besorgnisse von Bürger*innen und ungeklärte Fragen auf: Als soziotechnisches System wirft KI Fragen des (sozialen) Risikos eines weitreichenden Einsatzes von KI-Anwendungen ebenso wie ethische Fragen auf, die damit in den Mittelpunkt der Debatte rund um die Gestaltbarkeit einer wünschenswerten KI-Zukunft rücken.

KI weckt Hoffnungen und Befürchtungen

Allerdings sind nicht nur die (gesellschaftlichen) Auswirkungen noch Gegenstand der Diskussion, sondern auch der gegenwärtige Stand der Technik von KI-Anwendungen selbst. So argumentieren manche Wissenschafter*innen dafür, dass künstliche Intelligenz weder gegenwärtig existiere noch nah sei. Stattdessen gebe es *„leistungsfähige Statistiksysteme, denen durch einen attraktiven Namen eine gewisse Magie zugesprochen werden soll. ‚Künstliche Intelligenz' ist nur ein Werbebegriff“* (Geuter 2018).

technischer Status Quo unklar

KI wird jedoch zunehmend Gegenstand der politischen Diskussion: Um KI zu regulieren legte die Europäische Kommission im April 2021 einen Entwurf des „Artificial Intelligence Act“ vor [AI-Act], um KI-Innovation unter größtmöglichem Schutz zu ermöglichen. Sie will damit eine Vorreiterrolle im Bereich der KI-Governance einnehmen. Eine zentrale Forderung des AI-Acts (genauso wie der breiteren IT- oder KI-Diskussion) ist die Forderung nach Transparenz. Während der AI-Act hauptsächlich auf Transparenz für Entwickler*innen, Produzent*innen und den kommerziellen Vertrieb abstellt, bleiben andere, potenziell (Mit-)Betroffene wie Konsument*innen, Verbraucher*innen oder Nutzer*innen und Bürger*innen marginalisiert.

KI als Gegenstand politischer Diskussion

Damit kommt dem Begriff der Transparenz eine tragende Rolle zu: Was bezeichnet der Begriff „Transparenz“? Welche Art von Transparenz gibt es und welche ist für welche Personen(gruppen) notwendig um Nachvollziehbarkeit von KI-basierten Entscheidungen zu gewährleisten? Welches Verständnis von Transparenz ist notwendig für (regulatorisches) Handeln?

Transparenz und Erklärbarkeit zentral für verantwortungsvolle KI-Entwicklung

Gleichzeitig spielen neben Transparenz noch andere Konzepte (wie beispielsweise Erklärbarkeit und Nachvollziehbarkeit) in der Diskussion um KI zunehmend eine Rolle. Daher widmet sich der vorliegende Bericht „Künstliche Intelligenz – Verstehbarkeit und Transparenz“ diesen grundlegenden Fragen und Konzepten: Angeregt durch die Publikation des Entwurfs des AI-Acts der Europäischen Kommission, werden zunächst definitorische Unklarheiten und begriffliche Abgrenzungen (technischer) Aspekte von KI zu bereinigen versucht. In einem zweiten Schritt widmet sich der Bericht den technischen Grenzen von Erklärbarkeit bzw. Transparenz, bevor er exemplarisch auf (mögliche) Wirkungen von KI-Systemen eingeht. Kapitel 5 stellt unterschiedliche, bereits entwickelte Ansätze zum (regulatorischen) Umgang mit KI vor, bevor in einem letzten Schritt Handlungsempfehlungen abgeleitet werden.

*Auswirkungen auf Konsument*innen und Nutzer*innen*

2 GRUNDLAGEN

Dieses Kapitel bietet einen kurzen Überblick über die definitorische Vielfalt und damit einhergehende Unklarheiten, mit denen sich der Forschungsbereich KI und entsprechende politische Entwicklungen konfrontiert sehen. Ausgangspunkt für diese Betrachtungen ist der Vorschlag für die Definition von KI im gegenwärtigen Entwurf des AI-Acts der Europäischen Kommission.

politische Vielfalt an Definitionen

Der Entwurf des AI Acts der EU-Kommission vom 21.4.2021 definiert KI als

„a) Konzepte des maschinellen Lernens, mit beaufsichtigtem, unbeaufsichtigtem und bestärkendem Lernen unter Verwendung einer breiten Palette von Methoden, einschließlich des tiefen Lernens (Deep Learning);
b) Logik- und wissensgestützte Konzepte, einschließlich Wissensrepräsentation, induktiver (logischer) Programmierung, Wissensgrundlagen, Inferenz- und Deduktionsmaschinen, (symbolischer) Schlussfolgerungs- und Expertensysteme;
c) Statistische Ansätze, Bayessche Schätz-, Such- und Optimierungsmethoden“
(Europäische Kommission 2021).

AI-Act (Entwurf) der Europäischen Kommission April 2021

Auch ohne die Ansätze im Einzelnen zu definieren, wird bereits auf den ersten Blick sichtbar, wie umfangreich die Definition von KI hier gedacht wird. Von hochkomplexen technischen Ansätzen bis hin zu etablierten Statistikverfahren soll der AI-Act vieles adressieren. Diese Breite der Definition wird auch kritisiert, da diese Vielfalt auf technischer Ebene weitreichende Regulierung in vielen gesellschaftlichen Bereichen nach sich ziehen würde und Innovationen erschweren könnte. Einige kritische Punkte betreffen grundlegende Definitionen und die Art des Geltungsbereichs bzw. die Breite und Spezifizität des Wirkungsbereichs dieser Definition. Aus Betroffenen-Perspektive zeigt sich jedoch, dass gerade die Breite der Definition notwendig ist, da das Regulierungsziel die zu erwartenden Wirkungen des Einsatzes von KI sind und diese durchaus mit verschiedenen zugrundeliegenden Technologien bzw. Methoden erreicht werden können.

Im Vergleich zu einigen früheren Definitionen, die von der EU-Kommission verwendet wurden, ist die oben genannte Version bereits spezifischer. Die Kommission hat in ihrer KI-Mitteilung 2018 ursprünglich die folgende, weit gefasste Definition für KI vorgeschlagen:

„Künstliche Intelligenz (KI) bezeichnet Systeme mit einem „intelligenten“ Verhalten, die ihre Umgebung analysieren und mit einem gewissen Grad an Autonomie handeln, um bestimmte Ziele zu erreichen. KI-basierte Systeme können rein softwaregestützt in einer virtuellen Umgebung arbeiten (z. B. Sprachassistenten, Bildanalysesoftware, Suchmaschinen, Sprach- und Gesichtserkennungssysteme), aber auch in Hardware-Systeme eingebettet sein (z. B. moderne Roboter, autonome Pkw, Drohnen oder Anwendungen des ‚Internet der Dinge‘)“
(Europäische Kommission 2018, S. 1).

Europäische Kommission 2018

Die unabhängige europäische Expert*innengruppe zu Künstlicher Intelligenz (High-Level Expert Group on Artificial Intelligence, HLEG AI) wurde später beauftragt, die KI-Definitionen genauer zu untersuchen. Die HLEG hat die Definition der EU-Kommission erweitert, um bestimmte Aspekte der KI als wissenschaftliche Disziplin und als Technologie zu klären. Das erklärte Ziel war es, Missverständnisse zu vermeiden, ein gemeinsames Wissen über KI zu ermöglichen, das auch von Nicht-KI-Experten genutzt werden kann, und nützliche Details zu liefern, die in der Diskussion über Ethikrichtlinien und Politikempfehlungen

verwendet werden können. Die vorgeschlagene Definition ist weit gefasst und konzentriert sich auf die Fähigkeit des KI-Systems, Schlussfolgerungen zu ziehen und Entscheidungen zu treffen:

> *„Künstliche Intelligenz (KI) bezieht sich auf Systeme, die von Menschen entwickelt wurden und mit einem komplexen Ziel in der physischen oder digitalen Welt agieren, indem sie ihre Umgebung wahrnehmen, die gesammelten strukturierten oder unstrukturierten Daten interpretieren, das aus diesen Daten abgeleitete Wissen schlussfolgern und die beste(n) Aktion(en) zur Erreichung des gegebenen Ziels (nach vordefinierten Parametern) entscheiden. KI-Systeme können auch so konzipiert werden, dass sie lernen, ihr Verhalten anzupassen, indem sie analysieren, wie die Umgebung durch ihre vorherigen Aktionen beeinflusst wird“* (HLEG AI 2019, S. 7).

High-Level Expert Group on Artificial Intelligence, HLEG AI

Diese beiden Definitionen waren nur einige der Grundlagen der im vorgeschlagenen AI-Act verwendeten Definition. Die derzeitige Fassung stellt eine spezifischere Formulierung darstellt (Europäische Kommission 2021). Die Definitionen der Kommission stützen sich auf weitreichende akademische und politische Diskussionen zur Fassung und Abgrenzung von KI. Wie dieser kurze Abriss zeigt, liegt in den Definitionen viel Gestaltungspielraum, der wiederum die daraus folgenden politischen und wirtschaftlichen Wirkungen beeinflusst. Um bestimmte Kritikpunkte besser einordnen zu können, wird im Folgenden ein Überblick über wissenschaftliche Definitionen grundlegender Begriffe gegeben.

2.1 DEFINITIONEN UND BEGRIFFLICHE ABGRENZUNG

In der Diskussion um Künstliche Intelligenz (KI) bzw. Artificial Intelligence (AI) werden unterschiedliche Begrifflichkeiten, manchmal auch synonym, verwendet. Deshalb wird im Folgenden versucht, etwas Klarheit in die Debatte zu bringen und wesentliche Begriffe zu erklären bzw. voneinander abzugrenzen. Dies ist umso schwieriger, als es keine umfassende, abschließende Definition von Intelligenz gibt.

So versuchen beispielsweise Legg/Hutter (2007) in der Debatte um „Maschinenintelligenz“ eine einheitliche und umfassende Definition von Intelligenz als Ausgangsbasis für eine formalistisch-mathematische Interpretation derselben zu finden. Dabei argumentieren sie, dass menschliche Intelligenz eine Vielfalt an Konzepten umfasst, die häufig ähnlich schwierig wie der Gesamtbegriff zu definieren sind, während sie gleichzeitig eine fundamentale Kategorie darstellt, wie wir Menschen bewerten (ibid., S. 392). Im Gegensatz dazu können Maschinen sich in ihren physischen Formen, Sensoren, Kommunikationsmittel oder Informationsverarbeitungsmöglichkeiten und Umgebungen komplett von menschlichen unterscheiden, womit eine universell gültige Definition von Intelligenz erschwert wird (ibid.).

Maschinen-“Intelligenz“ vs. menschliche Intelligenz

Generell kann Intelligenz empirisch oder theoretisch definiert werden. Definiert man sie empirisch über Testungen, so lässt sich feststellen, dass Intelligenztests am Menschen grundsätzlich verlässlich bestimmte kognitive Eigenschaften vorhersagen. Je nach Test werden unterschiedlichen Fähigkeiten abgefragt, z. B.

verbale und non-verbale Fähigkeiten (ibid., S. 395).[3] Unklar bleibt allerdings weiterhin, ob diese Tests einen Teil bzw. einen bestimmten Typ von Intelligenz messen, ob sie eine bestimmte Gruppe bzw. ein Set von geistigen Eigenschaften priorisieren oder inwiefern Vorstellungen von Geschlecht, Kultur oder sozialer Klasse in solche Tests Eingang finden (ibid., S. 392). So bemühte sich beispielsweise John Raven um einen kulturell neutralen Test, in dem jedes Problem aus einer kurzen Folge von grundlegenden Formen besteht (ibid., S. 395). Die Komplexität von Intelligenztests erhöht sich, wenn man versucht, Intelligenzbewertungen bei Tieren durchzuführen, da diese weniger eindeutig sind als beim Menschen (ibid., S. 396). Beispielsweise kann das Ziel eines Tests nicht direkt erklärt werden, ein Problem, das sich aufgrund eingeschränkten Sprachverständnisses ähnlich in der Testung von Maschinenintelligenz zeigt (ibid., S. 397).

Nähert man sich dem Problem der menschlichen Intelligenz von theoretischer Seite, so unterscheiden Legg und Hutter (2007) danach, ob Intelligenz als eine gesamte oder viele einzelne Eigenschaften definiert wird. Im ersteren Fall geht es um generelle geistige Fähigkeit, die durch den g-Faktor die Korrelation zwischen verschiedenen Arten von geistigen Fähigkeiten angibt. Im zweiten Fall werden unterschiedliche Fähigkeiten einzeln untersucht. Letztendlich definieren Legg und Hutter (2007) Intelligenz informell so: „*Intelligence measures an agent's ability to achieve goals in a wide range of environments*" (S. 405). Entsprechend besteht Intelligenz aus drei grundlegenden Komponenten: dem Akteur, dem Umfeld und den Zielen, wobei Akteur und Umfeld fähig sein müssen, miteinander zu interagieren (gegenseitig Signale zu senden und zu empfangen; aus Sicht des Akteurs: Aktivitäten zu setzen und Wahrnehmungen zu empfangen). Zusätzlich benötigt die Definition von Intelligenz ein bestimmtes Ziel, das erreicht werden soll. Legg und Hutter (2007) sehen diese Definition als allgemein genug, um auch Maschinenlernen miteinzuschließen.

Intelligenz als Fähigkeit von Akteur*innen, in verschiedenen Umgebungen Ziele zu erreichen

Im Gegensatz dazu hat die ISO im Jahre 2015 für technische Entwicklungen eine Definition für „artificial intelligence" (AI) festgelegt. In diesem Verständnis ist AI jene *„capability of a functional unit to perform functions that are generally associated with human intelligence such as reasoning and learning"* (ISO/IEC JTC 1 2015). Allerdings greifen „logisches Denken" und „Lernen" zu kurz, um menschliche Intelligenz umfassend zu beschreiben.

Intelligenz: logisches Denken und Lernen

Neben diesen kognitiven Definitionen von Intelligenz umfassen landläufige Definitionen von menschlicher Intelligenz auch Bewusstsein oder Fähigkeiten wie soziale Fähigkeiten, Empathie oder soziales Lernen, die von Maschinen (bislang) nicht bewältigt werden können.

Empathie und soziales Lernen als Teil von Intelligenz

[3] So fragt der WAIS-III-Test nach verbalen und nicht-verbalen Fähigkeiten, inklusive Wissen, grundlegende Arithmetik, Verständnis, Vokabular und Kurzzeitgedächtnis (verbal) sowie Bildergänzung, räumliche Wahrnehmung, Problemlösung, Symbolsuche und Objektaufbau (non-verbal).

2.1.1 KÜNSTLICHE INTELLIGENZ

Das Konzept der „künstlichen Intelligenz" lässt sich weit zurückverfolgen: Bereits die klassischen Philosophen versuchten, menschliches Denken als symbolisches System zu konzipieren. Während es einzelne Anwendungen schon früher gab, entwickelte sich die Forschungsdisziplin der „künstlichen Intelligenz" formal erst 1956, als der Begriff auf einer Konferenz des Darthmouth-Colleges in Hanover, New Hampshire erstmals verwendet wurde (Dick 2019).

wissenschaftliche Definitionenvielfalt

In der KI-Forschung wird versucht, Maschinen Verhalten beizubringen, das menschlichem Verhalten und Entscheidungsprozessen nachgebildet ist. Ziel ist dabei seit langem, diesen möglichst nahezukommen. Ob etwas als KI gilt oder nicht, hängt häufig von der Komplexität der ausgeführten Aufgaben und der Komplexität des Umfelds ab, in dem diese Aufgaben ausgeführt werden (Schneier 2021, S. 12). Entsprechend finden sich in der Literatur Definitionen von KI, die diese Eigenschaften abbilden und zum Teil weiter ausbuchstabieren.

Künstliche Intelligenz bezeichnet beispielsweise die Fähigkeit eines Computersystems, ein menschenähnlich intelligentes Verhalten an den Tag zu legen, das durch Kernkompetenzen wie Wahrnehmung, Verständnis, Aktivität und Lernen gekennzeichnet ist (Wirtz et al. 2019; Wirtz et al. 2020). Dabei spielen datengetriebene Ansätze der Problemlösung unter der Nutzung von Algorithmen und häufig in Beziehung zu Maschinellem Lernen, eine grundlegende Rolle (Herm et al. 2021).

starke und schwache – generelle und spezifische KI

Eine grundlegende Unterscheidung innerhalb der KI betrifft die Einordnung in so genannte schwache und starke KI. Bei der schwachen KI, oder auch „spezifischen KI" (Schneier 2021), lösen Algorithmen einzelne, spezifische Aufgaben, diese jedoch schnell und, abhängig vom Gegenstand, in sehr hoher Qualität. Beispiele dafür sind das Analysieren großer Datenmengen, die Mustererkennung und Vorhersagen zukünftiger Zustände/Entwicklungen aufgrund erkannter Muster. Die starke KI, oder auch „generelle KI" (ibid.), dagegen bezeichnet eine Entwicklung, in der Maschinen dem Menschen vergleichbare intellektuelle Fertigkeiten bekommen sollen, womit letztendlich auch ein Bewusstsein ähnlich dem menschlichen gemeint ist. Allerdings handelt es sich dabei vornehmlich um ein visionär philosophisches Konzept, dessen Realisierung auf absehbare Zeit vielfach angezweifelt wird (Apt/Priesack 2019). Vielmehr noch stellt sich die Frage, ob es aus gesellschaftlicher Sicht überhaupt erstrebenswert wäre, einen derartigen Zustand zu erreichen, oder ob es nicht auch gute Gründe gibt, Entwicklungen in diese Richtungen einer strengen Regulierung bis hin zum Verbot zu unterziehen (vgl. Zweig 2019, S. 267ff). In diesen Bereich fallen auch Visionen einer „Superintelligenz", die sich als der menschlichen überlegen erweisen und den Menschen beherrschen könnte. Trotz kontroverser Diskussionen ist eine solche Intelligenz (aktuell) aber eher dem Bereich der Science-Fiction zuzuordnen. Aus heutiger Sicht zählen alle verfügbaren KI-Systeme zur „schwachen KI", die auf bestimmte Aufgaben spezialisiert ist (Schneier 2021, S. 12). Sie weisen neben den genannten Vorteilen und Fähigkeiten auch einige grundlegende Defizite auf. Dazu zählen ein geringes Abstraktionsvermögen, insbesondere bei der Übertragung von Erfahrungen und gelerntem Wissen auf andere Kontexte, hohe Anforderungen an die Vorstrukturierung von Daten, Informationen und Umgebungen sowie ein mangelhaftes Verstehen und Schlussfolgern im empathischen Sinne. Das führt dazu, dass KI-Systemen Erfahrungen, implizites Wissen, Urteilsfähigkeit, Empathie und Verbindlichkeit sowie soziales Lernen und Emotionen fehlen, die den Menschen und die menschliche Intelligenz auszeichnen (Apt/Priesack 2019).

2.1.2 MASCHINELLES LERNEN

Ähnlich wie bei der Definition von KI, existiert auch für maschinelles Lernen noch keine konsistente Definition. Meist wird maschinelles Lernen als ein Teilbereich der KI verstanden, der datengetriebene automatisierte „Lernprozesse" von Algorithmen in den Mittelpunkt stellt. Algorithmen analysieren Daten, lernen davon und wenden das Gelernte an, um informierte Entscheidungen zu treffen. Sie verwenden von Menschen extrahierte Funktionen aus Daten und verbessern sich mit der Erfahrung (Helm et al. 2020).

Im Kontext von algorithmischen Entscheidungssystemen (s. u.) leitet maschinelles Lernen Entscheidungsregeln über bisherige Nutzer*innen (eines Systems) aus Daten ab, während Expertensysteme mit von Menschen festgelegten Entscheidungsregeln agieren. Auf maschinellem Lernen basierende algorithmische Entscheidungssysteme bestehen aus zwei Algorithmen: einem, der Entscheidungsregeln aus der Vergangenheit ableitet (Lernverfahren), und einem (meist einfacheren) Entscheidungsalgorithmus. Entscheidungen und Annahmen im Designprozess sind vielfältig und erschweren Transparenz und Kontrolle (ebenso wie die Nachvollziehbarkeit von Entscheidungen) (Krafft/Zweig 2019).

Maschinelles Lernen gibt es in verschiedenen Formen, die ein sehr ganz unterschiedliches Maß an menschlicher Beteiligung aufweisen können. Die am weitesten verbreiteten Ansätze sind beaufsichtigtes Lernen *(supervised learning)*, unbeaufsichtigtes Lernen (*unsupervised learning*) und bestärkendes Lernen (*reinforced learning*) (HLEG AI 2019). Diese Methoden umfassen auch fortgeschrittene Unterkategorien wie *Deep Learning* und ermöglichen es einem KI-System zu lernen, wie es Probleme lösen kann, die nicht genau spezifiziert werden können oder deren Lösungsmethode nicht durch symbolische Argumentationsregeln beschrieben werden kann.

beaufsichtigtes, unbeaufsichtigtes und bestärkendes maschinelles Lernen

Beim beaufsichtigten maschinellen Lernen geben die Entwickler dem System keine Verhaltensregeln vor, sondern stellen ihm Beispiele für das Eingabe-Ausgabe-Verhalten zur Verfügung, in der Hoffnung, dass es in der Lage ist, aus den Beispielen zu verallgemeinern und sich auch in Situationen gut zu verhalten, die nicht in den Beispielen gezeigt werden. Im Gegensatz dazu verwenden unbeaufsichtigte maschinelle Lernmethoden Algorithmen, die in nicht kategorisierten und nicht benannten Daten nach bisher unbekannten Mustern und Beziehungen suchen um Daten zu erforschen und zu gruppieren: z. B. in der medizinischen Bildgebung zur Unterscheidung zwischen verschiedenen Gewebearten oder in Empfehlungssystemen, die bessere Kaufvorschläge oder Filmempfehlungen geben. Dieser Prozess funktioniert mit minimalen menschlichen Eingaben in Bezug auf die verwendeten Variablen. Einige Ansätze des maschinellen Lernens verwenden Algorithmen, die auf dem Konzept der neuronalen Netze beruhen. Das Konzept ist vom menschlichen Gehirn inspiriert, da es ein Netzwerk aus kleinen Verarbeitungseinheiten (analog zu unseren Neuronen) mit vielen gewichteten Verbindungen zwischen ihnen hat. Es gibt verschiedene Arten von neuronalen Netzen und Ansätzen des maschinellen Lernens, von denen das Deep Learning derzeit einer der erfolgreichsten ist. Deep Learning bedeutet, dass das neuronale Netz mehrere Layer (Ebenen) zwischen dem Input und dem Output hat, die es ermöglichen, die gesamte Input-Output-Beziehung in aufeinanderfolgenden Schritten zu lernen. Das macht den Ansatz insgesamt genauer und erfordert weniger menschliche Anleitung (HLEG AI 2019).

neuronale Netze und Deep Learning

Eine wichtige (erhoffte) Eigenschaft von maschinellem Lernen ist die Mustererkennung aus Daten. Entsprechend definieren Zuber et al. (2020, S. 155) maschinelles Lernen als *„eine Methode, um in Trainingsdaten Muster zu erkennen und neue unbekannte Datensets entsprechend dieser Muster zu klassifizieren. Diese Muster sind algorithmisch konstruierte, statistische Modelle, die Daten extrahieren und sortieren“* (ibid.). In Anlehnung an Mohri et al. (2018) argumentieren Zuber et al. (2020), dass alle individuellen Ansätze des maschinellen Lernens zu einer Sammlung von Informationsmethoden gehören, welche zukünftige Datenpunkte basierend auf vergangenen Datenpunkten prognostizieren, indem sie diese korrekt Klassen zuordnen (ibid., S. 156). Bisher vielversprechende Anwendungsfelder umfassen die Klassifizierung von Daten (z. B. Bild- oder Spracherkennung), Erkennung von Regressionen (z. B. Preisentwicklung von Produkten) und Sequenzen (z. B. wichtigste Webseite für eine Suchanfrage), sowie Dataclustering (z. B. Personen in spezifische Gruppen zu klassifizieren).

Mustererkennung als zentrale Fähigkeit

In Bezug auf ethische Dimension von Entscheidungsfindung verstehen Zuber et al. (ibid., S. 155) Prozesse des maschinellen Lernens als rein statistisch und sprechen ihnen daher jede Form von (künstlicher) Intelligenz ab: Da Ergebnisse maschinellen Lernens häufig nicht einmal von Entwickler*innen selbst verstanden werden, sind sie der Meinung, dass maschinelles Lernen nicht in der Lage ist, ethische Entscheidungen zu treffen (ibid).

KI bleibt Statistik ohne ausreichende Fähigkeit zu ethischen Entscheidungen

2.1.3 ALGORITHMISCHE ENTSCHEIDUNGSSYSTEME

Während Begriffe wie KI und maschinelles Lernen Phänomene beschreiben, die auf konzeptioneller und methodischer Ebene verankert sind, integriert der Begriff des „algorithmischen Entscheidungssystems“ auch pragmatische Dimensionen. So sieht Zweig (2018) technisch umgesetzte Algorithmen als *„informatische Werkzeuge, um mathematische Probleme automatisiert zu lösen. Sie berechnen zuverlässig eine Lösung für ein Problem, wenn sie die dafür nötigen Informationen bekommen, den sogenannten Input. Das mathematische Problem definiert, welche Eigenschaften der dazugehörige Output, also das Resultat der Berechnung, haben soll”* (ibid, S. 11). Ein solcher Lösungsweg eines Algorithmus wird jedoch erst durch die Implementierung von Software in Computern wirksam. Die daraus entstehenden algorithmischen Entscheidungssysteme (Algorithmic Decision-Making Systems, ADM-Systeme) dienen dann der Lösung eines spezifischen Problems. Auf Software-Ebene umfassen diese Systeme Ein- und Ausgabedaten, eine Operationalisierung des zu lösenden Problems sowie Modelle für die Anwendung der Algorithmen zur Entscheidungsfindung. *„A fully configured algorithm will incorporate the abstract mathematical structure that has been implemented into a system for analysis of tasks in a particular analytic domain“* (Mittelstadt et al. 2016, S. 2).

ein Algorithmus agiert nicht allein: Einbindung von Algorithmen in Systeme

Ein algorithmisches System umfasst einen oder mehrerer Algorithmen, die in Software implementiert wurden, um Daten zu erfassen, zu analysieren und Schlüsse zu ziehen und zur Lösung eines vorher definierten Problems beizutragen. Das System kann dabei selbstlernend sein oder vorprogrammierten Entscheidungsregeln folgen (Algo.rules. 2019, S. 3). Um ein algorithmisches System abschätzen und bewerten zu können, spielt auch die Einbettung der Software in den soziotechnischen Gesamtkontext eine wichtige Rolle. Der soziotechnische Gesamtkontext umfasst beispielsweise Deutung, Interpretation des Ergebnisses und die Ableitung einer Entscheidung durch Anwender*innen des Systems. (ibid.) Der Fokus liegt hierbei vorrangig auf Systemen die *„signifikanten Einfluss auf das Leben der Menschen oder die Gesellschaft haben"* (ibid.). Damit sind algorithmische Systeme direkt auf Entscheidungsfindung mit Auswirkungen auf den Menschen bezogen. Sie *„bekommen Informationen über Personen und deren Verhalten und benutzen eine klar definierte Handlungsanweisung (einen Algorithmus), um aus dieser Information eine einzige Zahl zu erzeugen. Diese Zahl ist die eigentliche Entscheidung"* (Krafft/Zweig 2019, S. 8).

algorithmische Entscheidungssysteme mit signifikantem Einfluss auf das Leben der Menschen oder die Gesellschaft

Wie die hier aufgeführten Definitionen von KI, maschinellem Lernen und algorithmischen Entscheidungssystemen zeigen, weisen Definitionen von KI (bzw. technische oder pragmatische Spielarten davon) eine gewisse Bandbreite auf. Der vorliegende Bericht fokussiert vorrangig auf entscheidende oder entscheidungsvorbereitende Algorithmen bzw. Systeme mit besonderem Augenmerk auf eine Konsument*innen- bzw. Nutzer*innen-Perspektive bzw. Auswirkungen auf den Menschen (vgl. Krüger/Lischka 2018; Zweig 2019). Da KI in unterschiedlichsten Lebensbereichen eingesetzt wird,[4] sind Menschen vielfältig mit KI konfrontiert bzw. von deren Entscheidungen betroffen. Haben diese Entscheidungen direkte und indirekte Auswirkungen auf das Leben der Bürger*innen oder Konsument*innen, muss es eine Möglichkeit geben, diese zu hinterfragen bzw. Klärung über mögliche Fehlentwicklungen zu bekommen. Damit ist ein fundamentaler Aspekt aus Sicht des Konsument*innenschutzes beim Einsatz von KI die Frage der Überprüf- bzw. Nachvollziehbarkeit und damit die Frage der Transparenz.

[4] Details siehe Kapitel 4.

3 TRANSPARENZ IN DER KI – GRENZEN DER ERKLÄRBARKEIT

In der Diskussion um die gesellschaftliche Rolle der KI ist die aktive und partizipative beziehungsweise inklusive Gestaltbarkeit von KI-Prozessen und der gesellschaftlichen Debatte ein zentraler Faktor.

gesellschaftliche Gestaltbarkeit von KI

> *„Es liegt in unserer Hand, gemeinsam dafür zu sorgen, dass algorithmische Systeme zum Wohle der Gesellschaft gestaltet werden. Die in den Menschenrechten abgebildeten individuellen und kollektiven Freiheiten und Rechte sollen durch algorithmische Systeme gestärkt, nicht eingeschränkt werden. Regulierungen zum Schutz dieser Normen müssen durchsetzbar bleiben“* (Algo.rules. 2019, S. 2).

Transparenz und Nachvollziehbarkeit zentral für ...

Entsprechend müssen Prinzipien berücksichtigt und Mechanismen entwickelt werden, um bestimmte Grundlagen zu erfüllen; häufig genannt in der Debatte werden Transparenz und Nachvollziehbarkeit. Doch welche Rollen spielen sie im Kontext einer auf Gemeinwohl ausgerichteten Gesellschaft?

Abbildung 1: Funktionen von Transparenz und Nachvollziehbarkeit algorithmischer Entscheidungssysteme (Quelle: Beining 2019)

Vertrauen und Akzeptanz

Beining (2019) sieht Transparenz als Grundlage für Nachvollziehbarkeit um Aufmerksamkeit für den Einsatz von algorithmischen Systemen zu stärken. Nachvollziehbarkeit bietet Schutz vor Informations-Asymmetrien und dient daher der informationellen Selbstbestimmung, und damit im weiteren Sinne der Wahrung von Autonomie und Menschenwürde. Transparenz und Nachvollziehbarkeit ermöglichen informierte Teilhabe am gesellschaftlichen Diskurs (etwa, um Fehler aufzudecken, algorithmische Entscheidungen anzufechten und zu korrigieren). Sie sollen die richtige Anwendung und den bewussten Umfang mit einem algorithmischen System und dessen Ergebnissen gewährleisten. Durch Kontrolle und Aufsicht kann in weiterer Folge Vertrauen und Akzeptanz in Technologien und algorithmische Systeme gestärkt werden (ibid.).

Im Zusammenhang mit Zielen des Konsument*innenschutzes sind Transparenzforderungen im Bereich der IT allgegenwärtig (vgl. z. B. Datenschutz-Grundverordnung der EU aus 2018). Während jedoch Funktionen von Transparenz im Sinne des Gemeinwohls hier aufgelistet sind, bleibt der Begriff selbst uneindeutig. Im Folgenden soll deshalb der Transparenzbegriff im Kontext des wissenschaftlichen KI-Diskurses näher beleuchtet werden.

3.1 TRANSPARENZBEGRIFF IN DER KI

Der Begriff Transparenz in der KI ist nicht eindeutig definiert. Je nach Zielgruppe stehen technische (siehe 3.1.2), politische (hinsichtlich Aufklärungspflichten) und juristische Unterscheidungen (konkrete Festschreibungen von Transparenzkriterien) im Vordergrund. Allen gemein ist, dass die umgangssprachliche Bedeutung von „Transparenz“ (im Sinne von „offensichtlich machen“) noch nicht zwingend ausreichend Effekt auf die Handhabe verspricht. Daher sind jeweils spezifischere Interpretationen vonnöten.

