4. Überlegungen zur Politikgestaltung

In diesem Kapitel werden Überlegungen zur Politikgestaltung erörtert, die es zu berücksichtigen gilt, damit Systeme der künstlichen Intelligenz (KI) vertrauenswürdig und menschenzentriert sind. Es behandelt Bedenken in Bezug auf Ethik und Fairness, die Achtung der Menschenrechte und der demokratischen Werte, darunter auch den Schutz der Privatsphäre. Außerdem geht es um die Gefahren, die mit der Übertragung bestehender Voreingenommenheiten bzw. Verzerrungen, sog. Biases, aus der analogen in die digitale Welt verbunden sind, insbesondere im Hinblick auf Geschlecht oder ethnische Herkunft. Betont wird die Notwendigkeit, Fortschritte auf dem Weg zu robusteren, sichereren und transparenteren KI-Systemen mit klaren Rechenschaftsmechanismen zu erzielen.

Die Förderung vertrauenswürdiger KI-Systeme erfolgt insbesondere über Politikmaßnahmen, die Anreize für Investitionen in verantwortungsvolle KI-Forschung und -Entwicklung setzen, die ein digitales Ökosystem ermöglichen, in dem der Schutz der Privatsphäre nicht durch einen breiteren Datenzugang infrage gestellt wird, die kleinen und mittleren Unternehmen Erfolgschancen eröffnen, die den Wettbewerb unterstützen und gleichzeitig geistiges Eigentum schützen und die Arbeitskräftemobilität bei sich verändernden Arbeitsmarktverhältnissen erleichtern.

Menschenzentrierte KI

Der Einfluss der künstlichen Intelligenz (KI) wächst zusehends. Je stärker sich diese Technologien durchsetzen, umso größere Auswirkungen haben ihre Vorhersagen, Empfehlungen oder Entscheidungen auf das Leben der Menschen. In Fachwelt, Wirtschaft und Politik wird aktiv untersucht, wie sich eine menschenzentrierte und vertrauenswürdige KI am besten gewährleisten lässt, wie ihr Nutzen maximiert und ihre Risiken minimiert werden können und wie ihre gesellschaftliche Akzeptanz erhöht werden kann.

Kasten 4.1. „Black-Box"-KI-Systeme lassen neue Herausforderungen entstehen

Neuronale Netze werden häufig als „Black Box" beschrieben. Obwohl das Verhalten solcher Systeme beobachtet werden kann, unterscheiden sie sich im Hinblick auf ihre Überwachbarkeit deutlich von bisherigen Technologien. Daher ist dieser Begriff durchaus passend. Neuronale Netze iterieren über die Daten, mit denen sie trainiert werden. Sie finden komplexe, probabilistische Korrelationen mit mehreren Variablen, die Teil des Modells werden, das sie aufbauen. Sie liefern jedoch keine Hinweise dazu, wie die Daten zusammenhängen könnten (Weinberger, 2018[1]). Die Daten sind viel zu komplex, als dass Menschen sie analysieren könnten. KI unterscheidet sich von früheren technologischen Entwicklungen u. a. durch folgende Merkmale, die sich auf Transparenz und Verantwortlichkeit auswirken:

- **Erkennbarkeit**: Regelbasierte Algorithmen können Regel für Regel gelesen und geprüft werden, sodass bestimmte Fehlerarten vergleichsweise einfach zu finden sind. Dagegen sind bestimmte Arten von Systemen des maschinellen Lernens (ML), insbesondere neuronale Netze, nur abstrakte mathematische Beziehungen zwischen Faktoren. Diese können äußerst komplex und schwer nachvollziehbar sein, selbst für diejenigen, die sie programmieren und trainieren (OECD, 2016).
- **Evolutiver Charakter**: Einige ML-Systeme sind iterativ und entwickeln sich im Lauf der Zeit weiter; sie können sogar ihr eigenes Verhalten auf unvorhergesehene Weise ändern.
- **Geringe Reproduzierbarkeit**: Es kann sein, dass ein ML-System nur dann eine bestimmte Prognose aufstellt oder eine bestimmte Entscheidung trifft, wenn bestimmte Bedingungen oder Daten vorliegen, welche nicht unbedingt reproduzierbar sind.
- **Zunehmende Zielkonflikte beim Schutz personenbezogener und sensibler Daten:**
  - **Inferenz**: Selbst wenn ihnen keine geschützten oder sensiblen Daten vorliegen, können KI-Systeme solche Informationen und Korrelationen aus Ersatzvariablen ableiten, die nicht personenbezogen oder sensibel sind, wie z. B. dem Einkaufsverhalten oder Standortdaten (Kosinski, M., D. Stillwell und T. Graepel, 2013[2]).
  - **Unerwünschte Proxy-Variablen**: Politische und technische Konzepte zum Schutz der Privatsphäre und zur Nichtdiskriminierung führen in der Regel dazu, dass die Datensammlung auf ein Minimum reduziert, die Verwendung bestimmter Daten verboten oder Daten gelöscht wurden, um ihre Verwendung zu verhindern. Ein KI-System könnte eine Prognose jedoch auf Ersatzvariablen

basieren, die zu den verbotenen und nicht gesammelten Daten in enger Verbindung stehen. Solche Proxys lassen sich nur erkennen, indem auch sensible oder personenbezogene Daten, beispielsweise zur ethnischen Zugehörigkeit, erhoben werden. Wenn solche Daten erfasst werden, muss sichergestellt werden, dass sie stets in angemessener Weise verwendet werden.

- **Das Datenschutz-Paradoxon:** Bei vielen KI-Systemen können mehr Trainingsdaten die Genauigkeit von KI-Prognosen verbessern und dazu beitragen, das Bias-Risiko aufgrund verzerrter Stichproben zu reduzieren. Je mehr Daten jedoch gesammelt werden, desto größer sind die Datenschutzrisiken für die betroffenen Personen.

Einige Arten von KI – die häufig als „Black Boxes“ beschrieben werden – bringen im Vergleich zu früheren technologischen Fortschritten neue Herausforderungen mit sich (Kasten 4.1). Daher hat die OECD – auf der Basis der Arbeit ihrer Sachverständigengruppe für KI (AIGO) – die wichtigsten Prioritäten für eine menschenzentrierte KI festgelegt: Erstens sollte sie zu inklusivem und nachhaltigem Wachstum und zur Lebensqualität beitragen. Zweitens sollte sie menschenzentrierte Werte und Fairness respektieren. Drittens sollten die Nutzung und die Funktionsweise von KI-Systemen transparent sein. Viertens sollten KI-Systeme robust und sicher sein. Fünftens sollte es eine Rechenschaftspflicht für die Ergebnisse von KI-Prognosen und die daraus resultierenden Entscheidungen geben. Bei Vorhersagen, bei denen besonders viel auf dem Spiel steht, werden diese Maßnahmen als entscheidend angesehen. Sie sind aber auch für betriebswirtschaftliche Empfehlungen oder KI-Anwendungsformen wichtig, die weniger starke Auswirkungen haben.

Inklusives und nachhaltiges Wachstum und Lebensqualität

KI verfügt über ein erhebliches Potenzial zur Förderung der Ziele für nachhaltige Entwicklung

KI kann zum Gemeinwohl und zur Erreichung der Ziele der Vereinten Nationen für nachhaltige Entwicklung (SDG) eingesetzt werden, u. a. in Bereichen wie Bildung, Gesundheit, Verkehr, Landwirtschaft und nachhaltige Städte. Viele öffentliche und private Organisationen, darunter die Weltbank, eine Reihe von Organisationen der Vereinten Nationen und die OECD, arbeiten daran, mithilfe von KI die Verwirklichung dieser Ziele voranzubringen.

Die Entwicklung gerechter und inklusiver künstlicher Intelligenz hat zunehmend Priorität

Die Entwicklung gerechter und inklusiver künstlicher Intelligenz hat zunehmend Priorität. Dies gilt insbesondere angesichts von Bedenken, dass KI Ungleichheit verstärken oder Unterschiede innerhalb und zwischen Industrie- und Entwicklungsländern vergrößern könnte. Solche Unterschiede sind auf die Konzentration von KI-Ressourcen – KI-Technologien, Kompetenzen, Datensätze und Rechenleistung – in einigen wenigen Unternehmen und Ländern zurückzuführen. Außerdem besteht die Sorge, dass KI Verzerrungen verfestigen könnte (Talbot et al., 2017[3]). So wird beispielsweise befürchtet, dass KI besondere Auswirkungen auf benachteiligte und unterrepräsentierte Bevölkerungsgruppen haben könnte, z. B. auf Menschen mit niedrigem Bildungsniveau, Geringqualifizierte, Frauen und ältere Menschen, insbesondere in Ländern der unteren und der mittleren Einkommensgruppe (Smith, M. und S. Neupane, 2018[4]). Das International

Development Research Centre in Kanada empfahl kürzlich die Einrichtung eines globalen Fonds für entwicklungsorientierte KI. Damit könnten in Ländern der unteren und der mittleren Einkommensgruppe Exzellenzzentren für künstliche Intelligenz eingerichtet werden, die die Gestaltung und Umsetzung evidenzbasierter, inklusiver Politikmaßnahmen unterstützen (Smith, M. und S. Neupane, 2018[4]). Ziel ist es sicherzustellen, dass durch KI entstehende Nutzeffekte gleichmäßig verteilt werden und zu gerechteren Gesellschaftsstrukturen führen. Mit inklusiven KI-Initiativen soll erreicht werden, dass die wirtschaftlichen Vorteile der KI möglichst breiten Bevölkerungskreisen zugutekommen und niemand den Anschluss verliert.

Inklusive und nachhaltige KI ist ein Bereich, dem Länder wie Indien (Indien, 2018[5]), Unternehmen wie Microsoft[1] und Hochschulinitiativen wie das Berkman Klein Center in Harvard besondere Aufmerksamkeit zuteilwerden lassen. So hat Microsoft beispielsweise Projekte wie Seeing AI lanciert, eine mobile Anwendung, die Menschen mit Sehbehinderung helfen soll. Diese Anwendung scannt und erkennt alle Elemente in der Umgebung der Person und stellt ihr eine Audiobeschreibung zur Verfügung. Microsoft investiert außerdem 2 Mio. USD in Initiativen, bei denen KI zur Sicherung der Nachhaltigkeit genutzt werden soll, z. B. zum Erhalt der Biodiversität oder im Kampf gegen den Klimawandel (Heiner, D. und C. Nguyen, 2018[6]).

Menschenzentrierte Werte und Fairness

Menschenrechte und Ethikrichtlinien

Ethische Normen sind in den internationalen Menschenrechtsbestimmungen verankert

Ethische Normen sind in den internationalen Menschenrechtsbestimmungen verankert. KI kann zur Gewährleistung der Menschenrechte beitragen, aber auch neue Risiken vorsätzlicher oder unbeabsichtigter Menschenrechtsverletzungen schaffen. Die Menschenrechtsbestimmungen und die an sie geknüpften rechtlichen und sonstigen institutionellen Strukturen könnten ein Instrument sein, um sicherzustellen, dass KI menschenzentriert ist (Kasten 4.2).

Kasten 4.2. Menschenrechte und KI

Bei den internationalen Menschenrechtsbestimmungen handelt es sich um verschiedene internationale Rechtsakte, etwa die Internationale Menschenrechtscharta,[1] sowie regionale Systeme zum Schutz der Menschenrechte, die in den letzten siebzig Jahren auf der ganzen Welt entwickelt wurden. Sie geben eine Reihe universeller Mindeststandards vor, die u. a. auf Werten wie Menschenwürde, Autonomie und Gleichheit beruhen und mit dem Prinzip der Rechtsstaatlichkeit im Einklang stehen. Aufgrund dieser Standards und der an sie geknüpften Rechtsgrundlagen sind Länder rechtlich verpflichtet, die Menschenrechte zu achten, zu schützen und zu gewährleisten. Sie verlangen zudem, dass Personen, denen Rechte verwehrt wurden oder deren Rechte verletzt wurden, Anspruch auf effektive Rechtsmittel haben.

Zu den Menschenrechten gehören das Recht auf Gleichheit, das Recht auf Nichtdiskriminierung, das Recht auf Vereinigungsfreiheit und das Vereinigungsrecht, das Recht auf Schutz der Privatsphäre und wirtschaftliche, soziale und kulturelle Rechte wie das Recht auf Bildung oder das Recht auf Gesundheit.

Neuere zwischenstaatliche Instrumente wie die „Leitprinzipien für Wirtschaft und Menschenrechte“ der Vereinten Nationen (OHCHR, 2011[7]) befassen sich auch mit der Rolle privater Akteure im Bereich der Menschenrechte. Schließlich sehen sie diese in der Verantwortung, die Menschenrechte zu achten. Darüber hinaus enthält die 2011 aktualisierte Version der *OECD-Leitsätze für multinationale Unternehmen* (OECD, 2011[8]) ein Kapitel über Menschenrechte.

Die Menschenrechte überschneiden sich mit weiterreichenden ethischen Fragen und anderen für die künstliche Intelligenz relevanten Regulierungsbereichen, wie etwa dem Schutz personenbezogener Daten oder dem Produktsicherheitsrecht. Allerdings haben diese anderen Anliegen und Fragen häufig eine andere Tragweite.

1. Die Internationale Menschenrechtscharta besteht aus der Allgemeinen Erklärung der Menschenrechte, dem Internationalen Pakt über bürgerliche und politische Rechte und dem Internationalen Pakt über wirtschaftliche, soziale und kulturelle Rechte.

KI könnte zur Förderung der Menschenrechte beitragen

Angesichts ihres potenziell breiten Anwendungs- und Nutzungsspektrums könnte KI den Schutz und die Gewährleistung der Menschenrechte verbessern. Zum Beispiel könnte KI eingesetzt werden, um Muster der Nahrungsmittelknappheit zur Hungerbekämpfung zu analysieren, medizinische Diagnosen und Behandlungen zu verbessern, das Angebot an Gesundheitsleistungen sowie den Zugang zur Gesundheitsversorgung auszuweiten und Diskriminierungen ans Licht zu bringen.

KI könnte Menschenrechte aber auch gefährden

Auf dem Gebiet der Menschenrechte kann KI aber auch eine Reihe von Herausforderungen mit sich bringen, was in Diskussionen über KI und Ethik häufig angesprochen wird. Bestimmte KI-Systeme könnten auf unbeabsichtigte Weise oder vorsätzlich Menschenrechte verletzen oder für Menschenrechtsverletzungen benutzt werden. Besonders über unbeabsichtigte Auswirkungen wird viel diskutiert. ML-Algorithmen, die Rückfallwahrscheinlichkeiten vorhersagen, können z. B. ein Bias aufweisen, das unerkannt bleibt. KI-Technologien können aber auch mit vorsätzlichen Menschenrechtsverletzungen in Verbindung gebracht werden. Beispiele hierfür sind der Einsatz von KI-Technologien, um politische Dissidenten ausfindig zu machen oder das Recht des Einzelnen auf freie Meinungsäußerung und politische Partizipation einzuschränken. In solchen Fällen ist der Verstoß selbst in der Regel nicht – oder nicht nur – auf die Nutzung von KI zurückzuführen. Allerdings könnte er durch die technische Raffinesse und Effizienz von KI verschärft werden.

Auch in Situationen, in denen die Auswirkungen auf die Menschenrechte unbeabsichtigt oder schwer zu erkennen sind, kann der Einsatz von KI besondere Herausforderungen darstellen. Dies kann auf die Verwendung qualitativ unzureichender Trainingsdaten, das Systemdesign oder komplexe Interaktionen zwischen dem KI-System und seiner Umgebung zurückzuführen sein. Ein Beispiel ist die algorithmusbedingte Verstärkung von Verhetzung oder Anstiftung zu Gewalt im Internet. Ein weiteres Beispiel ist die unbeabsichtigte Verbreitung sogenannte Fake News, die das Recht auf Teilhabe am politischen und öffentlichen Geschehen beeinträchtigen könnte. Welches Ausmaß und welche Auswirkungen die dadurch verursachten Schäden haben, hängt von der Tragweite der Entscheidungen des betreffenden KI-Systems ab. Zum Beispiel haben die Entscheidungen

eines KI-Systems, das Nachrichten empfiehlt, potenziell geringere Auswirkungen als Entscheidungen eines Algorithmus, der das Rückfallrisiko von zur Bewährung Verurteilten vorhersagt.

Menschenrechtsbestimmungen können durch KI-Ethikrichtlinien ergänzt werden

Mit Ethikrichtlinien kann dem Risiko begegnet werden, dass KI nicht menschenzentriert ist oder nicht mit menschlichen Werten im Einklang steht. Sowohl private Unternehmen als auch staatliche Stellen haben zahlreiche Ethikrichtlinien verabschiedet, die sich mit künstlicher Intelligenz befassen.

So hat beispielsweise das zu Google gehörende Unternehmen DeepMind im Oktober 2017 eine Ethik-Abteilung (DeepMind Ethics & Society) gegründet.[2] Diese Abteilung soll Technologieexperten helfen, die ethischen Implikationen ihrer Arbeit zu verstehen, und dazu beitragen, dass die Gesellschaft insgesamt besser entscheiden kann, wie KI für sie nutzbringend sein kann. Sie wird auch externe Forschung zu Themen wie algorithmische Verzerrungen, Zukunft der Arbeit und letale autonome Waffensysteme finanzieren. Google selbst hat eine Reihe von Ethikgrundsätzen aufgestellt, die seine Forschung, Produktentwicklung und Geschäftsentscheidungen leiten sollen.[3] Das Unternehmen hat ein Weißbuch zur KI-Governance veröffentlicht, in dem Fragen aufgezeigt werden, die in Zusammenarbeit mit Staat und Zivilgesellschaft geklärt werden müssen.[4] Microsoft verfolgt in Bezug auf KI die Vision, „die menschliche Genialität mit intelligenter Technologie zu verstärken“ (Heiner, D. und C. Nguyen, 2018[6]). Das Unternehmen hat Projekte ins Leben gerufen, die eine inklusive und nachhaltige Entwicklung gewährleisten sollen.