3.1.1 AMBIGUITÄT DES TRANSPARENZBEGRIFFS IN DER KI

Empirische Studien zeigen, dass der Begriff der „Transparenz“ in Bezug auf KI kein eindeutiges Prinzip beschreibt. Jobin et al. (2019) fanden in Soft-Law-Literatur zu ethischen Prinzipien und Guidelines für KI kein einziges universell vorkommendes ethisches Prinzip.[5] Transparenz kam (vor Gerechtigkeit und Fairness, Schadensvermeidung, Verantwortung und Privacy) am häufigsten vor (73 von 84 Dokumente), umfasste aber ein weites Spektrum von Konzepten, von Erklärbarkeit (im Sinne von explainability und explicability), über Verständlichkeit, Interpretierbarkeit, Kommunikation, bis hin zu Offenlegung (Jobin et al. 2019, S. 7). Darin zeigt sich die Unschärfe in der Definition von Transparenz und die daraus folgende Verwendung einer Vielzahl von Konzepten, die zum Teil nicht scharf voneinander getrennt verwendet werden.

[5] Die in der Literatur gefundenen Prinzipien waren: Transparenz, Gerechtigkeit und Fairness, Schadensvermeidung, Verantwortung, Privacy, Wohltätigkeit, Freiheit und Autonomie, Vertrauen, Würde, Nachhaltigkeit und Solidarität.

Transparenz ist vielschichtig, aber notwendig für den gesellschaftlichen Diskurs zu KI

Anwendungen von KI umfassen dabei die Datenverwendung, die Mensch-KI-Interaktion, automatisierte Entscheidungen und den Zweck der Datenverwendung. Hier soll Transparenz dazu beitragen, Systeme zu verbessern und mögliche von ihnen ausgehende Gefährdungen zu minimieren. In der untersuchten Literatur wird Transparenz aber auch hinsichtlich ihrer Vorteile aus juristischer Sicht, zur Herstellung von Vertrauen bzw. ihre Wichtigkeit für Dialog, Partizipation und demokratischen Prinzipien erörtert.

Unklar bleibt auch hier, wie die Offenlegung von KI-Systemen genau ausgestalten werden soll: Vorschläge umfassten die Verwendung von KI an sich, Quellcodes, Arten der Datenverwendung, Grenzen, Gesetze, Verantwortung für KI, sowie Investitionen in KI und mögliche Impacts. Auch Überprüfungen und Überprüfbarkeit wurden variabel diskutiert und vorrangig von Datenschutzbehörden und Non-Profit-Organisationen eingefordert. Daneben wurden technische Lösungen zur Transparenz sowie alternative Zugänge zu Aufsicht, Interaktion und Mediation mit Stakeholdern und der Öffentlichkeit und die Ermöglichung von „Whistleblowing" diskutiert (Jobin et al. 2019, S. 9).

Im Kontext von Governance ist neben Transparenz auch das Prinzip Verantwortung und Rechenschaftspflicht (responsibility and accountability) hervorzuheben: Trotz weitreichender Verweise bleibt „verantwortungsvolle" KI häufig wenig klar definiert (ibid., S. 10). In der Literatur wurde „Verantwortung/Verantwortlichkeit" durch die Verbindung von Transparenz und Fairness häufig auch indirekt propagiert, während gleichzeitig auf die Diskrepanz zwischen ethischen Prinzipien und durchsetzbaren Bedingungen hingewiesen wurde. (Jobin et al. 2019, S. 15). Weitere ethische Prinzipien sind in der Literatur häufig unterrepräsentiert.[6]

Transparenz allein reicht nicht, zusätzlich notwendig: Veränderbarkeit und Eingreifbarkeit

Transparenz allein reicht nicht, zusätzlich notwendig sind: Veränderbarkeit und Eingreifbarkeit In manchen Dokumenten wird Transparenz nicht als ethisches Prinzip, sondern als „proethische Bedingung" für KI charakterisiert. Das erklärt einerseits, warum sie allgegenwärtig im KI-Diskurs anzutreffen ist und gefordert wird, zeigt aber auch, dass mit Transparenz allein noch wenig gewonnen ist, sondern zusätzliche Aspekte (Veränderbarkeit, Eingreifbarkeit etc.) systemisch gestärkt werden sollten (Zweig 2018). Die empirisch nachgewiesene Ambiguität und Variabilität in der Interpretation von Transparenz zeigen, wie weitreichend Folgen für ihre Implementierung sein können.

3.1.2 TECHNISCHE TRANSPARENZ: INTERPETIERBARKEIT UND ERKLÄRBARKEIT DER KI

Erklärbare KI (Explainable AI oder XAI) ist ein neuer Forschungsbereich, der Lösungen für die technische Umsetzung von KI-Transparenz und das Recht auf Erklärung bieten soll. Darüber hinaus bezieht sich XAI auf die Methoden und Techniken, die KI-Technologie und ihre Ergebnisse für den Menschen verständ-

[6] Beispielsweise werden ethische Prinzipien wie Solidarität und Nachhaltigkeit im Mainstream-Diskurs zu KI stark marginalisiert. ibid. Ihre Implementierung bleibt ibid.Ibid. häufig unklar, vor allem vor dem Spannungsfeld zwischen supranationaler Harmonisierung (Standardisierung) und dem Erhalt kultureller Diversität und moralischem Pluralismus. Hier wird auf Stakeholder-Deliberation verwiesen, um diese genauer zu bestimmen (Jobin et al. 2019).

lich machen. Es gibt viele Konzepte und Begriffe im Zusammenhang mit den technischen und methodischen Lösungen aus dem Bereich der XAI, deren Definitionen und Verwendungen sich überschneiden.

Eine Recherche in der XAI-Literatur zeigt, dass die Kernbegriffe austauschbar verwendet werden: Transparenz, Verständlichkeit, Interpretierbarkeit und Erklärbarkeit. In den folgenden Abschnitten wird ausführlich erklärt, wie der Einsatz fortschrittlicher KI-Systeme die Transparenznormen in Frage stellt, und weiters werden die wichtigsten Begriffe erläutert, die in diesen Debatten im Bereich der KI eine zentrale Rolle spielen. Der folgende Abschnitt fasst die am häufigsten verwendete Terminologie zusammen und klärt die Unterschiede und Gemeinsamkeiten zwischen den Begriffen, die in der ethischen KI- und XAI-Community häufig verwendet werden (Arrieta et al. 2020).

Aktuelle Studien zeigen, wie das Konzept der Erklärbarkeit von KI in vielen Forschungsbereichen verwendet wurde, darunter Mathematik, Physik, Informatik, Ingenieurwissenschaften, Psychologie, Medizin und Sozialwissenschaften (Abdul et al. 2018; Vilone/Longo 2020). Während die KI-Forschungsgemeinschaft traditionell „Erklärbarkeit" als Synonym für „Interpretierbarkeit" verwendet, deuten neuere Studien zu den fortschrittlichsten Deep-Learning-Modellen auf signifikante Unterschiede zwischen den beiden Begriffen hin, was auf einen Missbrauch der Terminologie hinweist („*interchangeable misuse*") (Arrieta et al. 2020, S. 84). Dies macht die Schaffung einer gemeinsamen Grundlage in Bezug auf Kriterien, Formalisierung und Umsetzung immer noch zu einer Herausforderung. Mit der aktuellen Zunahme der XAI-Literatur (Abbildung 3) rückt diese Unterscheidung jedoch zunehmend in den Mittelpunkt der Debatten über vertrauenswürdige Deep-Learning-Modelle, insbesondere bei Hoch-Risiko-Anwendungen.

Erklärbarkeit vs. Interpretierbarkeit

Einerseits ist die Interpretierbarkeit ein passives Merkmal und bezieht sich auf das Ausmaß, in dem ein Modell für einen menschlichen Beobachter Sinn ergibt *„[…] interpretability refers to a passive characteristic of a model referring to the level at which a given model makes sense for a human observer"* (Arrieta et al. 2020, S. 84). Erklärbarkeit hingegen wird im weitesten Sinne als eine aktive Eigenschaft eines Modells verstanden, die seine internen Funktionen beschreibt: *„explainability can be viewed as an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions"* (ibid.). Zusammengefasst kann die Interpretierbarkeit im weitesten Sinne als Prozess des Verstehens dessen definiert werden, was ein Modell getan hat (oder getan haben könnte). Die Erklärbarkeit bezieht sich im weitesten Sinne auf das Verständnis der Ursachen für die Entscheidungen des Modells (Gilpin et al. 2018). Das bedeutet auch, dass die Erklärbarkeit nicht zentral für die Nutzung eines KI-Systems ist, da es möglich ist, ein Modell zu nutzen, ohne die Ursachen für seine Funktionen zu verstehen. Erklärbare Modelle können also standardmäßig interpretierbar sein, aber das Gegenteil ist nicht immer der Fall (ibid.).

Terminologie im Bereich der KI-Forschung derzeit noch im Wandel begriffen

Obwohl die wichtigsten Begriffe in der Literatur oft austauschbar verwendet werden, sind im Folgenden die gebräuchlichsten Begriffe innerhalb der ethischen KI- und XAI-Gemeinschaften aufgeführt. Es gibt noch viele offene Fragen, wie einige dieser Begriffe zu definieren sind, und viele sich überschneidende Definitionen, die manchmal zu Verwirrung führen können. Zusammenfassend ist festzustellen, dass die Terminologie im Bereich der KI-Forschung derzeit noch im Wandel begriffen ist, und (XAI-)Autor*innen weisen auf die Notwendigkeit hin, die am häufigsten verwendete Terminologie zu klären und zu unterscheiden (Arrieta et al. 2020). Ziel ist es, die relevanten Begriffe hervorzuheben und

zu definieren. Dabei ist zu beachten, dass es Überschneidungen zwischen den Begriffen gibt, da einige in den weiter gefassten Begriffen enthalten sind. Die folgenden Beispiele gehen von einem weit gefassten Begriff der „Transparenz" aus und setzen ihn in Beziehung zu den darin enthaltenen spezifischeren Begriffen (Abbildung 2).

TRANSPARENZ

Wie oben bereits gezeigt, gibt es je nach Forschungsbereich viele Definitionen von Transparenz. Im Bereich des Deep Learning gilt ein Modell als transparent, wenn es aus sich heraus verständlich ist. Transparenz bedeutet im weitesten Sinne „erklärt, wie das System funktioniert", und ein Modell mit dem niedrigsten Transparenzgrad wird deshalb als „Black-Box" bezeichnet (Lipton 2018; Tomsett et al. 2018).

Drei Unterbegriffe, die oft mit Transparenz in Verbindung gebracht werden und die prinzipiell von jedem Lernmodell erreicht werden sollten (Lipton, 2018; Vilone & Longo, 2020), sind:

Transparenz bedingt Simulierbarkeit, Zerlegbarkeit und Vertrauen in den Lernalgorithmus

- *Simulierbarkeit* – die Fähigkeit eines Modells, Nutzer*innen zu ermöglichen, seine Struktur und Funktionsweise vollständig zu verstehen;
- *Zerlegbarkeit* – der Grad, in dem ein Modell in seine einzelnen Komponenten (Input, Parameter und Output) zerlegt werden kann und deren intuitive Erklärbarkeit;
- *Algorithmische Transparenz* – der Grad des Vertrauens in einen Lernalgorithmus, dass er sich im Allgemeinen „vernünftig" verhält.

VERSTÄNDLICHKEIT (GLEICHBEDEUTEND MIT VERSTEHBARKEIT)

Funktionsweisen eines Modells verständlich machen

Verständlichkeit ist ein etwas spezifischerer Begriff und bezieht sich auf die Eigenschaft eines KI-Modells, einem Menschen seine Funktion – die Funktionsweise des Modells – verständlich zu machen, ohne dass seine interne Struktur oder die algorithmischen Mittel, mit denen das Modell intern Daten verarbeitet, erklärt werden müssen (Montavon et al. 2018). Dieses Konzept steht im Zusammenhang mit der Herausforderung, Menschen einen komplexen Berechnungsprozess zu vermitteln (Bellotti/Edwards 2001). Gemäß dem Diagramm in Abbildung 2 kann ein KI-System auf verschiedene Weise verständlich werden, z. B. durch Erklärungen (z. B. in natürlicher Sprache) und/oder durch Interpretationen.

NACHVOLLZIEHBARKEIT (COMPREHENSIBILITY)

Nachvollziehbarkeit oder Comprehensibility ist ein weiterer Begriff, der häufig anstelle von Verständlichkeit verwendet wird. Der Begriff ist definiert als die Fähigkeit eines Lernalgorithmus, das gelernte Wissen in einer für den Menschen verständlichen Weise darzustellen (Craven 1996; Gleicher 2016; Fernandez et al. 2019). In Standards für Nachvollziehbarkeit sollten die Ergebnisse des KI-Modells strukturell dem ähneln, was menschliche Expert*innen produzieren könnten, das heißt sie sollten in natürlicher Sprache interpretierbar sein (Michalski 1983).

INTERPRETIERBARKEIT

Interpretierbarkeit ist kein monolithisches Konzept. Es bestehen mehrere konzeptionelle Definitionen in der Literatur (vgl. Doshi-Velez/Kim 2017; Lipton 2018). Der Begriff kann jedoch allgemein als die Fähigkeit definiert werden, die Funktionsweise eines KI-Systems in einer für einen Menschen verständlichen Weise zu beschreiben (Gilpin et al. 2018). Ein System ist interpretierbar, wenn es in dem Maße verstanden werden kann, in dem ein Mensch vorhersagen kann, was bei einer Änderung der Eingabe oder der algorithmischen Parameter passieren wird. Eine Interpretation ist die Abbildung eines abstrakten Konzepts (z. B. einer vorhergesagten Klasse) auf einen Bereich, den der Mensch verstehen kann (Montavon et al. 2018). Beispiele für interpretierbare Domänen sind Bilder (Anordnungen von Pixeln) oder Texte (Abfolgen von Wörtern). Ein Mensch kann sie betrachten bzw. lesen. Beispiele für Domänen, die nicht interpretierbar sind, sind abstrakte Vektorräume (z. B. Worteinbettungen) oder Domänen, die aus nicht dokumentierten Eingabemerkmalen bestehen (z. B. Sequenzen mit unbekannten Wörtern oder Symbolen).

Interpretierbarkeit fokussiert auf das Verständnis der prinzipiellen Funktionsweise eines Systems

ERKLÄRBARKEIT (EXPLAINABILITY)

Der Begriff „Erklärbarkeit" wird oft gleichbedeutend mit „Interpretierbarkeit" verwendet, bezeichnet jedoch das Verständnis der Gründe für eine Modellentscheidung oder -vorhersage (Došilović et al. 2018; Montavon et al. 2018). Während sich Interpretierbarkeit im engeren Sinne auf das Verständnis der prinzipiellen Funktionsweise eines Systems, seiner Mechanik, bezieht, ohne notwendigerweise zu verstehen, warum, konzentriert sich die Erklärbarkeit darauf, wie das Modell eine Entscheidung getroffen hat. Eine Erklärung kann z. B. eine Heatmap sein, die aufzeigt, welche Pixel des Eingabebildes die Klassifizierungsentscheidung am stärksten unterstützen. Die Erklärbarkeit wird daher oft mit Post-hoc-Prozessen in Verbindung gebracht. Sie bezieht sich auf den Begriff der Erklärung als Schnittstelle zwischen Menschen und einem Entscheidungsträger (Guidotti et al. 2018). Erklärbare Modelle sind in der Lage: *„able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions"* (Gilpin et al. 2018, S. 80). Dies bedeutet zum Beispiel, dass die meisten formalen Initiativen (z. B. DSGVO) die Erklärbarkeit abdecken und Elemente der Interpretierbarkeit (z. B. Modelldesign, Kausalarchitekturen) nicht berücksichtigen (oder mindestens keinen Unterschied beinhalten). Das kann aus mindestens zwei Gründen problematisch sein. Erstens ist, wie wir auf der Grundlage der neueren XAI-Literatur argumentieren, die Unterscheidung und Qualifizierung zwischen den beiden Begriffen wichtig, um spezifische Lösungen für die verschiedenen Fragen zu finden, die mit jedem Begriff verbunden sind (Interpretierbarkeit und die Frage nach dem „Wie"; Erklärbarkeit und die Frage nach dem „Warum"). Zweitens können Gesetzesinitiativen, die sich allein auf die Erklärbarkeit konzentrieren, wichtige Dimensionen der Modellgestaltungspraktiken im Zusammenhang mit interpretierbarer KI übersehen (z. B. die Gestaltung von Modellen, deren Funktionsweise für viele Akteur*innen leicht verständlich ist).

Erklärbarkeit konzentriert sich darauf, wie eine Entscheidung zustande kam

Erklärbarkeit – Was?

Interpretierbarkeit – Wie?

Auch wenn die Terminologie derzeit noch im Fluss ist, ist es wichtig festzuhalten, dass je nach Teilbereich die KI-Transparenz im Allgemeinen den spezifischeren Begriff der Verständlichkeit umfasst, der wiederum technischere Methoden und Typen im Zusammenhang mit der Interpretierbarkeit und Erklärbarkeit

von Modellen einschließt (Abbildung 2). Auf der Grundlage der Terminologie und der Beziehungen zwischen diesen Konzepten findet sich eine klarere Definition von XAI in Arrieta et al. (2020), wo XAI folgendermaßen definiert wird: *„eine erklärbare Künstliche Intelligenz ist eine Intelligenz, die Details oder Gründe liefert, um ihre Funktionsweise klar oder leicht verständlich zu machen“* (S. 85).

Definitionen von XAI

Weitere Definitionen von XAI konzentrieren sich auf die zentrale Bedeutung der Verständlichkeit:

> *„XAI is a research field that aims to make AI systems results more understandable to humans“* (Adadi/Berrada 2018, S. 52139).
>
> *„Explainable AI can present the user with an easily understood chain of reasoning from the user's order, through the AI's knowledge and inference, to the resulting behavior”* (Van Lent et al. 2004, S. 900).

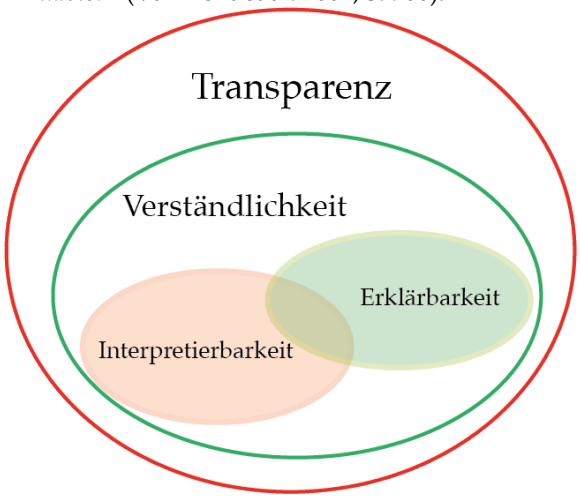

Abbildung 2: XAI-Terminologie (Quelle: nach Clinciu/Hastie 2019)

3.1.3 SYSTEMIMMANENTE GRENZEN DER ERKLÄRBARKEIT VON KI

In den letzten Jahrzehnten haben fortgeschrittene KI-Anwendungen ein hohes Potenzial zur Revolutionierung von Industrie, Handel, öffentlicher Dienste und der Gesellschaft insgesamt gezeigt. KI ist in der Lage, (selbst) zu lernen, zu entscheiden und sich anzupassen, wobei sie eine beispiellose Leistung bei der Lösung immer komplexerer Rechenaufgaben erzielt, was sie zu einem Schlüssel für die künftige Entwicklung der Gesellschaft macht (West 2018; Arrieta et al. 2020). Parallel zu diesem Trend werden die fortschrittlichsten, flexibelsten und genauesten Anwendungen wie Deep Learning von Entwickler*innen, Nutzer*innen und KI-Forscher*innen als „Black-Box“ betrachtet und behandelt (Castelvecchi 2016; Gershgorn 2017; Scarlett 2017; Yu/Alì 2019).

Ein Black-Box-Modell ist undurchsichtig, indem Entwickler*innen und Nutzer*innen nicht vollständig verstehen können, wie die Algorithmen funktionieren. Fortgeschrittene Deep-Learning-Systeme programmieren sich selbst um, treffen Entscheidungen und finden Muster in einem Maße, dass selbst die Entwickler*innen nicht immer in der Lage sind, die interne Logik hinter den KI-Entscheidungen zu verstehen (Yu/Alì 2019). Ein fortgeschrittenes maschinelles Lernsystem passt seine Parameter auf eine Art und Weise an, die von seinen Programmierer*innen nicht ausdrücklich vorgegeben wird (auch als *unbeaufsichtigtes System*

Black-Box

bekannt). In solchen Fällen bleibt unklar, wie das System genau zu seinen Vorhersagen oder Empfehlungen kommt (Deeks 2019). Dies macht es schwierig, versteckte Biases aufzuspüren und noch schwieriger (Allhutter/Berendt 2020), deren Ursachen zuverlässig zu erkennen: etwa fehlerhafte oder nicht repräsentative Trainingsdaten oder Annahmen, die in die Klassifizierung von Daten oder in die Modell einfließen. Ein Black-Box Neural Network empfängt also eine Eingabe und erzeugt eine Ausgabe, ohne Informationen darüber zu liefern, wie und warum es zu diesem Ergebnis kommt (Gershgorn 2017; Yu/Alì 2019). Es ist daher nur begrenzt möglich zu verstehen, wie einige fortgeschrittene KI-Systeme funktionieren und wie einige ihrer Entscheidungen getroffen werden. Kurz gesagt: Es gibt entscheidende Herausforderungen bei der Erklärung, wie und warum eine bestimmte Vorhersage oder ein bestimmtes Ergebnis zustande gekommen ist.

Fortgeschrittene KI-Anwendungen nehmen derzeit rasch zu und umfassen die Bereiche (digitale) Gesundheit, Soziales, Recht, Industrie, Finanzen und Verteidigung. Viele dieser Sektoren und Anwendungen sind risikoreich und erfordern ein hohes Maß an Modellverständlichkeit, Rechenschaftspflicht und somit Transparenz. Das Verständnis des Systems und die Erklärung der Entscheidungen und Vorhersagen von KI-Systemen sind von zentraler Bedeutung, um ihre Zuverlässigkeit rechtfertigen zu können. Dies wiederum erfordert eine bessere Interpretierbarkeit, d. h. das Verständnis der Funktionsmechanismen, die diesen Algorithmen zugrunde liegen. Angesichts der Forderungen nach ethischer KI (Goodman/Flaxman 2017) werden die Themen Transparenz, Erklärbarkeit und Interpretierbarkeit (Angelov et al. 2021) von vielen Stakeholdern zunehmend als wichtig erkannt (Preece et al. 2018).

Breite der Anwendungen macht Transparenz, Erklärbarkeit und Interpretierbarkeit zunehmend wichtig

Überlegungen bezüglich Innovationsdynamik und Datenschutz führen dazu, dass einige Autor*innen (z. B. Weller 2017; Lipton 2018; Vilone/Longo 2020) betonen, dass Transparenz als Anforderung mit anderen Überlegungen abgewogen werden sollte. Die Forderung, dass Daten und Modelle für die Endnutzer*innen vollständig sichtbar sind, verhindert die Schaffung von geistigem Eigentum; dies kann die Entwicklung neuer Technologien erheblich verlangsamen (Vilone/Longo 2020). Daten enthalten oft vertrauliche persönliche Informationen, die nicht ohne Datenschutzbedenken öffentlich gemacht werden können. Ein weiteres, technisches Problem besteht darin, dass die Bereitstellung von mehr Informationen über einen Algorithmus die Forscher*innen in bestimmten Fällen dazu veranlassen könnte, ein Modell für bestimmte Instanzen zu optimieren, was jedoch zu einer Verschlechterung der Gesamtleistung und des Grads der Verallgemeinerbarkeit führen würde (z. B. durch die Einführung eines persönlichen Bias). Dieses Gleichgewicht macht die Formalisierung solcher Fragen zu einer großen Herausforderung. Wichtig ist, dass Transparenz kein Selbstzweck ist, sondern ein Mittel zur Erreichung anderer Ziele (Weller 2017; Weller 2019).

Transparenz kein Selbstzweck, sondern ein Mittel zur Erreichung anderer Ziele

Als Reaktion auf diese Entwicklungen und die Herausforderungen beim Verständnis, wie und warum KI funktioniert, ist das Feld der erklärbaren KI (Explainable AI, XAI) mit dem ausdrücklichen Ziel entstanden, die algorithmische Intransparenz zu lösen. XAI ist zwar kein völlig neues Forschungsgebiet, da der Begriff seit den späten 1970er-Jahren verwendet wird (Moore/Swartout 1988; Clinciu/Hastie 2019). Das Feld hat aber in letzter Zeit zusammen mit fortgeschrittenen Anwendungen des maschinellen Lernens und Black-Box-Modellen erheblich an Bedeutung gewonnen (siehe Abbildung 3).

Obwohl die XAI-Literatur nicht einheitlich ist, gibt es bedeutende Initiativen zur Identifizierung und Klärung der wichtigsten Terminologie. Diese Konzepte

erklärbare KI im Aufwind

und Begriffe sind von zentraler Bedeutung, um die Grundlage für neue (technische) Normen und Rahmenbedingungen für ethische und verantwortungsvolle KI zu schaffen.

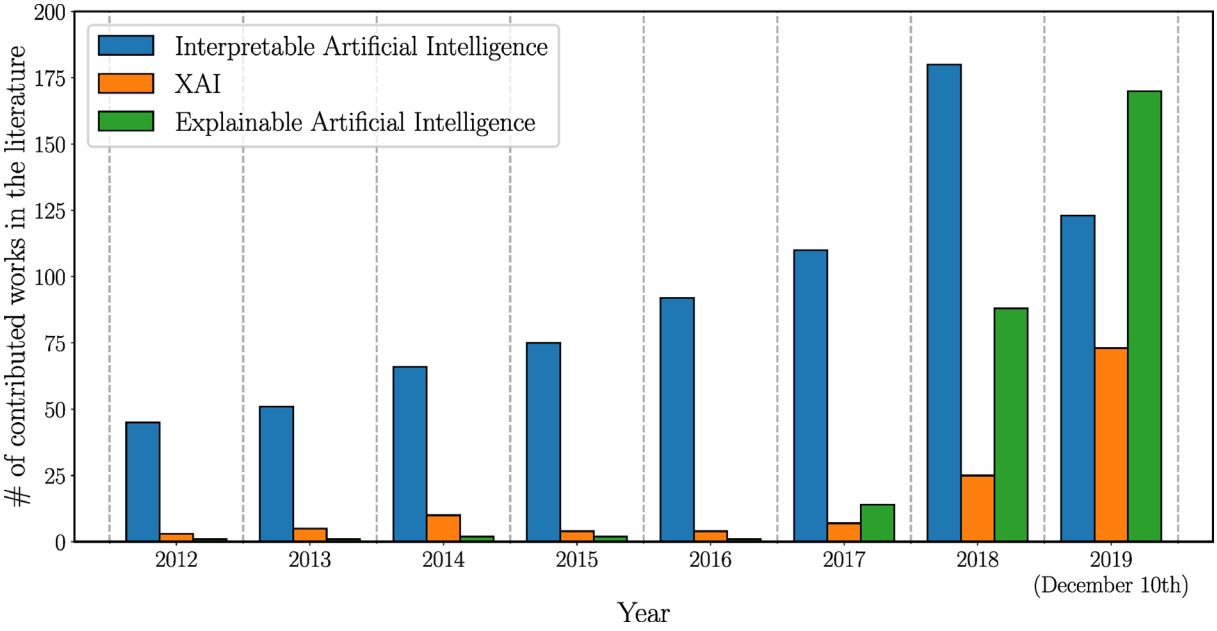

Abbildung 3: Entwicklung der Publikationen im Bereich „Explainable AI“
(Quelle: Arrieta et al. 2020. Daten abgerufen von Scopus: 10. Dezember 2019)

3.1.4 WELCHE BLACK-BOX? DEEP LEARNING UND DER TRADE-OFF ZWISCHEN TRANSPARENZ UND GENAUIGKEIT/LEISTUNGSFÄHIGKEIT (ACCURACY)

Zu den fortgeschrittenen Anwendungen der KI, die als so komplex gelten, dass sie mit dem Begriff „Black-Box“ in Verbindung gebracht werden, gehören komplexe Algorithmen für Machine Learning (ML) und Deep Learning (DL). Deep Learning als Teilbereich der KI (Abbildung 4) und die fortgeschrittenen Anwendungen wie Deep Neural Networks kombinieren effiziente (selbst-)lernende Algorithmen und einen riesigen Raum von Parametern und Variablen.

Künstliche neuronale Netze sind eine Form der KI, die dem menschlichen Gehirn nachempfunden ist und die derzeit (im Vergleich zu weniger komplexe Algorithmen) bei der Bewältigung komplexer realer Problemlösungen und der Mustererkennung sehr erfolgreich ist. Allerdings sind neuronale Netze ebenso undurchsichtig wie das Gehirn. Anstatt das Gelernte in ordentlichen Blöcken des digitalen Gedächtnisses zu organisieren, verbreiten sie die Informationen auf eine Art und Weise, die äußerst schwer zu entschlüsseln ist (Castelvecchi 2016). Solche Modelle haben eine Architektur, die auf einer riesigen Anzahl von Schichten und Parametern (Millionen oder sogar Milliarden) basiert, die Informationen aus den Eingabedaten enthalten und miteinander in Beziehung setzen. Abgesehen von der riesigen Anzahl von Parametern, Variablen und Interaktionen ist ihre

neuronale Netze sind ebenso undurchsichtig wie das Gehirn

Verbindung zur physischen Umgebung des Problems extrem schwer zu isolieren (Angelov et al. 2021). Dies macht die Erläuterung solcher KI-Systeme für Nutzer*innen und Verbraucher*innen höchst problematisch und nicht zuletzt zu einer Herausforderung.

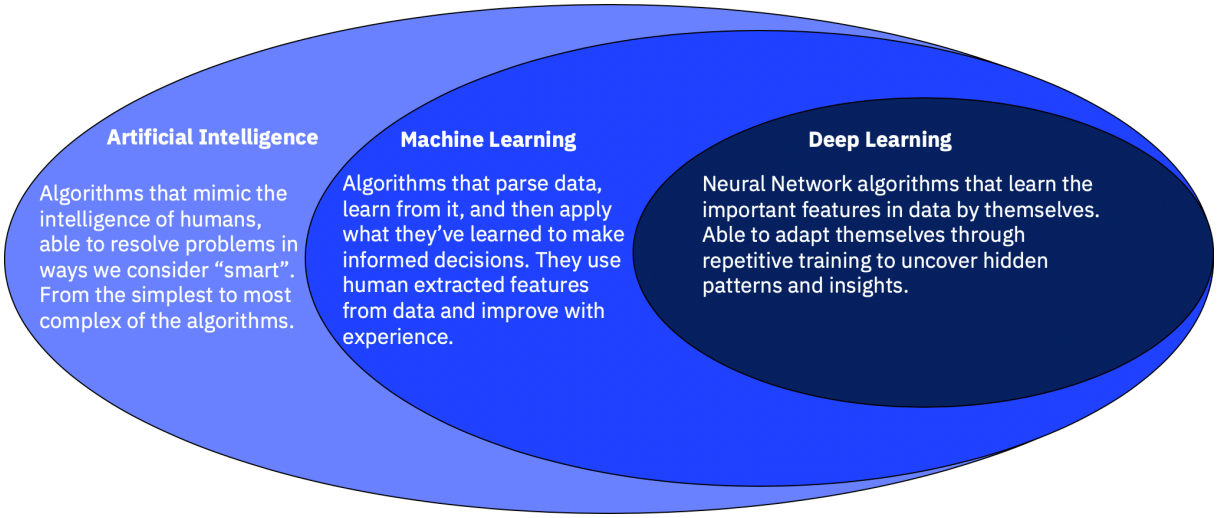

Abbildung 4: Das Forschungsfeld KI und seine Unterfelder
(Quelle: *https://www.ibm.com/blogs/systems/ai-machine-learning-and-deep-learning-whats-the-difference*)

vollständige Modelltransparenz derzeit technisch nicht möglich

Als Blackbox-Modell oder opakes KI-System werden Modelle bezeichnet, die aufgrund ihrer Komplexität verhindern, dass Nutzer*innen die Logik hinter ihren Vorhersagen und Ergebnissen nachvollziehen können (Došilović et al. 2018). Obwohl sich der Bereich weiterentwickelt, ist eine vollständige Modelltransparenz derzeit technisch nicht möglich, weil Expert*innen (noch) nicht in der Lage sind, den Schlussfolgerungsprozess dieser Modelle zu verstehen und zu beweisen, dass sie bei neuen, unbekannten Beobachtungen korrekt funktionieren (Vilone/Longo 2020, S. 11).

Trade-off zwischen Leistungsfähigkeit und Interpretierbarkeit

Signifikant ist, dass in der Literatur ein inverses Verhältnis zwischen der Modellgenauigkeit und seiner Transparenz festgestellt wird. Es besteht also ein Trade-off zwischen der Leistung eines Modells und der Interpretierbarkeit und damit der Transparenz eines Modells, wie Abbildung 5 zeigt (Došilović et al. 2018; Arrieta et al. 2020). Es handelt sich um einen empirisch nachgewiesenen Kompromiss: Je komplexer und genauer die Vorhersagen eines Algorithmus für maschinelles Lernen sind – er passt seine Parameter selbst an und findet bestimmte Muster, ohne dass die Programmierer*innen dies ausdrücklich angeben –, desto schwieriger ist es, die interne Logik hinter den KI-Entscheidungen zu verstehen (Ha et al. 2018; Caruana et al. 2020). Dies wird auch als „Trade-off-Problem" zwischen der berechenbaren Vorhersagegenauigkeit und der menschlichen Interpretierbarkeit und Erklärbarkeit bezeichnet (Mori/Uchihira 2019; Izumo/Weng 2021).

Wie Abbildung 5 zeigt, können Deep-Learning-Algorithmen beispielsweise sehr viel genauere Vorhersagen und Ergebnisse liefern (z. B. bei der Bilderkennung oder bei selbstfahrenden Autos) als simplere regelbasierte Algorithmen, die die von den Programmierer*innen festgelegten Regeln nicht selbst anpassen und deshalb nicht das gleiche Leistungsniveau erreichen können. Die letzteren sind

jedoch viel einfacher zu verstehen und ihre interne Entscheidungslogik ist viel besser erklärbar (Hacker et al. 2020). Daher gibt es erhebliche technische Herausforderungen, um bestimmte KI-Modelle verständlicher zu machen, ohne die Qualität ihrer Ergebnisse zu beeinträchtigen (Arrieta et al. 2020).

Die Forschung im Bereich des interpretierbaren maschinellen Lernens und der erklärbaren KI konzentriert sich auf die Minimierung oder sogar Vermeidung dieses Trade-offs durch die Entwicklung genauer interpretierbarer Modelle und durch die Entwicklung neuer Techniken zur Erklärung von Black-Box-Modellen.

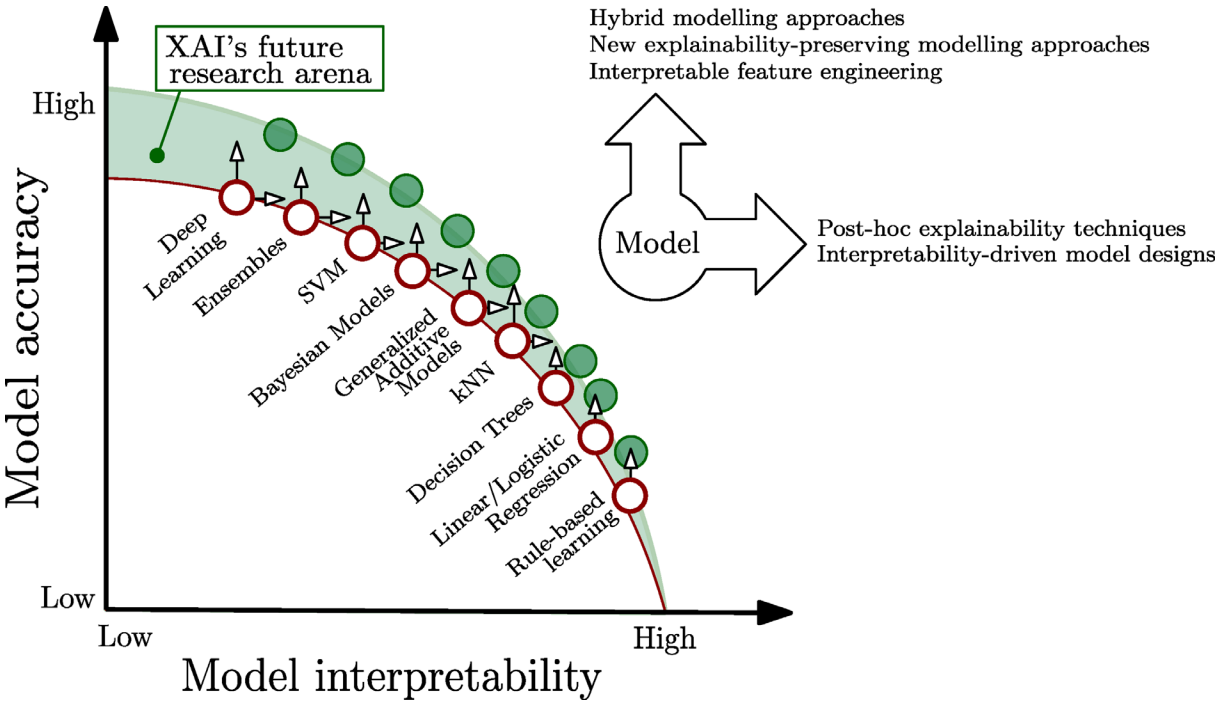

Abbildung 5: Trade-off zwischen Modellinterpretierbarkeit und Genauigkeit
(Quelle: Arrieta et al. 2020)

Es ist dieser Fokus auf Vorhersagegenauigkeit und Interpretierbarkeit/Erklärbarkeit, der das XAI-Feld vorantreibt, in dem Post-Hoc-Techniken mit hybriden Ansätzen kombiniert werden, die darauf abzielen, die allgemeine Interpretierbarkeit (und Erklärbarkeit) der fortgeschrittensten und komplexesten Deep-Learning-Systeme zu verbessern. Die Verknüpfung dieser Herausforderungen und ihre Formalisierung innerhalb bestehender Rahmenwerke ist noch problematisch.

Die Erklärbarkeit umfasst mehrere Fragen, um eine angemessene Erklärung für die interne Funktionsweise und das Verhalten eines Algorithmus zu liefern. Die beiden häufigsten dieser Fragen lauten: *Warum* erzeugt das untersuchte Modell seine Vorhersagen bzw. Schlussfolgerungen. (Vilone/Longo 2020). Die XAI-Forschung hat verschiedene Formen von Erklärungen identifiziert, die für fortgeschrittene KI-Algorithmen in Abhängigkeit von spezifischen Modelleigenschaften relevant sind (Lipton 2018). Es gibt zahlreiche Kategorien und Methoden für KI-Erklärungen (Vilone/Longo 2020), aber zwei große Typen stechen hervor: *Ex-ante-* und *Ex-post-* (oder Post-hoc-)Erklärungen (Lipton 2018; Tsakalakis et al. 2021). Der erste Typ konzentriert sich auf Erklärungen über die Logik des Algorithmus, die Trainingsdaten und das erwartete Verhalten vor der Datenverarbei-

tung. Ex-post-Erklärungen informieren über bestimmte Entscheidungen und bieten spezifische Informationen, um die getroffene Entscheidung zu begründen. (Tsakalakis et al. 2021).

In dieser Hinsicht gibt es auch technische Merkmale, die potenziell in die KI-Modelle eingebaut werden können und die Antworten auf diese Fragen der Erklärbarkeit liefern können. Abbildung 6 illustriert das Konzept solcher erklärbaren Schnittstellen. Es gibt Möglichkeiten, den Nutzer*innen Erklärungen zur Verfügung zu stellen: „*that enable them to understand the system's overall strengths and weaknesses, convey an understanding of how it will behave in future or different situations, and perhaps permit users to correct the system's mistakes*" (Gunning/Aha 2019, S. 45).

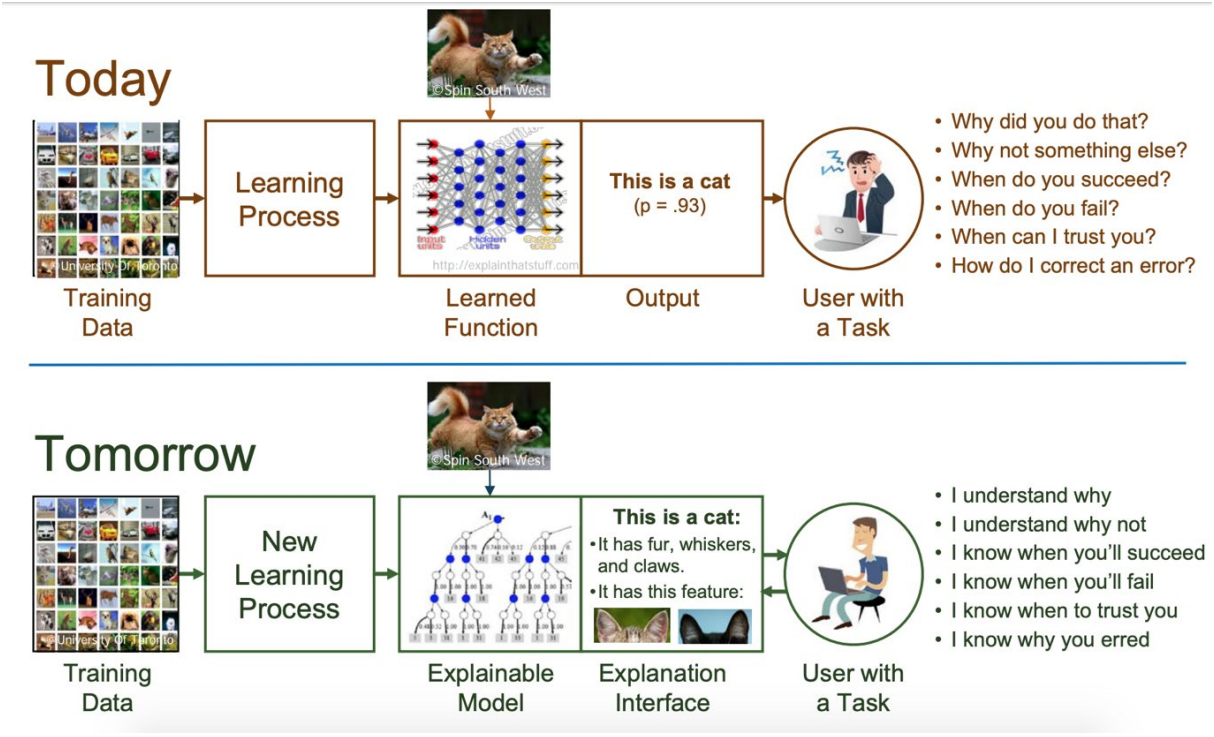

Abbildung 6: Erklärbare KI-Nutzer*innenzentriertes Konzept zur Erstellung von Modellen mit Erklärungsfunktionen (Quelle: Gunning & Aha 2019)

Obwohl solche Lösungen in zahlreichen Teilbereichen der KI noch eine offene Frage sind, wird die Ermöglichung dieser Art von Interaktion zwischen Nutzer*innen und dem Modell für die menschlichen Nutzer*innen und die Gesellschaft von zentraler Bedeutung sein, um das für eine groß angelegte Anwendung erforderliche Vertrauen in die KI zu entwickeln (vgl. Hacker et al. 2020).

Vertrauenswürdige und verantwortungsvolle KI sind Begriffe, mit denen wir uns auf die systemische Übernahme von KI-Prinzipien beziehen, die beim Einsatz von KI in der Praxis *(real use-cases)* unbedingt eingehalten werden müssen (Arrieta et al. 2020, S. 83, 84). Dazu gehören Kriterien wie Sicherheit, Zuverlässigkeit, Fairness (Unbiasedness), Datenschutz, Verantwortlichkeit (oder Rechtfertigungsfähigkeit), Benutzerfreundlichkeit usw. Diese allgemeinen Kriterien sind jedoch schwer zu formalisieren und zu quantifizieren, wes-

halb verschiedene Ansätze (aus den Bereichen Informatik, KI und Governance) viele (Proxy-)Konzepte als Zwischenziele verwenden (Došilović et al. 2018).[7] Eine der ersten Governance-Initiativen in diesem Sinne waren die Betroffenen Rechte in der EU-Datenschutz-Grundverordnung (DSGVO).

Wie bereits erwähnt, sollte ein erklärbares KI-System in der Lage sein, seine Entscheidungen und Erkenntnisse zu erklären. Es sollte darlegen, was es getan hat, was es jetzt tut und was als Nächstes geschehen wird, und die wichtigsten Informationen offenlegen, auf deren Grundlage es handelt (Bellotti/Edwards 2001). Weiterführend sollte KI in diesem Sinne transparent, verständlich, erklärbar und interpretierbar sein. Wie bereits gezeigt, überschneiden sich alle diese Begriffe und beziehen sich manchmal aufeinander; gleichzeitig sind sie kontextabhängig und können sich in wichtigen Aspekten unterscheiden. Darüber hinaus kann die Klärung der Verwendung dieser Begriffe die Grundlage für die Umsetzung von (rechtlichen) Rahmenbedingungen für vertrauenswürdige KI bilden.

ein KI-System sollte erklären, was es getan hat, was es jetzt tut und was als Nächstes geschehen wird, und die wichtigsten Informationen offenlegen, auf deren Grundlage es handelt

In der (z. B. politischen oder administrativen) Praxis werden häufig weniger komplexe gegenüber komplexeren Algorithmen bevorzugt, weil sie Nachvollziehbarkeit gewährleisten. Anwender*innen bevorzugen also „White-Box-" gegenüber „Black-Box-"ML-Algorithmen, da diese das Vertrauen in die resultierenden Ergebnisse und Empfehlung erhöhen. XAI versucht, die Intransparenz von Black-Box-ML-Algorithmen zu überwinden und gleichzeitig deren hohe Modell-Leistungsfähigkeit zu erhalten. Anstatt von vornherein auf White-Box-Modelle zurückzugreifen, machen die hierfür angewandten Transfertechniken (von Black-Box-Modellen zu XAI) jedoch das Trainieren zweier Modelle notwendig (Herm et al. 2021). Solche Ex-post-Erklärungsansätze („Grey-Box-Modelle"), bei denen XAI-Methoden auf bereits trainierte Modelle angewandt werden, um ihre interne Logik und Vorhersagen transparent zu machen, werden immer häufiger angewandt (ibid., S. 2).

In Hinblick auf Nachvollziehbarkeit stellen Ansätze der erklärbaren KI eine Möglichkeit dar, Transparenz herzustellen, sofern ihre Entscheidungen *von Menschen* als nachvollziehbar eingestuft werden und diese Systeme Erläuterungen abgeben (können), warum bestimmte Entscheidungen getroffen wurden (z. B. Zugänge zu bestimmten Leistungen, wie Sozialleistungen, verwehrt blieben) (Krafft/Zweig 2019).

[7] Siehe auch Kapitel 5.

4 WIRKUNGEN VON KI-SYSTEMEN

KI wird von vielen Wissenschafter*innen als „disruptive" Technologie gesehen, die in nahezu allen Lebensbereichen eingesetzt wird und so zu einem neuen Paradigma an der Schnittstelle von Mensch und Maschine führen kann. Die Universalität der Einsatzmöglichkeiten führt auch dazu, die KI als zukünftige „General Purpose Technology" oder Universaltechnik zu bezeichnen. Die erwarteten ökonomischen Vorteile sollen zu Wettbewerbsvorteilen für Entwickler*innen und Anwender*innen führen und auch allgemein zum Wirtschaftswachstum beitragen. Aus diesen Gründen werden in allen Industrieländern umfangreiche Förderprogramme für KI aufgelegt. Die großen Player dabei sind vor allem die USA, China und die EU. Eine Übersicht zu (inter-)nationalen Programmen zur Förderung der KI findet sich in der OECD-Datenbank für KI-Politik.[8] Hier sind über 600 Einträge zu Politiken, Förderprogrammen, Statistiken und Bewertungen zu ökonomischen und sozialen Auswirkungen gelistet. Diese teilen sich auf etwa zwanzig Politikbereiche auf und erlauben einen Blick auf KI-Anwendungen in unterschiedlichsten Bereichen. Dabei zeigt sich eine große Abhängigkeit der KI-Entwicklung von Daten als zentraler Ressource. Diese Abhängigkeit von Daten, sowohl für Trainingszwecke als auch für die tatsächliche Ausführung ihrer Funktion, führt direkt zur Diskussion potenzieller Risiken einer weit verbreiteten KI-Anwendung.

KI als disruptive Technologie

Was sind nun die Erwartungen, die mit dem verbreiteten Einsatz von KI verbunden werden? Es sind Visionen von schnelleren und besseren Entscheidungen durch derartige Systeme bis selbst zur Übernahme vieler Funktionen durch eine autonom handelnde KI. Bereiche, in denen sich diese Visionen zuerst realisieren sollen, sind die industrielle Automatisierung, intelligente Roboter, autonomes Fahren, Verarbeitung natürlicher Sprache, Kund*innen-Kommunikation und -kategorisierung, aber auch die Roboter-Medizin. Die Anwendungsbreite zeigt sich unter anderem auch darin, dass im aktuellen Monitoring von Zukunftsthemen für das Österreichische Parlament (Nentwich et al. 2021) mehr als 25 der 130 dargestellten sozio-technischen Entwicklungen direkt mit KI zu tun haben bzw. diese einsetzen.[9]

mit breitem Einsatzspektrum

[8] OECD.AI Policy Observatory, *https://oecd.ai/en/*.

[9] Beispielhaft etwa folgende Themen: Zukunft der Bewertungsplattformen, Robo-Journalismus, Micro-Targeting, Algorithmische Diskriminierung, Deep-Fakes, KI im Gesundheitswesen, Deep-Reading, KI-Kunst, Social (Ro-)Bots, Dezentrale KI-Lernen, KI-Kriegsführung, Fortgeschrittene Gesichtserkennung, Existentielle Risiken der KI, Sicherheits-Robotik, Algorithmische Polizeiarbeit, Autonomer öffentlicher Verkehr, Fernerkundung mit KI, Robotik in der Landwirtschaft, Automatisierung in der Rechtsberatung, Smart Spaces, Virtuelle & Augmentierte Realität, Zukunft der Mensch-Maschine-Interaktion: Spracherkennung und -steuerung, Roboterautos, Affective Computing – Emotionale Künstliche Intelligenz, Gamification von Wissenschaft, Arbeit und Politik?
Der gesamte Bericht findet sich hier: *https://fachinfos.parlament.gv.at/wp-content/uploads/2021/11/000_Bericht_gesamt_aktualisiert_November2021_korr.pdf*.

In den positiven Visionen zeigen sich sowohl quantitative wie auch qualitative Aspekte. Die Schnelligkeit von Computern bei einem dedizierten bzw. eingeschränkten Aufgabenspektrum steht tatsächlich weit über der von Menschen. Inwieweit tatsächlich qualitative Fortschritte möglich sind, lässt sich jedoch nicht pauschal, sondern nur in der Beurteilung einer konkreten Nutzung und ihres Anwendungskontextes beurteilen. KI führt jedoch auch zu massiven Bedenken hinsichtlich der Risiken für die Gesellschaft. Diese reichen von grundlegenden ethischen Überlegungen, die sich auf eine Beschränkung der menschlichen Autonomie oder die Bedrohung der Menschenwürde durch KI-Anwendungen beziehen, bis zu grundrechtlichen Problemen wie etwa diskriminierenden Verzerrungen (Bias) in den Ergebnissen. Darüber hinaus werden auch negative Auswirkungen auf die gesellschaftliche Diskursqualität und damit die Demokratie befürchtet. Dies fußt vor allem auf Überlegungen bezüglich der Bildung so genannter Echokammern oder Filterblasen bei der Nutzung von Social-Media-Plattformen. Besonders relevant sind auch Befürchtungen von Auswirkungen auf den Arbeitsmarkt und die zukünftigen Beschäftigten, sowohl in quantitativer als auch qualitativer Hinsicht.

Auswirkungen von Digitalisierung, Automatisierung und KI auf den Arbeitsmarkt

Dazu gibt es seit einigen Jahren eine lebhafte Diskussion über die Auswirkungen von KI, Automatisierung, Robotern und Digitalisierung im Allgemeinen auf den Arbeitsmarkt. Eine scharfe Unterscheidung zwischen den verschiedenen Bereichen ist kaum möglich. KI kann aber als Basistechnologie sowohl für Roboter, Automatisierung als auch für die Digitalisierung von Arbeitsplätzen jenseits des Fließbandes angesehen werden. Ausgelöst wurde die Debatte durch die Studie von Frey/Osborne (2013), die u. a. aufzeigte, dass ca. 47 % aller Arbeitsplätze in den USA ein hohes Risiko tragen, in einem Zeitraum von ca. 20 Jahren computerisiert zu werden. Als Haupteinwand gegen diese Studie wurde die historische Analogie ins Treffen geführt, die behauptet, dass alle technologischen Fortschritte in der Geschichte auch zu einer Zunahme neuer (anderer) Arbeitsplätze geführt hätten. Dieses Argument ignoriert jedoch die Tatsache, dass die wirtschaftlichen Bedingungen in früheren Perioden andere waren (Čas/Krieger-Lamina 2020). Mit der Digitalisierung stehen moderne Gesellschaften vor völlig neuen Herausforderungen. IKT und KI überfluten alle Lebensbereiche, nicht nur den Produktions- und Dienstleistungssektor. Der Analyseansatz von Frey/ Osborne wurde mehrfach wiederholt, führte jedoch nicht zu ähnlich dramatischen Ergebnissen. Dennoch wurden dadurch Diskussion über zukünftige Arbeitsmarktentwicklungen und mögliche Arbeitslosigkeit angeregt. Eine proaktive Auseinandersetzung mit möglichen auch dystopischen Zukünften, wie sie hier passierte, könnte dazu beitragen, die aktive Gestaltung eben dieser Zukunft frühzeitig anzuregen.

*KI als Herausforderung für die Qualifikation der Anwender*innen …*

Gleichzeitig weist die Oxford Commission on AI & Good Governance (2021) darauf hin, dass die Anwendung von KI auch massive Herausforderungen in Bezug auf (Aus-)Bildung der Verwender*innen im jeweiligen sozialen Umfeld bereithält. So identifizieren sie für die öffentliche Verwaltung die Beschaffung von KI und die Sammlung und Analyse von Trainingsdaten als Herausforderung. Vor allem vor dem Hintergrund, dass Personen der öffentlichen Verwaltung kaum über Expertise und Fähigkeiten, sowie praktische Toolkits für gute Entscheidungsfindung verfügen und damit entsprechende technische und praktische Fähigkeiten benötigen um KI für Good Governance einsetzen zu können:

… und in der Beschaffung

> *„Public servants lack expertise and skills, but also practical toolkits to make good decisions. It is clear that powerful technology companies have superior bargaining*

power and expertise in comparison with governments and public administrators. Public servants require technical and practical capacities for the adoption of AI for good governance" (Oxford Commission on AI & Good Governance 2021, S. 5).

4.1 GRUNDLEGENDE ÜBERLEGUNGEN

Bevor beispielhaft auf mögliche konkrete Auswirkungen in unterschiedlichen gesellschaftlichen Bereichen bzw. wirtschaftlichen Sektoren eingegangen wird, sollen die grundlegenden Fragen eines breiten und unregulierten Einsatzes von KI kurz dargestellt werden. Dazu zählen die Frage nach

- der Möglichkeit, eine moralische KI (oder Maschinen allgemein) zu schaffen;
- ethischen Überlegungen zu KI;
- der Mächtigkeit von KI, im Sinne einer schwachen (spezifischen) oder allgemeinen (universellen) KI und der Entwicklung einer „Superintelligenz";
- der konkreten Verantwortungsverteilung und -übertragung auf KI.

Viele Entscheidungen von algorithmischen Entscheidungssystemen greifen direkt in das Leben von Menschen ein. Oft wird das Beispiel des Verhaltens eines autonomen Fahrzeuges in unausweichlichen Unfallsituationen angesprochen und die Frage gestellt, wie sich das System entscheidet, wen es bevorzugt schützen solle, etwa Tiere, junge oder alte Menschen? Das so genannte Trolley-Problem[10] verweist auf moralische Fragen im Zusammenhang von Entscheidungen und Handlungen. In der Debatte ethischer Aspekte von KI kommen daher auch Fragen der Verantwortung in den Blick. Die Frage, wer die Verantwortung für Aktionen von KI-Systemen tragen soll, verweist auch auf zwei unterschiedliche Dimensionen der Ethik-Debatte. Jene, die die Ansicht vertreten, dass die Entwickler*innen und Anwender*innen von KI-Systemen in erster Linie Verantwortung für deren Einsatz tragen, beschäftigen sich mit *Maschinenethik* und denken über Regeln im Bereich KI und Robotik nach. Mögliche ethisch begründete Regeln für KI-Systeme, Roboter und deren Entwickler*innen werden in Kapitel 5 genauer beleuchtet. Andere, die auch den Maschinen Verantwortung übertragen wollen, befassen sich mit der so genannten Ethik der Maschinen (*machine morality*). Die moralische Maschine, die sich selbst Regeln gibt und nach diesen dann handelt, kann jedoch erst mit der Umsetzung der sogenannten starken KI relevant werden, ist daher dzt. nicht umsetzbar (Schaber et al. 2019).

Ausgangspunkt vieler Überlegungen zu einer Ethik der KI bilden die frühen Robotergesetze von Isaac Asimov. Diese drei Gesetze, die Asimov 1942 entwickelte, sollen dem Gefährdungspotential des Roboters den Menschen gegenüber begegnen, indem sie Handlungen (oder Unterlassungen von Handlung), die Menschen zu Schaden kommen lassen, verbieten (erstes Gesetz), die Befehlsgewalt über den Roboter beim Menschen belassen (zweites Gesetz) und die Selbsterhaltung des Roboters gewährleisten (drittes Gesetz). Die Gesetze sind hierarchisch strukturiert, d. h. das dritte Gesetz darf nur befolgt werden, solange die ersten zwei Gesetze dadurch nicht verletzt werden. So einflussreich Asimovs Arbeiten auch anfangs waren, stellen sie jedoch keine ausreichende und effektive Grundlage für das Design von Robotern im Allgemeinen dar (Clarke 1993; Clarke 1994; Čas et al. 2017).

Ethik für KI

[10] *https://de.wikipedia.org/wiki/Trolley-Problem*.

Die Beschwörung einer möglicherweise entstehenden „Superintelligenz" haben den medialen Hype um ethische (und oft auch dystopische) Fragen der KI sicherlich befördert. Doch auch zum derzeitigen Stand der Technik und der erwartbaren Weiterentwicklung schwacher KI bieten sich genügend Fragen, die in einem breiten gesellschaftlichen Diskurs erörtert und gelöst werden sollten – dies umso mehr, als KI schon heute eine wichtige Rolle im Alltag von Menschen spielt und diese in naher Zukunft noch viel mehr beeinflussen wird. Eine zentrale Forderung in diesem Diskurs ist jene nach Transparenz – jene über den Einsatz von KI, aber auch über die Funktionsweisen, die eingebauten Algorithmen und deren Wirkungsweise.[11] Denn nur durch Transparenz lassen sich Systeme kontrollieren und so Vertrauen aufbauen, was auch gesellschaftliche Akzeptanz fördern kann.

Transparenz als zentrale Forderung

In der sozialwissenschaftlichen Literatur werden auch mögliche Limitierungen von KI diskutiert. Eine davon ist die Frage von Verantwortlichkeit von KI und ihr Auftreten als „moralischer Akteur". KI kann auf unterschiedliche Arten als moralischer Akteur auftreten (nach Weber/Zoglauer 2019, S. 149). Einerseits als Maschinen, die sich moralisch relevant verhalten (z. B. indem sie Menschen verletzen) und deren Verhalten deshalb von moralischen Normen und Werten geleitet wird (durch entsprechende Programmierung) (Siegetsleitner 2020). Dies bedeutet, dass Maschinen keine moralische Verantwortung übernehmen können. Andererseits gibt es Beschreibungen von Maschinen als „moralische Akteure", die selbst die kognitiven Fähigkeiten haben könnten, absichtlich moralischen Normen und Werten zu folgen. Dies ist in näherer Zukunft jedoch nicht absehbar, da Maschinen keine (eigenständige) Entscheidungsfähigkeit aufweisen. In der Praxis bedeutet dies, dass Maschinen sich nur innerhalb eines Frameworks für ethisch akzeptable Prinzipien bewegen sollten (Weber/Zoglauer 2019, S. 125). Die Frage der Verantwortung bei selbstlernenden Algorithmen wird allerdings unterschiedlich beantwortet, wobei die Vielfältigkeit der involvierten Akteur*innen im Bereich KI eine Herausforderung darstellt: *„Authors like Hellström (2013) are however of the opinion that the less predictable and controllable the behavior of AI systems is, the less responsible their developers can be held"* (Siegetsleitner 2020, S. 128).

Maschinen können keine Verantwortung tragen

Im Gegensatz dazu argumentieren Dieter und Birnbacher (2016) am Beispiel autonomes Fahren, dass zwar KI selbst keine (oder nur eingeschränkt) Verantwortung übertragen werden kann, dafür aber zusätzliche auf den Entwickler*innen lastet:

diese bleibt bei den gestaltenden Menschen

> *„How the machine 'decides' in conflict situations is determined by those individuals who program the behavior of the machine in hypothetical situations; these are ultimately the instances that decide on their approval for specific applications. At the same time the decision-makers at this higher level also bear an additional responsibility. They cannot excuse themselves from guilt and responsibility by referring to the typical psychological stress factors in conflict situations"*
> (Dieter/Birnbacher 2016, S. 129).

[11] Details dazu in Kapitel 3.

Neben den grundsätzlichen Erwägungen bezüglich Moralität von Maschinen bzw. der besonderen Verantwortung von Entwickler*innen und den mit Entwicklung, Einsatz und Kontrolle von KI befassten Institutionen gibt es noch weitere grundsätzliche Fragen, die vor dem Einsatz von KI bedacht werden sollten. Dazu gehört, dass mit Hilfe von KI versucht wird, aus Daten der Vergangenheit anhand eines Modells mit einer gewissen Wahrscheinlichkeit die Zukunft vorherzusehen bzw. in der Gegenwart Handlungen vorzuschlagen bzw. zu setzen, die diese Zukunft beeinflussen. Das führt zur Frage, ob es grundsätzlich sinnvoll und ethisch vertretbar ist, ein so komplexes Gemenge, wie eine bestimmte Lebenssituation von Menschen in einer Zahl (oder Zahlen) auszudrücken? Weiters zeigt sich die zentrale Rolle der verwendeten Modelle, die definitionsgemäß eine Komplexitätsreduktion gegenüber der realen Welt darstellen. Zu fragen bleibt in jedem konkreten Fall: Was geht aufgrund der Komplexitätsreduktion verloren?

Soll man das Leben von Menschen auf eine Zahl reduzieren?

Doch auch die zentrale Fähigkeit der Vorausschau durch „Lernen aus der Vergangenheit" ist zu hinterfragen. Die Betonung des „Erfahrungsschatzes" in Daten der Vergangenheit führt tendenziell zu konservativen, fortschreibenden Einschätzungen und verringert so die Flexibilität und Offenheit für Neues. Damit werden andere Optionen schwächer gewichtet bzw. ausgeschlossen, was wiederum Veränderungen, Fortschritt und möglichen sozialen Wandel verlangsamt oder gar verunmöglicht. Schließlich spielen in der KI Statistik und Wahrscheinlichkeiten eine zentrale Rolle. Damit läuft die Gesellschaft aber Gefahr, ein zentrales Paradigma zu verändern, kommt es doch zu einem Wechsel von Kausalität zur Probabilität. Statt konkrete Zusammenhänge und „Wenn-Dann"-Beziehungen zu untersuchen, wird vermehrt auf Korrelationen gesetzt, die neben der o. a. angeführten Problematik der Vergangenheitsabhängigkeit grundsätzlich auch als Fehlerquelle angesehen werden können. Es ist also genau zu beachten, in welchen Bereichen diese eingesetzt werden und ob die gefundenen Korrelationen tatsächlich in einem nachvollziehbaren Zusammenhang stehen. Besonders prägnante Beispiele für hohe Korrelationen, die jedoch inhaltlich nicht in Zusammenhang stehen, können bei den Scheinkorrelationen (spurious correlations)[12] gesehen werden.

die Begründung von Entscheidungen auf Basis der Geschichte führt tendenziell zu konservativem Verhalten

Korrelationen statt Kausalität

Neben der Qualität des Modells und der Beachtung inhaltlicher Zusammenhänge in den Korrelationen ist es vor allem aber die Qualität der Daten, die auf die Qualität der Entscheidungen wesentlichen Einfluss nehmen. KI-Systeme werden oft als „unbestechlich" und „objektiv" angesehen, dennoch gibt es eine große Anzahl von Beispielen, dass KI-Systeme inhärente Diskriminierungen nicht vermeiden, sondern in vielen Fällen sogar verstärken. Schlechte Trainingsdaten und unreflektierte bzw. nicht qualitativ, inhaltlich hinterfragte Qualitätsmaße – Ansprüche an das Entscheidungssystem – führen zu Verzerrungen (biases) und möglicherweise zu verstärkter Diskriminierung (Allhutter 2019). Deshalb erscheint es besonders wichtig, dass grundsätzlich alle Schritte einer Entwicklung eines algorithmischen Entscheidungssystems hinterfragt und offen diskutiert werden (Allhutter/Berendt 2020). Die Entwicklung eines algorithmischen Entscheidungssystems ist sehr komplex und es sind auch viele unterschiedliche Akteur*Innen daran beteiligt. Dass dabei eine Vielzahl von Fehlermöglichkeiten entstehen, zeigt die untenstehende Grafik (Zweig 2018).