Die Menschenrechtsbestimmungen geben mit ihren institutionellen Mechanismen und ihrer globalen Architektur die Richtung vor und schaffen die Basis für eine ethische und menschenzentrierte Entwicklung und Nutzung von KI in der Gesellschaft.

Menschenrechtsbestimmungen sind im KI-Kontext ein wichtiges Instrument

Im KI-Kontext ist die Nutzung von Menschenrechtsbestimmungen insbesondere deshalb von Vorteil, weil sie es ermöglicht, sich auf etablierte Institutionen, eine bestehende Rechtsprechung und eine universelle Sprache zu stützen, und weil diese Bestimmungen große internationale Akzeptanz genießen:

- **Etablierte Institutionen**: Auf dem Gebiet der Menschenrechte ist im Lauf der Zeit eine umfassende internationale, regionale und nationale Infrastruktur entwickelt worden. Sie besteht aus zwischenstaatlichen Organisationen, Gerichten, Nichtregierungsorganisationen, wissenschaftlichen Einrichtungen und anderen Institutionen und Organen, in denen Menschenrechte geltend gemacht und Rechtsmittel eingelegt werden können.
- **Rechtsprechung**: Als Rechtsnormen werden die durch die Menschenrechte geschützten Werte in konkreten Situationen durch die Rechtsprechung und die Auslegungsarbeit internationaler, regionaler und nationaler Institutionen umgesetzt und rechtsverbindlich gemacht.
- **Universelle Sprache**: Die Menschenrechte bieten eine universelle Sprache für ein globales Thema. Zusammen mit der allgemeinen Menschenrechtsinfrastruktur kann sie einer größeren Zahl von Akteuren die Möglichkeit geben, an der Debatte über den Platz von KI in der Gesellschaft teilzuhaben. So kann diese Debatte über den Kreis der Akteure, die direkt an der KI-Entwicklung beteiligt sind, hinaus ausgedehnt werden.

- **Internationale Akzeptanz**: Die Menschenrechte genießen hohe internationale Akzeptanz und Legitimität. Bereits wenn der Eindruck entsteht, dass ein bestimmter Akteur Menschenrechte verletzt, kann dies für ihn erhebliche Folgen haben: Die Kosten des Reputationsverlusts sind hoch.

Ein an den Menschenrechten ausgerichteter KI-Ansatz kann dazu beitragen, Risiken, Prioritäten und benachteiligte Gruppen zu identifizieren und Abhilfe zu schaffen

- **Risikoerkennung**: Menschenrechtsbestimmungen können helfen, Schadensrisiken zu erkennen. Insbesondere bieten sie eine Grundlage, um menschenrechtlichen Sorgfaltspflichten (Due Diligence) nachzukommen, z. B. im Rahmen von Menschenrechtsverträglichkeitsprüfungen (Kasten 4.3).
- **Kernanforderungen**: Als Mindeststandards legen die Menschenrechte unantastbare Kernanforderungen fest. Zum Beispiel hilft die Menschenrechtsprechung, Regeln für die Meinungsäußerung in sozialen Netzwerken festzulegen: z. B. indem Verhetzung klar als eine Grenze gekennzeichnet wird, die nicht überschritten werden darf.
- **Ermittlung von Risikobereichen**: Menschenrechte können ein nützliches Instrument sein, um Bereiche oder Aktivitäten mit hohem Risiko zu erkennen. In diesen Bereichen ist dann erhöhte Vorsicht geboten oder muss u. U. auf die Nutzung von KI verzichtet werden.
- **Ermittlung benachteiligter Gruppen oder Gemeinden**: Menschenrechte können helfen, in Bezug auf KI benachteiligte oder gefährdete Gruppen oder Gemeinden zu ermitteln. Manche Personen oder Gruppen können z. B. aufgrund fehlender Möglichkeiten zur Nutzung von Smartphones unterrepräsentiert sein.
- **Abhilfe**: Als Rechtsnormen mit daraus erwachsenden Verpflichtungen können die Menschenrechte für diejenigen Abhilfe schaffen, deren Rechte verletzt werden. Beispiele für solche Abhilfemaßnahmen sind die Einstellung der Tätigkeit, die Entwicklung neuer Verfahren oder Maßnahmen, eine Entschuldigung oder eine Schadenersatzzahlung.

Kasten 4.3. Menschenrechtsverträglichkeitsprüfungen

Menschenrechtsverträglichkeitsprüfungen (Human Rights Impact Assessments – HRIA) können helfen, Risiken zu erkennen, die die am KI-Lebenszyklus beteiligten Akteure sonst möglicherweise übersehen würden. Bei solchen Prüfungen richtet sich der Blick stärker auf unbeabsichtigte Auswirkungen auf den Menschen als auf die Optimierung der Technologie oder ihrer Ergebnisse. Mit solchen Prüfungen oder ähnlichen Verfahren kann sichergestellt werden, dass die Menschenrechte während des gesamten Lebenszyklus der Technologien geachtet werden.

Bei Menschenrechtsverträglichkeitsprüfungen werden Technologien umfassend in Bezug auf ein breites Spektrum an Auswirkungen bewertet, die sie auf die Menschenrechte haben könnten, was recht ressourcenintensiv ist. Dabei kann es einfacher sein, mit dem KI-System selbst zu beginnen und sich nach außen vorzuarbeiten. So kann eine begrenzte Anzahl von Bereichen untersucht werden, in denen rechtliche Herausforderungen am

wahrscheinlichsten sind. Branchenverbände können bei der Durchführung von Menschenrechtsverträglichkeitsprüfungen in kleinen und mittleren Unternehmen (KMU) oder Nicht-Technologieunternehmen helfen, die KI-Systeme zwar einsetzen, sich mit solchen Technologien aber möglicherweise nicht genügend auskennen. In Bezug auf die Meinungsfreiheit und den Schutz der Privatsphäre ist die Global Network Initiative ein gutes Beispiel für solche Organisationen. Sie hilft Unternehmen, vorausschauend zu planen und Menschenrechtsverträglichkeitsprüfungen in ihre Pläne für neue Produkte einzubeziehen (https://globalnetworkinitiative.org/).

Menschenrechtsverträglichkeitsprüfungen haben den Nachteil, dass sie im Allgemeinen auf der Ebene der einzelnen Unternehmen durchgeführt werden. An KI-Systemen können jedoch viele Akteure beteiligt sein. Es kann daher ineffektiv sein, nur einen Teil dieser Akteure zu prüfen. Microsoft war das erste große Technologieunternehmen, das 2018 eine Menschenrechtsverträglichkeitsprüfung zu KI durchführte.

Auch die Umsetzung eines an den Menschenrechten ausgerichteten KI-Ansatzes ist mit erheblichen Herausforderungen verbunden. Dabei geht es z. B. darum, dass sich die Menschenrechtsbestimmungen an staatliche Instanzen richten, dass ihre Durchsetzung an eine Gebietshoheit geknüpft ist, dass sie eher geeignet sind, bei schweren Beeinträchtigungen einer geringen Zahl von Personen Abhilfe zu schaffen, und dass sie für Unternehmen kostspielig sein können:

- **Menschenrechtsbestimmungen richten sich an staatliche Instanzen, nicht an private Akteure.** Akteure des privaten Sektors spielen jedoch eine Schlüsselrolle bei der Erforschung, Entwicklung und Einführung von KI. Dies ist ein Problem, das nicht nur im KI-Bereich besteht. In mehreren zwischenstaatlichen Initiativen wird versucht, den Graben zwischen öffentlichem und privatem Sektor zu schließen. Außerdem setzt sich zunehmend die Erkenntnis durch, dass eine gute Menschenrechtsbilanz auch gut fürs Geschäft ist.[5]
- **Die Durchsetzung der Menschenrechte ist an die Gebietshoheit geknüpft**. Im Allgemeinen müssen Kläger nachweisen, dass sie in einem bestimmten Staat Klagebefugnis haben. In Fällen, in denen es um große multinationale Unternehmen und KI-Systeme geht, die in mehreren Staaten eingesetzt werden, ist dies möglicherweise schwierig.
- **Menschenrechtsbestimmungen eignen sich besser, um bei schweren Beeinträchtigungen kleiner Gruppen von Personen Abhilfe zu schaffen,** als wenn einer großen Gruppe von Menschen ein weniger erheblicher Schaden entstanden ist. Zudem können Menschenrechtsbestimmungen und ihre Strukturen auf Außenstehende undurchsichtig wirken.
- **Menschenrechtsbestimmungen stehen teilweise im Ruf, kostspielig für die Unternehmen zu sein.** Daher dürften Ansätze, die Ethik, Verbraucherschutz oder verantwortungsvolles unternehmerisches Handeln sowie wirtschaftliche Argumente für die Einhaltung der Menschenrechte in den Mittelpunkt rücken, besonders vielversprechend sein.

Einige der allgemeineren Herausforderungen, die im Zusammenhang mit KI bestehen, z. B. die Frage der Transparenz und der Nachvollziehbarkeit, wirken sich auch im Menschenrechtsbereich aus (vgl. Abschnitt „Transparenz und Nachvollziehbarkeit“). Ohne Transparenz ist es schwierig, festzustellen, wann Menschenrechte verletzt wurden, oder eine Beschwerde über eine Menschenrechtsverletzung zu begründen. Dasselbe gilt für die

Einlegung von Rechtsmitteln, die Bestimmung des ursächlichen Zusammenhangs und die Rechenschaftslegung.

Schutz personenbezogener Daten

KI stellt das Konzept der „personenbezogenen Daten" und der Einwilligung infrage

KI ist zunehmend in der Lage, verschiedene Datensätze miteinander zu verknüpfen und verschiedene Arten von Daten abzugleichen, was tiefgreifende Konsequenzen hat. Separat gespeicherte Daten (bzw. Daten, die von persönlichen Identifikatoren befreit, d. h. „de-identifiziert" wurden) galten früher nicht als personenbezogen. Mit KI können jedoch nicht personenbezogene Daten mit anderen Daten korreliert und bestimmten Personen zugeordnet werden, wodurch sie zu personenbezogenen Daten (bzw. „re-identifiziert") werden. Durch die algorithmische Korrelation verschwimmt die Grenze zwischen personenbezogenen und nicht personenbezogenen Daten. Nicht personenbezogene Daten können zunehmend verwendet werden, um Personen zu re-identifizieren oder über sie sensible Informationen abzuleiten, die über das hinausgehen, was diese Personen ursprünglich wissentlich preisgegeben hatten (Cellarius, 2017[9]). 2007 hatten Forscher z. B. bereits angeblich anonyme Daten verwendet, um die Liste der auf Netflix ausgeliehenen Filme mit den in der Internet Movie Database (IMDb) veröffentlichten Bewertungen zu verknüpfen. Auf diese Weise konnten sie die Personen identifizieren, die Filme ausgeliehen haben, und auf ihre vollständige Ausleihhistorie zugreifen. Da die Menge der erfassten Daten zunimmt und die Technologien immer ausgereifter werden, wird es zunehmend möglich sein, solche Verknüpfungen herzustellen. Damit wird es schwierig zu beurteilen, welche Daten tatsächlich als nicht personenbezogen betrachtet werden können und dies auch bleiben werden.

Die Unterscheidung zwischen sensiblen und nicht sensiblen Daten wird immer schwieriger, wie beispielsweise die Datenschutz-Grundverordnung (DSGVO) der Europäischen Union zeigt. Einige Algorithmen können sensible Daten aus „nicht sensiblen" Daten ableiten, z. B. kann der emotionale Zustand von Personen anhand der Art, wie sie auf ihrer Tastatur tippen, bestimmt werden (Privacy International and ARTICLE 19, 2018[10]). Die Verwendung von KI zur Identifikation oder Re-Identifikation von Daten, die ursprünglich nicht personenbezogen bzw. de-identifiziert waren, stellt auch ein rechtliches Problem dar. Verschiedene Texte befassen sich mit dem Schutz personenbezogener Daten, so z. B. die Empfehlung des Rats der OECD über Leitlinien für den Schutz des Persönlichkeitsbereichs und den grenzüberschreitenden Verkehr personenbezogener Daten (*Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, „Leitlinien zum Datenschutz") (Kasten 4.4). Es ist jedoch nicht klar, ob oder ab wann sie für Daten gelten, die unter bestimmten Umständen mit bestimmten Personen verknüpft werden oder verknüpft werden könnten (OVIC, 2018[11]). Eine extreme Auslegung könnte dazu führen, dass der Umfang des Datenschutzes erheblich ausgeweitet würde, was seine Anwendung erschweren würde.

Kasten 4.4. Die OECD-Leitlinien zum Datenschutz

Die Empfehlung des Rats über Leitlinien für den Schutz des Persönlichkeitsbereichs und den grenzüberschreitenden Verkehr personenbezogener Daten („Leitlinien zum Datenschutz“) wurde 1980 angenommen und 2013 aktualisiert (OECD, 2013[12]). Sie enthält Definitionen einschlägiger Begriffe. Insbesondere werden „personenbezogene Daten“ als „alle Informationen, die sich auf eine bestimmte oder bestimmbare Person (Datensubjekt) beziehen“ definiert. Außerdem legt sie Grundsätze fest, die bei der Verarbeitung personenbezogener Daten zu beachten sind. Dabei handelt es sich um die Grundsätze der begrenzten Datenerhebung (die gegebenenfalls nur mit Einwilligung als Mittel zur Gewährleistung dieses Grundsatzes erfolgen darf), der Datenqualität, der Zweckbestimmung, der Nutzungsbegrenzung, der Sicherung, der Offenheit, des Mitspracherechts und der Rechenschaftspflicht. Sie sehen auch vor, dass Mitgliedstaaten bei der Umsetzung der Leitlinien zum Datenschutz sicherzustellen haben, dass Datensubjekte nicht in unfairer Weise diskriminiert werden. Die Umsetzung der Leitlinien zum Datenschutz sollte 2019 überprüft werden, um u. a. den jüngsten Entwicklungen, auch im KI-Bereich, Rechnung zu tragen.

KI stellt auch die Datenschutzgrundsätze der begrenzten Datenerhebung, der Nutzungsbegrenzung und der Zweckbestimmung infrage

Zum Trainieren und Optimieren von KI-Systemen benötigen ML-Algorithmen große Datenmengen. Dadurch entstehen Anreize, die Datenerhebung zu maximieren, anstatt sie auf ein Minimum zu reduzieren. Mit der zunehmenden Nutzung von KI-basierten Geräten und dem Internet der Dinge (Internet of Things – IoT) werden immer mehr Daten erhoben. Die Datenerhebung erfolgt immer häufiger und ist insgesamt einfacher. Die erhobenen Daten werden zudem mit anderen Daten verknüpft, z. T. ohne Wissen oder Einwilligung der betroffenen Personen.

Die identifizierten Muster und die Abläufe des „Lernprozesses“ lassen sich nur schwer antizipieren. Daher kann der Umfang der Datenerhebung und -nutzung über das hinausgehen, was den Datensubjekten ursprünglich bekannt war, was sie selbst an Daten preisgegeben und wozu sie ihre Einwilligung gegeben hatten (Privacy International and ARTICLE 19, 2018[10]). Dies könnte mit den in den Leitlinien zum Datenschutz festgelegten Grundsätzen der begrenzten Datenerhebung, der Nutzungsbegrenzung und der Zweckbestimmung unvereinbar sein (Cellarius, 2017[9]). Die ersten beiden dieser Grundsätze beruhen z. T. auf der Einwilligung des Datensubjekts (wobei anerkannt wird, dass eine Einwilligung in einigen Fällen eventuell nicht möglich ist). Die Einwilligung ist die Basis für die Erhebung personenbezogener Daten oder ihre Nutzung zu anderen als den ursprünglich angegebenen Zwecken. KI-Technologien wie z. B. Deep Learning, deren Funktionsweise schwer nachvollziehbar und überwachbar ist, lassen sich den betroffenen Personen auch nur schwer erklären. Dies ist eine Herausforderung für die Unternehmen. Aus Unternehmenskreisen wurde verlautet, dass sich die Datenschutzgrundsätze angesichts des rasanten Tempos, mit dem KI Zugang zu Daten erhält, diese analysiert und nutzt, zunehmend schwer umsetzen lassen (OECD, 2018[13]).

Diese Schwierigkeiten verstärken sich durch die Verquickung von KI-Technologien und Entwicklungen im IoT-Bereich, d. h. der Anbindung einer immer größeren Zahl von Geräten und Gegenständen an das Internet. Die immer stärkere Verflechtung von KI- und IoT-Technologien (z. B. IoT-Geräte, die mit KI oder KI-Algorithmen zur Analyse von IoT-Daten ausgestattet sind) führt dazu, dass ständig mehr Daten, auch personenbezogene,

erhoben werden. Diese können zunehmend miteinander verknüpft und dann zusammen analysiert werden. Während die Zahl der Geräte wächst, die Daten sammeln (z. B. Überwachungskameras oder autonome Fahrzeuge), verbessert sich zugleich die KI-Technologie (z. B. Gesichtserkennung). Die Kombination dieser beiden Trends birgt die Gefahr eines stärkeren Eingriffs in die Privatsphäre als jeder dieser beiden Faktoren für sich genommen (OVIC, 2018[11]).

KI kann die Prinzipien der Beteiligung und Zustimmung des Einzelnen aber auch stärken

KI könnte den Datenschutz aber auch verbessern. Beispielsweise laufen in einer Reihe von technischen Normungsorganisationen Initiativen zum Aufbau von KI-Systemen nach den Grundsätzen *Privacy by design* (Datenschutz durch Technikgestaltung) und *Privacy by default* (Datenschutz durch datenschutzfreundliche Voreinstellungen). Die meisten dieser Organisationen verwenden und adaptieren Datenschutzrichtlinien, darunter auch die OECD-Leitlinien zum Datenschutz. Darüber hinaus wird KI eingesetzt, um den Nutzern personalisierte, auf ihre Bedürfnisse zugeschnittene Dienste anzubieten, die auf ihren im Lauf der Zeit erfassten persönlichen Datenschutzpräferenzen basieren (OVIC, 2018[11]). Diese Dienste können dem Einzelnen helfen, sich im Dickicht der Maßnahmen zur Verarbeitung personenbezogener Daten zurechtzufinden und sicherzustellen, dass seine Präferenzen insgesamt berücksichtigt werden. Auf diese Weise stärkt KI die informierte Zustimmung und Beteiligung des Einzelnen. Ein Forscherteam entwickelte z. B. Polisis, ein automatisiertes Framework, das mithilfe von Klassifikatoren neuronaler Netze Datenschutzrichtlinien analysiert (Harkous, 2018[14]).

Fairness und Ethik

ML-Algorithmen können implizit in ihren Trainingsdaten enthaltene Verzerrungen reproduzieren

Politikinitiativen im KI-Bereich befassen sich aktuell intensiv mit Fragen der Ethik, Fairness und/oder Gerechtigkeit. Es gibt erhebliche Bedenken, dass ML-Algorithmen in ihren Trainingsdaten implizit enthaltene Voreingenommenheiten, z. B. gegenüber bestimmten ethnischen Gruppen, und stereotype Vorstellungen übernehmen und reproduzieren. Da technologische Artefakte häufig gesellschaftliche Werte verkörpern, muss in der Debatte um Fairness deutlich gemacht werden, welcher Art von Gesellschaft die Technologien dienen sollen, wer Schutz genießen soll und welche Grundwerte dabei zu beachten sind (Flanagan, M., D. Howe und H. Nissenbaum, 2008[15]). Disziplinen wie Philosophie, Recht und Wirtschaft setzen sich seit Jahrzehnten aus unterschiedlichen Perspektiven mit unterschiedlichen Gerechtigkeitskonzepten auseinander. Sie verdeutlichen das breite Spektrum möglicher Interpretationen des Prinzips der Fairness und der daraus erwachsenden Implikationen für die Politik.

Philosophie, Recht und Informatik haben unterschiedliche Auffassungen von Fairness und ethischer KI

Die Philosophie befasst sich mit Konzepten von richtigem und falschem Verhalten, von Gut und Böse und von Moral. Im Kontext ethischer KI sind drei große philosophische Theorien relevant (Abrams et al., 2017[16]):

- Das **Konzept der menschlichen Grundrechte**, bei dem auf Immanuel Kant Bezug genommen wird, stellt auf die formalen Grundsätze der Ethik ab. Dabei geht es um

konkrete Rechte wie das Recht auf Schutz der Privatsphäre oder auf Freiheit. Diese Grundsätze werden durch Vorschriften geschützt, die KI-Systeme einhalten sollten.

- Der **utilitaristische Ansatz**, der von Jeremy Bentham und John Stuart Mill propagiert wurde, beruht auf dem Prinzip, dass staatliches Handeln das menschliche Wohlergehen auf der Grundlage wirtschaftlicher Kosten-Nutzen-Analysen maximieren soll. Für KI wirft der utilitaristische Ansatz die Frage auf, *wessen* Wohlergehen maximiert werden soll (das Wohlergehen des Einzelnen, der Familie, der Gesellschaft oder des Staats). Die Antwort auf diese Frage kann Auswirkungen auf das Design der Algorithmen haben.
- Der **Ansatz der Tugendethik**, der sich auf Aristoteles gründet, richtet den Blick auf die Werte und ethischen Normen, die eine Gesellschaft braucht, um die Menschen in ihren täglichen Bemühungen um ein lebenswertes Leben zu unterstützen. Dies wirft die Frage auf, welche Werte und welche ethischen Normen schützenswert sind.

In der Rechtswissenschaft werden Konzepte der Fairness häufig mit den Begriffen „Gleichheit" und „Gerechtigkeit" umrissen. Zwei zentrale Konzepte sind hier die individuelle Fairness und die Gruppenfairness.

- Die **Individuelle Fairness** bzw. Gerechtigkeit für den Einzelnen entspricht dem Konzept der Gleichheit vor dem Gesetz. Sie impliziert, dass alle Personen gleichbehandelt und nicht aufgrund besonderer Merkmale diskriminiert werden sollten. Gleichheit ist als internationales Menschenrecht anerkannt.
- Bei der **Gruppenfairness** geht es um die Fairness der Ergebnisse. Es gilt sicherzustellen, dass sich die Situation für Personen, die unterschiedlichen, durch bestimmte Merkmale (z. B. ethnische Herkunft oder Geschlecht) gekennzeichneten Gruppen angehören, im Ergebnis nicht systematisch anders darstellt. Dem liegt der Gedanke zugrunde, dass bestimmte Unterschiede und historische Umstände dazu führen können, dass verschiedene Gruppen auf bestimmte Situationen unterschiedlich reagieren. Um Gruppenfairness zu gewährleisten, werden unterschiedliche Ansätze verfolgt. Einer davon ist die positive Diskriminierung.

Entwickler von KI-Systemen haben darüber nachgedacht, wie man Fairness in KI-Systemen gewährleisten kann. Unterschiedliche Definitionen von Fairness schlagen sich dabei in verschiedenen Herangehensweisen nieder (Narayanan, 2018[17]):

- Das **Konzept der Unkenntnis**, bei dem ein KI-System keine identifizierbaren Faktoren kennen sollte, beruht auf dem rechtlichen Prinzip der individuellen Fairness. Das KI-System darf in diesem Fall keine Daten über sensible Attribute wie Geschlecht, ethnische Herkunft und sexuelle Orientierung berücksichtigen (Yona, 2017[18]). Allerdings können zahlreiche andere Faktoren mit dem geschützten Attribut (z. B. Geschlecht) korreliert sein. Zudem könnte sich die Entfernung dieser Daten negativ auf die Genauigkeit des KI-Systems auswirken.
- Beim Konzept der **kenntnisbasierten Fairness** wird Unterschieden zwischen bestimmten Gruppen Rechnung getragen. Ziel ist es, vergleichbare Personen gleich zu behandeln. Die Herausforderung besteht jedoch darin festzulegen, wer mit wem gleichbehandelt werden soll. Um zu verstehen, wer für eine bestimmte Aufgabe als vergleichbar anzusehen ist, bedarf es der Kenntnis sensibler Merkmale.

- Mit **Konzepten der Gruppenfairness** soll sichergestellt werden, dass die Ergebnisse für Personen, die verschiedenen Gruppen angehören, sich nicht systematisch unterscheiden. Es besteht die Befürchtung, dass KI-Systeme nicht fair sind und traditionelle Voreingenommenheiten verfestigen oder verstärken, da sie sich oft auf Datensätze aus der Vergangenheit stützen.

Unterschiedliche Auffassungen von Fairness führen für unterschiedliche gesellschaftliche Gruppen und unterschiedliche Akteure zu unterschiedlichen Ergebnissen. Nicht alle Ziele können gleichzeitig erreicht werden. Daher sollten bei technologischen Designentscheidungen, die sich nachteilig auf bestimmte Gruppen auswirken könnten, auch politische Erwägungen bzw. gegebenenfalls Entscheidungen berücksichtigt werden.

Im Personalwesen zeigen sich die Chancen und Risiken von KI besonders deutlich

Im Personalwesen können durch künstliche Intelligenz Verzerrungen bei der Einstellung von Mitarbeitern entweder verfestigt oder im Gegenteil aufgedeckt und verringert werden. Eine von Carnegie Mellon durchgeführte Studie, in der die Muster von Online-Stellenausschreibungen untersucht wurden, ergab z. B., dass eine Annonce für eine hoch bezahlte Führungsposition Männern 1 816 Mal, Frauen hingegen nur 311 Mal angezeigt wurde (Simonite, 2018[19]). Die Zusammenarbeit zwischen Mensch und KI ist daher sinnvoll, um sicherzustellen, dass KI-Anwendungen für Personaleinstellungen und -bewertungen transparent sind. Es muss gewährleistet sein, dass Verzerrungen nicht in den Code eingehen können, dass es also z. B. nicht möglich ist, dass Kandidaten aus bestimmten Kulturkreisen bei der Besetzung von Posten, die ihnen in der Vergangenheit nicht offenstanden, automatisch ausgeschlossen werden (OECD, 2017[20]).

Diskriminierung in KI-Systemen kann auf verschiedene Weise verringert werden

Um das Diskriminierungsrisiko in KI-Systemen zu verringern, wurden bereits verschiedene Ansätze vorgeschlagen, so z. B. Sensibilisierungsmaßnahmen, diversitätsfördernde organisatorische Maßnahmen, Normen, technische Lösungen zur Erkennung und Korrektur algorithmischer Verzerrungen sowie Selbstregulierungs- bzw. Regulierungsansätze. Im Bereich des Predictive Policing wurde beispielsweise die Einführung von Algorithmus-Folgenabschätzungen oder entsprechenden Erklärungen empfohlen. Dazu müsste die Polizei die Wirksamkeit, den Nutzen und die möglichen diskriminierenden Auswirkungen der verschiedenen technologischen Optionen, die ihr für die vorausschauende Polizeiarbeit zur Verfügung stehen, bewerten (Selbst, 2017[21]). Rechenschaftspflicht und Transparenz sind wichtig, um Fairness zu erreichen. Aber selbst zusammengenommen sind sie noch keine Garantie für Fairness (Weinberger, 2018[22]); (Narayanan, 2018[17]).

Anstrengungen zur Gewährleistung von Fairness in KI-Systemen können Zielkonflikte mit sich bringen

KI-Systeme sollen „fair“ sein. Ihre Vorhersagen sollen z. B. gewährleisten, dass nur Angeklagte, bei denen ein hohes Risiko besteht, nicht auf Kaution freigelassen werden oder dass einem Kreditkunden das je nach seiner Rückzahlungsfähigkeit günstigste Finanzierungsangebot gemacht wird. Bei **falsch-positiven Fehlern** wird irrtümlicherweise ein negativer Ausgang vorausgesagt. Zum Beispiel können KI-Systeme fälschlicherweise vorhersagen, dass ein Angeklagter erneut straffällig wird, der in Wirklichkeit keine neuen Straftaten begehen wird. Sie können auch fälschlicherweise eine Erkrankung vorhersagen, an der die untersuchte Person gar nicht leidet. **Falsch-negative Fehler** sind hingegen Fälle,

bei denen ein KI-System fälschlicherweise einen positiven Ausgang vorhersagt, beispielsweise dass ein Angeklagter nicht erneut straffällig wird oder dass ein Patient eine bestimmte Krankheit nicht hat, an der er in Wirklichkeit leidet.

Konzepte der Gruppenfairness versuchen der Tatsache Rechnung zu tragen, dass die Ausgangssituation nicht für alle Gruppen die gleiche ist. Sie berücksichtigen solche Unterschiede mathematisch, indem sie gewährleisten, dass die Richtigkeitsquote bzw. die Fehlerquote für alle Gruppen gleich hoch ist. Dies bedeutet z. B., dass der Anteil der fälschlicherweise als Wiederholungstäter eingestuften Personen unter den Männern gleich hoch sein muss wie unter den Frauen (oder dass die falsch-positiven und die falsch-negativen Ergebnisse einander ausgleichen).

Der Ausgleich von falsch-positiven und falsch-negativen Ergebnissen stellt eine Herausforderung dar. Falsch-negative Ergebnisse werden häufig als weniger wünschenswert und risikoreicher angesehen als falsch-positive Ergebnisse, weil sie kostspieliger sind (Berk, R. und J. Hyatt, 2015[23]). Zum Beispiel sind die Kosten, die einer Bank entstehen, wenn sie einer Person einen Kredit gewährt, die den Prognosen eines KI-Systems zufolge den Kredit zurückzahlen wird, ihren Zahlungsverpflichtungen dann aber nicht nachkommt, größer als der Gewinn aus diesem Kredit. Eine Person, die der Diagnose zufolge frei von einer bestimmten Krankheit ist, diese Krankheit aber hat, wird möglicherweise schwer leiden. Der Ausgleich von echt-positiven und echt-negativen Ergebnissen kann ebenfalls zu unerwünschten Ergebnissen führen: So könnte es z. B. geschehen, dass Frauen inhaftiert bleiben, die keine Sicherheitsbedrohung darstellen, weil gewährleistet werden soll, dass ein gleicher Anteil an Männern und Frauen aus der Haft entlassen wird (Berk, R. und J. Hyatt, 2015[23]). Einige Ansätze zielen darauf ab, sowohl falsch-positive als auch falsch-negative Ergebnisse auszugleichen. Allerdings ist es schwierig, gleichzeitig verschiedenen Konzepten von Fairness gerecht zu werden (Chouldechova, 2016[24]).

Politikverantwortliche sollten über den angemessenen Umgang mit sensiblen Daten im KI-Kontext nachdenken

Es könnte angebracht sein, erneut darüber nachzudenken, wie mit sensiblen Daten umzugehen ist. In einigen Fällen müssen Unternehmen möglicherweise sensible Daten speichern und verwenden, um sicherzustellen, dass ihre Algorithmen diese Daten nicht unbeabsichtigt rekonstruieren. Eine weitere Priorität der Politik ist die Überwachung unbeabsichtigter Rückkopplungseffekte. Wenn sich die Polizei beispielsweise in Viertel begibt, die von Algorithmen als Orte mit hohem Kriminalitätsrisiko eingestuft wurden, könnte dies zu einer verzerrten Datenerhebung führen und in der Folge die Voreingenommenheit des Algorithmus – und der Gesellschaft – gegenüber diesen Vierteln verstärken (O'Neil, 2016[25]).

Transparenz und Nachvollziehbarkeit

Transparenz ist beim Einsatz von KI und bei der Funktionsweise von KI-Systemen von entscheidender Bedeutung

Der Begriff „Transparenz" hat technisch und politisch gesehen nicht die gleiche Bedeutung. In der Politik bezieht sich Transparenz traditionell darauf, wie eine Entscheidung getroffen wird, wer an diesem Prozess beteiligt ist und welche Faktoren in die Entscheidung einfließen (Kosack, S. und A. Fung, 2014[26]). Transparenzmaßnahmen können hier darin

bestehen, offenzulegen, wie KI für Prognosen, Empfehlungen oder Entscheidungen eingesetzt wird. Gegebenenfalls könnte der Nutzer auch darauf aufmerksam gemacht werden, wenn er mit einem KI-System interagiert.

Für die Technologieexperten geht es bei der Transparenz eines KI-Systems in erster Linie um prozessbezogene Fragen. Menschen sollen verstehen können, wie ein KI-System entwickelt, trainiert und eingesetzt wird. Das kann auch bedeuten, dass Einblick in die Faktoren gegeben wird, die eine bestimmte Prognose oder Entscheidung beeinflussen. Um eine gemeinsame Nutzung von bestimmtem Code oder bestimmten Datensätzen geht es dabei in der Regel nicht. In vielen Fällen sind die Systeme zu komplex, als dass dies wirklich Transparenz schaffen könnte (Wachter, S., B. Mittelstadt und C. Russell, 2017[27]). Darüber hinaus könnten durch die gemeinsame Nutzung von Code oder Datensätzen Geschäftsgeheimnisse oder sensible Nutzerdaten preisgegeben werden.

Generell wird es als wichtig erachtet, das Bewusstsein für die in der KI verwendeten Schlussfolgerungsprozesse zu schärfen und deren Verständnis zu fördern, damit diese Technologien allgemein akzeptiert werden und allen Nutzen bringen.

Transparenzansätze in KI-Systemen

Experten der Working Group on Explanation and the Law des Berkman Klein Center der Harvard-Universität haben drei Ansätze zur Erhöhung der Transparenz von KI-Systemen aufgezeigt und festgestellt, dass es bei jedem dieser Ansätze zu Zielkonflikten kommen kann (Doshi-Velez et al., 2017[28]). Zusätzlich zu diesen drei Ansätzen gibt es noch das Konzept der Optimierungstransparenz, d. h. der Transparenz in Bezug auf die Ziele eines KI-Systems und die in Verbindung mit diesen Zielen erhaltenen Ergebnisse. Die drei in Harvard untersuchten Ansätze sind: a) theoretische Garantien, b) empirische Evidenz und c) Erklärung (Tabelle 4.1).