Bias und Diskriminierung in KI-Systemen

[12] *https://www.tylervigen.com/spurious-correlations* und *https://scheinkorrelation.jimdofree.com/*.

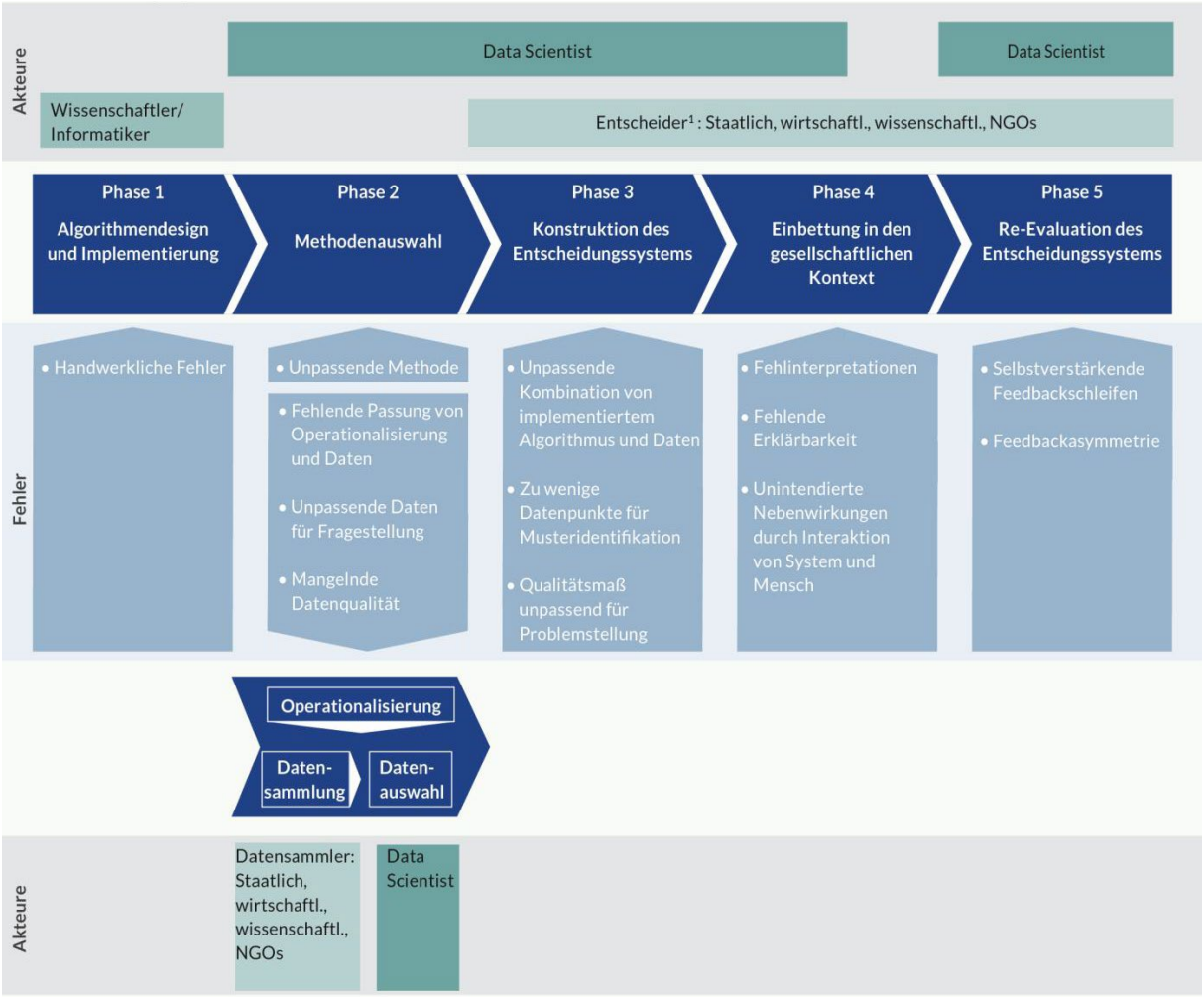

Abbildung 7: Übersicht über mögliche Fehler im Entwicklungs- und Einbettungsprozess von algorithmischen Entscheidungssystemen (Quelle: Zweig 2018)

In diesem Bild wird deutlich, dass spätestens ab Phase 2 der Entwicklung Wertentscheidungen fallen, die sich auf die Funktionsweise des algorithmischen Entscheidungssystems auswirken und deshalb auch hinterfragt und offen diskutiert werden sollten.

In der KI-Entwicklung stellt sich die grundlegende Frage, welche Werte in KI-Anwendungen eingeschrieben werden sollen. Unter Umständen existieren Diskrepanzen zwischen menschlicher und maschineller Logik im Handeln nach bestimmten Werten. Beispielsweise könnten automatisierte Entscheidungen nach zuvor festgesetzten Werten (z. B. Rationalität oder Effizienz) absolut gesetzt (und auch missbräuchlich uneingeschränkt exekutiert) werden und damit massive negative Konsequenzen nach sich ziehen (z. B. die systematische Ausnutzung von bestimmten Systemen, wie Steuersysteme; siehe „AI Hackers“ vgl. Schneier (2021)).

4.2 ANWENDUNGSKONTEXTE

Abseits der wissenschaftlichen Debatten um KI werden algorithmische Entscheidungssysteme schon in vielen Bereichen eingesetzt. KI-Algorithmen berechnen Hochschulrankings, Kreditwürdigkeitsprüfungen und die Bearbeitung von Jobbewerbungen, sie entscheiden, welche Online-Werbung man sieht oder welche Beiträge im Facebook-Newsfeed erscheinen; sogar die Polizei nutzt Big Data, um vorherzusagen, wo Verbrechen geschehen könnten. Es ist also eine offenkundige Tatsache, dass Algorithmen eine immer wichtigere Rolle in der Gesellschaft spielen. In ihrem oft zitierten Buch „Weapons of Math Destruction" zeigt O'Neil (2016) wortgewandt, dass diese so genannten Algorithmen oft nicht in der Lage sind, die reale Welt widerzuspiegeln: *„mathematical models should be our tools, not our masters"* (S. 164). Das Hauptargument ist einfach, dass Vorhersagemodelle niemals neutral sind, sondern die Ziele und die Ideologie derer widerspiegeln, die sie erstellen und entwickeln. Das Kernproblem liegt im Fehlen von Verantwortlichkeit, Kontrolle und Erklärungsmechanismen. Wie O'Neil es formuliert, sind Algorithmen undurchsichtig und ihre Urteile oft unanfechtbar:

> *„Like gods, these mathematical models were opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists. Their verdicts, even when wrong or harmful, were beyond dispute or appeal. And they tended to punish the poor and the oppressed in our society, while making the rich richer"* (ibid, S. 14).

Die Frage nach den Auswirkungen kann nicht ohne Beachtung des Einsatzkontextes erfolgen. Im Folgenden deshalb eine kurze Darstellung von einigen exemplarischen Anwendungsfällen, die die unterschiedlichen Dimensionen der möglichen Auswirkungen deutlich machen sollen.

NAVIGATIONSSYSTEME

nützliche Helfer mit geringem konkreten Schadens- aber Missbrauchspotenzial

Eines der am weitesten verbreiteten KI-Systeme im Alltag von Konsument*innen dürfte die Navigationssoftware im Auto, fürs Radfahren oder Wandern sein. Hier wird ein logistisches Problem – der effizienteste, schnellste oder schönste Weg von A nach B – unter Einbeziehung bestehender Daten (Karten) und aktueller Informationen aus der Umwelt (Verkehrsdichte, Unfälle etc.) dynamisch gelöst. In der Regel wird von solchen Systemen nur ein geringes Schadenspotential für Konsument*innen ausgehen. Diese Systeme sind, abgesehen von punktuellen Fehlfunktionen, in erster Linie hilfreich und erleichtern den Alltag. Sie können aber für andere Zwecke missbraucht werden – etwa die Verhaltensbeeinflussung (Nudging) oder auch im Zusammenhang mit Arbeitsverhältnissen. Hier besteht durch Navigationssysteme ein enormes Steuerungs- und Überwachungspotential mit gekoppeltem Autonomieverlust für die Arbeitnehmer*innen.

ONLINEHANDEL

Weniger bewusst, aber umso breiter angewandt wird KI im Bereich Verhaltensmonitoring und Vorschlagsysteme, etwa beim Online-Handel und Online-Dienstleistungen wie z. B. Musik- oder Filmstreaming sowie auch bei digitalen Assistenten wie Alexa, Siri und Google Assistent (Schaber et al. 2019). Dabei werden aufgrund des bisherigen Verhaltens der Konsument*innen Profile angelegt, mit anderen Diensten abgeglichen und so eine möglichst genaue Vorhersage des zukünftigen Verhaltens zu errechnen versucht. Konsument*innen werden auf diese Weise zu berechenbaren Einheiten und nicht in ihrer Gesamtheit als Mensch wahrgenommen. Sie tragen somit zur Optimierung von Marketing und Lagerhaltung und damit zu verbesserten Erlösen bei. In diesen Bereichen ist das Schadenspotential von KI durch mögliche Übervorteilung auf Seiten der Konsument*innen und vor allem ökonomischer Natur. Allerdings kommt es aufgrund der Informationsungleichheit – viele Konsument*innen wissen gar nichts über den Einsatz von KI im Hintergrund und die Möglichkeiten der Beeinflussung – zu einer weiteren Verfestigung von Ungleichheit und damit auch zu ungleichen Marktbedingungen, die die Konsument*innen benachteiligen und ein „rationales" Vorgehen der Konsument*innen erschweren.

zielgruppengenaue Produktvorschläge vs. (geringes) ökonomisches Schadenspotenzial durch Beeinflussung

Personalisierte Werbung kann zwar nützlich und zeitgemäß sein, aber sie kann auch anfällig für unlautere Werbung machen und weitreichendere Konsequenzen nach sich ziehen (z. B. Verleiten zu übermäßigem Konsum und damit zusammenhängende Armutsgefährdung, Verschuldung und Nachhaltigkeitsprobleme). Vor allem, wenn soziale und wirtschaftliche Ungleichheit eine Rolle spielen, etwa Werbeanzeigen, die Menschen in großer Not ausfindig machen und ihnen falsche Versprechungen machen: *„Anywhere you find the combination of great need and ignorance, you'll likely see predatory ads"* (O'Neil 2016, S. 64).

SUCHMASCHINEN UND SOCIAL MEDIA

Ein ganz wesentlicher Dienst im Internet, ohne dessen Nutzung ein Alltag kaum mehr vorstellbar ist, sind Suchmaschinen (Mager 2014a; Mager 2014b). Diese speichern alle Anfragen der Nutzer*innen und gewinnen so ein Bild von den Interessen und zu einem guten Teil auch der persönlichen Lebensumstände der Nutzer*innen. Diese Profile bilden die Basis für „personalisierte" und leicht akzeptierbare Ergebnisse (Lewandowski 2014; Mager 2014a). Dass dies keine „neutralen" Ausschnitte aus der Informationsflut des Internet sind, sondern damit eine bestimmte Weltsicht unterstützt bzw. hergestellt wird, ist den meisten Nutzer*innen nicht bewusst. Individuell besteht dadurch in den wenigsten Fällen ein existenzielles Risiko.

Da aber beispielsweise auf den Plattformen der Sozialen Medien ähnliche Profile generiert und von Algorithmen ausgewertet werden, können mittlerweile einschneidende gesellschaftliche Risiken und massive demokratiepolitische Auswirkungen festgestellt werden. Stichworte dazu sind etwa das Microtargeting, die gezielte Ansprache von potentiellen (Nicht-)Wählern und die Darstellung völlig unterschiedlicher Botschaften im Zuge von Wahlkämpfen oder anderen Kampagnen (Wahl des US-Präsidenten 2016 und Referendum zum Brexit 2016, vgl. Howard (2020)). Die Auswahl-Algorithmen führen auch dazu, dass neben den bewusst gelenkten Informationen im Zuge von Kampagnen die Plattformen der Sozialen Medien auch im Normalbetrieb versuchen, ihre Nutzer*innen möglichst lange auf ihren Seiten zu halten, was den Wert der geschalteten Werbeeinnah-

Filterblasen und Demokratie

men erhöht. Die Verweildauer wird durch erhöhte Aufmerksamkeit erreicht, diese wiederum mittels prononcierter Stellungnahmen, sodass eine positive Zustimmung erzeugt werden kann. Da man durch die Plattformen vor allem Meldungen von Menschen bekommt, die die eigene Meinung ebenfalls vertreten, kommt es zu sogenannten Filterblasen (Pariser 2011) und einer tendenziell abnehmenden Bereitschaft und Fähigkeit, mit Menschen außerhalb dieser einen Blase Dialog zu führen. Die abnehmende Diskursqualität führt zu Polarisierung und gefährdet damit grundlegende Aspekte des Zusammenlebens in demokratischen offenen Gesellschaften (Falkner 2022).

RECRUITING UND ARBEITSSUCHE

Im Bereich der Arbeitssuche spielt KI eine zunehmende Rolle. Da auch für Unternehmen sehr viel davon abhängt, die richtigen Mitarbeiter*innen einzustellen, wird mit Hilfe von KI versucht, den Personalbeschaffungsprozess (recruiting) effizienter zu gestalten und eine möglichst perfekte Besetzung für konkrete Aufgaben bzw. Positionen zu finden. Software für diesen Bereich umfasst sowohl die Textanalyse von Bewerbungsschreiben und die darauf aufbauende Kategorisierung und Priorisierung von Bewerber*innen als auch bereits von Robotern durchgeführte Einstellungsgespräche.[13] Wie in anderen Fällen auch, wird als eine der Chancen des KI-Einsatzes gesehen, dass Vorurteile ausgeblendet und eine neutrale Beurteilung von Bewerber*innen erfolgen kann. Tatsächlich dürfte eine der Chancen in der Nutzung von KI-Systemen sein, dass sich die für den Recruiting-Prozess Verantwortlichen viel präziser mit den Anforderungen an die Beurteilung der Bewerber*innen auseinandersetzen müssen: Ein Merkmal dieser Systeme ist, dass sie schneller und effizienter arbeiten; da dadurch der Durchsatz erhöht (skaliert) wird, führt das dazu, dass sowohl positive, wie auch negative Folgen sehr große Auswirkungen annehmen können (Knobloch/Hustedt 2019). KI im Personalverfahren kann also dazu führen, dass intensiver über Anforderungen und (versteckte) Vorurteile nachgedacht wird, sie kann aber auch zur Fortschreibung und Verstärkung, von diskriminierenden Ergebnissen führen, wenn der Gestaltung des jeweiligen Systems keine hohe Aufmerksamkeit geschenkt wird bzw. unpassende Trainingsdaten Anwendung finden. KI-Systeme in diesem Bereich greifen direkt in individuelle Lebenschancen ein und haben insofern auch für Individuen ein recht hohes Schadenspotential, zumal mit Entscheidungen bezüglich des Arbeitsplatzes konkret Lebenschancen von Personen von KI mitbestimmt werden. Aus diesem Grund gibt es Stimmen, die vor einer „Entmenschlichung“ (Beining 2019) des Bewerbungsprozesses warnen und darauf hinweisen, dass das persönliche Gespräch und der menschliche Eindruck wohl nicht so schnell aus den Bewerbungsprozess verschwinden werden (Knobloch/Hustedt 2019).

Effizienz im Recruiting vs. potenziell unfaire oder falsche Kategorisierung von Arbeitssuchenden

Einen Spezialfall stellt das „Arbeitsmarktchancen-Assistenz-System“ (AMAS) in Österreich dar (eine umfassende Analyse dazu bieten Allhutter et al. (2020a); Allhutter et al. (2020b)). Damit sollen auf Basis von Statistiken vergangener Jahre zukünftige Chancen von Arbeitssuchenden am Arbeitsmarkt vorhergesagt werden. Arbeitssuchende werden anhand ihrer „Integrationschance“ in drei Gruppen eingeteilt (ITA 2019). Zu den offenen Fragen zählen einerseits die problematische Verkürzung individueller Erwerbsbiographien auf eine Zahl und der stark

Arbeitsmarktchancen-Assistenz-System

[13] Z.B. *https://furhatrobotics.com*.

determinierende Effekt des „objektiven" Ergebnisses der Kategorisierung durch das System auf die individuellen Entscheidungen der AMS-Berater*innen, andererseits die Verlagerung der Beratungspraxis vom persönlichen Förderbedarf einer Einzelperson hin zu Arbeitsmarktchancen auf Basis einer Populationsberechnung. Besonders deutlich wird anhand dieses Systems, wie problematisch die starke Abhängigkeit von Daten aus der Vergangenheit sein kann (vgl. Allhutter et al. 2020a), was insbesondere bei disruptiven Ereignissen, wie Wirtschaftskrisen oder der Covid-19-Pandemie, deutlich wird (ITA 2021).

MONITORING IM ARBEITSALLTAG

KI spielt jedoch nicht nur im Zusammenhang mit Kategorisierungsverfahren im Jobwechsel eine Rolle, sondern nimmt auch subtil Einfluss auf den Arbeitsalltag. Durch automatisierte Aufgabenverteilung und Monitoring von Arbeitnehmer*innen wird vor allem eine effizientere Produktion erhofft, wobei mittelfristige Effekte auf die Arbeitsplatzqualität häufig ignoriert werden. So erhöht sich der Stresslevel, wenn Reihenfolge, Geschwindigkeit und Methoden von zu erledigenden Aufgaben, wie durch KI angedacht, nicht autonom bestimmt werden können. Um negative Gesundheitsauswirkungen zu minimieren, schlägt Nurski (2021) eine Verankerung solcher Indikatoren in der KI-Regulierung vor, auf die sich das Arbeitsrecht beziehen kann.

Effizienz vs. Autonomie und mental health im Arbeitsalltag

HOCHSCHULRANKINGS

Im Hochschulwesen werden Zulassungen und Rankings in einigen Fällen zunehmend mithilfe von KI durchgeführt. Klare und gültige Rankings können ein nützliches Instrument sein, das, wenn es erfolgreich ist, Millionen von jungen Menschen bei dieser wichtigen Lebensentscheidung helfen könnte. Da die Daten jedoch Konzepte wie Lernen, Vertrauen oder Zufriedenheit der Studierenden nicht direkt messen können, werden Ersatzwerte (Proxies) verwendet, die mit dem Erfolg korrelieren. Das Problem besteht darin, dass die Rankings selbstverstärkend sein können. Wenn eine Hochschule in der Rangliste schlecht abschneidet, leidet ihr Ruf und die Bedingungen verschlechtern sich. Dadurch würde das Ergebnis des Algorithmus zu seinem Schicksal werden (ibid). Während wichtige Aspekte wie beispielsweise Studiengebühren in solchen Rankings nicht berücksichtigt werden, besteht eine weitere negative Auswirkung des starken Fokus auf Rankings darin, dass sich die Hochschulen vor allem auf die Verbesserung der Proxies(-Messungen)[14] und der zur Berechnung der Rankings verwendeten Metriken konzentrieren würden, was die Feedbackschleifen des Modells weiter verstärken würde.

Rankings als „self-fulfilling prophecies"

STAATLICHE TRANSFERLEISTUNGEN

Besondere Verantwortung trifft den Staat, wenn in großem Stil mittels KI die öffentliche Verwaltung effizienter gestaltet werden soll und z. B. Sozialhilfe oder andere Transferleistungen davon betroffen sind. 2016 bzw. 2020 führten die australische und auch die niederländische Regierung[15] eine Software ein, die abschätzen sollte, ob Bürger*innen zu viel Sozialleistungen bezogen haben. Der Vergleich

Fehlbeurteilungen im Sozialsystem

[14] So ist es z. B. einfach, die Abschlussquoten zu erhöhen, indem man die Standards senkt.

[15] *https://www.dutchnews.nl/news/2020/02/governments-fraud-algorithm-syri-breaks-human-rights-privacy-law/*.

von Steuererklärungen und Leistungsbezugsdaten sollte durch das System schneller vor sich gehen und damit mehr Transparenz in das System bringen. Allerdings zeigte sich, dass es zu massenhaften Fehlbeurteilungen gekommen war. Nach Recherchen des TV-Senders ABC hat die Regierung in den ersten Monaten des Einsatzes 200.000 Schreiben wegen Widersprüchen zwischen Steuererklärungen und bezogener Sozialleistungen verschickt. Bei etwa 80 Prozent dieser Fälle lautete das Ergebnis der algorithmischen Entscheidungsfindung: Menschen schulden dem Staat Geld (Rohde 2017). Die Vorteile derartiger Systeme – Schnelligkeit und Skalierbarkeit – bringen aber auch Nachteile mit sich. Die Einzelfallbeurteilung wird günstiger und schneller, weshalb auch mehr Beurteilungen durchgeführt werden können. Sind nun ein oder mehrere Fehler im System, werden auch diese beschleunigt und vermehrt. Aus Einzelfällen können so Massenphänomene werden.

Es ist aber auch nicht überraschend, dass algorithmische Systeme bei ungewöhnlichen Fällen problematische Ergebnisse liefern. Der Software fehlt oft die Flexibilität, auf relevante, aber unerwartete Details adäquat zu reagieren. Algorithmische Systeme, arbeiten eine vorgegebene Entscheidungslogik konsistent in jedem Einzelfall ab. Sie tun genau das viel zuverlässiger als Menschen. Im Gegensatz zu menschlichen Entscheider*innen ist Software nicht tagesformabhängig und wendet nicht willkürlich in Einzelfällen neue, unter Umständen ungeeignete Kriterien an. Aber wenn ein Einzelfall von typischen Mustern abweicht, kann die algorithmische Konsistenz zum Nachteil werden. Das ist gerade im Sozialsystem problematisch, wo bei ungewöhnlichen Einzelfällen oft Unterstützung am nötigsten gebraucht wird (Müller-Eiselt/Lischka 2018). Die Sensitivität von Entscheidungen gegenüber Details des Algorithmus oder ausgewerteter Daten und die daraus folgende starke Abhängigkeit von verwendeten Prozeduren, Algorithmen oder Datensätzen, könnten dies jedoch relativieren. Sie können dazu führen, dass menschliche Willkür durch maschinelle Willkür, also (scheinbare) Beliebigkeit in Entscheidungen, ersetzt wird (vgl. Helbing et al. 2021, S. 2).

keine Willkür aber fehlende Flexibilität

JUSTIZSYSTEM

Eine öffentliche Aufgabe, im Rahmen derer KI eingesetzt wird, ist das Justizsystem. In den USA wird seit dem Jahr 2012 unter anderem die Software COMPAS[16] genutzt. Mit ihrer Hilfe soll die Rückfallswahrscheinlichkeit von Straftätern beurteilt werden. Diese wurde bisher bei mehr als einer Million Straftäter angewendet. COMPAS nutzt 137 Merkmale einer Person, um auf die Wahrscheinlichkeit eines möglichen zukünftigen Verhaltens zu schließen (O'Neil 2016; Yong 2018). Obwohl Hautfarbe kein der Analyse zugrundeliegendes Merkmal ist, stellte sich bei einer Überprüfung heraus, dass die Software bei Weißen eher zugunsten der Angeklagten irrte, bei Schwarzen dagegen prognostizierte sie doppelt so oft fälschlicherweise einen Rückfall. Bei genauerer Betrachtung zeigte sich, dass es sich um unterschiedliche Verständnisse von „Fairness" handelt. So ist beispielsweise der Anteil der Rückfallstäter mit einem bestimmten Score in beiden Populationen gleich, was aus Sicht der Firma als fair zu bezeichnen wäre. Der Vorwurf (s. o.) lautete jedoch, dass unter jenen, die nicht rückfällig geworden waren doppelt so

Merkmale um zukünftiges Verhalten (falsch) zu prognostizieren

[16] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) *https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/*.

oft Schwarze mit einem höheren Risiko bewertet wurden, was von Kritiker*innen als unfair eingestuft wurde. Da die allgemeine Rückfallwahrscheinlichkeit bei People of Color (PoC) in den USA höher ist, ergibt sich bei „fairer" Berechnung ein höherer Anteil an Rückfallstäter*innen, was im Umkehrschluss zu einer Fehleinschätzung jener People of Color (PoC) führt, die keine Rückfalltäter sind (Corbett-Davies et al. 2016). Es ist mathematisch unmöglich beide Fairnesskriterien zugleich zu erfüllen – das jeweils angestrebte Kriterium müsste dementsprechend deutlich gemacht und diskutiert werden. Die Ungleichbehandlung von People of Color (PoC) durch den Algorithmus wurde nicht durch ein diskriminierendes Merkmal eingeführt, sondern durch die Beobachtung einiger anderer Merkmale, die in der statistischen Realität bei People of Color (PoC) häufiger vorkommen und aus dieser historischen Evidenz wurde auf die Zukunft geschlossen. Interessant an diesem konkreten Anwendungsfall ist auch, dass in einem Experiment herausgefunden wurde, dass die KI in der Prognose nicht besser als Zufallsnutzer*innen aus dem Internet sei (Dressel/Farid 2018). Auch wenn nur eine geringe Anzahl von Menschen von derartigen (Fehl-)Einschätzungen betroffen ist, so wird deutlich, dass die genaue Kenntnis der Algorithmen und der zugrundeliegenden Annahmen unabdingbar ist und die Zukunft eines Menschen nicht einfach einer KI überlassen werden darf.

PREDICTIVE POLICING

In den USA zum Beispiel befindet sich die Polizeiarbeit an einem kritischen Punkt. Prädiktive Programme wie PredPol, die sich auf drei Variablen stützen (Art der Straftat, Datum und Uhrzeit sowie Ort), werden zunehmend eingesetzt, insbesondere in Gebieten mit geringeren Ressourcen und Budgets (Shapiro 2017). Trotz der weiten Verbreitung befindet sich die vorausschauende Polizeiarbeit noch in einem frühen Stadium, ist anfällig für Verzerrungen und schwer zu bewerten. Täterbasierte Modelle erstellen Risikoprofile für Personen im Strafrechtssystem auf der Grundlage von Alter, Vorstrafen, beruflichem Werdegang und sozialer Zugehörigkeit. Polizeidienststellen, Gerichte oder Bewährungsausschüsse verwenden diese Profile – wie z. B. die Einschätzung der Wahrscheinlichkeit, dass eine Person in eine Schießerei verwickelt ist –, um zu entscheiden, ob die Person inhaftiert, an Sozialdienste verwiesen oder überwacht werden sollte. Durch räumliche Modellierung werden Risikoprofile für Orte erstellt und Algorithmen, die anhand von Kriminalitäts- und Umweltdaten trainiert wurden, sagen voraus, wo und wann die Polizei Streife fahren sollte, um Verbrechen aufzudecken oder zu verhindern. Die Algorithmen können jedoch durch Verzerrungen und Rückkopplungsschleifen beeinträchtigt werden, da z. B. unschuldige Menschen in der Nähe von Kriminellen ins Visier genommen werden, während Kriminelle in der Nähe von gesetzestreuen Bürger*innen verschont bleiben. Da es einen starken Zusammenhang zwischen Armut und angezeigter Kriminalität gibt, geraten die Armen immer wieder in die Fänge dieser digitalen Überwachungssysteme (O'Neil 2016; Shapiro 2017).

verfälschte Risikoprofile durch Zusammenhang von Armut und angezeigter Kriminalität

Besonders intensiv mit KI arbeiten moderne Überwachungstechnologien auf Basis biometrischer Daten. Insbesondere die Gesichtserkennung stellt ein beachtliches Gefahrenpotential dar. Den erwarteten Sicherheitsgewinnen aus flächendeckender Überwachung stehen noch Technologien mit hohen Fehlerraten und vor allem das drohende Ende der Anonymität im öffentlichen Raum – und damit eine Bedrohung grundlegender Freiheitsrechte und der Basis der Demokratie – entgegen (Schaber et al. 2020). Unzählige Beispiele belegen, dass diese Gefahren

Gesichtserkennung

nicht nur theoretischer Natur sind. Zahlreiche nationale und internationale Menschenrechtsorganisationen und Datenschutzeinrichtungen sowie der Europäische Datenschutzbeauftragte (EDSB) beurteilen Gesichtserkennungstechnologie als gefährlichen Schritt in Richtung Massenüberwachung mit enormen Gefahren für Freiheit und Menschenrechte und fordern daher seit längerem Verbote (Stolton 2020).

Neben den hoheitlichen Strafverfolgungsbehörden nutzen auch eine Vielzahl von privaten Sicherheitsfirmen KI und Gesichtserkennung. Basis der Überwachungstechnologie sind unter anderem riesige Datenbanken mit Bildern aus unterschiedlichen Quellen. Gesichtsbilder können mit geringem Aufwand und einfachen Mitteln erfasst werden, sie sind nahezu unerschöpflich auf Social-Media-Plattformen vorhanden. Neben der US-Firma Clearview AI, die hauptsächlich Behördenanwendungen bereitstellt (Hill 2020), gibt es auch in Europa ähnliche Angebote. Die polnische Suchmaschine PimEyes verfügt etwa über rund 900 Millionen Gesichtsbilder, die mit geringem Aufwand durchsuchbar sind (Dachwitz et al. 2020). Großes Schadenspotential entsteht im Bereich Biometrie und Gesichtserkennung vor allem durch einen möglichen Identitätsdiebstahl. Da biometrische Daten nicht, wie etwa Passwörter, geändert werden können, führt ein Verlust bzw. die missbräuchliche Verwendung derartiger Daten oft zu massiven Beeinträchtigungen für die Opfer. Gleichzeitig macht KI es möglich, so genannte Deep Fakes herzustellen. Dabei werden echten Bildern bzw. Videos falsche Töne und Aussagen unterlegt, die die Opfer kompromittieren können. Die Unsicherheit, ob man Videos „noch trauen kann" führt zu Verunsicherung und könnte sogar zu politischen Verwicklungen führen. Diese Technologie ist mittlerweile leicht zu beschaffen und schon recht weit fortgeschritten.

Biometrie als Basis für Identitätsdiebstahl und Deep Fakes

DIAGNOSE UND THERAPIE

Im Bereich Gesundheit wird KI bereits bei der Diagnose und für Therapieempfehlung eingesetzt, um Empfehlungen effizienter zu gestalten. Hierdurch kann eine direkte existenzielle Abhängigkeit von KI entstehen, indem sie auf unterschiedlicher Ebene in Entscheidungsprozesse eingreift. Ein aktuelles Beispiel stellt der in manchen Staaten angedachte Einsatz von Triage im Zuge einer durch die Covid-19-Pandemie bedingten Überlastung der Spitäler dar. Insbesondere bei der Auslagerung von Triage-Entscheidungen an eine KI stellen sich grundlegende ethische sowie soziale Probleme. Helbing et al. (2021) listen eine Reihe von Fragen und Problemen für die Verwendung von KI im Medizinsektor auf: fehlende Algorithmentransparenz, während denselben Deutungshoheit zugeschrieben wird; die Gefahr der (scheinbaren) Beliebigkeit durch geringe Nachvollziehbarkeit von Entscheidungen durch die Sensitivität von datengetriebenen Entscheidungen gegenüber Algorithmus-Details und ausgewerteten Daten; auf Mittelwerte gestützte, statt an den Einzelfall angepasste Maßnahmen; eine zunehmende Unterwerfung der Menschenwürde unter ökonomische Nutzenerwägungen; soziale Selektion durch individuellen Versicherungsschutz; Verstärkung von eugenischen Selektionseffekten durch Rückgriff auf Gesundheitsdaten (gesundheitliche Verfassung einer Person ist zum Teil durch Verhalten, Arbeit und Umwelt, zum Teil aber auch genetisch bedingt); Auslagerung von problematischen Entscheidungen an datengetriebene Algorithmen und Maschinen. Damit wäre der Mensch zum Objekt einer maschinellen Entscheidung geworden, was mit der Menschenwürde unvereinbar sei.

Auslagerung von Verantwortung im Gesundheitswesen

SOCIAL SORTING DURCH SCORING

Eine besondere Bedrohungskonstellation für Bürger*innen und Gesellschaft insgesamt stellt die Verbindung mehrerer bereits vorgestellter Anwendungen von KI dar. Wenn mittels Big-Data-Analysen sowohl das Kaufverhalten, die Interessensgebiete und Vorlieben, das Verhalten im Cyberspace und auch im realen Leben zusammengefügt und daraus individuelle Profile sowie soziale Kategorisierungen werden, kann sowohl die Verhaltenssteuerung als auch die Überwachung der Bürger*innen bzw. Konsument*innen freiheitsbeschränkende Ausmaße annehmen. Begonnen haben diese Praktiken noch eindimensional mit der Schuldner*innenbewertung zur Absicherung von Kreditgeschäften (Peissl/Krieger-Lamina 2017). Mittlerweile werden derartige Beurteilungen der Kreditwürdigkeit in vielen Lebensbereichen angewandt. Sie entscheiden, ob man im Internet mittels Rechnung, Kreditkarte oder nur per Vorauszahlung einkaufen kann oder ob man in der Hotline früher oder später drankommt. Sie wirken aber existenziell, wenn die KI entscheidet, ob man einen Kredit beispielsweise für eine Wohnung bekommt oder nicht. In seiner extremsten Ausformung stellt sich „Social Sorting" (Lyon 2003) in den Visionen des chinesischen Sozialkreditsystems dar. Auch wenn sich das System derzeit eher als eine heterogene Ansammlung von fragmentierten und dezentralisierten Systemen darstellt und nicht so umfassend und koordiniert ist, wie es dargestellt wird, könnten seine Logik und seine Methoden, immer mehr Informationen über Daten-Silos hinweg auszutauschen, um Verhaltensweisen zu beeinflussen, bald konkreter werden (Adelmant 2020).

Beschränkung der Lebenschance durch Datenanalyse

5 GOVERNANCE VON KI-SYSTEMEN

Wie in Kapitel 4 ausgeführt, bieten KI-Systeme vielfältige Chancen und teilweise nicht oder nur schwer vorherzusehende Risiken. Chancen werden häufig auf kohärentere und konsistentere (und damit fairere) Entscheidungen zurückgeführt, während Risiken potenzielle Ungerechtigkeit durch (unbemerkte) Fehleinschätzungen oder Skalierungseffekte in der Entscheidungsfindung und entsprechende Konsequenzen betreffen. Eine große Unsicherheit in Bezug auf algorithmische Entscheidungssysteme liegt in der Gefahr der „Verselbstständigung" von algorithmischen Entscheidungen (Schneier 2021): Ohne entsprechende Kontrollen könnten die unterschiedlichen Funktionsweisen von menschlicher und künstlicher Intelligenz dazu führen, dass KI-Systeme (bzw. Anwendungen) „unwissentlich" und unbemerkt menschliche Systeme (auch „unbeabsichtigt") unterlaufen (z. B. indem das KI-typische Optimierungsparadigma auf Systeme übertragen wird, denen andere Werte zugrunde liegen, wie z. B. auf die Gesetzgebung) (Schneier 2021). Entsprechend werden aktuell unterschiedliche Governance-, Kontroll- und Aufsichtsmechanismen entwickelt und implementiert, die unterschiedlich formalisiert und dem Politikprozess eingegliedert sind (vgl. Jobin et al. 2019).

Der sogenannte AI-Act, der im Entwurf der EU-Kommission vorliegt (Europäische Kommission 2021) ist der Versuch, eine grundlegende Rechtsstruktur für KI über die Mitgliedsstaaten hinweg zu schaffen, und verfolgt den Anspruch, international eine Art „Goldstandard" für KI-Regulierung zu etablieren.

AI-Act

Governance-Ansätze zu KI berufen sich auf unterschiedliche Arten von Kriterien, um zwischen verschiedenen KI-Systemen zu unterscheiden. Häufig wird mit Hilfe von technischen bzw. system-, anwendungsfeld- und auswirkungsbezogenen Kriterien unterschieden. Der aktuelle Entwurf des AI-Acts der EU-Kommission bildet sowohl system- als auch anwendungsbezogene Kriterien ab: Während er die technischen bzw. systembezogenen Kriterien möglichst umfassend definiert (inkl. statistischer Verfahren), beschränkt er KI in „Hochrisiko-Bereichen" auf acht Anwendungsfelder, in denen KI gesellschaftlich besonders problematische Auswirkungen zugeschrieben wird. Rein technische Kriterien zur Beurteilung von KI-Systemen sind freilich unzureichend, da deren potenzielle (oder antizipierte zukünftige) Lernfähigkeit nicht vorhersagbare Auswirkungen haben kann (Schneier 2021); auch technisch wenig komplexe (z. B. klassische statistische) Verfahren können massive Auswirkungen auf bestimmte Gruppen (Verbraucher*innen, Nutzer*innen, Betroffene, z. B. Kategorisierte) haben.

Problematik technischer Einordnung

Ein Beispiel für die Schwierigkeit, Algorithmen nach technischen Gesichtspunkten zu unterscheiden, zeigt das AMAS-System des österreichischen Arbeitsmarktservices (vgl. Kapitel 4; ITA 2021). Hierbei handelt es sich um ein statistisches Verfahren, keine KI im engeren Sinne, was zeigt, dass auch „althergebrachte" Systeme problematisch sein können. Um Ressourcen zielgerichteter zu verteilen, werden hier Biografien von Arbeitssuchenden durch einen Algorithmus auf einen scheinbar „objektiven" Wert (Integrationschancen-Wert, IC-Wert) reduziert, der auf Statistiken vergangener Jahre basiert. Er ermöglicht die Einteilung in drei Gruppen, wobei der „mittleren" am meisten Förderung zukommen soll. Daher kategorisiert AMAS Personen mit ähnlichen Eigenschaften in einer

Konstellation, die Chancenhomogenität suggeriert. Dabei fußt das System aber auf einem groben Kategoriensystem und einem reduktionistischen Verhältnis zu bestimmten Eigenschaften (z. B. körperliche Einschränkungen, unbeeinflusst vom angestrebten Tätigkeitsfeld). Die Funktion des Algorithmus besteht in einer zusätzlichen Informationsquelle zu persönlicher Beratung, dessen Einsatz aber durch bestimmte Rahmenbedingungen (beschränktes Zeitbudget der Angestellten des Arbeitsmarktservice, Argumentationspflicht bei Ablehnung der computergenerierten Empfehlung, etc.) bedingt ist. Das ITA empfiehlt in diesem Fall Transparenz und Nachvollziehbarkeit (Einsichts- und Einspruchsrechte), die Durchführung öffentlicher Konsultationen zum Thema, die Vermittlung kritischer Kompetenzen für AMS-Berater*innen, Anti-Diskriminierungsmaßnahmen in der Entwicklung, und einen Schwerpunkt (auch) auf nicht-technischen Lösungen (z. B. Betreuungsschlüssel der Berater*innen zu Arbeitssuchenden). Gleichzeitig werden Rechenschaftspflichten (System- und Datentransparenz), die Wahrung der Grundrechte, gesetzliche Grundlagen, die Einrichtung von Gremien und Aufsichtsorgane, sowie Auditing-Verfahren empfohlen (ITA 2021).

Dieses Beispiel illustriert die Wichtigkeit eines breiten KI-Begriffs auf technischer Ebene (inklusive klassischer statistischer Verfahren) um direkte Auswirkungen auf Leben und sozioökonomische Situation von Menschen zu minimieren, während Unterscheidungskriterien nach Anwendungsbereichen und Auswirkungen prioritär zu sehen sind.

breiter KI-Begriff notwendig

5.1 ETHIK-CODES, TOOLKITS, CHECKLISTEN – GRUNDLEGENDES UND BEISPIELE

Um KI bzw. Algorithmen und algorithmische Entscheidungssysteme sicher, ethisch und transparent zu gestalten, findet sich in der Literatur eine ganze Bandbreite an unterschiedlichen Gütekriterien, Ethik-Codes, Toolkits oder Checklisten, die jeweils unterschiedliche Schwerpunktsetzungen aufweisen.

So argumentieren Zuber et al. (2020) beispielsweise, dass gegenwärtige Softwaresysteme keine Abwägungen zwischen Begründungen vornehmen, nicht unabhängig zwischen verschiedenen Denkmodi wechseln und ihre Handlungen nicht begründen können. Sie können daher nicht als autonome moralische Akteure auftreten, sondern lediglich ihre Programmierung umsetzen (ibid., S. 157), da der aktuelle Forschungsstand keine „reflexiven" und „theoriebildenden" Algorithmen erlaubt. Daher liegt es an den Entwickler*innen, darüber zu reflektieren (und letztendlich zu entscheiden), ob Beiträge mit ihrer professionellen Ethik übereinstimmen (ibid., S. 158). Allerdings kann die Vielfalt benötigter Expertise für die Entwicklung datenbasierter Technologien in unterschiedlichen wichtigen Arbeitsprozessen zu einer breiten Streuung von Verantwortung und Verantwortlichkeiten führen (Siehe dazu auch Abbildung 7) wobei User*innen miteingeschlossen sind (z. B. durch missverständliche Anwendung). Existierende Verhaltenskodizes (Codes of Conducts) basieren auf individuell ausdifferenzierten Werten, die jeweils unterschiedlich im Zusammenhang mit ethisch wünschenswertem Design und nach Anwendungsfeld priorisiert werden sollten. Zuber et al. (2020) betonen damit, wie technische Artefakte in unterschiedlichen Kontexten jeweils

zu Verzerrungen oder begründeten Konflikten zwischen unterschiedlichen Zielsetzungen führen können (ibid, S. 161).

So verglich die Bertelsmann-Stiftung (Rohde 2018) Stärken und Schwächen von Gütekriterien-Katalogen dreier Institutionen für Algorithmen (ACM US Public Policy Council, FAT/ML Organisation, Future of Life Institute) und analysiert diese, um Schlussfolgerungen für allgemeine Gütekriterien zu ziehen. Dabei werden zwei Aspekte als Prämissen festgesetzt: dass die ausschließliche Verantwortung über algorithmische Prozesse bei menschlichen Akteure*innen bleibt um einen aktiven Kontrapunkt zu Szenarien zu setzen, in denen Menschen Maschinen machtlos ausgeliefert sind; und, dass die gesellschaftliche Bedeutung von algorithmischen Prozessen skizziert wird um die Teilhaberelevanz für die Rechtfertigung für Gestaltungskodizes hervorzuheben.

Vergleich unterschiedlicher Kriterienkataloge

Stärken der ausgewählten Dokumente liegen vor allem in der Betonung der Vielschichtigkeit des algorithmischen Gestaltungsprozesses (Programmierung, vor- und nachgelagerte Schritte), die Berücksichtigung der Datengrundlage als Fehlerquelle (und damit die Wichtigkeit der Qualität der Daten), die Größe des Adressat*innenkreises (abhängig von der dargestellten Vielschichtigkeit des Prozesses), der Programmierer*innen, Politiker*innen, Auftraggeber*innen, Institutionen und z. T. Anwender*innen anspricht. Zusätzlich werden Forschung und die Öffentlichkeit besonders hervorgehoben, da diese eine Sonderrolle einnehmen (technische Machbarkeit zukünftiger Lösungen (Future of Life Institute), Förderung eines öffentlichen Diskurses (ACM US Public Policy Council, FAT/ML Organisation). Zusätzlich sollen Social-Impact-Statements (FAT/ML Principles) dafür sorgen, dass Konkurrenz zwischen Anbieter*innen nicht zu Lasten der Sicherheit ausgelebt wird. Die Einbindung der Öffentlichkeit in die Gestaltung berge außerdem den Vorteil, dass Fehler im System früher erkannt würden (sofern die Einbindung nicht erst im Produktiveinsatz erfolgt). Gleichzeitig sprechen alle Dokumente Konflikte zwischen technischen, ökonomischen und sozialen Interessen an, wobei häufig das Abwägen zwischen vollkommener Transparenz eines Systems und dem Schutz eines Daten- und Industriegeheimnisses genannt wird (z. B. FAT/ML principles). In Bezug auf ethische Fragen weist vor allem das Future of Life Institute darauf hin, dass komparative Konzepte wie Fairness durch weitere ergänzt werden sollten (z. B. Einhaltung der Menschenrechte, Würde, Rechte, Freiheiten und kulturelle Diversität) bzw. „*Metafragen zur Einstellung zu künstlicher Intelligenz bzw. deren grundsätzliche Ziele*“ (Rohde 2018, S. 21) Beachtung finden sollen. Weiters spricht sich das Dokument für ein detailliertes Verständnis einzelner Gütekriterien (z. B. Verantwortlichkeit/Zurechenbarkeit) aus, um die Verständlichkeit und Implementierbarkeit zu erhöhen (z. B. die Unterscheidbarkeit zwischen juristischer und Fehlertransparenz in den Asilomar-Prinzipien des Future of Life Instituts (2017)). Insgesamt werden zwei Strategien für die Erstellung solcher und ähnlicher Gütekriterien vorgeschlagen: ein gesetzesähnlicher Forderungskatalog (idealistisch ausgerichtet) oder aber eine Sammlung konkreter Schritte.

Stärken ...

Schwächen der Dokumente werden in fehlenden Anhaltspunkten zur praktischen Implementierung gesehen sowie in der vernachlässigten Rolle der Politik. Damit im Zusammenhang steht die Frage der Verbindlichkeit der Prinzipien und in weiterer Folge auch der Verbote. Die Dokumente (genauso wie der weitere Diskurs) spart diese tendenziell aus. Entsprechend empfiehlt die Bertelsmann-Stiftung (Beining 2019) die Möglichkeit eines Verbots, wo die ethische Angemessenheit eines Softwaresystems nicht gewährleistet werden kann.

... und Schwächen

Um die Nachvollziehbarkeit von Algorithmen für Betroffene zu stärken, definiert die Bertelsmann-Stiftung fünf zentrale Dimensionen (Beining 2019, S. 11f):

- Kommt ein algorithmisches System zum Einsatz?
- Wie funktioniert das algorithmische System? (*Systemtransparenz*: verwendete Daten und Trainingsdaten, Festlegung und Gewichtung von Entscheidungskriterien und Grenzen des Systems)
- Warum gibt es das algorithmische System? (*Kontexttransparenz*, z. B. zugrundeliegende Wertentscheidungen, erwartete Wirkungen, Stellenwert des Systems im Entscheidungsprozess)
- Wie kommt die konkrete Entscheidung über die jeweilige Person zustande? (*Entscheidungstransparenz*, Erklärung des Ergebnisses)
- Über welche Handlungsoptionen verfügt die jeweilige Person? (z. B. Beschwerde einlegen, Korrektur verlangen)

Herausforderungen bei der Herstellung von Nachvollziehbarkeit und Transparenz liegen in der Komplexität der Systeme (z. B. maschinelles Lernen) oder im Zielkonflikt zwischen verständlichen Modellen und weniger akkuraten Ergebnissen sowie Geschäftsgeheimnissen oder Regulierungen (z. B. Datenschutzrecht) und fehlenden offensichtlichen Anreizen, diese tatsächlich zu erhöhen (negatives Kosten-Nutzen-Verhältnis für die einsetzende Stelle). Weiters zeigt sich die Ambivalenz von mehr Wissen im Bereich der öffentlichen Wahrnehmung: Eine Erhöhung von Transparenz führt nicht zwingend zu mehr Vertrauen in die Technologie. In Bezug auf Herausforderungen im sozialen Kontext verweist die Studie einerseits auf den sogenannten Automation-Bias, also die Tendenz, Ergebnissen von ADM-Systemen blind zu vertrauen, und auf mögliche Auswirkungen auf menschliches Verhalten, z. B. Anpassungen an ADM-System-Logiken. Konkrete Ausgestaltung von Transparenz in der Praxis wird in drei teilhaberelevanten Bereichen (Recruiting, Gesundheitswesen und Polizeiarbeit) diskutiert.

Auf Basis dieser Diskussion zieht die Bertelsmann-Stiftung die Schlussfolgerung nach anwendungsbereichsspezifischen Anforderungen an Transparenz und Nachvollziehbarkeit, die anhand der folgenden Checkliste jeweils spezifiziert werden sollen (Beining 2019, S. 28f):

- Wie hoch ist die Teilhaberelevanz? (höhere Teilhaberelevanz = höherer Anspruch an Transparenz und Nachvollziehbarkeit)
- Was soll mit Transparenz erreicht werden? (Funktionen und Ziele)
- Welchen Kontakt hat die/der Betroffene zum algorithmischen System?
- Wird eine Entscheidung über eine ganze Gruppe getroffen oder spezifisch über ein Individuum?
- Wie viel Zeit, Wissen und Bereitschaft bringen Betroffene mit, um sich mit dem ADM-Prozess auseinanderzusetzen? (Bedürfnisse der Adressat*innen)
- Welche regulatorischen Anforderungen gibt es? (DSGVO, etc.)

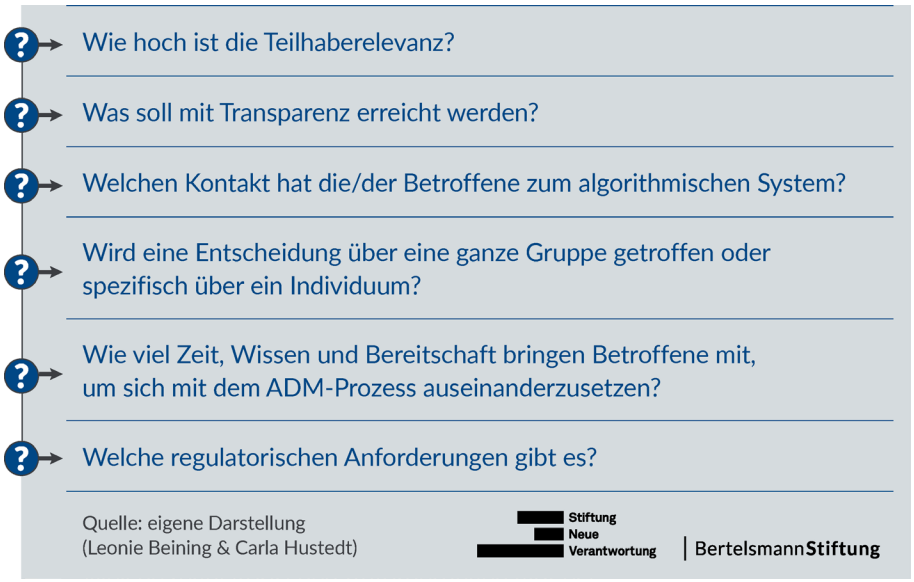

Abbildung 8: Checkliste für mehr Nachvollziehbarkeit von algorithmischen Systemen (Quelle: Beining 2019, S. 30)

Weiters entwickelte die Bertelsmann-Stiftung gemeinsam mit dem iRights.Lab auch einen dynamisch angelegten Regelkatalog, die sogenannten Algo.Rules, um algorithmische Ethik zu gewährleisten. Die Algo.Rules wurden in einem partizipativen und interdisziplinären Prozess mit (bisher) ca. 500 Beteiligten erarbeitet. Sie umfassen mehrere Prinzipien, nach denen Algorithmen gestaltet werden sollen, um wünschenswerte KI zu ermöglichen. Dabei soll (Algo.rules. 2019):

Algo.Rules

- Kompetenz über die Funktionsweise und mögliche Auswirkungen eines Algorithmus aufgebaut,
- Verantwortung durch eine natürliche oder juristische Person definiert,
- Ziele und erwartete Wirkung dokumentiert,
- Sicherheit gewährleistet,
- Kennzeichnung als algorithmisches System durchgeführt,
- Nachvollziehbarkeit über die dem System zugrundeliegenden Daten und Modelle hergestellt,
- seine Architektur sowie die möglichen Auswirkungen veröffentlicht und leichte Verständlichkeit sichergestellt (sowie etwaige Substitution durch weniger komplexe Algorithmen geklärt),
- Beherrschbarkeit und Kontrolle durch Menschen abgesichert,
- Wirkung durch eine externe Prüfstelle unter Wahrung legitimer Geschäftsgeheimnisse überprüft und
- gegebenenfalls Beschwerden (z. B. von Betroffenen) ermöglicht werden.

Zusätzlich zum Wissen über und Ansprüchen an den konkreten Einsatz von bestimmten Algorithmen, empfiehlt die Literatur auch, darauf zu achten, wie entsprechende Dokumente zu kontextualisieren sind (vgl. Jobin 2020). So formuliert Jobin (2020) (in Anlehnung an Mittelstadt (2019)) nicht nur Empfehlungen zu Algorithmen selbst, sondern auch Fragen, die an die Leitfäden- und Prinzipien-Dokumente selbst gestellt werden sollen, um sie besser zu kontextualisieren (Jobin 2020):

1. Wer schrieb diese Angaben und wie?
2. An wen richten sie sich, und was ist ihr Zweck?
3. Warum soll ich ihnen folgen?
4. Wie kann ich ihnen folgen oder sie umsetzen?
5. Wie soll ich widersprüchliche Auslegungen von grundsätzlich kontroversen Konzepten klären?
6. Wie kann festgestellt werden, ob ich mich an die Angaben halte?
7. Was geschieht, wenn ich mich nicht daran halte?
8. Wie kann ich Einwände oder Fragen nach Klarstellungen ansprechen?

Ansprüche an Leitfäden und Prinzipiendokumente

Daher wird die Wichtigkeit der Rolle, die Kontextfaktoren in Zusammenhang mit einer ethischen Akzeptabilität von KI einnehmen, hervorgehoben (Jobin et al. 2019, S. 144).

5.1.1 UMSETZUNG VON KI-ETHIK

Anwendungen von KI-Systemen durchdringen zunehmend alle sozioökonomischen Sektoren, und ein breites Spektrum von Stakeholdern hat bereits verschiedene Leitlinien und Initiativen für die KI-Ethik formuliert, die auf Unternehmensebene eingesetzt werden sollen. In den letzten Jahren ist eine zunehmende Vielfalt an Grundsätzen und Richtlinien für den Einsatz von KI-Technologien entstanden (z. B. Jobin et al. 2019). Forscher*innen, Organisationen des öffentlichen Sektors und private Unternehmen schlagen unterschiedliche Anforderungen und Standards dafür vor, was „ethische KI" ausmacht. Aktuelle Studien – wie der Vorschlag der AI Ethics Impact Group (2020) für ein Ethik-Label für KI – zeigen jedoch, dass sich ein breiter globaler Konsens über Es haben sich fünf ethische Grundsätze herauskristallisiert, wobei es erhebliche Unterschiede in der Art und Weise ihrer Umsetzung gibt. Die fünf Prinzipien sind Transparenz, Gerechtigkeit und Fairness, Nicht-Schädigung (non-maleficence), Verantwortung und Datenschutz (Jobin et al. 2019, S. 389).

5 Prinzipien:
- Transparenz
- Gerechtigkeit
- Nicht-Schädigung
- Verantwortung
- Datenschutz

Die bestehende KI-Ethikforschung reicht von Metastudien (Boddington 2017; Goldsmith/Burton 2017; Prates et al. 2018; Vakkuri/Abrahamsson 2018; Greene et al. 2019; Hagendorff 2020) bis hin zu Algorithmen-Entwicklung und Entscheidungsroutinen von autonomen Systemen (z. B. Anderson et al. 2016; Etzioni/Etzioni 2017; Yu et al. 2018), sowie umfassenden Richtlinien für KI (z. B. Chatila/Havens 2019). Floridi/Cowls (2021) schlagen einen einheitlichen Rahmen vor, in dem die Erklärbarkeit mit den oben genannten Prinzipien kombiniert wird, um einige der technischen Aspekte der Rechenschaftspflicht beim Einsatz von KI abzudecken.

In diesem Spannungsfeld müssten einzelne Unternehmen, die KI-Systeme und Entscheidungsalgorithmen entwickeln, ethische Leitlinien selbst anpassen, entwickeln und umsetzen. Mindestens drei Kategorien von Herausforderungen sind für Unternehmen bei der Entwicklung und Umsetzung praktischer KI-Ethikrahmen von Bedeutung: (1) der Kontext (KI-Anwendungen sind kultur- und gesellschaftsabhängig), (2) die soziale Einbettung (die sozialen Auswirkungen hängen von der Entwicklung und Anwendung ab, die sowohl von den Entwickler*innen als auch von den Nutzer*innen gesteuert werden) und (3) die Flexibilität der Stakeholder (KI-Ethik-Frameworks müssen verschiedene Akteure und die Rollen von Entwickler*innen und Nutzer*innen gleichermaßen berücksichtigen) (AI Ethics Impact Group 2020).

Governance-Ansätze, die sich auf soziotechnische Gesamtsysteme beziehen, werden zum Teil in der Literatur auch durch – auf den Algorithmus bezogene (soziale) – Qualitätskriterien ergänzt. So plädiert das AI Ethics Impact Group (2020) für ein zusätzliches Labelsystem, das die dahinterliegenden Werte bzw. Anforderungen jeweils sichtbar machen soll. In einer aktuellen Initiative wird versucht, jede dieser Herausforderungen auf Unternehmensebene anzugehen. Die AI Ethics Impact Group hat dazu ein umfassendes System zur Integration von „Ethik in KI" entwickelt (ibid.). Das durch die Verknüpfung einer Systemperspektive (spezifische ethische Anforderungen an KI) mit einer Prozessperspektive (Design- und Implementierungsprozesse) gekennzeichnet ist. Das Ethik-Label orientiert sich an dem Kennzeichnungssystem für Energieeffizienz (AbbilAbbildung 9).

Ethik-Label

Ein entsprechendes Checklisten- und Kennzeichnungssystem adressiert verschiedene Interessengruppen. Unternehmen, die KI-Systeme einsetzen und implementieren, können Risikostufen bestimmen und Bewertungsrunden entwickeln. Entwickler*innen von KI-Systemen können das Kennzeichnungssystem und die Risikomatrix im Prinzip auch nutzen, um das Spektrum der zu erwartenden Anwendungen zu überprüfen und zu entscheiden, die Implementierung auf Bereiche mit geringen ethischen Risiken zu beschränken (ibid., S. 41). Schließlich können Regulierungsbehörden solche Systeme grundsätzlich auch nutzen, um Anforderungen für den Kontext von KI-Anwendungen und -bereichen zu stellen.

Abbildung 9: KI-Ethik-Label System (Verwendungsbeispiel)
(Quelle: nach AI Ethics Impact Group 2020)

Die Entwicklung eines umfassenden und flexiblen Bewertungsinstruments für KI-Unternehmen bleibt jedoch eine Herausforderung. Solche Labelling-Systeme erscheinen zwar unkompliziert und sind von den Unternehmen selbst einfach zu handhaben, verfügen aber bei weitem nicht über die notwendige Flexibilität, um die riesigen KI-Anwendungsbereiche abzudecken. In der bestehenden Literatur gibt es auch starke Kritik an solchen Checklisten- und Labelling-Initiativen. Im Wesentlichen, bezieht sich die Kritik auf das Fehlen jeglicher Durchsetzungsmechanismen, die keine unmittelbare Änderung des Status quo bewirken (z. B. Hagendorff 2020).

Vor allem, wenn es sich um industriegeführte Richtlinien handelt, müssten diese im Rahmen breiterer Initiativen entwickelt werden, die auch staatliche und demokratische Institutionen einbeziehen. Wenn dieser inklusive Ansatz nicht verfolgt wird, warnen Forscher*innen davor, dass KI-Ethik lediglich dazu dient, kritische öffentliche Diskurse zu beruhigen und gleichzeitig den Status quo der Industrie zu stärken (Hagendorff 2020). In einer Metastudie wird als prominentes Beispiel der Verein „Partnership for AI" genannt (Heer 2018; Hagendorff 2020), dem die Spitzenindustrie angehört (z. B. Amazon, Apple, Baidu, Facebook, Google, IBM). Unternehmen würden mitunter ihre Mitgliedschaft in solchen Verbänden immer dann betonen, wenn es darum geht, den Eindruck zu erwecken, dass sie sich ernsthaft für eine gesetzliche Regulierung ihrer Geschäftstätigkeit einsetzen (Hagendorff 2020, S. 100). Nicht zuletzt muss die Schnittstelle zwischen Ethik- und Technikdiskursen weiter überbrückt werden, indem soziale, öffentliche und persönliche Aspekte in einem öffentlichen und integrativen Rahmen stärker in den Mittelpunkt gestellt werden.

5.1.2 SEKTORALE GOVERNANCE: KI IN DER VERWALTUNG

Neben universalen Governance-Ansätzen beschäftigt sich die Literatur auch mit sektoralen Governance-Lösungen zu KI-Systemen, wie z. B. in der Verwaltung. Der Verwaltung als Repräsentantin des Staates und direktes Gegenüber der Bürger*innen kommt hierbei besondere Verantwortung zu. So stellt die Oxford Commission on AI & Good Governance (2021) im Bereich öffentlicher Verwaltung drei übergeordnete Fragen in den Raum, die im Zusammenhang von KI und öffentlicher Verwaltung grundlegend diskutiert werden sollten: Wer soll KI in der öffentlichen Verwaltung anleiten? Wie können Kapazitäten für Good Governance für KI in der öffentlichen Verwaltung aufgebaut werden? Wie können wir garantieren, dass KI in der öffentlichen Verwaltung vertrauenswürdig ist und Vertrauen bekommt? Laut Oxford Commission (2021) sind Bedingungen für den Einsatz von KI im Bereich der öffentlichen Verwaltung inklusive Gestaltung die Anleitung durch eine informierte öffentliche Stelle und zweckdienlicher und durchgängig verantwortlich gestalteten Einsatz inkl. Monitoring (Oxford Commission on AI & Good Governance 2021, S. 7). Um dies zu erreichen, empfiehlt die Kommission drei umfassende Maßnahmen (ibid., S. 1):

1. die Einrichtung eines wissenschaftlichen Gremiums gemeinsam mit einem Schiedsgericht,
2. den entsprechenden Kapazitätsaufbau innerhalb von Organisationen, um sich mit Design, Einkauf, Implementierung und Verantwortlichkeit von KI in der öffentlichen Verwaltung informiert auseinander zu setzen bzw. entsprechende Instrumente für Mitarbeiter*innen zur Verfügung zu stellen, und
3. Förderung entsprechender Bildung der Öffentlichkeit zu (absehbaren) Verwendungen, deren Auswirkungen und Risiken sowie Information über den beabsichtigten Einsatz von KI-Technologien und die Entwicklung eines grundlegenden Zertifizierungssystems durch eine multisektorale Behörde. Dazu sollen Machbarkeitsstudien zu den beiden intendierten Gremien, Konsultationsprozesse mit bereits existierenden nationalen und multilateralen Behörden, sowie „Stakeholder Engagement" durchgeführt werden.

Besonders der Vorschlag der Kommission, ein Beratungssystem mit zwei Kammern einzusetzen, ist aus Sicht der Technikfolgenabschätzung interessant. Ein Gremium soll durch wissenschaftliche und technische Diskussionen (ähnlich dem IPCC) Evidenz für ökonomische, kulturelle und politische Ergebnisse von technischen Entscheidungen herstellen. Dieses multidisziplinäre Projekt soll dabei helfen, die Interaktion zwischen KI und sozialen Systemen zu verstehen: Es soll die politischen Wege von Mitgliedsstaaten evaluieren, Auswirkungen auf soziale Gleichheit bewerten und Zertifizierungen und Standards entwickeln. Der IPCC dient dabei als Vorbild, wie wissenschaftliches Lernen koordiniert werden kann (ibid., S. 9). Daneben soll ein unabhängiges Schiedsgericht dazu dienen, Konflikte und Meinungsverschiedenheiten zwischen Industrie, zivilen Stakeholdern und staatlichen Akteuren zu lösen. Dafür werden ein leistungsfähiges Sekretariat, unabhängige Finanzierung und die Möglichkeit, fallbezogene Mediation in Anspruch nehmen zu können, sowie die Institutionalisierung an einer politisch neutralen Institution als Bedingung gesehen (ibid.). Die Berücksichtigung weiterer (zusätzlicher) Expertise soll punktuell garantiert werden.

etablierte Gremien als Vorbild

und

Stakeholder Beteiligung

Andere Ansätze fokussieren in Ergänzung dazu auf konkrete Schritte, wie die Verwendung von KI in der öffentlichen Verwaltung adäquat in bereits existierende Abläufe und Mechanismen eingebettet werden kann (Wirtz et al. 2020). Der hierzu entwickelte Rahmen basiert auf zwei unterschiedlichen Vorschlägen zu Governance-Frameworks und verbindet ethische Kriterien und Prinzipien auf unterschiedlichen Ebenen mit einem „Society-in-the-loop"-Ansatz.