Tabelle 4.1. Ansätze zur Erhöhung der Transparenz und Rechenschaftspflicht von KI-Systemen

Ansatz	Beschreibung	Gut geeignete Kontexte	Schlecht geeignete Kontexte
Theoretische Garantien	In einigen Situationen ist es möglich, theoretische Garantien für ein KI-System zu geben, die durch Beweise gestützt sind.	Die Umgebung ist vollständig beobachtbar (ein Beispiel ist das Go-Spiel) und sowohl Problem als auch Lösung können formalisiert werden.	Die Situation kann nicht eindeutig beschrieben werden (wie die meisten realen Situationen).
Statistische Evidenz/ Wahrscheinlichkeit	Anhand von empirischer Evidenz wird die Gesamtleistung eines Systems gemessen und der durch dieses System entstehende Nutzen oder Schaden aufgezeigt, ohne bestimmte Entscheidungen zu erklären.	Ergebnisse können vollständig formalisiert werden; es ist möglich, negative Ergebnisse abzuwarten, um sie zu messen; Probleme sind u. U. nur aggregiert sichtbar.	Das Ziel kann nicht vollständig formalisiert werden; es ist möglich, für eine bestimmte Entscheidung die Verantwortlichkeiten festzulegen (Schuld oder Unschuld).
Erklärung/ Nachvollziehbarkeit	Menschen können Informationen über die Logik interpretieren, nach der ein System mit einem bestimmten Satz von Eingangsdaten zu einer bestimmten Schlussfolgerung gelangt ist.	Probleme sind nicht vollständig spezifiziert, Ziele sind nicht klar und Eingangsdaten könnten falsch sein.	Andere Formen der Rechenschaftslegung sind möglich.

Quelle: Nach Doshi-Velez et al. (2017[28]), „Accountability of AI under the law: The role of explanation“, https://arxiv.org/pdf/1711.01134.pdf.

Einige Systeme bieten theoretische Garantien für die Einhaltung bestimmter Grenzen

In einigen Fällen ist es möglich, **theoretische Garantien** zu geben, dass ein System nachweislich innerhalb enger Grenzen (*constraints*) operiert. Dies ist in Situationen möglich, in denen die Umgebung vollständig beobachtbar ist und sowohl das Problem als auch die Lösung vollständig formalisiert werden können, wie z. B. beim Go-Spiel. In solchen Fällen können bestimmte Ergebnisse nicht eintreten, selbst wenn ein KI-System neue Arten von Daten verarbeitet. Beispielsweise könnte ein System entwickelt werden, das vereinbarte Prozesse für Abstimmungen und Stimmauszählungen nachweislich einhält. In diesem Fall ist möglicherweise keine Erklärung oder Evidenz erforderlich: Das System muss nicht erklären, wie es zu einem bestimmten Ergebnis gekommen ist, weil die Arten von Ergebnissen, die Anlass zur Sorge geben, mathematisch unmöglich sind. Es kann bereits frühzeitig geprüft werden, ob die festgelegten einschränkenden Bedingungen ausreichend sind.

In einigen Fällen kann statistische Evidenz für die Gesamtleistung eines Systems erbracht werden

In einigen Fällen kann es ausreichend sein, sich auf **statistische Evidenz** für die Gesamtleistung eines Systems zu verlassen. Evidenz dafür, dass ein KI-System einen bestimmten Nutzen oder Schaden für die Gesellschaft insgesamt oder für den Einzelnen messbar erhöht, kann zur Erfüllung der Rechenschaftspflicht ausreichen. Zum Beispiel könnte statistisch belegt werden, dass es mit einem autonomen Flugzeuglandesystem zu weniger sicherheitsrelevanten Zwischenfällen kommt als mit menschlichen Piloten oder dass ein klinisches Instrument zur Diagnoseunterstützung die Sterblichkeit verringern kann. Statistische Evidenz könnte ein geeigneter Rechenschaftsmechanismus für viele KI-Systeme sein, weil bei ihrer Verwendung Geschäftsgeheimnisse geschützt werden können und weil sie es gestattet, häufig auftretende, aber wenig risikoträchtige Schäden zu erkennen, die erst aggregiert sichtbar werden (Barocas, S. und A. Selbst, 2016[29]; Crawford, 2016[30]). Bias- oder Diskriminierungsfragen können statistisch festgestellt werden: Ein Kreditgenehmigungssystem würde z. B. ein Bias aufweisen, falls es (bei Ausklammerung anderer Faktoren) mehr Kredite für Männer als für Frauen genehmigt. Die zulässige Fehlerrate und die tolerierte Unsicherheit sind je nach Anwendung unterschiedlich. Die für eine Übersetzungssoftware als akzeptabel erachtete Fehlerquote könnte für autonomes Fahren oder medizinische Untersuchungen z. B. nicht akzeptabel sein.

Optimierungstransparenz ist die Transparenz in Bezug auf die Ziele und Ergebnisse eines Systems

Ein anderer Ansatz zur Erhöhung der Transparenz von KI-Systemen besteht darin, den Fokus von den Instrumenten des Systems auf seine Ziele zu verlagern: Es geht dann nicht mehr darum, die Nachvollziehbarkeit der Funktionsweise eines Systems zu fordern, sondern seine Ergebnisse zu messen – d. h. das, wofür das System „optimiert" wurde. Dies setzt voraus, dass erklärt wird, wofür das betreffende KI-System optimiert wurde, wobei zu berücksichtigen ist, dass Optimierungen unvollkommen sind, mit Zielkonflikten verbunden sind und „kritischen Erfordernissen" wie dem der Sicherheit und der Fairness unterliegen sollten. Nach diesem Ansatz sollen KI-Systeme dafür eingesetzt werden, wofür sie optimiert wurden. Dabei wird gegebenenfalls auf bestehende Ethik- und Rechtsvorschriften sowie auf gesellschaftliche Diskussionen und politische Prozesse Bezug genommen, um klarzustellen, wofür die KI-Systeme optimiert werden sollen (Weinberger, 2018[1]).

Das Konzept der Nachvollziehbarkeit bezieht sich auf ein bestimmtes Ergebnis eines KI-Systems

Nachvollziehbarkeit ist in Situationen unerlässlich, in denen konkret die Verantwortung für ein bestimmtes Ereignis zugewiesen werden muss – Situationen also, wie sie immer häufiger auftreten dürften, wenn KI-Systeme eingesetzt werden, um Empfehlungen abzugeben oder Entscheidungen zu treffen, die derzeit noch im menschlichen Ermessen liegen (Burgess, 2016[31]). Die DSGVO schreibt vor, dass die betroffenen Personen aussagekräftige Informationen über die zugrunde liegende Logik, die Bedeutung und voraussichtlichen Konsequenzen automatisierter Entscheidungssysteme erhalten. In der Regel muss die Nachvollziehbarkeit nicht für den gesamten Entscheidungsprozess des Systems gewährleistet sein. Meistens reicht die Beantwortung einer der folgenden Fragen aus (Doshi-Velez et al., 2017[28]):

1. **Hauptfaktoren für eine Entscheidung**: Bei vielen Arten von Entscheidungen, z. B. bei Sorgerechtsstreitigkeiten, Kreditprüfungen und Untersuchungshaftentlassungen, müssen eine Reihe von Faktoren berücksichtigt werden (oder ist es im Gegenteil ausdrücklich verboten, bestimmte Faktoren zu berücksichtigen). Mit einer Auflistung der Faktoren, die für eine KI-Prognose wichtig waren – idealerweise in der Reihenfolge ihrer Bedeutung –, kann sichergestellt werden, dass die richtigen Faktoren berücksichtigt wurden.
2. **Maßgebliche Faktoren, d. h. Faktoren, die das Ergebnis entscheidend beeinflussen**: Manchmal ist es wichtig zu wissen, ob ein bestimmter Faktor das Ergebnis beeinflusst hat. Durch die Änderung einer bestimmten Variablen, z. B. der ethnischen Herkunft bei der Vergabe von Studienplätzen, kann aufgezeigt werden, ob der jeweilige Faktor richtig eingesetzt wurde.
3. **Warum kam es in zwei scheinbar ähnlichen Fällen zu unterschiedlichen Ergebnissen oder umgekehrt warum kam es in zwei unterschiedlichen Fällen zu einem gleichen Ergebnis?** Es ist möglich, die Konsistenz und Integrität von KI-basierten Prognosen zu bewerten. Beispielsweise sollte das Einkommen bei der Entscheidung über eine Darlehensgewährung berücksichtigt werden, aber es sollte in ansonsten ähnlichen Fällen nicht mal entscheidend und mal unerheblich sein.

Im Bereich der Nachvollziehbarkeit wird aktiv geforscht, dies ist aber mit Kosten und möglichen Zielkonflikten verbunden

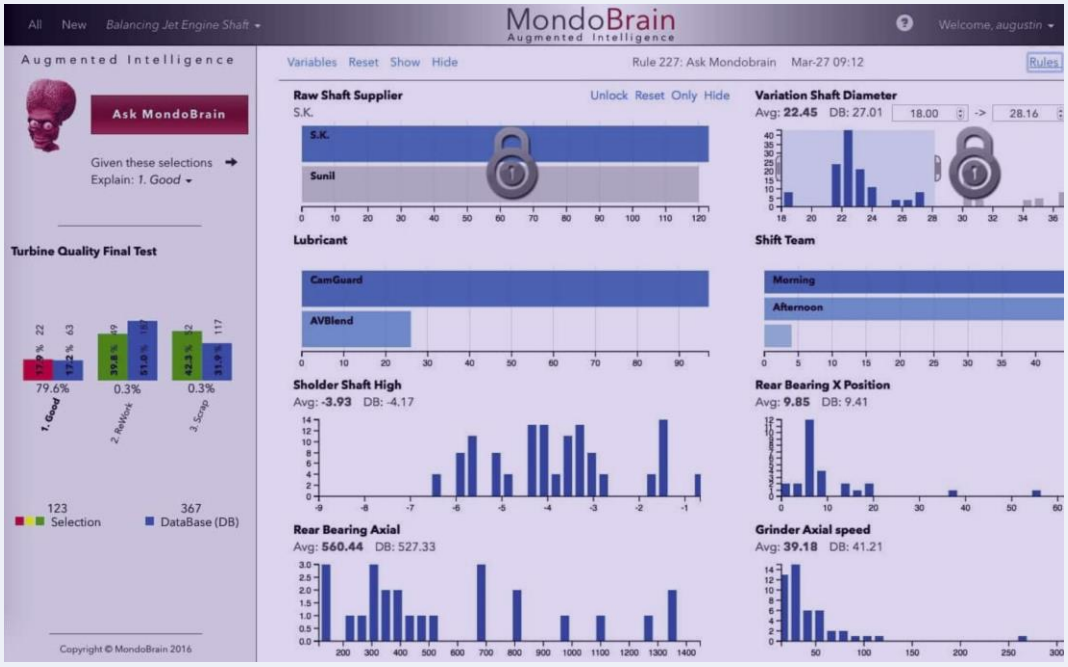
Einzelne Unternehmen, Normungsgremien, gemeinnützige Organisationen und öffentliche Einrichtungen betreiben technische Forschung, um KI-Systeme zu schaffen, die ihre Prognosen erklären können. Unternehmen in stark regulierten Bereichen wie der Finanzbranche, dem Gesundheitswesen und dem Personalwesen sind besonders aktiv bemüht, potenzielle finanzielle, rechtliche und Reputationsrisiken anzugehen, die sich aus den Prognosen von KI-Systemen ergeben. Zum Beispiel hat die US-amerikanische Bank Capital One 2016 ein Forschungsteam gegründet, um Wege zu finden, die Nachvollziehbarkeit von KI-Techniken zu verbessern (Knight, 2017[32]). Unternehmen wie MondoBrain haben Benutzeroberflächen entworfen, die helfen, wichtige Faktoren zu erklären (Kasten 4.5). Gemeinnützige Organisationen wie OpenAI erforschen Ansätze zur Entwicklung erklärbarer KI und zur Prüfung von KI-Entscheidungen. Es wird auch öffentlich finanzierte Forschung betrieben. Die Defense Advanced Research Projects Agency (DARPA) finanziert beispielsweise 13 verschiedene Forschungsgruppen, die an einer Reihe von Ansätzen arbeiten, um die Erklärbarkeit von KI zu verbessern.

Kasten 4.5. Nachvollziehbarkeitsprobleme durch besser gestaltete Benutzeroberflächen angehen

Einige Unternehmen haben begonnen, Nachvollziehbarkeit in ihre Lösungen einzubauen, damit Benutzer die im Hintergrund ablaufenden KI-Prozesse besser verstehen. Eines dieser Unternehmen ist MondoBrain. Das in Frankreich ansässige Unternehmen kombiniert menschliche, kollektive und künstliche Intelligenz, um Unternehmen eine Augmented-Reality-Lösung anzubieten. Mit interaktiven Dashboards zur Datenvisualisierung werden alle in einem Unternehmen vorhandenen Daten ausgewertet (z. B. Daten aus der Enterprise-Resource-Planning-, Business-Programme-Management- oder Customer-Relationship-Management-Software) und auf der Basis von Kundenanfragen präskriptive Empfehlungen abgegeben (Abbildung 4.1). Mit einem besonderen ML-Algorithmus werden dabei betriebswirtschaftliche Variablen ausgeklammert, die für die Abfrage uninteressant sind, und die Variablen mit dem größten Effekt extrahiert.

Eine einfache Ampelsymbolik führt den Benutzer durch alle Schritte der Abfrage und erleichtert ihm so das Verständnis des Entscheidungsprozesses. Jede einzelne Entscheidung wird automatisch dokumentiert, wodurch sie überprüfbar und nachvollziehbar wird. Damit werden alle Schritte, die zur endgültigen Geschäftsempfehlung führten, vollständig, aber einfach dargestellt.

Abbildung 4.1. Datenvisualisierungsinstrumente zur Verbesserung der Erklärbarkeit



Quelle: www.mondobrain.com.

In vielen Fällen ist es möglich, eine oder mehrere Erklärungen zu den Ergebnissen von KI-Systemen zu erstellen. Solche Erklärungen sind jedoch mit Kosten verbunden. Die Konzeption eines Systems, dessen Entscheidungen nachvollziehbar sind, kann komplex und teuer sein. Für alle Systeme Nachvollziehbarkeit zu fordern, ist angesichts des unterschiedlichen Zwecks der verschiedenen Systeme möglicherweise nicht angemessen und

kann insbesondere KMU benachteiligen. KI-Systeme müssen oft ex ante konzipiert werden, um eine bestimmte Art von Erklärung zu liefern. Die nachträgliche Suche nach Erklärungen erfordert in der Regel zusätzliche Arbeit; möglicherweise muss das gesamte Entscheidungssystem neu gestaltet werden. So kann ein KI-System beispielsweise nicht alle wichtigen Faktoren erklären, die sich auf ein Ergebnis ausgewirkt haben, wenn es aufgrund seines Designs nur auf die Erklärung eines dieser Faktoren ausgelegt ist. Ein KI-System zur Erkennung von Herzerkrankungen kann z. B. nicht nach dem Einfluss des Geschlechts auf eine Diagnose abgefragt werden, wenn es nicht mit Geschlechtsdaten trainiert wurde. Dies ist selbst dann der Fall, wenn das KI-System das Geschlecht tatsächlich über Ersatzvariablen berücksichtigt, wie z. B. andere Erkrankungen, die bei Frauen häufiger auftreten als bei Männern.

In einigen Fällen gibt es einen Zielkonflikt zwischen Nachvollziehbarkeit und Genauigkeit. Damit sie erklärbar sind, müssen die Lösungsvariablen möglicherweise auf eine hinreichend kleine Zahl reduziert werden, um von Menschen verstanden zu werden. Dies könnte bei komplexen, hochdimensionalen Problemen suboptimal sein. Einige ML-Modelle, die in der medizinischen Diagnostik verwendet werden, können beispielsweise die Wahrscheinlichkeit einer Krankheit genau vorhersagen, sind aber zu komplex, als dass sie von Menschen verstanden werden könnten. In solchen Fällen sollte der Schaden, der durch ein weniger präzises System entstehen könnte, das klare Erklärungen bietet, gegen den Schaden abgewogen werden, den ein präziseres System verursachen würde, bei dem Fehler schwieriger zu erkennen sind. Zum Beispiel kann die Rückfallprognose einfache, erklärbare Modelle erfordern, bei denen Fehler erkennbar sind (Dressel, J. und H. Farid, 2018[33]). In Bereichen wie der Klimavorhersage hingegen werden komplexere Modelle, die zwar bessere Prognosen liefern, aber weniger erklärbar sind, u. U. eher akzeptiert. Dies ist insbesondere dann der Fall, wenn es andere Mechanismen zur Rechenschaftslegung gibt, z. B. statistische Daten zur Erkennung möglicher Verzerrungen oder Fehler.

Robustheit und Sicherheit

Was unter Robustheit und Sicherheit zu verstehen ist

Robustheit kann als die Fähigkeit verstanden werden, widrigen Bedingungen, einschließlich digitaler Sicherheitsrisiken, zu widerstehen oder sie zu überwinden (OECD, 2019[34]). Als sichere KI-Systeme gelten Systeme, die bei normalem oder vorhersehbarem Gebrauch oder Missbrauch während ihres gesamten Lebenszyklus keine unvertretbaren Sicherheitsrisiken darstellen (OECD, 2019[35]). Die Fragen der Robustheit und Sicherheit von KI-Systemen sind eng miteinander verknüpft. Die digitale Sicherheit kann sich beispielsweise auf die Produktsicherheit auswirken, wenn vernetzte Produkte wie selbstfahrende Fahrzeuge oder KI-fähige Haushaltsgeräte nicht ausreichend sicher sind; Hacker könnten die Kontrolle über sie übernehmen und Einstellungen aus der Ferne ändern.

Risikomanagement in KI-Systemen

Die Bestimmung des erforderlichen Sicherheitsniveaus sollte auf einer Nutzen-Risiko-Abwägung beruhen

Der Schaden, der durch ein KI-System verursacht werden könnte, sollte gegen die Kosten für die Integration von Transparenz und Rechenschaftspflicht in KI-Systeme abgewogen werden. Dabei geht es beispielsweise um Risiken in Bezug auf Menschenrechte, Datenschutz, Fairness und Robustheit. Allerdings birgt nicht jede Nutzung von KI die gleichen

Risiken, und die Forderung nach Nachvollziehbarkeit ist z. B. auch mit Kosten verbunden. Im Risikomanagement scheint ein breiter Konsens darüber zu bestehen, dass Kontexte, in denen viel auf dem Spiel steht, ein höheres Maß an Transparenz und Verantwortlichkeit erfordern, insbesondere wenn es um das Leben oder die Freiheit des Einzelnen geht.

Anwendung von Risikomanagementkonzepten während des gesamten Lebenszyklus von KI-Systemen

Unternehmen nutzen Konzepte des Risikomanagements, um potenzielle Risiken, die das Verhalten und die Ergebnisse eines Systems negativ beeinflussen können, zu erkennen, zu bewerten, zu priorisieren und anzugehen. Solche Konzepte können auch genutzt werden, um Risiken für verschiedene Akteure zu bestimmen und um festzulegen, wie diesen Risiken während des gesamten Lebenszyklus des KI-Systems begegnet werden kann (vgl. Abschnitt „Lebenszyklus eines KI-Systems“ in Kapitel 1).

KI-Akteure – d. h. Personen, die eine aktive Rolle im Lebenszyklus eines KI-Systems spielen – bewerten und mindern Risiken in diesem System als Ganzem sowie in jeder Phase seines Lebenszyklus. Das Risikomanagement in KI-Systemen untergliedert sich in folgende Schritte, deren Bedeutung je nach Lebenszyklusphase des KI-Systems variiert:

1. **Ziele**: Definition von Zielen, Funktionen oder Eigenschaften des KI-Systems im Kontext. Diese Funktionen und Eigenschaften können sich je nach Phase des KI-Lebenszyklus ändern.
2. **Beteiligte bzw. betroffene Akteure**: Bestimmung der Akteure, die in den verschiedenen Phasen des Lebenszyklus des Systems direkt oder indirekt von dessen Funktionen oder Eigenschaften betroffen sind.
3. **Risikobewertung**: Bewertung der potenziellen Auswirkungen (Vorteile und Risiken) für die beteiligten bzw. betroffenen Akteure. Diese hängen von den Akteuren selbst sowie von der Lebenszyklusphase ab, die das KI-System erreicht hat.
4. **Risikominderung**: Bestimmung von Strategien zur Risikominderung, die dem Risiko angemessen sind und im Verhältnis zu ihm stehen. Dabei sollten Faktoren wie die Ziele des Unternehmens, die beteiligten bzw. betroffenen Akteure, die Eintrittswahrscheinlichkeit der Risiken und der potenzielle Nutzen berücksichtigt werden.
5. **Umsetzung**: Umsetzung von Strategien zur Risikominderung.
6. **Überwachung, Bewertung und Rückmeldung**: Überwachung, Bewertung und Rückmeldung zu den Ergebnissen der Umsetzung.

Der Einsatz von Risikomanagementkonzepten im Lebenszyklus von KI-Systemen und die Dokumentation der Entscheidungen in jeder Lebenszyklusphase können dazu beitragen, die Transparenz von KI-Systemen und die Rechenschaftspflicht der Unternehmen für diese Systeme zu verbessern.

Das Ausmaß des Gesamtschadens und der unmittelbare Risikokontext sollten nebeneinander betrachtet werden

Isoliert betrachtet stellen einige Anwendungen von KI-Systemen ein geringes Risiko dar. Aufgrund ihrer gesellschaftlichen Auswirkungen können sie jedoch ein höheres Maß an Robustheit erfordern. Ein System, durch dessen Betrieb einer großen Zahl von Menschen ein leichter Schaden entsteht, könnte insgesamt einen erheblichen Schaden verursachen.

Dies könnte beispielsweise geschehen, wenn eine kleine Zahl von KI-Tools in mehrere Dienste integriert ist und in verschiedenen Branchen genutzt wird, etwa für Kreditanträge, den Abschluss von Versicherungsverträgen oder Hintergrundprüfungen. Ein einziger Fehler oder eine einzige Verzerrung in einem System könnte dann zu einer ganzen Kaskade von Ablehnungen führen (Citron, D. und F. Pasquale, 2014[36]). Für sich genommen dürften die einzelnen Ablehnungen kaum von Bedeutung sein. Zusammengenommen könnten sie jedoch einen disruptiven Effekt haben. In Politikdiskussionen sollte daher neben dem unmittelbaren Risikokontext auch das Ausmaß des Gesamtschadens berücksichtigt werden.

Robustheit gegenüber den mit KI verbundenen digitalen Sicherheitsrisiken

KI ermöglicht raffiniertere Angriffe potenziell größeren Ausmaßes

Je kostengünstiger und je leichter einsetzbar KI wird, umso mehr dürfte neben ihrer Nutzung zur Verbesserung der digitalen Sicherheit auch ihre Nutzung zu böswilligen Zwecken zunehmen (vgl. Unterabschnitt „KI und digitale Sicherheit“ in Kapitel 3). Cyberkriminelle arbeiten daran, ihre KI-Fähigkeiten auszubauen. Schnellere und raffiniertere Angriffe stellen eine zunehmende Gefahr für die digitale Sicherheit dar.[6] Vor diesem Hintergrund weiten sich bestehende Bedrohungen aus, während neue Bedrohungen entstehen und sich die Art der Bedrohungen selbst verändert.

Heutige KI-Systeme weisen eine Reihe von Schwachstellen auf. Böswillige Akteure können z. B. die Daten manipulieren, mit denen ein KI-System trainiert wird („Datenvergiftung“). Sie können auch die Eigenschaften identifizieren, die in einem digitalen Sicherheitsmodell zur Erkennung von Malware verwendet werden. Mit diesen Informationen können sie nicht erkennbaren bösartigen Code entwickeln oder Informationen absichtlich falsch klassifizieren (Adversarial Examples bzw. „feindliche Beispiele“) (Kasten 4.6) (Brundage et al., 2018[37]). Mit der zunehmenden Verfügbarkeit von KI-Technologien können immer mehr Menschen KI nutzen, um raffiniertere Angriffe potenziell größeren Ausmaßes durchzuführen. Die Häufigkeit und die Effizienz arbeitsintensiver digitaler Sicherheitsangriffe wie z. B. gezieltes Spear-Phishing könnten mit deren Automatisierung durch Algorithmen des maschinellen Lernens zunehmen.

Kasten 4.6. Die Gefahr von Adversarial Examples für maschinelles Lernen

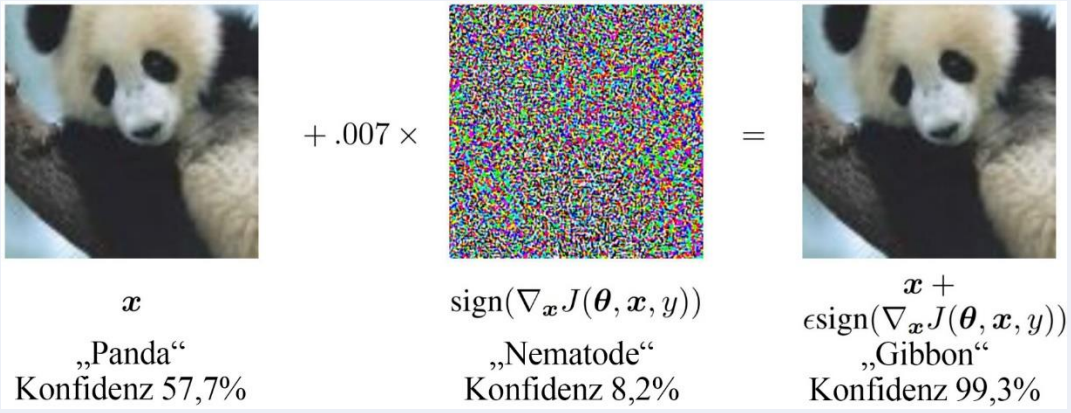
Adversarial Examples bzw. „feindliche Beispiele“ sind Informationen, die von Angreifern absichtlich in ML-Modelle eingegeben werden, damit das Modell einen Fehler macht, dabei aber zugleich ein hohes Konfidenzniveau anzeigt. Feindliche Beispiele sind ein echtes Problem für die Robustheit und Sicherheit von KI-Systemen, da verschiedene ML-Modelle, darunter auch dem neuesten Stand der Technik entsprechende neuronale Netze, für sie anfällig sind.

Diese feindlichen Beispiele können sehr subtil funktionieren. In Abbildung 4.2 wurde dem Bild eines Pandas eine unmerklich kleine Veränderung – ein „feindlicher Input“ – hinzugefügt. Dieser Input wurde speziell entwickelt, um das Bildklassifikationsmodell zu täuschen. Der Algorithmus klassifiziert den Panda daraufhin mit einem Konfidenzniveau von nahezu 100 % als Gibbon.

Neuere Forschungsarbeiten haben zudem gezeigt, dass sich feindliche Beispiele erstellen lassen, indem ein Bild auf Normalpapier gedruckt und mit einem Smartphone mit Standardauflösung abfotografiert wird. Solche Bilder könnten gefährlich sein: Ein gezielt

aufgebrachter Aufkleber auf einem Stoppschild könnte dazu führen, dass ein selbstfahrendes Auto es als Vorfahrts- oder anderes Schild interpretiert.

Abbildung 4.2. Durch eine kleine Veränderung wird ein Algorithmus so getäuscht, dass er einen Panda als Gibbon klassifiziert



Quelle: Goodfellow, Shlens und Szegedy (2015[38]), „Explaining and harnessing adversarial examples“, https://arxiv.org/pdf/1412.6572.pdf; Kurakin, Goodfellow und Bengio (2017[39]), „Adversarial examples in the physical world“, https://arxiv.org/abs/1607.02533.

Sicherheit

Lernende und autonome KI-Systeme haben sicherheitsrechtliche Konsequenzen

Das Spektrum von Produkten mit integrierter künstlicher Intelligenz nimmt rasant zu – von der Robotik und selbstfahrenden Fahrzeugen bis hin zu gängigen Konsumgütern und verbraucherorientierten Dienstleistungen, wie z. B. intelligenten Haushaltsgeräten und Haussicherheitssystemen. KI-Produkte bieten erhebliche Sicherheitsvorteile, lassen im Bereich der Produktsicherheitsbestimmungen aber auch neue praktische und rechtliche Herausforderungen entstehen (OECD, 2017[20]). Sicherheitsbestimmungen sind in der Regel eher auf „fertige“ Hardware-Produkte als auf Software ausgerichtet, zudem lernen viele KI-Software-Produkte während ihres gesamten Lebenszyklus und entwickeln sich weiter.[7] KI-Produkte können auch „autonom“ oder „halbautonom“ sein, d. h. Entscheidungen ohne oder mit wenig menschlichem Input treffen und ausführen.

Verschiedene Arten von KI-Anwendungen dürften unterschiedliche politische Maßnahmen erfordern (Freeman, 2017[40]). Ganz allgemein sind für KI-Systeme vier Erwägungen von Bedeutung. Erstens muss geklärt werden, wie am besten gewährleistet werden kann, dass Produkte sicher sind bzw. – anders ausgedrückt – während ihres gesamten Lebenszyklus bei normalem oder vorhersehbarem Gebrauch oder Missbrauch kein unvertretbares Sicherheitsrisiko darstellen. Dies gilt auch für Fälle, in denen nur wenige Daten vorliegen, mit denen das System trainiert werden kann (Kasten 4.7). Zweitens stellt sich die Frage, wer in welchem Umfang für Schäden, die durch ein KI-System verursacht wurden, haftbar gemacht werden sollte. Gleichzeitig ist zu überlegen, welche Akteure zur Sicherheit von autonomen Maschinen beitragen können. Dabei könnte es sich um Nutzer, Produkt- und Sensorhersteller, Softwarehersteller, Designer, Infrastrukturanbieter und Datenanalysefirmen handeln. Drittens ist es wichtig, über die Art(en) der Haftung nachzudenken – ob es sich um eine verschuldensunabhängige oder verschuldensabhängige Haftung handeln soll

und welche Rolle die Versicherung spielen soll. Die mangelnde Transparenz einiger KI-Systeme verschärft das Problem der Haftung. Viertens müssen politische Entscheidungsträger überlegen, wie geltendes Recht durchgesetzt werden kann, was ein „Fehler" in einem KI-Produkt ist, wie die Beweislast verteilt wird und welche Rechtsmittel zur Verfügung stehen.

Kasten 4.7. Synthetische Daten für eine sicherere und genauere KI: autonome Fahrzeuge

Im Bereich des maschinellen Lernens werden zunehmend synthetische Daten verwendet, da sie die Simulation von Szenarien ermöglichen, die sich unter realen Bedingungen nur schwer beobachten oder reproduzieren lassen. Philipp Slusallek, wissenschaftlicher Direktor am Deutschen Forschungszentrum für Künstliche Intelligenz, nennt als Beispiel hierfür den Fall eines Kindes, das über die Straße rennt: Es muss sichergestellt werden, dass autonome Fahrzeuge in einer solchen Situation richtig reagieren.

Eine „digitale Realität" – d. h. eine simulierte Umgebung, die die relevanten Merkmale der realen Welt nachbildet – könnte vier Dinge ermöglichen. Erstens könnte sie synthetische Eingabedaten generieren, aus denen KI-Systeme lernen können, mit komplexen Situationen umzugehen. Zweitens könnte sie Ergebnisse validieren und synthetische Daten gegenüber realen Daten neu kalibrieren. Drittens könnte sie für die Durchführung von Tests, z. B. eine Führerscheinprüfung für autonome Fahrzeuge, verwendet werden. Viertens bietet sie die Möglichkeit, den Entscheidungsprozess des Systems und die möglichen Effekte alternativer Entscheidungen zu untersuchen. Dank eines solchen Ansatzes konnte Google z. B. seine selbstfahrenden Fahrzeuge mit mehr als 4,8 Millionen simulierten Kilometern pro Tag trainieren (dies entspricht mehr als 500 Hin- und Rückfahrten zwischen New York City und Los Angeles).

Quelle: Golson (2016[41]), „Google's self-driving cars rack up 3 million simulated miles every day", https://www.theverge.com/2016/2/1/10892020/google-self-driving-simulator-3-million-miles; Slusallek (2018[42]), *Artificial Intelligence and Digital Reality: Do We Need a CERN for AI?*, https://www.oecd-forum.org/channels/722-digitalisation/posts/28452-artificial-intelligence-and-digital-reality-do-we-need-a-cern-for-ai.

Die Produkthaftungsrichtlinie der Europäischen Union (Richtlinie 85/374/EWG) von 1985 legt das Prinzip der „verschuldensunabhängigen Haftung" (Gefährdungshaftung) fest. Wenn ein fehlerhaftes Produkt einem Verbraucher Schaden zufügt, haftet der Hersteller nach diesem Prinzip auch, wenn keine Fahrlässigkeit oder Verschulden vorliegt. Diese Richtlinie wird gegenwärtig von der Europäischen Kommission überarbeitet. Den ersten Schlussfolgerungen zufolge ist das derzeitige Modell im Großen und Ganzen angemessen (Ingels, 2017[43]). Aktuelle und in Zukunft zu erwartende KI-Technologien stellen allerdings die Konzepte von „Produkt", „Sicherheit", „Fehler" und „Schaden" infrage, was die Last der Beweisführung schwieriger macht.

Im Bereich des autonomen Fahrens ist Sicherheit ein sehr wichtiges Thema für die Politik. Beispielsweise muss festgelegt werden, wie autonome Fahrzeuge getestet werden können, damit gewährleistet ist, dass ihr Einsatz sicher ist. Dies beinhaltet z. B. Zulassungsregelungen, die die Möglichkeit der Vorerprobung von AF-Systemen vorsehen, oder Anforderungen an die Systeme, die die Wachsamkeit menschlicher Fahrer überwachen müssen, wenn diese als Rückfallebene dienen. In einigen Fällen stellt die Zulassung für Firmen, die Fahrzeuge testen möchten, ein echtes Problem dar. Zudem stehen staatliche Stellen solchen Tests mehr oder weniger offen gegenüber. Von verschiedener Seite wurden Forderungen nach einer verschuldensunabhängigen Haftung von Herstellern autonomer Fahrzeuge laut. Eine solche Haftung würde auf der Kontrollierbarkeit des Risikos beruhen.

Damit würde beispielsweise anerkannt, dass einen Mitfahrer in einem selbstfahrenden Fahrzeug kein Verschulden treffen kann oder er nicht gegen eine Sorgfaltspflicht verstoßen kann. Juristen zufolge könnte selbst das Konzept des Fahrzeughalters hier nicht greifen, weil der Halter in der Lage sein muss, das Risiko zu kontrollieren (Borges, 2017[44]). Einige schlagen vor, dass Versicherungen das Risiko von durch autonome Fahrzeuge verursachten Schäden übernehmen könnten. Grundlage dafür könnte eine auf Risikobewertungen beruhende Klassifikation der zugelassenen autonomen Fahrzeuge sein.

Möglicherweise müssen die Standards für sichere Arbeitsbedingungen aktualisiert werden

Zu den direkten Auswirkungen von KI auf die Arbeitsbedingungen kann auch die Notwendigkeit neuer Sicherheitsprotokolle gehören. So wird immer deutlicher, dass neue bzw. überarbeitete Branchenstandards und Technologievereinbarungen zwischen Geschäftsleitung und Beschäftigten erforderlich sind, um zuverlässige, sichere und produktive Arbeitsplätze zu gewährleisten. Der Europäische Wirtschafts- und Sozialausschuss (EWSA) riet den „Interessenträgern, sich gemeinsam für komplementäre KI-Systeme und ihre Ko-Kreation am Arbeitsplatz einzusetzen“ (EWSA, 2017[45]).