Society-in-the-loop

Diese beiden Ansätze zeigen exemplarisch, wie breit selbst innerhalb der Verwaltung Ansätze gestreut sein können und welche Herausforderungen sich jeweils zeigen. Wie in allen Bereichen, in denen der Einsatz von KI angedacht wird, ist auch hier eine Einbindung von Stakeholdern unerlässlich, wie sie sich auch in umfassenden Governance-Frameworks findet (vgl. Kapitel 5.2). In Deutschland wurde Ende 2021 bereits ein entsprechender erster Schritt in Richtung einer Koordinationsstelle in die Wege geleitet. Um die Interessen der Verbraucher*innen zu stärken, verkündete das deutsche Bundesministerium der Justiz und für Verbraucherschutz (BMJV) am 20. Oktober 2021 die Einrichtung eines „Zentrums für vertrauenswürdige Künstliche Intelligenz" (Förderhöhe: 4,5 Mio. Euro bis Ende 2023), das Akteur*innen aus Wissenschaft, Wirtschaft, Politik und Zivilgesellschaft zusammenbringen und über verbraucher*innenrelevante Aspekte von KI informieren, bundesweite Informationskampagnen, ein Zertifizierungsschema für vertrauenswürdige KI und Handlungsempfehlungen für die Politik entwickeln und entsprechende wissenschaftliche Studien initiieren und durchführen soll. (Bundesministerium der Justiz 2021). Eine besondere Einladung zur Teilnahme erfolgte in der Presseaussendung an zivilgesellschaftliche Organisationen (Bundesministerium der Justiz 2021), was deren Rolle in der Sichtbarmachung von Fehlentwicklungen im Bereich KI betont (vgl. Jobin et al. 2019).

Zentrum für vertrauenswürdige Künstliche Intelligenz

5.2 EUROPÄISCHE ANSÄTZE

5.2.1 DER RISIKOBASIERTE GOVERNANCE ANSATZ DER EUROPÄISCHEN KOMMISSION

In Bezug auf konkrete, in der Umsetzung (oder kurz vor der Umsetzung) befindlichen Governance-Ansätze ist der Entwurf des AI-Acts der EU-Kommission

vom April 2021 zentral. Die EU-Kommission propagiert hier einen risiko-basierten Ansatz zur Governance von KI[17], bei dem Anwendungen entsprechend bestimmten Kriterien in vier unterschiedliche Risikoklassen eingeteilt werden:

1. *Unzulässige* KI-Systeme umfassen „alles, was als eindeutige Bedrohung für EU-Bürger angesehen wird […] von der behördlichen Bewertung des sozialen Verhaltens (Social Scoring) bis hin zu Spielzeug mit Sprachassistent, das Kinder zu riskantem Verhalten verleitet."
2. Hohes Risiko bergen KI-Systeme in den Bereichen
   - „Kritische Infrastrukturen (z. B. im Verkehr), in denen das Leben und die Gesundheit der Bürger gefährdet werden könnten
   - Schul- oder Berufsausbildung, wenn der Zugang einer Person zur Bildung und zum Berufsleben beeinträchtigt werden könnte (z. B. Bewertung von Prüfungen)
   - Sicherheitskomponenten von Produkten (z. B. eine KI-Anwendung für die roboterassistierte Chirurgie)
   - Beschäftigung, Personalmanagement und Zugang zu selbstständiger Tätigkeit (z. B. Software zur Auswertung von Lebensläufen für Einstellungsverfahren)
   - Zentrale private und öffentliche Dienstleistungen (z. B. Bewertung der Kreditwürdigkeit, wodurch Bürgern Darlehen verwehrt werden)
   - Strafverfolgung, die in die Grundrechte der Menschen eingreifen könnte (z. B. Überprüfung der Echtheit von Beweismitteln)
   - Migration, Asyl und Grenzkontrolle (z. B. Überprüfung der Echtheit von Reisedokumenten)
   - Rechtspflege und demokratische Prozesse (z. B. Anwendung der Rechtsvorschriften auf konkrete Sachverhalte)"
3. *KI-Systeme mit begrenzten Risiken* benötigen minimale Transparenzverpflichtungen (z. B. Hinweise auf Chatbots für Nutzer*innen).
4. *KI-Systeme mit minimalen Risiken* dürfen frei verwendet werden (z. B. KI-gestützte Videospiele oder Spamfilter). Gegenwärtig fallen die meisten KI-Anwendungen innerhalb der EU in diese Kategorie.

risikobasierter Ansatz im AI-Act

[17] *https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_de*.

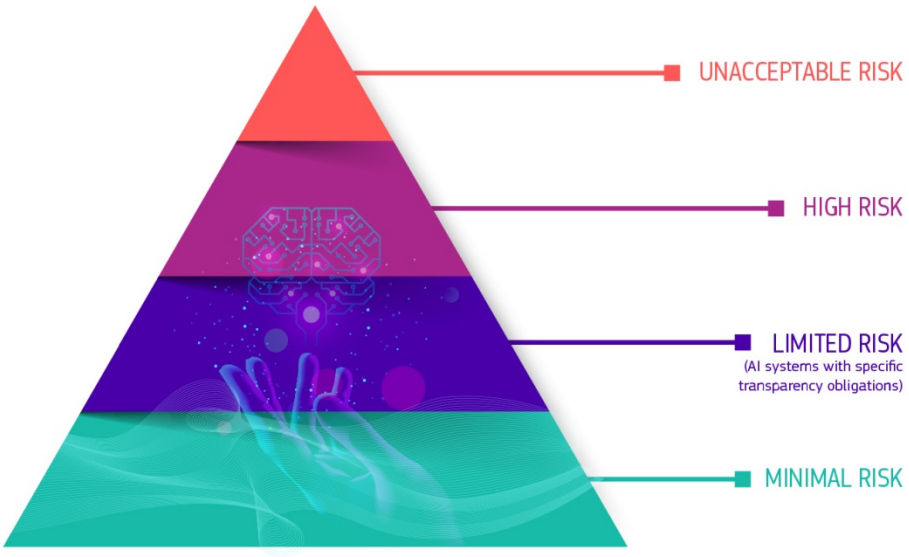

Abbildung 10: Der risikobasierte Ansatz der Europäischen Kommission
(Quelle: *https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_de*)

Besondere Aufmerksamkeit bekommen im Entwurf Hochrisiko-KI-Systeme, die bestimmte Voraussetzungen erfüllen müssen, bevor sie auf den Markt gebracht werden dürfen (Übersetzung durch Autor*innen[18])

Anforderungen an Hochrisiko-KI-Systeme

- Adäquate Risikoabschätzung und Schadensminderungssysteme
- Hohe Qualität von Datensets, die das System füttern um Risiken und diskriminierende Ergebnisse zu minimieren
- Protokollierung von Aktivitäten, um die Nachvollziehbarkeit von Ergebnissen zu ermöglichen
- Detaillierte Dokumentation über alle notwendigen Informationen über das System und seine Zwecke für Behörden, um seine Übereinstimmung abzuschätzen
- Klare und adäquate Informationen für Nutzer*innen
- Angemessene menschliche Überblicksmaßnahmen, um Risiken zu minimieren
- Hohe Levels von Robustheit, Sicherheit und Präzision

Vor allem biometrische Fern-Identifikationssysteme gelten als Hochrisiko-Anwendungen und werden streng reguliert. Ihre Live-Verwendung in öffentlich zugänglichen Plätzen zur Strafverfolgung ist aus Prinzip verboten – Ausnahmen werden genau definiert (z. B. zur Suche von verlorenen Kindern, zur Vermeidung von Terroranschlägen oder um Verdächtige zu verfolgen) und geografisch, zeitlich und in Bezug auf die durchsuchten Daten beschränkt (übersetzt durch Autor*innen[19]).

[18] *https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.*
[19] ibid

5.2.2 GOVERNANCE-ANSATZ AUS VERBRAUCHER*INNEN-PERSPEKTIVE

Kategorisierung nach Schadenspotential und Re-Evaluierungs-möglichkeit

Während der AI-Act-Entwurf der EU-Kommission den ersten europaweit umfassenden Versuch einer KI-Regulierung darstellt, gibt es daneben weitere Governance-Ansätze. Einer, der besonders auf die Perspektive und den Schutz von Verbraucher*innen abzielt, ist jener von Krafft/Zweig (2019). Dieser Ansatz konzeptualisiert KI als sozioinformatisches System und legt so besondere Betonung auf den sozialen Kontext, in dem bestimmte algorithmische Entscheidungssysteme eingesetzt werden und der die Regulierungstiefe (mit-)bestimmt. Kraft und Zweig (2019) schlagen eine Kategorisierung von algorithmischen Entscheidungssystemen anhand zweier Dimensionen vor: *„dem möglichen Schadenspotenzial auf individuellem und gesamtgesellschaftlichem Level“* und *„nach der Möglichkeit der Re-Evaluation bei einer Fehlbeurteilung“* (S. 18) (vgl. Abbildung 11).

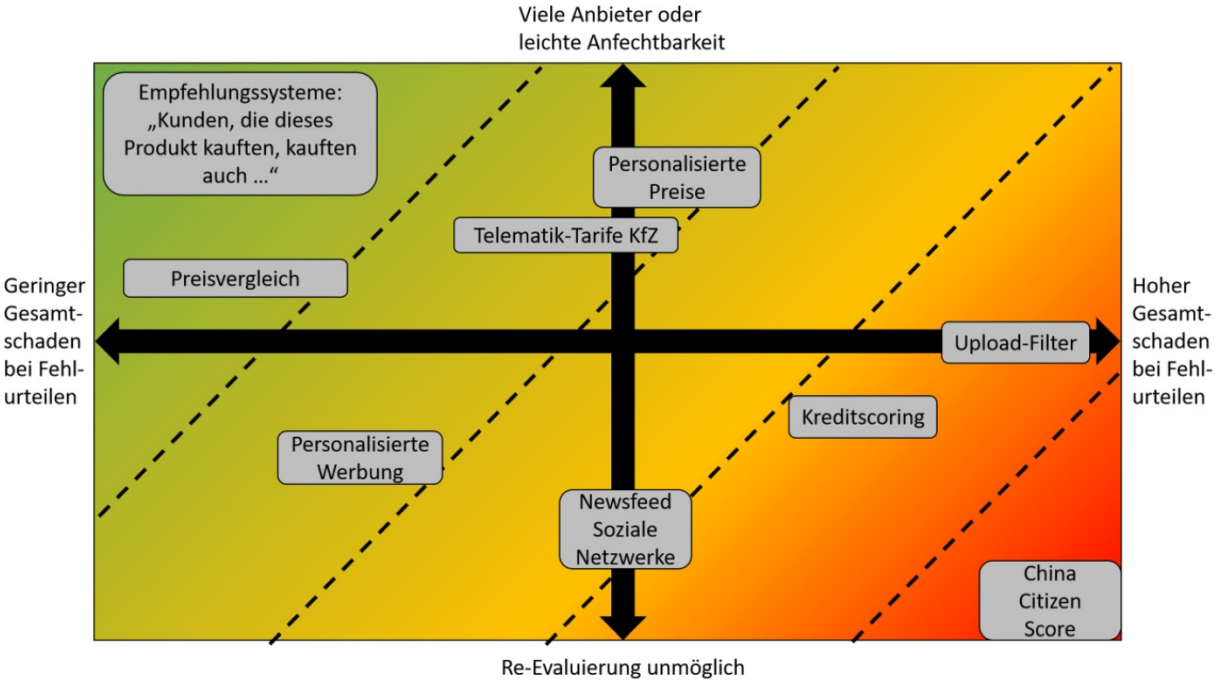

Abbildung 11: Risikomatrix um Anwendungsszenarien zu verorten (Quelle: Krafft/Zweig 2019)

Das Schema teilt KI-Systeme in fünf Regulierungsklassen ein, wobei jede höhere jeweils Nachvollziehbarkeitsforderungen niedrigerer Regulierungsklassen auf technischer Ebene beinhaltet. Die Kriterien, nach denen sich die einzelnen Risikoklassen konstituieren, wurden durch Diskussionen mit Expert*innen konsolidiert und erwiesen sich als belastbarer Versuch einer solchen Einteilung. Eine empirische Validierung der Matrix in einem Multi-Stakeholder-Verfahren wird von den Autor*innen angeregt, steht aber gegenwärtig noch aus (Krafft/Zweig 2019, S. 42).

Klasse 0

Klasse 0 fordert weder Transparenzpflichten noch dauerhafte Kontrollprozesse; gegebenenfalls werde eine Post-hoc-Analyse durchgeführt und die Risikobewertung muss unter Umständen wiederholt werden.

Klasse 1 fordert eine *Schnittstelle* zur ständigen Überwachung des Systems als Black-Box-Analyse (vgl. Diakopoulos 2015), mithilfe derer das System auf auffällige Effekte überprüft werden kann (z. B. mögliche Diskriminierung von einzelnen Personengruppen) sowie die Beschreibung der *Einbettung* des algorithmischen Entscheidungssystems in den sozialen Entscheidungsprozess. Daher fordern Kraft und Zweig (2019) für Klasse-1-Systeme die Erfüllung folgender Transparenzpflichten: (a) Nennung der Qualitätsmaße; (b) statistische Auswertung; (c) Wahl des Optimierungskriteriums; (d) Einblicke in die vom Betreiber gewählten Ziele des algorithmischen Entscheidungssystems, wobei die Angabe des Qualitätsmaßes und des Verfahrens Auskunft über die Angemessenheit der Qualitätsgüte (auch die erreichte Güte für den Anwendungsfall) gibt; (e) Nennung des Lernverfahrens; (f) Beschreibung der Einbettung des algorithmischen Entscheidungssystems in den sozialen Entscheidungsprozess; und (g) notwendige Interfaces für Blackbox-Analysen, die die Fütterung des Systems mit variablen Daten ermöglichen und diese mit den gelieferten Ergebnissen des Entscheidungssystems auswerten können.

Klasse 1

In **Klasse 2** müssen die *Eingangsdaten* vollständig beschrieben werden und die Angaben zur Qualität des Entscheidungssystems müssen überprüfbar sein, wobei das Zielpublikum jeweils zu bestimmen ist. Die Ziele des algorithmischen Entscheidungssystems müssen überprüft und klar kommuniziert werden ebenso wie die verwendeten *Trainings- und Eingabedaten,* um die Nachvollziehbarkeit der Ergebnisse zu gewährleisten. Kraft und Zweig (2019) verlangen hier, „*die verwendeten Eigenschaften eines zu bewertenden Objekts zu nennen und zu kommunizieren, woher die tatsächlichen Instanzen dieser Daten kommen. Zusätzlich kann auf offensichtliche Proxy-Variablen hin überprüft werden, die eine fragwürdige Eigenschaft, nach der sich die Entscheidung nicht richten sollte, abbilden*“ (Krafft/Zweig 2019, S. 38). Auch müssen für die Nachvollziehbarkeit der Qualitätsbewertung Ergebnisse von den Betreibern eines algorithmischen Entscheidungssystems so bereitgestellt werden, „*dass überwachende Instanzen die angegebenen Qualitätsmaße berechnen*“ können, und der „*gesamte Berechnungs- und Bestimmungsweg*“ offengelegt wird (Krafft/Zweig 2019, S. 38).

Klasse 2

In **Klasse 3**, „*wenn ein ADM-System in seinem sozioinformatischen Gesamtsystem bei Fehlentscheidungen ein von der Gesellschaft gesetztes Risikopotential überschreitet*“ (Krafft/Zweig 2019, S. 39), müssen alle Angaben von Betreibern *mindestens für ein Expertengremium* in angemessener Zeit nachvollziehbar und überprüfbar sein. Dazu sind verschiedene Interfaces zu den *Eingangsdaten und den Resultaten* der Maschine notwendig. In Klasse 3 wird vollständige Nachvollziehbarkeit gefordert, die sich auf drei Bereiche bezieht: (a) Nachvollziehbarkeit der Daten, indem gegenüber einer überwachenden Instanz die verwendeten Trainingsdaten zur Verfügung gestellt werden (etwa ob Messfehler gefiltert/beachtet wurden); (b) Nachvollziehbarkeit des Lernverfahrens, um Fehlerquellen und unerwünschte Entscheidungsmuster auszuschließen. Dies geschieht durch Einblicke in den Lernprozess des algorithmischen Entscheidungssystems und den darin wirkenden Code, inklusive Aussagen zum Systemverhalten (z. B. Kommunikation der verwendeten Hyperparameter wie Ebenen-Anzahl des neuronalen Netzes oder die maximal erreichbare Tiefe des Entscheidungsbaums) bzw. der Trainingsdaten und Lernverfahren (um eine Wiederholung des Lernprozesses zu ermöglichen). Alternativ können auch Code-Audits mit einer beglaubigten Kopie des Programmcodes zu Prüfzwecken (und Angaben zum trainierten statistischen Modell) durchgeführt werden. Als letzter Bereich findet sich in der Abhandlung von Krafft und

Klasse 3

Zweig (2019) die (c) Nachvollziehbarkeit der Entscheidung, was den ausschließlichen Einsatz von erklärbaren Methoden des maschinellen Lernens für die lernende Komponente beinhaltet. Eine Einteilung von Methoden in „erklärbare" (z. B. lineare und logistische Regressionen sowie einzelne Entscheidungsbäume, solange sie nicht auf zu hochdimensionalen Daten trainiert wurden) und „nicht-erklärbare" (z. B. aktuell neuronale Netze) ist jedoch noch nicht abgeschlossen.

Letztendlich bergen lernende algorithmische Entscheidungssysteme, die in **Klasse 4** eingeteilt werden, ein zu hohes Risiko, um eingesetzt zu werden, bzw. sollten nur dann eingesetzt werden, wenn ihnen beweisbar ein genügend hoher Gesamtnutzen entgegensteht. *Klasse 4*

Dieses Schema soll durch Stakeholder-Einbindung weiter validiert und weiterentwickelt werden (Krafft/Zweig 2019).

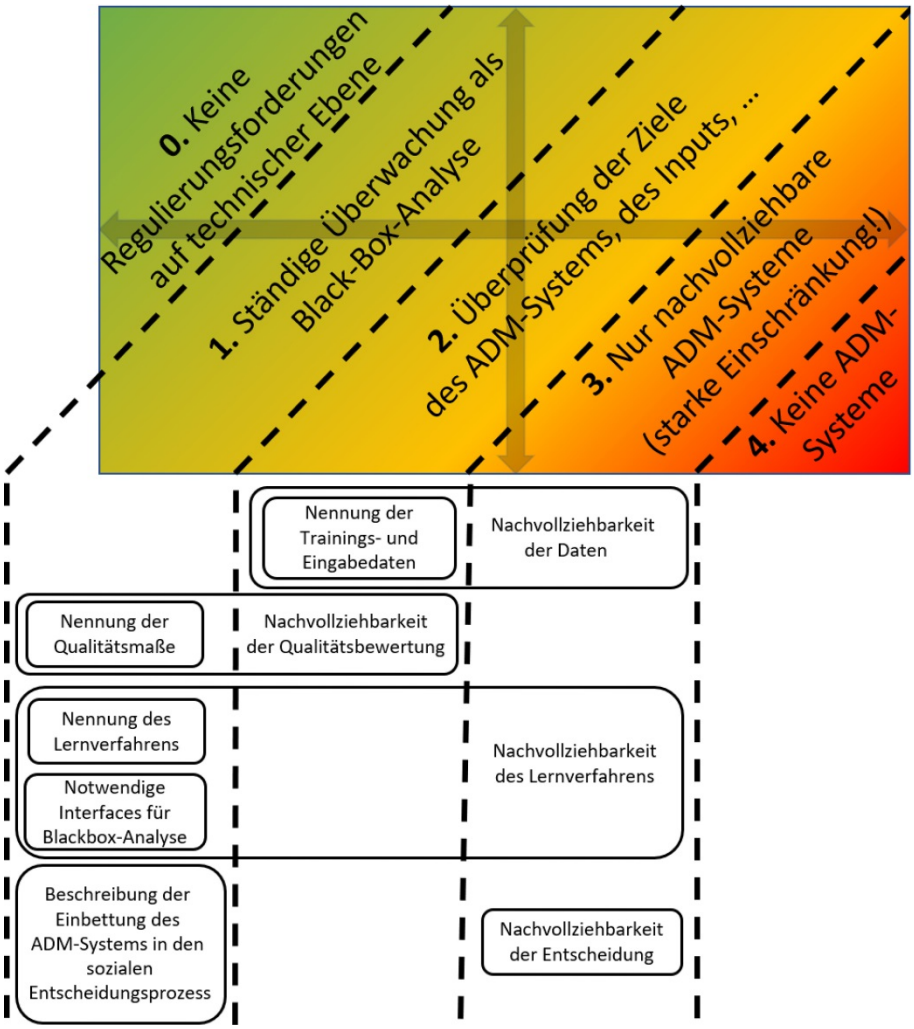

Abbildung 12: Transparenz- und Nachvollziehbarkeitsforderungen nach Regulierungsklassen (Quelle: Krafft/Zweig 2019)

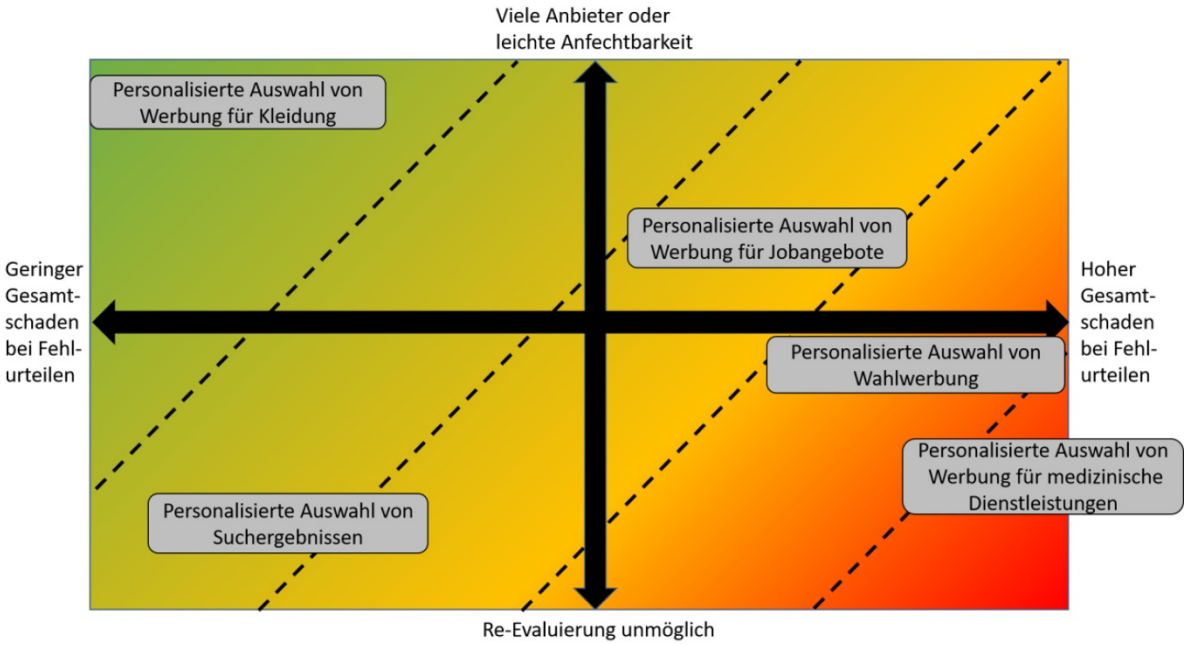

Abbildung 13: Anwendungsgebiete von Empfehlungssystemen (Quelle: Krafft/Zweig 2019)

Diese beiden hier vorgestellten Ansätze – der Ansatz der EU-Kommission, auf dem der AI Act fußt und das Modell von Krafft/Zweig (2019); und Zweig (2019) – zeigen grundlegende Gemeinsamkeiten, aber auch Unterschiede in ihrer Ausgestaltung, die sich auf die Perspektive, sowie die Kriterien zur Einteilung von KI-Anwendungen und den Umfang der angestrebten Regelung beziehen.

Grundsätzlich sind beide Ansätze risikobasiert und nehmen daher Auswirkungen – vorrangig auf Mensch und Gesellschaft – mit in den Blick; beide sehen (prinzipiell) mehrere Risikoklassen für KI-Anwendungen vor, wobei die Kriterien je nach Modell und Klasse unterschiedlich genau definiert sind, insgesamt tendenziell aber vage bleiben. Unterschiede finden sich in hinsichtlich der Perspektive, aus denen der jeweilige Ansatz verfasst ist. Krafft/Zweig (2019) präsentieren ein aus Vorsorgeüberlegungen verfasstes Modell, das vorrangig Auswirkungen auf Konsument*innen und (passiv) Betroffene von KI-Systemen in den Blick nimmt und aus dieser Perspektive argumentiert (unter Berücksichtigung limitierter Ressourcenverfügbarkeit). Entsprechend definieren Krafft/Zweig (2019) fünf Risikoklassen (Klasse 0 bis 4) mit unterschiedlichen, abgestuften Transparenzanforderungen, die Nachvollziehbarkeit für unterschiedliche Akteure gewährleisten sollen. Im Gegensatz dazu zeigt der AI Act nicht nur eine stärkere grundsätzliche Innovationsorientierung, sondern konsolidiert auch bereits eine Vielzahl von unterschiedlichen Interessen. Er definiert lediglich zwei Risikoklassen (Verbotene Anwendungen und Hochrisiko-Systeme) und entsprechende Transparenzanforderungen bzw. Verbote bestimmter Anwendungen.

Entsprechend müssen Ansätze zur Einteilung von KI-Anwendungen und die Entwicklung von entsprechenden Kriterien in einem inklusiven, deliberativen und partizipativen Prozess entwickelt und validiert werden. Dies soll sicherstellen, dass auch in der Politik (genauso wie im Entwicklungsbereich) eine breite Palette an Perspektiven berücksichtigt wird und insbesondere die Perspektive von Konsument*innen bzw. Betroffenen (die ja potenziell omnipräsent ist und jede*n betrifft) noch stärker in den Blick genommen wird.

5.3 DIE ÖSTERREICHISCHE SITUATION

5.3.1 AI-GOVERNANCE-ANSÄTZE IN ÖSTERREICH

In Österreich wurden mehrere politische Strukturen und Initiativen verfolgt, um eine offizielle nationale Strategie für KI zu schaffen. Diese legt den Schwerpunkt auf klare Empfehlungen für die Entwicklung und Förderung von Innovationen und KI-basierten Technologien. Zurzeit gibt es vier wesentliche politische Initiativen, die die nationale KI-Strategie Österreichs bilden (European Commission 2018; OECD 2020; Europäische Kommission 2021). Die wichtigste ist die KI-Strategie des Bundes bzw. *Artificial Intelligence Mission Austria 2030*, die 2019 von den Ministerien BMK (damals BMVIT) und Bundesministerium für Digitalisierung und Wirtschaftsstandort (BMDW) gestartet wurde. Im Jahr 2021 präsentierten die Bundesministerinnen Margarete Schramböck und Leonore Gewessler gemeinsam die Strategie der Bundesregierung für künstliche Intelligenz (Artificial Intelligence Mission Austria 2030 – AIM AT 2030) mit ihren Zielen und Handlungsfeldern (BMDW 2021). Das Hauptziel ist die Schaffung eines (rechtlichen) Rahmens zur Förderung eines sicheren und verantwortungsvollen Einsatzes von KI-Technologien im öffentlichen Interesse auf der Grundlage europäischer Grundwerte und Rechtsvorschriften. Die Verfolgung der folgenden drei Ziele steht im Mittelpunkt der nationalen Strategie (ibid.):

> *„(1) Es wird ein am Gemeinwohl orientierter, breiter Einsatz von KI angestrebt, der in verantwortungsvoller Weise auf Basis von Grund- und Menschenrechten, europäischen Grundwerten und des kommenden europäischen Rechtsrahmens erfolgt.*
> *(2) Österreich soll sich als Forschungs- und Innovationsstandort für KI in Schlüsselbereichen und Stärkefeldern positionieren und*
> *(3) mittels der Entwicklung und des Einsatzes von KI soll die Wettbewerbsfähigkeit des österreichischen Technologie- und Wirtschaftsstandorts gesichert werden."*

Vor dem Launch der nationalen Strategie hat das Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK, damals BMVIT) im Jahr 2017 den Österreichischen Rat für Robotik und Künstliche Intelligenz (ACRAI) gegründet. ACRAI besteht aus Expert*innen zu Robotik und KI aus Forschung, Lehre und Wirtschaft. Die Aufgabe des Rates ist, aktuelle und künftige Chancen, Risiken und Herausforderungen, die sich aus dem Einsatz von Robotern und autonomen Systemen sowie KI ergeben, zu identifizieren und zu diskutieren (ACRAI 2022). Im November 2018 veröffentlichte ACRAI ein *White Paper*, das die Grundlagen für Österreichs Robotik- und KI-Strategie festlegte. Diese Strategie umfasst mögliche Anwendungen von KI-Technologien, Governance-Herausforderungen sowie Empfehlungen zu den ethischen und gesellschaftlichen Herausforderungen von KI. Basierend auf diesen Initiativen beauftragte das BMK den ACRAI mit der Entwicklung der *nationalen KI-Strategie Österreichs* (Van Roy et al. 2021, S. 20).

Hauptakteur*innen

Österreich verfügt auch über eine vom BMDW entwickelte *Digitale Roadmap* (2016), die Leitprinzipien und Szenarien für 2025 zu bestimmten Technologien enthält: 5G, IoT, Big Data-Analytik, KI, offenes Wissen, Augmented und Virtual Reality, 3D-Druck, advanced materials, Blockchain und intelligente Energienetze.

Die Österreichische Gesellschaft für Mess-, Automatisierungs- und Robotertechnik hat 2015 im Rahmen des BMK die *Nationale Technologieplattform Robotertechnik* (GMAR) gegründet. GMAR hat das Ziel, KI-Technologien zu fördern, die Politik zu beraten und den Informationsaustausch zwischen den Akteuren zu erleichtern (GMAR 2022).

Die vom BMK gemeinsam mit einer Reihe von Industrieorganisationen gegründete Plattform Industrie 4.0 Österreich ist der größte Non-Profit-Verein. Die Plattform ist eine zentrale Koordinierungsstelle für die Politik zwischen den relevanten Stakeholdern.

5.3.2 DER KI-SEKTOR IN ÖSTERREICH

Nach einem Bericht des BMK über österreichischen Unternehmen sind in Österreich rund 600 Unternehmen im Themenkomplex KI aktiv (von rund 300.000 Unternehmen). Etwa ein Drittel davon zählt zur Branche der Software-Entwicklung bzw. ist Anwender (eigener) Lösungen und Anbieter entsprechender Datenverarbeitungen (von Business Intelligence bis Analyse bildgebender Verfahren aus dem medizinischen Bereich) oft in Kombination mit Beratungsleistungen (BMK 2019, S. 5). Viele dieser Unternehmen sind hochspezialisiert auf die Analyse von Unternehmens- und Finanzdaten; teilweise spiegeln sich hier große österreichische Branchen wie der Fahrzeug- und Maschinenbau wider. Die höchste Konzentration von KI-Firmen (gemessen am Anteil aller Unternehmen in den jeweiligen Sektoren) findet sich im Pharmasektor, in der Mineralölverarbeitung, im Versicherungswesen und in der Herstellung von Datenverarbeitungsgeräten in elektronischen und optischen Produkten (ibid.).

ca. 600 Firmen in Österreich

Während früher die Entwicklung von KI-Lösungen relativ selten war, erwähnen heute viele Unternehmen KI-Kompetenzen und führen KI-Projekte auf ihren Websites auf. Die Universitäten und Forschungsinstitute in Österreich haben eine beachtliche Bandbreite an Spezialisierungen in KI-Bereichen, allerdings sind die Forschungsteams, die daran arbeiten, recht klein. Besonders ausgeprägt sind die Forschungskompetenzen im Bereich des maschinellen Lernens, aber auch symbolische Methoden, Robotik und autonome Systeme sind gut vertreten. Geforscht wird fast österreichweit, mit starken Schwerpunkten in Wien, Graz, aber auch in Linz und Klagenfurt. Weitere einschlägige Unternehmen finden sich in Innsbruck, St. Pölten, Klosterneuburg und Salzburg (ibid.).