Rechenschaftspflicht

Die zunehmende Nutzung von KI verlangt nach einer Rechenschaftspflicht für die Funktionsweise von KI-Systemen

Rechenschaftspflicht bedeutet im Wesentlichen, dass die Beweislast für die ordnungsgemäße Funktionsweise eines KI-Systems auf die entsprechenden Organisationen oder Personen übertragen werden kann. Zu den Verantwortlichkeitskriterien, auf denen sie beruht, gehören die Achtung menschlicher Werte sowie Fairness, Transparenz, Robustheit und Sicherheit. Die Rechenschaftspflicht hängt von der Rolle der einzelnen KI-Akteure, dem Kontext und dem Stand der Technik ab. Für politische Entscheidungsträger ist sie an Mechanismen geknüpft, die verschiedene Funktionen erfüllen. Diese Mechanismen bestimmen, welche Akteure für eine bestimmte Empfehlung oder Entscheidung verantwortlich sind. Sie korrigieren die Empfehlung oder Entscheidung, bevor sie umgesetzt wird. Sie könnten die Entscheidung auch nachträglich anfechten und sie könnten sogar das für die Entscheidung verantwortliche System anfechten (Helgason, 1997[46]).

In der Praxis hängt die Verantwortlichkeit von KI-Systemen häufig davon ab, wie gut ein bestimmtes System nach Indikatoren der Genauigkeit oder Effizienz abschneidet. Zunehmend werden dabei auch Indikatoren für die Ziele Fairness, Sicherheit und Robustheit herangezogen. Allerdings werden solche Indikatoren immer noch seltener eingesetzt als Messgrößen der Effizienz oder Genauigkeit. Wie bei allen Messgrößen können Monitoring und Evaluierung kostspielig sein. Daher müssen Art und Häufigkeit der Erhebungen in einem angemessenen Verhältnis zu den potenziellen Risiken und Vorteilen stehen.

Das erforderliche Maß an Rechenschaftslegung hängt vom Risikokontext ab

Welcher Politikansatz angemessen ist, hängt vom Kontext und vom Anwendungsfall ab. Beispielsweise können die Rechenschaftserwartungen bei der Nutzung von KI im öffentlichen Sektor höher sein. Dies gilt insbesondere für die Ausübung hoheitlicher Aufgaben, wie Sicherheit und Rechtsvollzug, bei denen ein erhebliches Schadenspotenzial besteht. Formale Rechenschaftsmechanismen sind häufig auch für privatwirtschaftliche Anwendungen in den stark regulierten Bereichen Verkehr, Finanzen und Gesundheitswesen

erforderlich. In Bereichen des privaten Sektors, die weniger stark reguliert sind, unterliegt der Einsatz von KI weniger formalen Rechenschaftsmechanismen. Technische Ansätze für Transparenz und Rechenschaftspflicht spielen in diesem Fall eine noch wichtigere Rolle. Sie müssen sicherstellen, dass die von privatwirtschaftlichen Akteuren konzipierten und betriebenen Systeme gesellschaftlichen Normen und rechtlichen Auflagen gerecht werden.

Einige Anwendungen oder Entscheidungen können das Eingreifen eines Menschen erfordern, der den sozialen Kontext und mögliche unbeabsichtigte Konsequenzen berücksichtigt. Wenn Entscheidungen erhebliche Auswirkungen auf das Leben von Menschen haben, besteht weitgehend Einigkeit darüber, dass sie nicht allein aufgrund von KI-basierten Ergebnissen (z. B. einem Score) getroffen werden sollten. Die DSGVO beispielsweise befürwortet in solchen Fällen ein menschliches Eingreifen. Menschen müssen z. B. informiert werden, wenn KI genutzt wird, um in einem strafrechtlichen Verfahren ein Urteil zu fällen, um Kreditvergabeentscheidungen zu treffen, Plätze in bestimmten Bildungsgängen zu vergeben oder Kandidaten für eine Stelle auszuwählen. Wenn viel auf dem Spiel steht, sind meist formale Rechenschaftsmechanismen erforderlich. Zum Beispiel ist ein Richter, der KI bei der Urteilsfällung in Strafverfahren nutzt, jemand, der direkt eingreifen kann („human-in-the-loop“). Andere Rechenschaftsmechanismen – darunter auch traditionelle gerichtliche Berufungsverfahren – helfen jedoch sicherzustellen, dass KI-Empfehlungen für Richter nur ein Kriterium unter anderen sind, die es zu berücksichtigen gilt (Wachter, S., B. Mittelstadt und L. Floridi, 2017[47]). In risikoarmen Kontexten, z. B. bei Restaurantempfehlungen, kann man sich möglicherweise ganz auf Maschinen verlassen. Ein mehrere Ebenen umfassender Ansatz, der unnötige Kosten nach sich ziehen kann, ist hier vermutlich nicht nötig.

Politikumfeld für künstliche Intelligenz

Es bedarf Maßnahmen auf nationaler Ebene zur Förderung vertrauenswürdiger KI-Systeme. Solche Maßnahmen können nutzbringende und faire Ergebnisse für die Menschen und den Planeten begünstigen, insbesondere in vielversprechenden Bereichen, in die nicht genügend marktorientierte Investitionen fließen. Zur Schaffung eines für vertrauenswürdige KI günstigen Politikumfelds gehört u. a., dass öffentliche und private Investitionen in KI-Forschung und -Entwicklung erleichtert und dass die Menschen mit den notwendigen Kompetenzen ausgestattet werden, um in einer sich wandelnden Arbeitswelt erfolgreich zu sein. In den folgenden Unterabschnitten werden verschiedene Politikbereiche erörtert, die für die Förderung und Entwicklung vertrauenswürdiger KI von entscheidender Bedeutung sind.

In KI-Forschung und -Entwicklung investieren

Langfristige Investitionen in öffentliche Forschung können maßgeblichen Einfluss auf die Innovationstätigkeit im KI-Bereich haben

Die OECD befasst sich in ihren Arbeiten auch mit dem Einfluss innovationspolitischer Maßnahmen auf den digitalen Wandel und die Einführung von KI (OECD, 2018[48]). Dazu untersucht sie u. a. die Rolle der öffentlichen Forschungspolitik, des Wissenstransfers und der Ko-Kreation von Maßnahmen zur Förderung der Entwicklung von Forschungsinstrumenten und Forschungsinfrastruktur für KI. Die Politikverantwortlichen müssen prüfen, welches Niveau staatlichen Engagements in der KI-Forschung angemessen ist, um gesellschaftliche Herausforderungen zu bewältigen (OECD, 2018[13]). Außerdem werden For-

schungseinrichtungen in allen Bereichen leistungsfähige KI-Systeme benötigen, um wettbewerbsfähig zu bleiben, insbesondere in der Biomedizin und den Biowissenschaften. Neue Instrumente wie Plattformen zur gemeinsamen Nutzung von Daten und Superrechner können sich auf die Forschung im KI-Bereich förderlich auswirken und möglicherweise neue Investitionen erforderlich machen. Japan investiert z. B. jährlich mehr als 120 Mio. USD in den Aufbau einer Hochleistungsrecheninfrastruktur für Universitäten und öffentliche Forschungszentren.

KI gilt als Universaltechnologie mit potenziellen Auswirkungen auf eine große Zahl von Wirtschaftszweigen (Agrawal, A., J. Gans und A. Goldfarb, 2018[49]) (Brynjolfsson, E., D. Rock und C. Syverson, 2017[50]). Deshalb wurde KI auch schon als „Erfindung einer Methode der Erfindung" bezeichnet (Cockburn, I., R. Henderson und S. Stern, 2018[51]). Als solche wird sie von Wissenschaftlern und Erfindern bereits umfassend genutzt, um Innovationen zu erleichtern. Zudem könnten dank der bahnbrechenden technologischen Entwicklungen, die KI ermöglicht, ganz neue Wirtschaftszweige entstehen. Das zeigt, wie wichtig Grundlagenforschung und eine langfristig ausgerichtete Forschungspolitik sind (OECD, 2018[52]).

Ein digitales Ökosystem für KI fördern

KI-Technologien und -Infrastruktur

In den letzten Jahren wurden bei KI-Technologien erhebliche Fortschritte erzielt. Dies ist auf die zunehmende Reife statistischer Modellierungstechniken, z. B. neuronaler Netze und insbesondere tiefer neuronaler Netze (bekannt als Deep Learning), zurückzuführen. Viele der Tools zur Verwaltung und Nutzung von KI sind gemeinfreie Open-Source-Ressourcen. Dies erleichtert ihre Einführung und ermöglicht die Behebung von Softwarefehlern durch Crowdsourcing. Beispiele solcher Tools sind TensorFlow (Google), Michelangelo (Uber) und Cognitive Toolkit (Microsoft). Einige Unternehmen und Forscher geben auch kuratierte Trainingsdatensätze und Trainingswerkzeuge öffentlich weiter, um die Verbreitung von KI-Technologien zu unterstützen.

Die jüngsten Fortschritte im KI-Bereich sind z. T. dem exponentiellen Anstieg der Rechenleistung zu verdanken, wobei sich auch der Effekt des Mooreschen Gesetzes bemerkbar macht (das besagt, dass sich die Anzahl der Transistoren in einem integrierten Schaltkreis etwa alle zwei Jahre verdoppelt). Zusammen führen diese beiden Entwicklungen dazu, dass KI-Algorithmen enorme Datenmengen schnell verarbeiten können. Wenn KI-Projekte die Konzeptphase verlassen und aus ihnen kommerzielle Anwendungen werden, steigt der Bedarf an spezialisierten und teuren Cloud-Computing- und Grafikprozessor-Ressourcen. Die bei KI-Systemen zu beobachtenden Trends gehen mit einem gewaltigen Anstieg der erforderlichen Rechenleistung einher. Einer Schätzung zufolge wurde für das größte Experiment der letzten Zeit, AlphaGo Zero, 300 000 Mal mehr Rechenleistung benötigt als für das größte Experiment sechs Jahre zuvor (OpenAI, 2018[53]). Die Erfolge von AlphaGo Zero im Go- und Schachspiel wurden mit einer Rechenleistung erzielt, die die Leistung der zehn leistungsstärksten Supercomputer der Welt zusammengenommen übertreffen dürfte (OECD, 2018[52]).

Datenzugang und -nutzung

Datenzugang und Datenaustausch können Fortschritte im KI-Bereich beschleunigen oder umgekehrt bremsen

Zum Training und zur Weiterentwicklung aktueller ML-Technologien werden kuratierte und genaue Daten benötigt. Der Zugang zu qualitativ hochwertigen Datensätzen ist für die KI-Entwicklung daher von entscheidender Bedeutung. Folgende Faktoren sind für Datenzugang und Datenaustausch relevant und können Fortschritte im KI-Bereich beschleunigen oder umgekehrt bremsen (OECD, erscheint demnächst[54]):

- **Standards**: Standards bzw. Normen sind erforderlich, um Interoperabilität sowie die Weiterverwendung von Daten in verschiedenen Anwendungen zu ermöglichen, den Zugang zu fördern und sicherzustellen, dass Daten auffindbar, katalogisiert und/oder durchsuchbar und effektiv weiterverwendbar sind.
- **Risiken**: „Data-Sharing", d. h. der Austausch bzw. die Weitergabe oder gemeinsame Nutzung von Daten, kann für Privatpersonen, Organisationen und Staaten mit verschiedenen Risiken verbunden sein, darunter Vertraulichkeits- und Datenschutzverletzungen, Risiken in Bezug auf Rechte geistigen Eigentums und kommerzielle Interessen sowie potenzielle Risiken für die nationale und die digitale Sicherheit.
- **Datenkosten**: Die Datengewinnung, der Datenzugang, die gemeinsame Nutzung oder Weiterverwendung von Daten erfordern Vorab- und Folgeinvestitionen. Neben den Investitionen für die Datenerfassung sind zusätzliche Investitionen für die Datenbereinigung, die Datenkuratierung, die Pflege der Metadaten, die Datenspeicherung und -verarbeitung sowie eine sichere IT-Infrastruktur erforderlich.
- **Anreize**: Marktbasierte Ansätze können Anreize schaffen, den Zugang zu Datenmärkten und -plattformen, die Daten kommerzialisieren und Mehrwertdienstleistungen wie Zahlungs- und Datenaustauschinfrastrukturen anbieten, zu ermöglichen und Daten mit ihnen auszutauschen.
- **Unsicherheiten bezüglich des Dateneigentums**: Verschiedene rechtliche Bestimmungen – Regelungen zum Schutz geistigen Eigentums, strafrechtliche Regelungen (auch in Bezug auf Cyberkriminalität), Wettbewerbsrecht und Datenschutzgesetze – haben in Verbindung mit der Vielzahl der an der Erzeugung von Daten beteiligten Akteuren zu Unsicherheiten in der Frage des Dateneigentums geführt.
- **Nutzerbefähigung, einschließlich KI-Agenten**: Nutzerbefähigung und eine leichtere Übertragbarkeit der Daten – sowie die Einführung wirksamer Einwilligungsmechanismen und Wahlmöglichkeiten für die betroffenen Personen – können Einzelne und Unternehmen dazu bewegen, personenbezogene oder geschäftliche Daten weiterzugeben. Teilweise wird auch die Ansicht vertreten, dass KI-Agenten, die die Präferenzen einer Person kennen, ihr helfen könnten, den komplexen Datenaustausch mit anderen KI-Systemen auszuhandeln (Neppel, 2017[55]).
- **Vertrauenswürdige Dritte:** Dritte können Vertrauen schaffen und den Austausch und die Weiterverwendung von Daten unter allen Akteuren erleichtern. Datenmittler können als Zertifizierungsstellen fungieren. Vertrauenswürdige Datenaustauschplattformen, wie z. B. Datentrusts, liefern qualitativ hochwertige Daten.

Institutionelle Prüfungskommissionen stellen zudem sicher, dass die legitimen Interessen Dritter respektiert werden.

- **Repräsentativität der Daten**: Die Prognosen von KI-Systemen beruhen auf Mustern, die in Trainingsdatensätzen festgestellt wurden. Daher müssen Trainingsdatensätze sowohl aus Gründen der Genauigkeit als auch der Fairness inklusiv, vielfältig und repräsentativ sein, damit bestimmte Gruppen nicht unter- oder falsch repräsentiert werden.

Bestimmte Politikmaßnahmen können den Datenzugang und -austausch für die KI-Entwicklung verbessern

Zur Verbesserung des Datenzugangs und Datenaustauschs bieten sich u. a. folgende Maßnahmen an (OECD, erscheint demnächst[54]):

- **Zugang zu Daten des öffentlichen Sektors gewähren**, einschließlich Open-Government-Daten, Geodaten (z. B. Karten) und Verkehrsdaten.
- **Datenaustausch im privaten Sektor erleichtern**, in der Regel auf freiwilliger Basis oder, im Fall verbindlicher Maßnahmen, beschränkt auf vertrauenswürdige Nutzer. Besondere Schwerpunktbereiche sind „Daten von öffentlichem Interesse", Daten in netzgebundenen Sektoren wie Verkehr und Energie (zur Sicherung der Interoperabilität der Dienste) sowie die Übertragbarkeit personenbezogener Daten.
- **Statistische/Datenanalyse-Kapazitäten entwickeln**, indem Technologiezentren eingerichtet werden, die Unterstützung und Orientierungshilfen bei der Nutzung und Analyse von Daten bieten.
- **Nationale Datenstrategien entwickeln**, um die Kohärenz nationaler Data-Governance-Rahmen und ihre Vereinbarkeit mit nationalen KI-Strategien zu gewährleisten.

Es werden technische Lösungen für Datenengpässe entwickelt

Die Leistungsfähigkeit einiger ML-Algorithmen, z. B. solcher zur Bilderkennung, übersteigt bereits die durchschnittlichen menschlichen Fähigkeiten. Um zu diesem Punkt zu gelangen, mussten sie jedoch mit großen Datenbanken mit Millionen von beschrifteten Bildern trainiert werden. Daher wird aktiv an der Entwicklung maschineller Lernverfahren gearbeitet, durch die sich der Datenbedarf zum Training von KI-Systemen verringert. Dazu bieten sich mehrere Methoden an:

- **Tiefes bestärkendes Lernen (*deep reinforcement learning*)** ist eine ML-Technik, die tiefe neuronale Netze mit bestärkendem Lernen kombiniert (vgl. Unterabschnitt „Cluster 2: ML-Techniken" in Kapitel 1). Die Netze lernen dabei, bestimmte Verhaltensweisen zu bevorzugen, die zum gewünschten Ergebnis führen (Mousave, S., M. Schukat und E. Howley, 2018[56]). Verschiedene KI-Agenten konkurrieren in einer komplexen Umgebung miteinander, in der sie für ihre Handlungen entweder belohnt oder bestraft werden, je nachdem, ob diese zum gewünschten Ergebnis führen oder nicht. Die Agenten passen ihr Handeln dann entsprechend diesem „Feedback" an.[8]
- Beim **Transferlernen oder Vortraining** (Pan, S. und Q. Yang, 2010[57]) werden Modelle wiederverwendet, die zur Ausführung verschiedener Aufgaben im selben Bereich trainiert wurden. Zum Beispiel könnten einige Schichten eines Modells,

das darauf trainiert ist, Bilder von Katzen zu erkennen, wiederverwendet werden, um Bilder von blauen Kleidern zu erkennen. Gelingt dies, wären wesentlich kleinere Bilderdatensammlungen erforderlich als für traditionelle AL-Algorithmen (Jain, 2017[58]).