Forschung und Entwicklung

Der Start-up-Sektor ist ein wichtiger Technologietreiber für KI-Innovationen. Start-ups verfügen oft über Wissen für spezialisierte Lösungen und haben die Flexibilität, wichtige Lösungen an lokale Bedürfnisse und Kontexte anzupassen und zu entwickeln (ibid., S. 6). Die Hauptmotivationen für Unternehmen, KI-Lösungen in diesen Bereichen zu implementieren, sind in erster Linie Qualitätsverbesserung und Optimierung. Im Bereich der Geschäftsmodellinnovation wird erwartet, dass die dynamische Preisgestaltung ein wichtiger Aspekt von KI-Anwendungen sein könnte. Es gibt auch einige Anzeichen dafür, dass die Entwicklung von Lösungen zu den Kunden verlagert wird, insbesondere im Bereich der Beratungstätigkeit, wo die Grenzen zwischen Beratungs- und KI-Entwicklungsunternehmen manchmal verschwimmen (ibid.).

Start-ups

Das Haupthemmnis für das Wachstum dieses Sektors in Österreich stellt laut BMK der Mangel an Personal dar, sowohl an KI-Generalist*innen als auch KI-Spezialist*innen in Themen wie z. B. neuronalen Netzwerken sowie KI spezialisierte Software-Ingenieur*innen. Außerdem ist der Know-how-Erwerb sehr ressourcenintensiv, die Entwicklungszeiten sind sehr lang und die Erwartungen an die Technologie, positive Ergebnisse für das Unternehmen zu bringen, werden oft nicht erfüllt. Es besteht hohe Unsicherheit, während die Erwartungen hoch sind (ibid., S. 7).

Mangel an qualifiziertem Personal

Mehrere Länder begegnen diesen Herausforderungen und anderen Risiken im Zusammenhang mit der Ausweitung von KI-Innovationen mit der Formulierung nationaler KI-Innovationsstrategien. Während solche Leitlinien in ihrem Ansatz recht unterschiedlich sind, gibt es Gemeinsamkeiten wie den Ausbau von Fähigkeiten, eine engere Zusammenarbeit zwischen Wissenschaft und Industrie, verbesserten Zugang zu Daten sowie die Rolle des Staates bei Pilotprojekten und Regulierung. Auf nationaler Ebene wird nach Einschätzung des BMK jedoch wenig Wert auf europäische und internationale Kooperationsinitiativen, die Rolle von KI in Entwicklungsfragen oder kreative und künstlerische Anwendungen von KI-Technologien gelegt (ibid., S. 8).

geringe internationale Orientierung

Die Ergebnisse der SWOT-Analyse des BMK (2019) legen nahe, dass Österreich sowohl über Forschungs- als auch über industrielle Innovationskompetenzen im Bereich KI verfügt, insbesondere in den dominierenden Branchen Fahrzeug- und Maschinenbau. Zahlreiche Forschungsinstitute beschäftigen sich mit KI-Forschung und diese sind über ganz Österreich verteilt, die meisten davon sind freilich sehr klein. Es gibt eine aktive „KI-Szene" innovativer Unternehmen, die sich selbst organisieren und eigene Initiativen wie spezialisierte Vernetzungsplattformen und Veranstaltungen haben. Es gibt ein Bewusstsein für neue Geschäftsmodelle, wie etwa *AI-as-a-Service*. Die größte Bedrohung ist, wie bereits erwähnt, der Mangel an Fachkräften in der Informatik und an KI-Anwendern.

kleine österreichische Szene

6 SCHLUSSFOLGERUNGEN

KI als Teil gestaltbarer sozio-technischer Systeme

„Künstliche Intelligenz“ wird in diesem Bericht – in Anlehnung an Forschungsfelder wie die Technikfolgenabschätzung, die Science and Technology Studies, oder die sozialwissenschaftliche Technikfolgenforschung – als sozio-technisches System verstanden. Das rückt nicht nur die technische Gestaltbarkeit bestimmter Anwendungen in den Mittelpunkt des Interesses, sondern auch die vielfältigen Möglichkeiten, die sich aus der Einbindung von KI in soziale Kontexte, das Zusammenwirken von sozialen und technischen Aspekten und die Umgestaltung von sozialen Praktiken durch KI ergeben. Dadurch wird auch die Möglichkeit betont, KI aktiv zu gestalten: Die Entwicklung von KI ist kein Prozess, der nach bekannten oder unbekannten Schemata autonom abläuft, sondern ein Feld, das bewusste Gestaltung erfordert.

Algorithmic Decision Making Systems (ADM)

Der Begriff „KI“ ist problematisch, weil er nur ungenau definiert und abgegrenzt werden kann. Im Zusammenhang mit konkreten Entscheidungsansprüchen eignet sich der breitere Begriff des algorithmischen Entscheidungssystems (Algorithmic Decision Making Systems, ADM) wesentlich besser. Er umfasst keine spezifischen Technologien oder Methoden, sondern stellt auf den Gesamtprozess und dessen Ziel ab, Entscheidungen zu unterstützen oder selbständige Entscheidungen durch ein System zu treffen (vgl. Krafft/Zweig 2019).

KI wird breit eingesetzt

Zur *Gestaltung von KI* als auch ihrer sozialen Einbettung ist eine Auseinandersetzung mit dem technischen Status-quo unerlässlich. Was können (aktuelle) KI-Systeme leisten, worin liegen ihre gegenwärtigen Beschränkungen? Während der aktuelle Status-quo hinsichtlich technischer Reife und flächendeckenden Einsatzes nicht überschätzt werden sollte – insbesondere was die Erwartungen oder auch Befürchtungen bezüglich so genannter genereller oder starker KI betrifft – können aufgrund rein technischer Spezifikationen keine Aussagen über soziale Auswirkungen von KI-Systemen getroffen werden.

hat noch viel Entwicklungspotenzial

Daher können Ansätze des (regulatorischen) Umgangs mit KI nicht ausschließlich über technische Aspekte definiert werden, sondern müssen Anwendungen bzw. Auswirkungen von KI auf den Menschen verstärkt in den Mittelpunkt des Interesses rücken. Aufgrund möglicher Auswirkungen in unterschiedlichsten Bereichen (vgl. Kapitel 4) ergibt sich die Notwendigkeit, KI zu regulieren. Auf die Frage, welche Anwendungen zu welchem Zweck reguliert werden sollen und wie eine Regulierung konkret ausgestaltet werden soll, gibt es eine Reihe von Vorschlägen. Diese reichen von Ethik-Richtlinien und Codes of Conduct für die Entwicklung von KI bis zu elaborierteren Konzepten, die die Beurteilung von Anwendungen zum Ziel haben. Festzuhalten ist, dass Schaden an Leib, Leben und Eigentum nicht erst durch KI-Anwendungen (oder auch ADMs) im engeren Sinn entstehen, sondern oft auch bereits durch den Einsatz weniger komplexer Technologien oder Ansätze (z. B. statistische Verfahren). Eine zu enge KI-Definition tendiert dann dazu, Risiken zu ignorieren, die durch den Einsatz von etablierten IT-gestützten Ansätzen bereits allgegenwärtig sind (z. B. Diskriminierung am Arbeitsmarkt durch den Einsatz statistischer Verfahren bei der Ressourcen-Zuteilung). Daher plädiert der Bericht für die Beibehaltung einer breiten KI-Definition. Über technische Spezifikationen von KI-Systemen hinaus ist es wichtig, auf die Merkmale oder Eigenschaften eines Systems zu fokussieren, die für die Regulierung von Bedeutung sind.

sollte jedenfalls risikobasiert und kontextbezogen reguliert werden

Neben einer Engführung der technischen Komponente der KI-Definition, kann auch ein zu starker Fokus auf bestimmte Technologien mit fundamentalen Auswirkungen dazu führen, dass Anwendungen mit weitreichenden Konsequenzen *per definitionem* unreguliert (oder unterreguliert) bleiben. Die Einführung eines risikobasierten abgestuften Ansatzes wäre aus pragmatischer Sicht sinnvoller als allein auf technische Spezifikationen zu fokussieren.
Risikobasierte Ansätze erlauben, Auswirkungen von KI-Systemen in den Blick zu nehmen. Je nach Ausführung sind Perspektive, Kriterien zur Einschätzung des Risikos oder das abgeleitete Schutzmaß der Regulierung grundlegend unterschiedlich, wobei fundamentale Werte wie Gesundheit, Sicherheit und Grundrechte den Kern des Schutzgutes bilden. Abhängig von der jeweiligen Perspektive (z. B. wer zu schützen ist) werden diese jedoch unterschiedlich konzeptualisiert.

Verantwortung bleibt beim Menschen und den beteiligten Institutionen

Weiters muss die *Verantwortung* gegenüber Betroffenen und Konsument*innen von KI-Anwendungen durch regulierende Stellen, aber auch Entwickler*innen, Produzent*innen und Verwender*innen hervorgehoben und gestärkt werden. Der aktuelle Fokus des AI Acts auf Verantwortung und Haftbarkeit allein gegenüber Nutzer*innen[20] hinterlässt eine Lücke im Umgang mit KI, da eine flächendeckende Betroffenheit von Bürger*innen (oder Konsument*innen) durch KI-Anwendungen zu befürchten ist. Wie im Fall der Datenschutz-Grundverordnung (DSGVO) müssen starke Rechtsbehelfe und Rechtsmittel für die Bürger*innen Teil jeder umfassenden EU-KI-Regulierung werden.

Die aktive Einbindung von potentiell Betroffenen und Konsument*innen in Gestaltungsprozesse von KI oder ADM durch inklusive partizipative und deliberative Ansätzen ist unerlässlich, um den grundlegenden Schritt in Richtung Sozialverträglichkeit von KI-Anwendungen zu machen. Hierbei geht es um eine breite gesellschaftliche Diskussion, welche Werte in KI-Anwendungen eingeschrieben werden (sollen) und gesellschaftlich akzeptabel erscheinen. Gleichzeitig ermöglichen partizipative Ansätze, möglichst vielfältig potenzielle Auswirkungen von KI-Anwendungen (oder ADMs) aufzuzeigen. Diese Debatte erlaubt es auch Kriterien für die Einteilung von KI-Systemen bezüglich ihrer potenziellen Gefährlichkeit zu konkretisieren und so bestehende Konzepte weiter zu verbessern.

unterschiedliche Dimensionen von Transparenz

Die gesellschaftliche Akzeptanz von KI-Systemen basiert zu einem guten Teil auf Vertrauen in KI-Anwendungen. Dabei kommt dem Begriff der *Transparenz* eine zentrale Rolle zu. Diese ist auch häufig in der Forderung nach verantwortungsvoller Entwicklung von KI inkludiert. Transparenz ist allerdings vielschichtig (bezogen auf den Algorithmus, den Prozess, den Kontext) und in manchen Bereichen auch ambivalent (z. B. Geschäftsgeheimnisse). Sie kann einerseits technisch (als die detaillierte Offenlegung von Codes), andererseits prozedural (als zielgerichtete Kommunikation, um mehr Verständnis bei der intendierten Zielgruppe zu erzeugen) verstanden werden. Durch diese inhärente Mehrdeutigkeit ist die bloße Forderung nach Transparenz nicht ausreichend, um KI verantwortungsvoll zu entwickeln, in Umlauf zu bringen und zu regulieren, auch, weil sie in einer rein technischen Interpretation auch zur Verschleierung von Verantwortung beitragen kann.

[20] „Nutzer“ im Sinne des AI Acts ist eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System in eigener Verantwortung verwendet, es sei denn, das KI- System wird im Rahmen einer persönlichen und nicht beruflichen Tätigkeit verwendet.

In Erweiterung des Begriffes betont ein Fokus auf *Nachvollziehbarkeit (*als Teil des breiteren Konzepts der *Transparenz*, einschließlich technischer *Erklärbarkeit* und sozialer *Verständlichkeit)* die soziale Dimension des Verstehens von KI-Anwendungen und damit der informierten Entscheidung von Nutzer*innen- und Betroffenen bzw. des verantwortungsvollen Umgangs mit KI. Hierfür ist es notwendig, genau zu definieren, was, bis zu welchem Grad und welcher Zielgruppe gegenüber nachvollziehbar gemacht werden muss. Abgestufte Transparenzregeln (je nach Adressatengruppe) erscheinen sinnvoll, wobei der institutionelle Rahmen entsprechend gestaltet sein muss. Je nach Transparenzstufe sind dann spezifische Kompetenzen auf Seite der Entwickler*innen und Produzent*innen, aber auch der Adressat*innen notwendig, um Nachvollziehbarkeit gewährleisten zu können. Der Erwerb oder die Bereitstellung solcher Kompetenzen sind durch verantwortliche (öffentliche) Stellen zu garantieren. Unabhängige Institutionen müssen eingerichtet und mit entsprechenden Kompetenzen besetzt werden, um Transparenzforderungen, Beschwerden oder Klagen zu überprüfen und gegebenenfalls adäquat handeln zu können. Sozialverträglichkeit wird neben der notwendigen Transparenz nur durch entsprechende Regulierung und Institutionen herzustellen sein.

Nachvollziehbarkeit und abgestufte Transparenzregeln

Die Ergebnisse von in KI-Anwendungen intendierten autonomen Lernprozessen können nicht immer vollständig antizipiert werden. Die Vermeidung der sogenannten „Black-Box", also der Uneinsichtigkeit und Unverständlichkeit bestimmter Prozesse, ist somit zentrales Thema im Bereich der Entwicklung (und Anwendung) von KI. Entsprechend wird in der Literatur argumentiert, dass Entwickler*innen (auch bei vollständiger Transparenz in Bezug auf Codes) nicht vollständig für Konsequenzen des Einsatzes von KI verantwortlich gemacht werden können. Damit ergibt sich die Gefahr eines Verantwortungsvakuums, das hinsichtlich der potenziellen weitreichenden Auswirkungen auf Produzent*innen, Anwender*innen, Betroffene und Konsument*innen problematisch werden kann. Aus Sicht von Betroffenen und Konsument*innen könnte es hier schwierig sein, das Recht auf Information, aber auch auf Widerspruch (gegen bestimmte Entscheidungen) oder Schadensersatz bei Fehlentscheidungen einzufordern. Daher ist menschliche Aufsicht („human oversight") als eine Grundbedingung zu sehen, um KI-Anwendungen auf den Markt zu bringen. Sollte dies nicht möglich sein, erscheint ein Moratorium für bestimmte Anwendungen sinnvoll, bis eine solche Aufsicht garantiert werden kann. Für bestimmte KI-Anwendungen ist ein grundsätzliches Verbot angemessen (siehe dazu die Diskussion zur ethischen KI). Der vorliegenden Entwurf des AI-Acts berücksichtigt dies, indem er unter anderem Systeme zur Bewertung oder Klassifizierung der Vertrauenswürdigkeit natürlicher Personen über einen bestimmten Zeitraum auf der Grundlage ihres sozialen Verhaltens oder bekannter oder vorhergesagter persönlicher Eigenschaften verbietet (Social Scoring).

solange menschliche Aufsicht in bestimmten Anwendungen nicht möglich ist, ist dort ein Moratorium vorzusehen

Auf technischer Ebene stellt sich die Frage, ob die in diesem Bericht angesprochenen Ansätze der „erklärbaren KI" (XAI) ausreichen um (technische) Transparenz – und in weiterer Folge soziale Nachvollziehbarkeit und Verantwortung – herzustellen. Aktuelle Forschungsansätze liefern interessante Ergebnisse, befinden sich aber noch in der Entwicklung. Da sie jedoch soziale Kontexte der Anwendungen und einen bewussten Umgang mit Wertentscheidungen nicht berücksichtigen, können XAI-Ansätze auch nur eingeschränkt Wirkung in Richtung Verantwortung entfalten.

„erklärbare KI" (XAI) kann zur Nachvollziehbarkeit beitragen

Eine realistische Einschätzung von KI (oder ADMs) basiert daher auf einer technischen und sozialen Komponente, die jeweils unterschiedlich zu bewerten sind. Während umfassende technische Innovationen (wie generelle KI) noch weit entfernt scheinen, gestalten weniger elaborierte Ansätze (z. B. schwache KI oder auch statistische Verfahren) bereits die soziale Realität um. Schon gegenwärtig verändern sich Sozialsysteme unter dem Einsatz von IT- und KI-Systemen (Jobsuche, Zuteilung von Sozialleistungen, Credit Scoring). Gleichzeitig zieht der flächendeckende Einsatz von KI-Systemen weitere Veränderungen nach sich. In der aktuellen Literatur wurden Unterschiede hinsichtlich Wissens- und Ausbildungsniveau bezüglich KI und dem Umgang damit zwischen privatem und öffentlichem Sektor festgestellt; dies bezieht sich einerseits auf konkrete Entwicklung und korrekte Anwendung von KI Systemen, andererseits auf die Durchsetzung von Rechten der Bürger*innen, wie beispielsweise von Informations- und Klagsrechten.

Transparenz ist notwendig aber nicht hinreichend, um nachvollziehbare und verantwortungsvolle KI-Entwicklung und -Anwendung zu garantieren

Daher lässt sich festhalten, dass Transparenz – im Gegensatz zum Untertitel dieses Berichts – nur einer unter mehreren relevanten Aspekten ist, der bei Entwicklung und Einsatz von KI bzw. ADMs eine wichtige Rolle spielt. Wichtiger als technisch-basierte Zugänge sind aus Sicht der Studienautor*innen regulatorische Ansätze, die gesellschaftliche Realitäten und Anwendungskontexte zumindest gleichwertig berücksichtigen, um eine aktive Gestaltung von KI-Systemen (und auch etablierter Methoden) zu ermöglichen. Transparenz ist notwendig aber nicht hinreichend, um nachvollziehbare und verantwortungsvolle KI-Entwicklung und -Anwendung zu garantieren.

7 HANDLUNGSEMPFEHLUNGEN

Aus den Schlussfolgerungen des Berichts lassen sich folgende Handlungsempfehlungen ableiten. Eine Grundvoraussetzung für deren Umsetzung ist neben dem politischen Willen vor allem die Bereitstellung ausreichender Ressourcen:

GESELLSCHAFTLICHE EINBINDUNG

Es braucht eine Berücksichtigung der Interessen von Nutzer*innen, Konsument*innen und Betroffenen im KI-Diskurs und in der Entwicklung. Die Entwicklung von KI ist stark wertebasiert und ihre Anwendungen lassen massive soziale Auswirkung erwarten. Daher ist eine umfassende Einbindung von Betroffenen-Interessen in die Entwicklung und Anwendung von KI unabdingbar. Ein Beispiel für einen solchen Einbindungsprozess ist z. B. die Entwicklung der Algo.Rules[21], bei der neben Workshops und Konsultationen mit Expert*innen der Informatik, Wissenschafter*innen und Praktiker*innen aus verschiedenen Bereichen auch Öffentlichkeitseinbindung in Form von Online-Umfragen und Diskussionen der Algo.Rules stattfinden.[22]

NUTZUNG EINER BREITEN DEFINITION

Die Frage der Definition von KI ist von zentraler Bedeutung für jede KI-Regulierung. Daher sollte eine weit gefasste Definition von KI, die auch mögliche Folgen von KI-Anwendungen berücksichtigt, angestrebt werden. Anstatt bestimmte technische Anwendungen ein- oder auszuschließen, sollte eine solche Definition neben neuartigen KI-Anwendungen auch etablierte IT-Systeme umfassen, die sich der Datenanalyse und -interpretation widmen und deren Vorschläge und Entscheidungen Menschen in ihrer physischen, psychischen oder ökonomischen Existenz betreffen. Unabdingbar ist, dass jede Regulierung das Wohlergehen der Menschen in den Mittelpunkt stellt, wobei die Grundrechte die nicht zu unterschreitende Mindestanforderung darstellen.

KONKRETISIERUNG DES TRANSPARENZ-BEGRIFFS IN ENTWICKLUNGS- UND REGULIERUNGSPROZESSEN

Angesichts des vielschichtigen Charakters des Konzepts reicht Transparenz als allgemeine Anforderung allein nicht aus, um sichere und vertrauenswürdige KI zu ermöglichen. Auf Grundlage der vorgestellten Analyse gibt es mehrere Dimensionen von Transparenz, die alle in den Entwicklungs- und Regulierungsprozessen entsprechend berücksichtigt werden sollten:

a. *Transparenz als Recht auf Information:* Transparenz im konkreten Sinn bezieht sich auf das grundlegende Recht auf Zugang zu Information über die Funktionsweise und Folgen von KI-Anwendungen auf individuelle Personen und darauf, wie dieses Recht umgesetzt wird. Dieses könnte in Analogie zu den Betroffenenrechten der DSGVO (Art. 12 bis 23) gestaltet werden.

---

21 *https://algorules.org/de/startseite*.
22 *https://algorules.org/de/startseite#c166111*.

b. *Transparenz als Stufe zur Nachvollziehbarkeit:* Das Konzept der Transparenz sollte durch das Konzept der Nachvollziehbarkeit ergänzt bzw. erweitert werden. Die Umsetzung von Transparenz muss über die technische Erklärbarkeit des Systems (z. B. Offenlegung technischer Details) und über den Anspruch der Vermeidung von „Black-Box"-KI hinausgehen. Abgestufte Transparenzregelungen wie jene von Krafft/Zweig (2019) (Kapitel 5.2) erscheinen zielführend, um Nachvollziehbarkeit für (je nach Stufe unterschiedliche) Adressatengruppen zu ermöglichen. Dies steht in engem Zusammenhang mit einer verfahrenstechnischen Dimension von Transparenz (Punkt c).
c. *Transparenz als Kernaufgabe von dazu bestimmten Institutionen:* Transparenz hat auch eine institutionelle (und verfahrenstechnische) Dimension in Bezug auf KI-Anwendungen. Da das Konzept der Transparenz in Beziehung zu anderen, technischen Konzepten wie Erklärbarkeit, Interpretierbarkeit usw., aber auch zu sozialen wie Nachvollziehbarkeit (z. B. Verständlichkeit in Sprache und Inhalt) steht, ist dieses für die jeweiligen Betroffenen zentral. Transparenzerfordernisse sollten demnach je nach Zielgruppe, Kategorisierung, potenziellem Schaden etc. abgestuft sein. Das bedeutet, dass Betroffene jedenfalls über den Einsatz von KI-Systemen (und welche) verständlich informiert werden müssen. Bei höherem potenziellen Schaden sind auch Informationen zum inneren Funktionieren (Trainingsdaten, Algorithmus, Code etc.) bereitzustellen. Um diese Rechte wahrzunehmen, könnten Betroffene durch fachkundige Vertreter*innen (z. B. von NGOs) oder in weiterer Folge bei staatlichen Beschwerde-/Schiedsstellen und Gerichten Unterstützung bekommen. Erst umfassende Transparenz erlaubt also mögliche Kontrolle bzw. Klagbarkeit auch durch Nutzer*innen, Konsument*innen und Betroffene über entsprechende unabhängige Stellen. Eine solche Institution müsste die gesellschaftliche Gefährlichkeit bzw. Wertigkeit konkreter Systeme oder Gruppen von Anwendungen bewerten – unter Einbeziehung aller gesellschaftlichen Perspektiven und Akteur*innen (vgl. Punkt 1).

ENGMASCHIGES MONITORING DER KI-ENTWICKLUNG

Jedes KI-System beinhaltet Fehlerpotential. Sobald Menschen betroffen sein können, besteht dadurch oft auch ein Diskriminierungspotential. Deshalb sind alle KI-Systeme die mit Menschen interagieren bzw. deren Entscheidungen sich auf das Leben und die Entfaltungsmöglichkeiten der Menschen auswirken, zu registrieren, einem grundlegenden Zertifizierungsprozess zu unterwerfen und im Verkehr kenntlich zu machen. Beispiele für Bewertungen und Zertifizierungen in diesem Sinn sind unabhängige Folgenabschätzungen (Impact Assessments) und Audits.

ENTWICKLUNG EINES BEURTEILUNGSRAHMENS FÜR KI-SYSTEME

Der gesellschaftliche Diskurs zur KI sollte in einen ausdifferenzierten, abgestuften und möglichst konkreten Kriterienkatalog münden, der es Entwickler*innen und auch Betroffenen möglich macht die Kategorisierung von KI-Systemen zu antizipieren bzw. auch nachzuvollziehen. Die Einstufungskriterien für KI-Systeme, die der Zertifizierung zugrunde liegen, müssen in diesem gesellschaftlichen Prozess ausgehandelt, konkretisiert und publiziert werden.

REGULIERUNGSMONITORING

Bei KI handelt es sich um ein sich schnell entwickelndes, multidisziplinäres Feld mit Anwendungen in einer Vielzahl von Bereichen und mit verschiedenen ethischen und grundrechtlichen Implikationen. Dieser dynamische Charakter der KI macht einen flexiblen Regulierungsansatz erforderlich, bei dem die kontinuierliche Kontrolle im Mittelpunkt steht. Die grundlegenden Werte, die die Basis für die Kriterienentwicklung (siehe oben) bilden, müssen periodisch evaluiert und möglichen neuen technischen, wie auch gesellschaftlichen Entwicklungen angepasst werden.

Der Europäische Ausschuss für Künstliche Intelligenz sollte um eine breite öffentliche und gesellschaftliche Beteiligung erweitert werden. Die Aufgabe einer solchen Multi-Stakeholder-Organisation (bestehend aus Wissenschafter*innen, Organisationen der Zivilgesellschaft, Interessensvertretungen, Unternehmen, die KI-Technologie entwickeln und nutzen, und anderen Akteur*innen) wäre die Ausarbeitung und Aktualisierung des Katalogs der grundlegenden Werte wie auch der steten Weiterentwicklung und Anpassung der konkreten Kriterien zur Einordnung von KI-Systemen.

FORSCHUNG ZU XAI STÄRKEN

Das Feld der XAI wird zu erklärbaren KI-Systemen führen, die Erklärungen ihrer Funktionsweise als Teil ihres Outputs enthalten. Da es sich noch im Aufbau befindet, ist es wichtig, solche Lösungen in die Anforderungen für vertrauenswürdige und ethische KI-Regelungen aufzunehmen und entsprechende Forschung zu fördern, ohne deren (gegenwärtige) Ergebnisse überzubewerten. Dazu gehören auch andere Ansätze der KI-Forschung, die sich mit Fairness, Gerechtigkeit, Rechenschaftspflicht und Verantwortlichkeit beschäftigen.

FORSCHUNG ZU GESELLSCHAFTLICHEN WIRKUNGEN VON KI

Die vielfältige Wirkung unterschiedlichster KI-Systeme kann derzeit schwer abgeschätzt werden. Deshalb erscheint die Stärkung der interdisziplinären Erforschung der Wirkungen eines breiten Einsatzes von KI notwendig.

VERBOTE AUS ETHISCHEN GRÜNDEN

Bestimmte KI-Systeme haben das Potenzial bestehende Grundrechte und Freiheiten sowie die Demokratie als solche schwer zu beschädigen, weshalb Anwendungen von KI, die aus grundlegend ethischen Grundsätzen (Menschenwürde, Gleichheit, die Unantastbarkeit des Lebens usw.) abzulehnen sind, verboten werden sollten. Dazu zählen unter anderem KI-gestützte Waffensysteme, Systeme der Massenüberwachung, wie z. B. Gesichtserkennung im öffentlichen Raum und unbemerkte Verhaltensbeeinflussung in allen Lebensbereichen. Auch massive Auswirkungen auf weitere potenziell betroffene, schützenswerte Bereiche, z. B. in Form von Umweltauswirkungen, sind zu berücksichtigen, da sie unmittelbar auf die Lebensgrundlagen von Menschen rückwirken.

MORATORIUM

Die Letztverantwortung für Entscheidungen muss weiterhin bei natürlichen oder juristischen Personen verbleiben. Diese ist aber nur sinnvoll zu übernehmen, wenn es möglich ist, die Mechanismen und Ergebnisse von KI-Systemen nachzuvollziehen. Solange XAI-Lösungen nicht elaboriert genug sind, um alle Arten von KI-Systemen vollständig erklären zu können und Transparenz und Nachvollziehbarkeit nicht ausreichen, um menschliche Kontrolle zu gewährleisten, sollten Moratorien für bestimmte KI-Systeme angedacht werden, sofern ihre Entscheidungen weitreichende soziale Konsequenzen nach sich ziehen oder eine starke moralische und ethische Komponente beinhalten.

LITERATUR

Alle in diesem Bericht verwendeten Links wurden in der Zeit von Oktober 2021 bis Jänner 2022 aufgerufen und vor der Veröffentlichung überprüft.

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y. und Kankanhalli, M., 2018, Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda, *Proceedings of the 2018 CHI conference on human factors in computing systems*.

ACRAI, 2022, *Die Zukunft Österreichs mit Robotik und Künstlicher Intelligenz positiv gestalten*, ***https://www.acrai.at***.

Adadi, A. und Berrada, M., 2018, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), *IEEE access 6*, 52138-52160.

Adelmant, V., 2020, Social Credit in China: Looking Beyond the „Black Mirror" Nightmare: center for human rights and global justice – nyu school of law, ***https://chrgj.org/2021/04/20/social-credit-in-china-looking-beyond-the-black-mirror-nightmare/***.

AI Ethics Impact Group, 2020, From Principles to Practice An interdisciplinary framework to operationalise AI ethics: VDE Bertelsmann Stiftung, ***https://www.ai-ethics-impact.org/en***.

Algo.rules., 2019, *Regeln für die Gestaltung algorithmischer Systeme*, im Auftrag von: Bertelsmann Stiftung (Hrsg.), ***https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/algorules***.

Allhutter, D., 2019, Of „working ontologists" and „high-Quality human components". The politics of semantic infrastructures," *DigitalSTS: A Field Guide for Science & Technology Studies*, 326-34.

Allhutter, D. und Berendt, B., 2020, Deconstructing FAT: using memories to collectively explore implicit assumptions, values and context in practices of debiasing and discrimination-awareness, *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Allhutter, D., Cech, F., Fischer, F., Grill, G. und Mager, A., 2020a, Algorithmic profiling of job seekers in Austria: how austerity politics are made effective, *frontiers in Big Data 3*, 5.

Allhutter, D., Mager, A., Cech, F., Fischer, F. und Grill, G., 2020b, *Der AMS-Algorithmus – Eine soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)*, November 2020, Wien: ITA.

Anderson, M., Anderson, S. L. und Berenz, V., 2016, Ensuring ethical behavior from autonomous systems, *AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments*, 2016.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I. und Atkinson, P. M., 2021, Explainable artificial intelligence: an analytical review, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11(5)*, e1424.

Apt, W. und Priesack, K., 2019, KI und Arbeit – Chance und Risiko zugleich, in: Wittpahl, V. (Hg.): *Künstliche Intelligenz*, Berlin, Heidelberg: Springer, 221–238, ***http://link.springer.com/10.1007/978-3-662-58042-4_14***.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D. und Benjamins, R., 2020, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion 58*, 82-115.

Beining, L., 2019, *Wie Algorithmen verständlich werden – Ideen für Nachvollziehbarkeit von algorithmischen Entscheidungsprozessen für Betroffene*: Stiftung Neue Verantwortung, Bertelsmann Stiftung.

Bellotti, V. und Edwards, K., 2001, Intelligibility and accountability: human considerations in context-aware systems, *Human–Computer Interaction 16(2-4)*, 193-212.

BMDW, 2021, *Strategie der Bundesregierung für Künstliche Intelligenz „AIM AT 2030"*, ***https://www.bmdw.gv.at/Themen/Digitalisierung/Strategien/Kuenstliche-Intelligenz.html***.

BMK, 2019, AI in Österreich. Eine Annäherung auf Basis wirtschaftsstatistischer Analysen. ***https://www.bmk.gv.at/dam/jcr:abf0cdc3-bd4c-4335-aef9-8e5b0a33c119/ai_potenzial_oesterreich.pdf***.

Boddington, P., 2017, *Towards a code of ethics for artificial intelligence*: Springer.

Bundesministerium der Justiz, 2021, Pressemitteilung: Zentrum für vertrauenswürdige Künstliche Intelligenz (KI) soll Verbraucherinteressen in der digitalen Welt stärken., ***https://www.datev-magazin.de/nachrichten-steuern-recht/recht/zentrum-fuer-vertrauenswuerdige-kuenstliche-intelligenz-ki-soll-verbraucherinteressen-in-der-digitalen-welt-staerken-65216***.