- Durch **erweitertes Datenlernen (*augmented data learning*)** oder „Datensynthetisierung" können Daten durch Simulationen oder Interpolationen aus vorhandenen Daten künstlich erzeugt werden. Dadurch erhöht sich das Datenvolumen und wird der Lernprozess effizienter. Diese Methode bietet sich vor allem an, wenn die Datennutzung durch Datenschutzbestimmungen eingeschränkt ist oder wenn Szenarien simuliert werden sollen, die in der Realität nur selten auftreten (Kasten 4.7).[9]
- **Hybride Lernmodelle** können Unsicherheit modellieren, indem sie verschiedene Arten tiefer neuronaler Netze mit probabilistischen oder Bayesschen Ansätzen kombinieren. Die Modellierung von Unsicherheit dient dazu, die Leistung und Nachvollziehbarkeit zu verbessern und die Wahrscheinlichkeit fehlerhafter Prognosen zu verringern (Kendall, 2017[59]).

Aufgrund von Datenschutz-, Vertraulichkeits- und Sicherheitsbedenken könnten Datenzugang und Datenaustausch eingeschränkt werden. Dies könnte zu einem Missverhältnis zwischen der Geschwindigkeit, mit der KI-Systeme lernen können, und den für ihr Training zur Verfügung stehenden Datensätzen führen. Jüngste Fortschritte in der Kryptografie, z. B. bei der sicheren Mehrparteienberechnung (*Multi-party computation – MPC*) und der homomorphen Verschlüsselung, könnten Datenanalysen ermöglichen, bei denen die Datenschutzrechte gewahrt werden. So könnte insbesondere erreicht werden, dass KI-Systeme operieren können, ohne sensible Daten erfassen oder auf sensible Daten zugreifen zu müssen (Kasten 4.8). KI-Modelle sind zunehmend in der Lage, mit verschlüsselten Daten zu arbeiten.[10] Da diese Lösungen jedoch rechenintensiv sind, lassen sie sich möglicherweise nur schwer skalieren (Brundage et al., 2018[37]).

Kasten 4.8. Neue kryptografische Werkzeuge ermöglichen datenschutzgerechte Berechnungen

Die im Bereich der Verschlüsselung erzielten Fortschritte ebnen den Weg für vielversprechende KI-Anwendungen. So könnten ML-Modelle beispielsweise mit kombinierten Daten verschiedener Organisationen trainiert werden. Dabei würden die Daten aller Beteiligten vertraulich bleiben. Dies könnte dazu beitragen, Hindernisse im Zusammenhang mit Datenschutz- oder Vertraulichkeitsbedenken zu überwinden. Die Verschlüsselungstechniken, die solche Berechnungen ermöglichen, sind nicht neu: homomorphe Verschlüsselung und sichere MPC wurden bereits vor Jahren bzw. Jahrzehnten entdeckt. Für die praktische Anwendung waren sie jedoch bislang zu ineffizient. Dank der jüngsten Fortschritte bei den Algorithmen und bei der Umsetzung entwickeln sie sich nun zunehmend zu praktischen Werkzeugen, die Datensätze aus der realen Welt produktiv analysieren können.

- Die **Homomorphe Verschlüsselung** ist eine Technik, mit der Berechnungen an verschlüsselten Daten durchgeführt werden können, ohne dass die unverschlüsselten Daten angezeigt werden müssen.
- Die **sichere MPC** ist die Berechnung einer Funktion von Daten, die aus vielen verschiedenen Quellen stammen, ohne dass Informationen über die Daten einer Quelle an eine andere Quelle weitergegeben werden. Sichere MPC-Protokolle

ermöglichen es mehreren Beteiligten, Algorithmen gemeinsam zu berechnen, wobei die Daten, die die einzelnen Beteiligten in den Algorithmus eingegeben haben, vertraulich bleiben.

Quelle: Brundage et al. (2018[37]), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf; Dowlin (2016[60]), *CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/CryptonetsTechReport.pdf.

KI-Modelle könnten auch Blockchain-Technologien nutzen, die ebenfalls kryptografische Werkzeuge zur sicheren Datenspeicherung einsetzen (Kasten 4.9). Lösungen, die KI- und Blockchain-Technologien kombinieren, könnten die Datenverfügbarkeit erhöhen. Gleichzeitig könnten sie die mit einer unverschlüsselten Datenverarbeitung verbundenen Datenschutz- und Sicherheitsrisiken auf ein Minimum reduzieren.

Kasten 4.9. KI-basierte datenschutzgerechte Identitätsprüfung dank Blockchain-Technologie

Kairos, ein Unternehmen, das Gesichtserkennungslösungen anbietet, hat in sein Portfolio Blockchain-Technologien aufgenommen. Durch die Kombination von Gesichtsbiometrie und Blockchain-Technologie ermöglicht es Nutzern einen besseren Schutz der Privatsphäre. Ein Algorithmus vergleicht das Bild einer Person mit charakteristischen Gesichtsmerkmalen (oder Identifizierungsmerkmalen), bis eine eindeutige Übereinstimmung hergestellt ist. Diese Übereinstimmung wird dann in eine eindeutige, zufällige Zahlenfolge umgewandelt. Das Originalbild kann danach verworfen werden. Diese „biometrische Blockchain“ wird unter der Prämisse aufgebaut, dass Unternehmen oder staatliche Stellen nicht die Identität einer Person kennen müssen, um zu überprüfen, dass sie es tatsächlich mit dieser Person zu tun haben.

Quelle: https://kairos.com/.

Wettbewerb

Die OECD hat sich mit der Frage der Auswirkungen des digitalen Wandels auf den Wettbewerb und die Wettbewerbspolitik auseinandergesetzt (OECD, 2020[61]). Einige dieser Auswirkungen hängen besonders mit KI zusammen. Sie sind Gegenstand dieses Unterabschnitts. Es wird allgemein anerkannt, dass KI einen wettbewerbsfördernden Effekt hat, da sie den Marktzutritt neuer Anbieter erleichtert. Ein Großteil der Aufmerksamkeit, die die Wettbewerbspolitik großen KI-Akteuren schenkt, ist darauf zurückzuführen, dass sie Betreiber von Online-Plattformen und Inhaber großer Datenmengen sind. Sie hängt nicht mit der Nutzung von KI als solcher zusammen.

Speziell im Zusammenhang mit KI stellt sich jedoch die Frage, ob es datengetriebene Netzwerkeffekte gibt. Von einem Netzwerkeffekt spricht man, wenn der einzelne Nutzer umso größeren Nutzen aus einer bestimmten Plattform zieht, je mehr andere Personen diese Plattform ebenfalls nutzen. Mit jedem neuen Nutzer erhält die Plattform z. B. mehr Informationen, mit deren Hilfe ihre Algorithmen lernen, wie sie die Nutzer besser bedienen können (OECD, 2019[62]). Dem steht die Annahme gegenüber, dass im Fall von Daten von abnehmenden Skalenerträgen auszugehen ist: Ist ein bestimmtes Datenvolumen erreicht, führt eine weitere Zunahme der Daten zu immer geringeren Verbesserungen der Prognosen. Daher ist nicht klar, ob KI langfristig Wettbewerbsprobleme hervorrufen könnte (Bajari et al., 2018[63]; OECD, 2016[64]; Varian, 2018[65]).

Im Hinblick auf den geschäftlichen Wert zusätzlicher Daten sind Skaleneffekte möglich. Wenn ein Unternehmen aufgrund seiner im Vergleich zur Konkurrenz etwas besseren Datenqualität deutlich mehr Kunden gewinnen kann, könnte dies zu einer positiven Feedbackschleife führen. Mehr Kunden bedeuten mehr Daten, wodurch sich dieser Kreislauf beschleunigt und im Lauf der Zeit eine immer größere Marktbeherrschung erzielt werden könnte. Zu Skaleneffekten könnte es auch bei dem für den Aufbau effektiver KI-Systeme erforderlichen Fachwissen kommen.

Zudem besteht die Befürchtung, dass Algorithmen durch die Überwachung von Marktbedingungen, Preisen und Reaktionen der Konkurrenz auf Preisänderungen Absprachen zwischen verschiedenen Anbietern erleichtern könnten. Dadurch könnten Unternehmen über neue und verbesserte Instrumente zur Koordinierung von Strategien, zur Festsetzung von Preisen und zur Durchsetzung von Kartellabsprachen verfügen. Etwas spekulativer ist die Sorge, dass im Fall ausgereifterer Deep-Learning-Algorithmen nicht einmal tatsächliche Absprachen zwischen den Unternehmen nötig wären, um eine kartellartige Situation herbeizuführen. Eine solche Situation könnte vielmehr ohne menschliches Eingreifen erreicht werden. Dies würde große Herausforderungen in Bezug auf die Rechtsdurchsetzung mit sich bringen. Nach den geltenden wettbewerbsrechtlichen Bestimmungen müssen Beweise für Absprachen oder „korrespondierende Willenserklärungen“ vorliegen, bevor ein Kartellverstoß festgestellt und geahndet werden kann (OECD, 2017[66]).

Geistiges Eigentum

In diesem Unterabschnitt werden einige mögliche Auswirkungen von KI auf geistiges Eigentum erörtert. Es handelt sich hier um einen Bereich, in dem viel Bewegung herrscht und man gerade erst beginnt, sich auf evidenzbasierte Analysen zu stützen. Regelungen zum Schutz des geistigen Eigentums im KI-Bereich erhöhen im Allgemeinen den Umfang und das Tempo von Neuentdeckungen, Erfindungen und Technologieverbreitung. Damit ähneln sie den Regelungen für andere Technologien, die durch Rechte des geistigen Eigentums geschützt sind. Regelungen zum Schutz des geistigen Eigentums sollen Erfinder, Autoren, Künstler und Markeninhaber belohnen. Die Politik in diesem Bereich muss aber auch das Potenzial von KI als Input für weitere Innovationen berücksichtigen.

Wenn zum Schutz von geistigem Eigentum im KI-Bereich andere Instrumente als Geschäftsgeheimnisse eingesetzt werden, stellen sich neue Fragen in Bezug darauf, wie man Innovatoren dazu anregen kann, KI-Innovationen offenzulegen, einschließlich Algorithmen und deren Training. Das Büro des Europäischen Parlaments hat auf einer Konferenz drei mögliche Arten der KI-Patentierung diskutiert (EPO, 2018[67]). Der erste Typ (*Core AI*) bezieht sich häufig auf Algorithmen, die als mathematische Methoden nicht patentierbar sind. Beim zweiten Typ – trainierte Modelle/ML – könnten Ansprüche auf Variationen und Parameterbereiche problematisch sein. Im dritten Fall könnte KI als ein Werkzeug in einem Anwendungsbereich patentiert werden, der über technische Effekte definiert ist. Andere internationale Organisationen und OECD-Länder untersuchen ebenfalls die Auswirkungen von KI im Bereich des geistigen Eigentums.[11]

Im Zusammenhang mit der Verbreitung von KI stellt sich zudem die Frage, ob die Systeme zum Schutz der geistigen Eigentumsrechte in einer Welt, in der KI-Systeme selbst Erfindungen machen können, angepasst werden müssen (OECD, 2017[68]). Bestimmte KI-Systeme können bereits patentierbare Erfindungen hervorbringen, insbesondere in der Chemie, der Pharmazie und der Biotechnologie. In diesen Bereichen bestehen viele Erfindungen darin, originale Molekülkombinationen für neue Verbindungen zu kreieren oder neue Eigenschaften bestehender Moleküle zu bestimmen. Zum Beispiel gelang es

KnIT, einem von IBM entwickelten ML-Tool, besondere Kinasen zu entdecken (Kinasen sind Enzyme, die als Katalysator für die Übertragung von Phosphatgruppen auf bestimmte Substrate wirken). Diese Kinasen wiesen in einer Gruppe bekannter Kinasen spezifische Eigenschaften auf, die experimentell getestet wurden. Die spezifischen Eigenschaften dieser Moleküle wurden von einer Software entdeckt, und die entsprechenden Erfindungen wurden zum Patent angemeldet. Diese und andere Aspekte von KI und geistigen Eigentumsrechten werden derzeit von den in diesem Bereich kompetenten Organisationen des OECD-Raums untersucht, wie dem Europäischen Patentamt, dem Patent- und Markenamt der Vereinigten Staaten und der Weltorganisation für geistiges Eigentum (WIPO). Fragen im Zusammenhang mit dem Urheberrechtsschutz von mittels KI verarbeiteten Daten könnten dabei ebenfalls geprüft werden.

Kleine und mittlere Unternehmen

Maßnahmen und Programme, die KMU bei der Einführung von KI unterstützen sollen, gewinnen zunehmend an Priorität. Dies ist ein sich rasch entwickelnder Bereich, der allmählich zum Gegenstand evidenzbasierter Analysen wird. Digitale Ökosysteme, die die Einführung und Nutzung von KI in KMU erleichtern, können u. a. durch folgende Arten von Maßnahmen gefördert werden:

- Weiterqualifizierung ist eine entscheidende Komponente, da es KMU oft schwerfällt, sich im Wettbewerb um KI-Fachkräfte zu behaupten.
- Förderung gezielter Investitionen in ausgewählte vertikale Branchen. In Frankreich laufen z. B. Maßnahmen zur Förderung von Investitionen in spezifische KI-Anwendungen in der Landwirtschaft. Solche Maßnahmen können helfen, wenn einzelne KMU es sich nicht leisten können, allein zu investieren (OECD, 2018[13]).
- Unterstützung von KMU beim Datenzugang, u. a. durch die Schaffung von Plattformen für den Datenaustausch.
- Förderung des Zugangs von KMU zu KI-Technologien – auch durch Technologietransfer von öffentlichen Forschungsinstituten – sowie zu Rechenkapazitäten und Cloud-Plattformen (Deutschland, 2018[69]).
- Verbesserung der Finanzierungsmechanismen, um auf KI spezialisierte KMU bei der Ausweitung ihrer Aktivitäten zu unterstützen, z. B. durch neue öffentliche Investitionsfonds sowie mehr Flexibilität und höhere Finanzierungsobergrenzen bei Programmen zur Unterstützung von Investitionen in wissensintensive Unternehmen (Vereinigtes Königreich, 2017[70]). Die Europäische Kommission engagiert sich insbesondere mit dem AI4EU-Projekt, einer On-Demand-KI-Plattform, für europäische KMU.

Günstige Rahmenbedingungen für KI-Innovationen schaffen

Die OECD analysiert derzeit, welche Veränderungen aufgrund von KI und anderen digitalen Transformationen in der Innovationspolitik und ähnlichen KI-relevanten Politikbereichen erforderlich sind (OECD, 2018[48]). Dabei wird geprüft, wie die Anpassungsfähigkeit, die Reaktionsfähigkeit und die Flexibilität der politischen Instrumente und Experimente verbessert werden können. Staatliche Stellen können Experimente nutzen, um kontrollierte Umgebungen zum Testen von KI-Systemen bereitzustellen. Solche Umgebungen sind z. B. Regulatory Sandboxes („regulatorische Sandkästen“) bzw. Reallabore oder Innovationszentren. Die Experimente können dort im „Start-up-Modus“ erfolgen. In

diesem Fall werden sie durchgeführt, ausgewertet und modifiziert und dann entweder ausgeweitet oder umgekehrt zurückgefahren oder schnell wieder aufgegeben.

Eine weitere Möglichkeit, schnellere und effektivere Entscheidungen zu treffen, besteht im Einsatz digitaler Werkzeuge zur Gestaltung von Politikmaßnahmen (einschließlich innovationsfördernder Maßnahmen) und zum Monitoring von Politikzielen. Einige Staaten greifen beispielsweise auf eine „agentenbasierte Modellierung" zurück, um die Auswirkungen verschiedener Politikoptionen auf verschiedene Arten von Unternehmen vorherzusehen.

Die zuständigen staatlichen Stellen können KI-Akteure ermutigen, Selbstregulierungsmechanismen wie Verhaltenskodizes, freiwillige Standards und Best Practices zu entwickeln. Diese Instrumente können den KI-Akteuren dann als Orientierungshilfen während des gesamten KI-Lebenszyklus dienen, d. h. auch bei Monitoring, Berichterstattung und Bewertung sowie bei der Verhinderung nachteiliger Auswirkungen oder des Missbrauchs von KI-Systemen.

Außerdem können die zuständigen staatlichen Stellen gegebenenfalls Kontrollmechanismen für KI-Systeme im öffentlichen und im privaten Sektor einrichten und fördern. Dazu könnten Compliance-Prüfungen, Audits, Konformitätsbewertungen und Zertifizierungsprogramme gehören. Solche Mechanismen könnten auch nützlich sein, um die besonderen Bedürfnisse von KMU und die Sachzwänge, mit denen sie konfrontiert sind, zu berücksichtigen.

Wandel der Arbeitswelt und Kompetenzentwicklung

Beschäftigung

KI dürfte die Menschen bei einigen Aufgaben ergänzen und bei anderen ersetzen und neue Arbeitsformen schaffen

Die OECD hat sich eingehend mit den Auswirkungen des digitalen Wandels auf die Beschäftigung und den daraus erwachsenden Konsequenzen für die Politik befasst (OECD, 2020[61]). Die Entwicklungen im Bereich KI schreiten rasch voran und die evidenzbasierte Analyse steht erst am Anfang. Sicher ist jedoch bereits, dass es mit der zunehmenden Verbreitung von KI zu erheblichen Veränderungen in der Arbeitswelt kommen wird. Künstliche Intelligenz dürfte die Menschen bei einigen Aufgaben ergänzen und bei anderen ersetzen und neue Arbeitsformen schaffen. In diesem Abschnitt werden einige der Veränderungen erörtert, die KI auf den Arbeitsmärkten auslösen dürfte. Dabei geht es auch um die Frage, wie die Politik den Übergang zu einer KI-Wirtschaft begleiten kann.