Caruana, R., Lundberg, S., Ribeiro, M. T., Nori, H. und Jenkins, S., 2020, Intelligible and explainable machine learning: Best practices and practical challenges, *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Čas, J. und Krieger-Lamina, J., 2020, KI und Arbeitswelt, in: Christen M.; Mader C.; Cas J.; Abou-Chadi T.; Bernstein A.; BraunBinder N.; Dell'Aglio D.; Fábián L.; George D.; Gohdes, A. (Hg.): *Wenn Algorithmen für und entscheiden: Chancen und Risiken der künstlichen Intelligenz*, Zürich: vdf, 144-164.

Čas, J., Rose, G. und Schüttler, L., 2017, *Robotik in Österreich: Kurzbericht – Entwicklungsperspektiven und politische Herausforderungen*, Wien: Institut für Technikfolgen-Abschätzung, ***http://epub.oeaw.ac.at/ita/ita-projektberichte/2017-03.pdf***.

Castelvecchi, D., 2016, Can we open the black box of AI?, *Nature News 538(7623)*, 20. ***https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731***.

Chatila, R. und Havens, J. C., 2019, The ieee global initiative on ethics of autonomous and intelligent systems: *Robotics and Well-Being*: Springer, 11-16.

Clarke, R., 1993, Asimov's laws of robotics: implications for information technology-Part I, *Computer Band 26 (Ausgabe 12)*, 53–61.

Clarke, R., 1994, Asimov's laws of robotics: Implications for information technology. 2, *Computer Band 27 (Ausgabe 12)*, 57–66.

Clinciu, M.-A. und Hastie, H., 2019, A survey of explainable AI terminology, *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*.

Corbett-Davies, S., Pierson, E., Feller, A. und Goel, S., 2016, A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear., *Washington Post*, ***https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/***.

Craven, M. W., 1996, *Extracting comprehensible models from trained neural networks*: The University of Wisconsin-Madison.

Dachwitz, I., Laufer, D. und Meineck, S., 2020, Gesichtserkennung ist eine Waffe, *netzpolitik.org*, ***https://netzpolitik.org/2020/npp-204-pimeyes-gesichtserkennung-ist-eine-waffe/***.

Deeks, A., 2019, The judicial demand for explainable artificial intelligence, *Columbia Law Review 119(7)*, 1829-1850.

Diakopoulos, N., 2015, Algorithmic accountability: Journalistic investigation of computational power structures, *Digital journalism 3(3)*, 398-415.

Dick, S., 2019, Artificial intelligence.

Dieter, B. und Birnbacher, W., 2016, Automatisiertes Fahren, *Dieter und Wolfgang Birnbacher über ethische Fragen an der Schnittstelle von Technik und Gesellschaft, Information Philosophie 4*, 8-15.

Doshi-Velez, F. und Kim, B., 2017, Towards a rigorous science of interpretable machine learning, *arXiv preprint arXiv:1702.08608*.

Došilović, F. K., Brčić, M. und Hlupić, N., 2018, Explainable artificial intelligence: A survey, *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*.

Dressel, J. und Farid, H., 2018, The accuracy, fairness, and limits of predicting recidivism, *Science Advances 4*. ***https://www.science.org/doi/epdf/10.1126/sciadv.aao5580***.

Etzioni, A. und Etzioni, O., 2017, Incorporating ethics into artificial intelligence, *The Journal of Ethics 21(4)*, 403-418; auch veröffentlicht in: The Journal of Ethics.

Europäische Kommission, 2018, *Mitteilung der Kommission an das europäische Parlament, den europäischen Rat, den Rat, den europäischen wirtschafts- und sozialausschuss und den Ausschuss der Regionen, Künstliche Intelligenz für Europa* 25.4.2018, ***https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52018DC0237&from=EN***.

Europäische Kommission, 2021, *Vorschlag für eine Verordnung des europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union*, 21.04.2021, ***https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_2&format=PDF***.

European Commission, 2018, *The European Artificial Intelligence Landscape*, 2018, ***https://ec.europa.eu/jrc/communities/sites/jrccties/files/reportontheeuropeanailandscapeworkshop.pdf***.

Falkner, G., 2022, Digitale Demokratie oder Digitale Diktatur? Warum disziplinäre Perspektiven verknüpft werden sollten, in: Bogner, A., Decker, M., Nentwich, M. und Scherz, C. (Hg.): *Digitalisierung und die Zukunft der Demokratie. Beiträge aus der Technikfolgenabschätzung*, Berlin: Nomos, 173-188.

Fernandez, A., Herrera, F., Cordon, O., del Jesus, M. J. und Marcelloni, F., 2019, Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?, *IEEE Computational intelligence magazine 14(1)*, 69-81.

Floridi, L. und Cowls, J., 2021, A unified framework of five principles for AI in society: *Ethics, Governance, and Policies in Artificial Intelligence*: Springer, 5-17.

Frey, C. B. und Osborne, M. A., 2013, *The Future of Employment: how susceptible are Jobs to Computerisation?*, im Auftrag von: Oxford Martin School, U. o. O., Oxford: Oxford Martin School, University of Oxford, ***https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf***.

Future of Life Institute, 2017, Asilomar AI Principles, ***https://futureoflife.org/ai-principles/***.

Gershgorn, D., 2017, *AI is now so complex its creators can't trust why it makes decisions*, ***https://qz.com/1146753/ai-is-now-so-complex-its-creators-cant-trust-why-it-makes-decisions/***.

Geuter, J., 2018, Nein, Ethik kann man nicht programmieren, *Zeit Online*, ***https://www.zeit.de/digital/internet/2018-11/digitalisierung-mythen-kuenstliche-intelligenz-ethik-juergen-geuter?utm_referrer=https%3A%2F%2Fwww.startpage.com%2F***.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. und Kagal, L., 2018, Explaining explanations: An overview of interpretability of machine learning, *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*.

Gleicher, M., 2016, A framework for considering comprehensibility in modeling, *Big data 4(2)*, 75-88.

GMAR, 2022, *Die österreichische Gesellschaft für Mess-, Automatisierungs- und Robotertechnik*; 2022], ***http://www.gmar.at***.

Goldsmith, J. und Burton, E., 2017, Why teaching ethics to AI practitioners is important, *Proceedings of the... AAAI Conference on Artificial Intelligence*, 2017.

Goodman, B. und Flaxman, S., 2017, European Union regulations on algorithmic decision-making and a „right to explanation", *AI magazine 38(3)*, 50-57.

Greene, D., Hoffmann, A. L. und Stark, L., 2019, Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning, *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. und Pedreschi, D., 2018, A survey of methods for explaining black box models, *ACM computing surveys (CSUR) 51(5)*, 1-42.

Gunning, D. und Aha, D., 2019, DARPA's explainable artificial intelligence (XAI) program, *AI Magazine 40(2)*, 44-58.

Ha, T., Lee, S. und Kim, S., 2018, Designing explainability of an artificial intelligence system: *Proceedings of the Technology, Mind, and Society*, 1-1.

Hacker, P., Krestel, R., Grundmann, S. und Naumann, F., 2020, Explainable AI under contract and tort law: legal incentives and technical challenges, *Artificial Intelligence and Law*, 1-25.

Hagendorff, T., 2020, The ethics of Ai ethics: An evaluation of guidelines, *Minds and Machines*, 1-22; auch veröffentlicht in: Minds and Machines.

Heer, J., 2018, The partnership on AI, *AI Matters 4(3)*, 25-26; auch veröffentlicht in: AI Matters.

Helbing, D., Beschorner, T., Frey, B., Diekmann, A., Hagendorff, T., Seele, P., Spiekermann, S., van den hoven, J. und Zwitter, A., 2021, Angesichts von Triage und „Todesalgorithmen": Ist die heutige daten-getriebene Medizin mit der Verfassung vereinbar?

Hellström, T., 2013, On the moral responsibility of military robots, *Ethics and information technology 15(2)*, 99-107.
Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I. und Ramkumar, P. N., 2020, Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions, *Current reviews in musculoskeletal medicine 13(1)*, 69-76.
Herm, L.-V., Wanner, J., Seubert, F. und Janiesch, C., 2021, I Don't Get It, but It Seems Valid! The Connection Between Explainability and Comprehensibility in (X)AI Research, *European Conference on Information Systems*, Marrakech.
Hill, K., 2020, The Secretive Company That Might End Privacy as We Know It, *The New York Times*, New York, ***https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html***.
HLEG AI, 2019, A definition of AI: Main capabilities and scientific disciplines, *URL: **https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf***.
Howard, M., 2020, Is diplomacy an Option? Finance Theory and Brexit.
ISO/IEC JTC 1, 2015, ISO 2382:2015 Information technology – Vocabulary, ***https://iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en:term:2123770***.
ITA, 2019, AMS-Algorithmus am Prüfstand. ITA-Dossier Nr. 43 (Juli 2019; AutorInnen: Doris Allhutter, Fabian Fischer, Astrid Mager), *Institut für Technikfolgen-Abschätzung (ITA)*, Wien, ***http://epub.oeaw.ac.at/ita/ita-dossiers/ita-dossier043.pdf***.
ITA, 2021, *Wie fair ist der AMS-Algorithmus? ITA-Dossier Nr. 52 (Jänner 2021; AutorInnen: Astrid Mager, Doris Allhutter)*; in Reihe: Institut für Technikfolgen-Abschätzung (ITA), Wien, ***http://epub.oeaw.ac.at/ita/ita-dossiers/ita-dossier052.pdf***.
Izumo, T. und Weng, Y.-H., 2021, Coarse ethics: how to ethically assess explainable artificial intelligence, *AI and Ethics*, 1-13.
Jobin, A., 2020, Ethische Künstliche Intelligenz – von Prinzipien zu Prozessen, in: Hengstschläger, M. (Hg.): *Digitaler Wandel und Ethik*: Ecowin, 144-159, ***https://www.nature.com/articles/s42256-019-0088-2***.
Jobin, A., Ienca, M. und Vayena, E., 2019, The global landscape of AI ethics guidelines, *nature – machine intelligence*. ***https://www.nature.com/articles/s42256-019-0088-2***.
Knobloch, T. und Hustedt, C., 2019, *Der maschinelle Weg zum passenden Personal – Zur Rolle algorithmischer Systeme in der Personalauswahl*: Stiftung Neue Verantwortung, Bertelsmann Stiftung.
Krafft, T. und Zweig, K., 2019, Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse, *Ein Regulierungsvorschlag*.
Krüger, J. und Lischka, K., 2018, *Damit Maschinen den Menschen dienen – Lösungsansätze, um algorithmische Prozesse in den Dienst der Gesellschaft zu stellen*, Gütersloh: Bertelsmann Stiftung.
Legg, S. und Hutter, M., 2007, Universal Intelligence: A Definition of Machine Intelligence, *Minds and machines (Dordrecht) 17(4)*, 391-444.
Lewandowski, D., 2014, Die Macht der Suchmaschinen und ihr Einfluss auf unsere Entscheidungen, *Information – Wissenschaft & Praxis 65*.
Lipton, Z. C., 2018, The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery, *Queue 16(3)*, 31-57.
Lyon, D. (Hg.), 2003, *Surveillance as Social Sorting. Privacy, Risk and Digital Discrimination*, London: Routledge.
Mager, A., 2014a, Defining algorithmic ideology: Using ideology critique to scrutinize corporate search engines, *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society 12(1)*, 28-39.
Mager, A., 2014b, Ideologie des Algorithmus. Wie der neue Geist des Kapitalismus Suchmaschinen formt, in: Stark B., Dörr D. und Aufenanger S. (Hg.): *Die Googleisierung der Informationssuche. Suchmaschinen zwischen Nutzung und Regulierung* Berlin/Boston: De Gruyter, 201-223.
Michalski, R. S., 1983, A theory and methodology of inductive learning: *Machine learning*: Elsevier, 83-134.
Mittelstadt, B., 2019, Principles alone cannot guarantee ethical AI, *Nature Machine Intelligence 1(11)*, 501-507.
Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. und Floridi, L., 2016, The ethics of algorithms: Mapping the debate, *Big Data & Society 3(2)*.
Mohri, M., Rostamizadeh, A. und Talwalkar, A., 2018, *Foundations of machine learning*: MIT press.

Montavon, G., Samek, W. und Müller, K.-R., 2018, Methods for interpreting and understanding deep neural networks, *Digital Signal Processing 73*, 1-15.

Moore, J. D. und Swartout, W. R., 1988, *Explanation in expert systemss: A survey*: UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Mori, T. und Uchihira, N., 2019, Balancing the trade-off between accuracy and interpretability in software defect prediction, *Empirical Software Engineering 24(2)*, 779-825.

Müller-Eiselt, R. und Lischka, K., 2018, Vorwort, in: Zweig, K. A. (Hg.): *Wo Maschinen irren können – Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung*: Bertelsmann Stiftung, 6.

Nentwich, M., Weber, M., Appelt, D., Capari, L., Gudowsky, N., Ornetzeder, M., Peissl, W., Buchinger, E., Filippova, E., Heller-Schuh, B., Kienegger, M., Kubeczko, K., Schaper-Rinkel, P., Wang, A. und Wasserbacher, D., 2021, *Foresight und Technikfolgenabschätzung: Monitoring für das Österreichische Parlament, neue Themen, Ausgabe November 2021*, Array, Wien, ***http://epub.oeaw.ac.at/ita/ita-projektberichte/ITA-AIT-15.pdf***.

Nurski, L., 2021, Algorithmic management is the past, not the future of work, *Dostupno na: **https://www.bruegel.org/2021/05/algorithmic-management-is-the-past-not-the-future-of-work/***.

O'Neil, C., 2016, *Weapons of math destruction: How big data increases inequality and threatens democracy*: Crown.

OECD, 2020, *STIP Compass database*, 2020: powered by EC/OECD (2020), ***http://oecd.ai***.

Oxford Commission on AI & Good Governance, 2021, *AI in the Public Service: From Principles to Practice*, im Auftrag von: Oxford Internet Institute, O. U., 12.2021, ***https://oxcaigg.oii.ox.ac.uk/wp-content/uploads/sites/124/2021/12/AI-in-the-Public-Service-Final.pdf***.

Pariser, E., 2011, *The filter bubble: What the Internet is hiding from you*: Penguin UK.

Peissl, W. und Krieger-Lamina, J., 2017, *The Scored Consumer – Privacy and Big Data*, hg. v. Bala, C. und Schuldzinski, W., Düsseldorf: Verbraucherzentrale Nordrhein-Westfalen e.V.

Prates, M., Avelar, P. und Lamb, L. C., 2018, On quantifying and understanding the role of ethics in AI research: A historical account of flagship conferences and journals, *arXiv preprint arXiv:1809.08328*; auch veröffentlicht in: arXiv preprint arXiv:1809.08328.

Preece, A., Harborne, D., Braines, D., Tomsett, R. und Chakraborty, S., 2018, Stakeholders in explainable AI, *arXiv preprint arXiv:1810.00184*.

Rohde, N., 2017, In Australien prüft eine Software die Sozialbezüge und erfindet Schulden für 20.000 Menschen, ***https://algorithmenethik.de/2017/10/25/in-australien-prueft-eine-software-die-sozialbezuege-und-erfindet-schulden-fuer-20-000-menschen/***.

Rohde, N., 2018, *Gütekriterien für algorithmische Prozesse – Eine Stärken- und Schwächenanalyse ausgewählter Forderungskataloge*: Bertelsmann Stiftung.

Scarlett, C., 2017, *The Future of Law: Artificial Intelligence?*, ***https://knowledge-leader.colliers.com/colin-scarlett/future-law-artificial-intelligence/***.

Schaber, F., Krieger-Lamina, J. und Peissl, W., 2019, *Digitale Assistenten – Endbericht*, Array, Wien: Institut für Technikfolgen-Abschätzung (ITA), ***https://epub.oeaw.ac.at/ita/ita-projektberichte/2019-01.pdf***.

Schaber, F., Strauß, S. und Peissl, W. (ITA), 2020, *Der Körper als Schlüssel? – Biometrische Methoden für Konsument*innen*, November 2020, Wien: Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften, ***https://epub.oeaw.ac.at/ita/ita-projektberichte/2020-03.pdf***.

Schneier, B., 2021, Invited Talk: The Coming AI Hackers, *International Symposium on Cyber Security Cryptography and Machine Learning*.

Shapiro, A., 2017, Reform predictive policing, *Nature news 541(7638)*, 458.

Siegetsleitner, A., 2020, Who Bears Moral Responsibility in the Case of Autonomous Artifical Intelligence?: *Digital Transformation and Ethics*: Hengstschläger, Markus/Austrian Council for Research and Technology Development (eds.), 118–133.

Stolton, S., 2020, Gesichtserkennung: EU-Datenschutzagentur will Kommission von Verbot überzeugen, *EURACTIV.com*, ***https://www.euractiv.de/section/digitale-agenda/news/gesichtserkennung-eu-datenschutzagentur-will-kommission-von-verbot-ueberzeugen/***.

Tomsett, R., Braines, D., Harborne, D., Preece, A. und Chakraborty, S., 2018, Interpretable to whom? A role-based model for analyzing interpretable machine learning systems, *arXiv preprint arXiv:1806.07552*.

Tsakalakis, N., Stalla-Bourdillon, S., Carmichael, L., Huynh, T. D., Moreau, L. und Helal, A., 2021, The dual function of explanations: Why it is useful to compute explanations, *Computer Law & Security Review 41*, 105527.

Vakkuri, V. und Abrahamsson, P., 2018, The key concepts of ethics of artificial intelligence, *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 2018.

Van Lent, M., Fisher, W. und Mancuso, M., 2004, An explainable artificial intelligence system for small-unit tactical behavior, *Proceedings of the national conference on artificial intelligence*.

Van Roy, V., Rossetti, F., Perset, K. und Galindo-Romero, L., 2021, *AI Watch-National strategies on Artificial Intelligence: A European perspective*: Joint Research Centre (Seville site).

Vilone, G. und Longo, L., 2020, Explainable artificial intelligence: a systematic review, *arXiv preprint arXiv:2006.00093*.

Weber, K. und Zoglauer, T., 2019, Maschinenethik und Technikethik: *Handbuch Maschinenethik*: Springer, 145-163.

Weller, A., 2017, Challenges for transparency.

Weller, A., 2019, Transparency: motivations and challenges: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*: Springer, 23-40.

West, D. M., 2018, *The future of work: Robots, AI, and automation*: Brookings Institution Press.

Wirtz, B. W., Weyerer, J. C. und Geyer, C., 2019, Artificial intelligence and the public sector—applications and challenges, *International Journal of Public Administration 42(7)*, 596-615.

Wirtz, B. W., Weyerer, J. C. und Sturm, B. J., 2020, The dark sides of artificial intelligence: An integrated AI governance framework for public administration, *International Journal of Public Administration 43(9)*, 818-829.

Yong, E., 2018, A Popular Algorithm Is No Better at Predicting Crimes Than Random People, *The Atlantic*, ***https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/***.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R. und Yang, Q., 2018, Building ethics into artificial intelligence, *arXiv preprint arXiv:1812.02953*; auch veröffentlicht in: arXiv preprint arXiv:1812.02953.

Yu, R. und Alì, G. S., 2019, What's inside the Black Box? AI Challenges for Lawyers and Researchers, *Legal Information Management 19(1)*, 2-13.

Zuber, N., Kacianka, S., Pretschner, A. und Nida-Rümelin, J., 2020, Ethische Deliberation für agile Softwareprozesse: EDAP-Schema, in: Hengstschläger, M. (Hg.): *Digital Transformation and Ethics*, 2020. Aufl.: Digital Transformation ***https://www.bidt.digital/wp-content/uploads/2021/04/Digital-Transformation-and-Ethics_Zuber-et-al_EN.pdf***.

Zweig, K. A., 2018, *Wo Maschinen irren können – Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung*: Bertelsmann Stiftung.

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

GLOSSAR

Algorithmus
Eine Formel oder ein Satz von Regeln (oder Verfahren, Prozessen oder Anweisungen) zur Lösung eines Problems oder zur Durchführung einer Aufgabe. Im Bereich der künstlichen Intelligenz gibt der Algorithmus der Maschine Anweisungen, wie sie Antworten auf eine Frage oder Lösungen für ein Problem finden kann. Beim maschinellen Lernen verwenden die Systeme viele verschiedene Arten von Algorithmen. Gängige Beispiele sind Entscheidungsbäume, Clustering-Algorithmen, Klassifizierungsalgorithmen oder Regressionsalgorithmen.

Algorithmische Voreingenommenheit (Algorithmic Bias)
Wenn die Trainingsdaten fehlerhaft sind, dann wird der Algorithmus auch fehlerhafte Ergebnisse liefern. Ein System ist nur so gut wie die Daten, aus denen es lernt, und die Datenbanken müssen größer werden, damit die KI sich weiterentwickeln kann. Vgl. auch Ethik der KI.

Allgemeine Künstliche Intelligenz oder „Starke KI" (AKI oder Artificial General Inteligence ANI)
im Gegensatz zu schwacher Intelligenz, auch bekannt als vollständige, starke, Superintelligenz, Maschinenintelligenz auf menschlichem Niveau, bezeichnet die Fähigkeit einer Maschine, die erfolgreich alle Aufgaben auf intellektuelle Weise wie der Mensch ausführen kann. Künstliche Superintelligenz ist ein Begriff, der sich auf den Zeitpunkt bezieht, an dem die Fähigkeiten von Computern den Menschen übersteigen werden.

Autonomie
Im Bereich der KI und der Robotik bezieht sich der Begriff „autonom" auf die Fähigkeit eines künstlichen Agenten, unabhängig von menschlicher Führung zu handeln. Es wird davon ausgegangen, dass der Agent ein festes Ziel oder eine Nutzenfunktion hinsichtlich der Angemessenheit seiner Handlungen hat.

Beaufsichtigtes Lernen
Die Aufgabe des maschinellen Lernens, eine Funktion zu lernen, die eine Eingabe auf eine Ausgabe abbildet, basierend auf beispielhaften Input-Output-Paaren.

Bias
Induktive Voreingenommenheit (Inductive Bias): die Annahmen, die der Lernende bei der Vorhersage von Outputs aufgrund von Inputs verwendet, die er noch nicht kennengelernt hat.
Bestätigungsverzerrung (Confirmation Bias): die Tendenz, Informationen so zu suchen, zu interpretieren, zu bevorzugen und abzurufen, dass die eigenen Überzeugungen oder Hypothesen bestätigt werden, während Informationen, die diesen widersprechen, unverhältnismäßig weniger Aufmerksamkeit geschenkt wird.

Black-Box KI
bedeutet, dass KI auf der Grundlage eines Datensatzes zu Erkenntnissen gelangt, ohne dass der Endnutzer weiß, wie. Programme für maschinelles Lernen ziehen Schlussfolgerungen aus den eingegebenen Daten, aber es ist nicht klar, wie das Programm zu ihnen gekommen ist. Deep-Learning-Algorithmen verwenden oft einen Black-Box-Ansatz. Diese neuronalen Netze können so komplex sein, dass Menschen die Ergebnisse nicht erklären können, selbst wenn sie sich als richtig erweisen. Sie können einige der bahnbrechendsten Ergebnisse aller KI-Typen erzielen, aber selbst ihre Entwickler wissen nicht, wie.

Comprehensibility (Verständichkeit der KI)
ist ein weiterer Begriff, der häufig anstelle von Verständlichkeit verwendet wird. Der Begriff ist definiert als die Fähigkeit eines Lernalgorithmus, das gelernte Wissen in einer für den Menschen verständlichen Weise darzustellen.

Confirmation Bias
die Tendenz, Informationen so zu suchen, zu interpretieren, zu bevorzugen und abzurufen, dass die eigenen Überzeugungen oder Hypothesen bestätigt werden, während Informationen, die diesen widersprechen, unverhältnismäßig weniger Aufmerksamkeit geschenkt wird.

Daten
sind eine Sammlung von qualitativen und quantitativen Variablen. Sie enthalten die Informationen, die numerisch dargestellt werden und analysiert werden müssen.

Deep Learning
ist ein Teilbereich des maschinellen Lernens, der sich mit Algorithmen befasst, die vom menschlichen Gehirn inspiriert sind, das auf hierarchische Weise arbeitet. Diese Daten werden verarbeitet, indem sie die Schichten eines neuronalen Netzes durchlaufen, um das gewünschte Ergebnis zu erhalten. Deep-Learning-Modelle, die meist auf (künstlichen) neuronalen Netzen beruhen, wurden in verschiedenen Bereichen wie Spracherkennung, Computersehen und Verarbeitung natürlicher Sprache eingesetzt.

Erklärbarkeit
Während sich Interpretierbarkeit im engeren Sinne auf das Verständnis der prinzipiellen Funktionsweise eines Systems, seiner Mechanik, bezieht, ohne notwendigerweise zu verstehen, warum, konzentriert sich die Erklärbarkeit darauf, wie das Modell eine Entscheidung getroffen hat. Die Erklärbarkeit wird daher oft mit Post-hoc-Prozessen in Verbindung gebracht. Sie wird mit dem Begriff der Erklärung als Schnittstelle zwischen Menschen und einem Entscheidungsträger in Verbindung gebracht. Erklärbare Modelle sind in der Lage, „die Gründe für das Verhalten neuronaler Netze zusammenzufassen, das Vertrauen der Nutzer*innen zu gewinnen oder Erkenntnisse über die Ursachen ihrer Entscheidungen zu liefern“

Erklärbare KI (oder Explainable artificial intelligence, XAI)
Künstliche Intelligenz, die so programmiert ist, dass sie dem Laien ihren Zweck, ihre Beweggründe und ihre Entscheidungsfindung beschreibt. Ethik-Befürworter drängen auf den verstärkten Einsatz von XAI, um mehr Transparenz und Fairness zu fördern und von „Black-Box-Algorithmen“ wegzukommen.

Ethische KI
ist künstliche Intelligenz, die sich an klar definierten ethischen Richtlinien bezüglich grundlegender Werte wie Menschenrechte, Privatsphäre, Nicht-Diskriminierung und Nicht-Manipulation hält. Ethische KI legt Wert auf ethische Aspekte bei der Bestimmung legitimer und illegitimer Anwendungen von KI

Die Ethik der KI
ist die Ethik der Technologie, die sich speziell mit Robotern und anderen künstlichen intelligenten Wesen befasst und sich in Roboterethik und Maschinenethik unterteilt. Erstere befasst sich mit dem moralischen Verhalten von Menschen, wenn sie KI-Systeme entwerfen, konstruieren und nutzen. Bei der letzteren geht es um das moralische Verhalten von KI-Agenten (→ **Algorithmische Voreingenommenheit**).

Interpretierbarkeit
die Funktionsweise eines KI-Systems in einer für einen Menschen verständlichen Weise zu beschreiben. Ein System ist interpretierbar, wenn es in dem Maße verstanden werden kann, in dem ein Mensch vorhersagen kann, was bei einer Änderung der Eingabe oder der algorithmischen Parameter passieren wird.

Künstliche Intelligenz (KI oder maschinelle Intelligenz)
bezieht sich auf Systeme, die intelligentes Verhalten zeigen, indem sie ihre Umgebung analysieren und – mit einem gewissen Grad an Autonomie – Maßnahmen ergreifen, um bestimmte Ziele zu erreichen. KI-basierte Systeme können rein softwarebasiert sein und in der virtuellen Welt agieren (z. B. Sprachassistenten, Bildanalysesoftware, Suchmaschinen, Sprach- und Gesichtserkennungssysteme) oder KI kann in Hardware-Geräte eingebettet sein (z. B. fortschrittliche Roboter, autonome Autos, Drohnen oder Anwendungen des Internets der Dinge). Der Begriff „KI” wurde formal erst 1956 verwendet.

Lernalgorithmus
Ein Lernalgorithmus ist ein Algorithmus, der beim maschinellen Lernen verwendet wird, um der Technologie zu helfen, den menschlichen Lernprozess zu imitieren. In Kombination mit Technologien wie neuronalen Netzen schaffen Lernalgorithmen komplexe, anspruchsvolle Lernprogramme.

Maschinelles Lernen (Machine Learning)
ist ein Bereich der Computerwissenschaft, in dem Computermodelle entwickelt werden, die die Fähigkeit haben, aus Daten zu „lernen“ und dann Vorhersagen zu treffen. Je nachdem, ob es ein übergeordnetes Signal gibt, kann maschinelles Lernen in drei Kategorien unterteilt werden: das beaufsichtigte Lernen, das unbeaufsichtigte Lernen und das verstärkende Lernen.

Neuronale Netzwerke (KNN)
Auch bekannt als künstliches neuronales Netz, neuronales Netz, tiefes neuronales Netz; ein Computersystem, das von lebenden Gehirnen inspiriert ist. Eine Architektur, die aus aufeinanderfolgenden Schichten einfacher verbundener Einheiten, so genannter künstlicher Neuronen, besteht, die mit nichtlinearen Aktivierungsfunktionen verwoben sind, was in gewisser Weise an die Neuronen in einem tierischen Gehirn erinnert.

Schwache Künstliche Intelligenz (oder Artificial Narrow Intelligence (ANI)
Auch als schwache oder angewandte Intelligenz bekannt, der Begriff steht für die meisten der derzeitigen KI Systeme, die sich in der Regel auf eine bestimmte Aufgabe konzentrieren. Schwache KI ist in der Regel viel besser als Menschen, wenn es um die Aufgabe geht, für die sie entwickelt wurde. Virtuelle Assistenten und AlphaGo sind Beispiele für künstliche Systeme mit schwacher Intelligenz.

Simulierbarkeit
Die Fähigkeit eines Modells, Nutzer*innen zu ermöglichen, seine Struktur und Funktionsweise vollständig zu verstehen. Ein Model ist simulierbar, wenn ein Mensch (für den die Interpretation gedacht ist) in der Lage ist, den gesamten Entscheidungsprozess intern zu simulieren und darüber nachzudenken (d. h. wie ein trainiertes Modell eine Ausgabe für eine beliebige Eingabe erzeugt).

Tiefes neuronales Netzwerk (DNN)
Eine neuronale Netzarchitektur mit vielen Schichten, normalerweise 5-100. Ein Netz mit nur wenigen Schichten wird als flaches neuronales Netzwerk bezeichnet.

Transparenz
kann auf vielfältige Weise definiert werden. Es gibt eine Reihe von benachbarten Konzepten, die manchmal als Synonyme für Transparenz verwendet werden – darunter „Erklärbarkeit“ (die KI-Forschung in diesem Bereich ist als „XAI“ bekannt), „Interpretierbarkeit“, „Verständlichkeit“ und „Blackbox“. Transparenz ist im Allgemeinen eine Eigenschaft einer Anwendung. Es geht darum, wie viel man über das Innenleben eines Systems „in der Theorie“ verstehen kann. Je nach konkreter Situation kann die genaue Bedeutung von „Transparenz“ variieren. Es ist eine offene wissenschaftliche Frage, ob es mehrere verschiedene Arten oder Typen von Transparenz gibt. Außerdem kann sich Transparenz auf unterschiedliche Dinge beziehen, wenn es beispielsweise darum geht, die rechtliche Bedeutung von ungerechtfertigten Verzerrungen zu analysieren oder sie in Bezug auf die Eigenschaften von maschinellen Lernsystemen zu diskutieren.

Unbeaufsichtigtes Lernen
ist eine Art von Algorithmus für maschinelles Lernen, der verwendet wird, um Schlussfolgerungen aus Datensätzen zu erstellen, die aus Eingabedaten ohne beschriftete Antworten bestehen, z. B. Clusteranalyse. Das bedeutet, dass das System einem völlig zufälligen und neuen Datensatz ausgesetzt wird und automatisch Muster und Beziehungen innerhalb dieses Datensatzes findet.

Verantwortungsvolle oder vertrauenswürdige KI
Es ist von zentraler Bedeutung, dass KI-Algorithmen grundlegende menschliche Werte berücksichtigen und ihre Analysen, Interpretationen und Entscheidungen auf zuverlässige und vertrauenswürdige Weise durchführen. Verantwortungsvolle KI entwickelt Werkzeuge, die wichtige Werte wie Verantwortlichkeit, Datenschutz, Sicherheit und Transparenz beachten, und führt ihre Operationen auf der Grundlage dieser Überlegungen aus. Zusammen mit erklärbarer KI ist dies ein Weg, KI in einer Weise zu entwickeln und einzusetzen, die menschliche Werte fördert.

White Box AI
macht transparent, wie sie zu ihren Schlussfolgerungen kommt. Ein Datenwissenschafler kann sich einen KI-Großrechner ansehen und verstehen, wie er sich verhält und welche Faktoren seine Entscheidungen beeinflussen.

Zerlegbarkeit
Der Grad, in dem ein Modell in seine einzelnen Komponenten (Input, Parameter und Output) zerlegt werden kann und deren intuitive Erklärbarkeit.

WWW.OEAW.AC.AT