KI dürfte zu Produktivitätssteigerungen führen

KI wird die Produktivität voraussichtlich auf zweierlei Weise erhöhen. Erstens werden einige Tätigkeiten, die bisher von Menschen ausgeführt wurden, automatisiert werden. Zweitens werden Systeme dank Maschinenautonomie mit reduzierter oder ohne menschliche Kontrolle funktionieren und sich an die jeweiligen Umstände anpassen (OECD, 2017[68]; Autor, D. und A. Salomons, 2018[71]). Eine in zwölf Industrieländern durchgeführte Untersuchung ergab, dass KI die Arbeitsproduktivität bis 2035 um bis zu 40 % gegenüber dem erwarteten Basisniveau steigern könnte (Purdy, M. und P. Daugherty, 2016[72]). Konkrete Beispiele gibt es viele. Watson von IBM etwa unterstützt Kundenberater der französischen Bank Crédit Mutuel dabei, Kundenanfragen um 60 % schneller zu beantworten.[12] Der Chatbot von Alibaba bearbeitete im Schlussverkauf 2017 mehr als

95 % der Kundenanfragen. Dadurch konnten sich die menschlichen Kundenbetreuer um die komplizierteren oder individuelleren Anfragen kümmern (Zeng, 2018[73]). Eine höhere Arbeitsproduktivität führt theoretisch zu höheren Löhnen, da die Wertschöpfung je Beschäftigten steigt.

Teams, in denen Menschen und KI zusammenwirken, können Fehler begrenzen helfen und menschlichen Arbeitskräften neue Möglichkeiten eröffnen. Es hat sich gezeigt, dass solche Teams produktiver sind als KI oder menschliche Arbeitskräfte allein (Daugherty, P. und H. Wilson, 2018[74]). Bei BMW führte die Bildung solcher gemischten Teams in der Fertigung beispielsweise zu einer Steigerung der Produktivität um 85 % im Vergleich zu nicht integrierten Teams. Bei Walmart verwalten Roboter den Lagerbestand, sodass sich das Ladenpersonal ganz auf die Kundenbetreuung konzentrieren kann. Und wenn Radiologen mit Hilfe von KI-Modellen Röntgenaufnahmen auf Tuberkulose prüfen, liegt die Genauigkeit der Diagnose bei 100 % – und ist damit höher, als wenn nur auf KI oder nur auf menschliche Urteilskraft vertraut wird (Lakhani, P. und B. Sundaram, 2017[75]).

KI kann auch helfen, bereits automatisierte Arbeitsschritte zu verbessern und zu beschleunigen. Dadurch können Unternehmen ihr Produktionsvolumen erhöhen und gleichzeitig ihre Kosten senken. Werden die niedrigeren Kosten an andere Unternehmen oder Privatpersonen weitergegeben, ist mit einer steigenden Nachfrage nach den produzierten Waren zu rechnen. Dies erhöht wiederum die Nachfrage nach Arbeitskräften sowohl innerhalb der betreffenden Unternehmen – z. B. in der Produktion – als auch in anderen Unternehmen, die Vorleistungen erbringen.

KI dürfte die Art der Aufgaben verändern, die automatisiert werden können, und diesen Prozess vielleicht beschleunigen

Automatisierung ist kein neues Phänomen, KI dürfte jedoch die Art der Aufgaben verändern, die automatisiert werden können, und diesen Prozess vielleicht insgesamt beschleunigen. Im Gegensatz zu Computern sind KI-Technologien nicht streng vorprogrammiert und regelbasiert. Computer haben in Routineberufen mit mittleren Qualifikationsanforderungen zu einem tendenziellen Rückgang der Beschäftigung geführt. Neue KI-gestützte Anwendungen sind zunehmend in der Lage, relativ komplexe Aufgaben auszuführen, bei denen Vorhersagen angestellt werden müssen (vgl. Kapitel 3). Zu diesen Aufgaben gehören Transkription, Übersetzung, das Führen von Fahrzeugen, die Diagnose von Krankheiten und die Beantwortung von Kundenanfragen (Graetz, G. und G. Michaels, 2018[76]; Michaels, G., A. Natraj und J. Van Reenen, 2014[77]; Goos, M., A. Manning und A. Salomons, 2014[78]).[13]

In einer von der OECD durchgeführten explorativen Studie wurde geschätzt, inwieweit Technologien in der Lage sind, die Fragen der OECD-Erhebung über die Kompetenzen Erwachsener (PIAAC) in den Bereichen Lesekompetenz und alltagsmathematische Kompetenz zu beantworten (Elliott, 2017[79]). Dabei zeigte sich, dass das Leistungsniveau von KI-Systemen im Bereich Lesekompetenz 2017 dem von 89 % der Erwachsenen im OECD-Raum entsprach. Im Umkehrschluss bedeutet dies, dass nur 11 % der Erwachsenen der KI überlegen waren. Der Studie zufolge ist daher damit zu rechnen, dass sich der wirtschaftliche Druck erhöhen wird, für bestimmte lesekompetenzbasierte und alltagsmathematische Aufgaben Rechnerkapazitäten einzusetzen. Dies dürfte dazu führen, dass die Nachfrage nach menschlichen Arbeitskräften für Aufgaben, die eine geringe bis mittlere Lesekompetenz erfordern, entgegen den Trends der jüngsten Vergangenheit zurückgehen wird. Die Studie wies auch auf die Schwierigkeit hin, bildungspolitische Maßnahmen für Erwachsene zu gestalten, mit denen das Kompetenzniveau dieser

Arbeitskräfte über das aktuelle Leistungsniveau der Computer angehoben werden kann. Neben neuen Instrumenten und Anreizen zur Kompetenzentwicklung von Erwachsenen wurde dabei auch vorgeschlagen, kompetenzfördernde Maßnahmen mit anderen Initiativen zu kombinieren, u. a. im Bereich der sozialen Sicherung und des sozialen Dialogs (OECD, 2018[13]).

Der Beschäftigungseffekt der KI wird davon abhängen, wie schnell sie sich in den verschiedenen Sektoren durchsetzen wird

Der Beschäftigungseffekt der künstlichen Intelligenz wird auch davon abhängen, wie schnell sich KI-Technologien in den kommenden Jahrzehnten in den verschiedenen Sektoren entwickeln und verbreiten werden. Allgemein wird erwartet, dass autonome Fahrzeuge einen disruptiven Effekt auf Arbeitsplätze in der Personen- und Warenbeförderung haben werden. Etablierte Lkw-Hersteller wie Volvo und Daimler konkurrieren beispielsweise mit Start-ups wie Kodiak und Einride bei der Entwicklung und Erprobung fahrerloser Lastkraftwagen (Stewart, 2018[80]). Dem Weltverkehrsforum zufolge könnten fahrerlose Lastkraftwagen in den nächsten zehn Jahren auf vielen Straßen immer häufiger anzutreffen sein. Etwa 50 %-70 % der 6,4 Millionen Arbeitsplätze für Berufskraftfahrer in den Vereinigten Staaten und Europa könnten bis 2030 wegfallen (ITF, 2017[81]). Allerdings werden parallel dazu neue Arbeitsplätze geschaffen werden, um die gestiegene Zahl der fahrerlosen Lastkraftwagen zu unterstützen. Fahrerlose Lastkraftwagen könnten die Betriebskosten im Straßengüterverkehr um rd. 30 % senken, vor allem durch Einsparungen bei den Arbeitskosten. Dies könnte traditionelle Speditionsunternehmen aus dem Geschäft drängen und in der Folge zu einem noch schnelleren Rückgang der Arbeitsplätze im Speditionsgewerbe führen.

KI-Technologien werden sich wahrscheinlich auf Tätigkeiten auswirken, die traditionell höhere Qualifikationen voraussetzen

KI-Technologien übernehmen auch Prognoseaufgaben, die traditionell von höher qualifizierten Arbeitskräften – von Juristen bis hin zu medizinischen Fachkräften – ausgeführt werden. So hat beispielsweise ein Roboter-Anwalt Autofahrern dabei geholfen, Strafzettel in Höhe von insgesamt 12 Mio. USD anzufechten (Dormehl, 2018[82]). Im Jahr 2016 übertrafen die Systeme Watson von IBM und DeepMind Health menschliche Ärzte bei der Diagnose seltener Krebsarten (Frey, C. und M. Osborne, 2017[83]). Bei der Vorhersage von Börsenschwankungen erwies sich KI besser als Finanzexperten (Mims, 2010[84]).

KI kann Menschen ergänzen und neue Arbeitsformen schaffen

KI ergänzt Menschen und wird wahrscheinlich auch neue Beschäftigungsmöglichkeiten für menschliche Arbeitskräfte schaffen, insbesondere in Bereichen, in denen komplementäre Kompetenzen zur Prognose und insbesondere menschliche Kompetenzen wie kritisches Denken, Kreativität und Einfühlungsvermögen erforderlich sind (Vereinigte Staaten, 2016[85]; OECD, 2017[20]).

- **Tätigkeiten von Datenwissenschaftlern und ML-Experten**: Zur Erzeugung und Bereinigung von Daten sowie zur Programmierung und Entwicklung von KI-Anwendungen werden Spezialisten benötigt. Doch selbst wenn Daten und maschinelles Lernen neue Aufgaben für Menschen entstehen lassen, ist nicht mit einem allzu großen Zuwachs zu rechnen.

- **Menschenzentrierte Tätigkeiten**: Einige Handlungen sind von Natur aus wertvoller, wenn sie von einem Menschen anstatt von einer Maschine durchgeführt werden (z. B. solche von Berufssportlern, Erziehern oder Verkäufern). Viele halten es für wahrscheinlich, dass sich Menschen zunehmend auf Aufgaben konzentrieren werden, die das Leben anderer Menschen verbessern, z. B. Kinderbetreuung, Sport-Coaching und Pflege von unheilbar Kranken.
- **Urteilstätigkeiten – Bestimmung des Prognosegegenstands**: Am wichtigsten ist vielleicht das Konzept der Beurteilung, d. h. der Entscheidung über den Nutzen einer bestimmten Handlung in einem bestimmten Umfeld. Wenn KI für Prognosen genutzt wird, muss ein Mensch entscheiden, was vorhergesagt werden soll und wozu die Vorhersage dienen soll. Es bedarf Menschen mit Eigenschaften wie Urteilsvermögen und Fairness, um Dilemmas zu formulieren, Situationen zu interpretieren oder einem Text den richtigen Sinn zu entnehmen (OECD, 2018[13]). In der Wissenschaft kann KI beispielsweise Menschen ergänzen. Deren Fähigkeit zum konzeptuellen Denken ist jedoch unerlässlich, etwa um den Forschungsrahmen abzustecken und den Kontext für bestimmte Experimente festzulegen.
- **Urteilstätigkeiten – Bestimmung der optimalen Vorgehensweise**: Eine Entscheidung kann nicht allein auf der Grundlage einer Prognose getroffen werden. Dies zeigt sich schon an einer alltäglichen Entscheidung wie der, ob man auf einen Spaziergang einen Regenschirm mitnimmt oder nicht. Für diese Entscheidung wird eine Prognose über die Regenwahrscheinlichkeit angestellt. Die eigentliche Entscheidung wird aber großenteils von persönlichen Vorlieben abhängen, z. B. davon, was einem unangenehmer ist: nass zu werden oder einen Regenschirm mit sich herumzutragen. So verhält es sich mit vielen Entscheidungen. In der Cybersicherheit muss bei einer Prognose über die potenziell feindliche Natur einer neuen Anfrage das Risiko, eine freundliche Anfrage abzulehnen, gegen die Gefahr abgewogen werden, einer feindlichen Anfrage unautorisierte Informationen zugänglich zu machen.

Die Prognosen über den Nettoeffekt von KI auf das Beschäftigungsvolumen gehen stark auseinander

In den letzten 5 Jahren wurden unterschiedliche Schätzungen über die Gesamtauswirkungen der Automatisierung auf das Beschäftigungsvolumen angestellt (Winick, 2018[86]; MGI, 2017[87]; Frey, C. und M. Osborne, 2017[83]). Frey und Osborne sagten z. B. voraus, dass 47 % der Arbeitsplätze in den Vereinigten Staaten in den nächsten 10-15 Jahren verschwinden könnten. Das McKinsey Global Institute schätzte 2017, dass bei 60 % der Arbeitsplätze rd. ein Drittel der erledigten Aufgaben automatisiert werden könnte. Die Automatisierung ist in diesen Arbeitsplätzen jedoch nicht nur auf die Entwicklung und den Einsatz von KI zurückzuführen, sondern auch auf andere technologische Entwicklungen.

Allerdings ist nicht nur der Umfang der Beschäftigungsverluste schwer vorauszusehen, sondern auch das Volumen der Beschäftigungsschaffung in neuen Bereichen. Einer Studie zufolge könnte KI bis 2025 zu einem Nettozuwachs an 2 Millionen Arbeitsplätzen führen (Gartner, 2017[88]). Dies ist jedoch nicht nur durch die Entstehung neuer Tätigkeitsfelder bedingt, sondern auch durch indirekte Effekte. Beispielsweise wird KI wahrscheinlich die Kosten für die Produktion von Waren und Dienstleistungen senken und deren Qualität erhöhen. Dies dürfte zu einem Anstieg der Nachfrage und damit auch der Beschäftigung führen.

Die jüngsten Schätzungen der OECD tragen der Aufgabenheterogenität in eng definierten Berufen Rechnung, wobei Daten der Internationalen Vergleichsstudie der Kompetenzen Erwachsener (PIAAC) verwendet werden. Beim aktuellen Stand der technologischen Entwicklung sind 14 % der Arbeitsplätze in den OECD-Ländern einem hohen Automatisierungsrisiko ausgesetzt. 32% der Arbeitskräfte werden sich wahrscheinlich mit erheblichen Veränderungen ihrer Arbeitsplätze konfrontiert sehen (Nedelkoska, L. und G. Quintini, 2018[89]). Jugendliche und ältere Arbeitskräfte sind vom Automatisierungsrisiko am stärksten bedroht. Neuere Analysen der OECD zeigen, dass die Beschäftigung in Berufen, die als „stark automatisierbar" eingestuft werden, in 16 europäischen Ländern in 82 % der Regionen zurückgeht. Gleichzeitig wird in 60 % der Regionen eine größere Zunahme der Zahl der Arbeitsplätze mit „geringem Automatisierungsgrad" festgestellt, die diesen Arbeitsplatzverlust ausgleicht. Diese Untersuchung bestätigt die Annahme, dass die Automatisierung das Verhältnis zwischen den verschiedenen Beschäftigungskategorien verschieben könnte, ohne die Gesamtbeschäftigung zu verringern (OECD, 2018[90]).

KI wird die Arbeitswelt verändern

Die Einführung von KI dürfte die Arbeitswelt verändern. KI kann dazu beitragen, Arbeit interessanter zu machen, indem sie Routineaufgaben automatisiert, flexiblere Arbeitsformen ermöglicht und u. U. auch die Vereinbarkeit von Berufs- und Privatleben verbessert. Menschliche Kreativität und Genialität können mit immer leistungsfähigeren Rechen-, Daten- und Algorithmusressourcen kombiniert werden, wodurch neue Aufgaben und Tätigkeiten entstehen können (Kasparov, 2018[91]).

Ganz allgemein könnte KI effizienzsteigernd wirken und so den Wandel an den Arbeitsmärkten beschleunigen. KI-Techniken in Verbindung mit Big Data können Unternehmen heute potenziell bei der Definition der Rollen unterstützen, die Arbeitskräfte übernehmen sollen – und ihnen helfen, für ihre Beschäftigten die jeweils optimalsten Einsatzmöglichkeiten zu finden und ihre Stellen mit den jeweils geeigneten Personen zu besetzen. IBM beispielsweise nutzt KI zur Optimierung innerbetrieblicher Weiterbildungsmaßnahmen und empfiehlt seinen Mitarbeitern auf der Grundlage ihrer bisherigen Leistung, ihrer Karriereziele und des Kompetenzbedarfs des Unternehmens passende Schulungsmodule. Unternehmen wie KeenCorp und Vibe haben Textanalyse-Techniken entwickelt, um Unternehmen bei der Analyse der Mitarbeiterkommunikation zu unterstützen und so die Bewertung von Messgrößen wie Arbeitsmoral, Arbeitskräfteproduktivität und Netzwerkeffekte zu erleichtern (Deloitte, 2017[92]). Mit diesen Informationen könnte KI Unternehmen helfen, die Produktivität ihrer Arbeitskräfte zu optimieren.

Es müssen Parameter für organisatorische Veränderungen festgelegt werden

Es wird zunehmend deutlich, dass neue bzw. überarbeitete Branchenstandards und Technologievereinbarungen zwischen Geschäftsleitung und Beschäftigten erforderlich sind, um zuverlässige, sichere und produktive Arbeitsplätze zu gewährleisten. Der Europäische Wirtschafts- und Sozialausschuss (EWSA) rät den „Interessenträgern, sich gemeinsam für komplementäre KI-Systeme und ihre Ko-Kreation am Arbeitsplatz einzusetzen" (EWSA, 2017[45]). Darüber hinaus ist es wichtig, die Beschäftigungsflexibilität zu erhöhen, ohne die Autonomie der Arbeitskräfte und die Beschäftigungsqualität zu beeinträchtigen, auch in Bezug auf die Gewinnbeteiligung. Der jüngste Tarifvertrag zwischen der deutschen IG Metall und dem Arbeitgeberverband Gesamtmetall verdeutlicht die wirtschaftliche Machbarkeit variabler Arbeitszeiten. Er zeigt, dass Arbeitgeber und Gewerkschaften heute Arbeitsbedingungen aushandeln können, die den organisatorischen Anforderungen der Unternehmen ebenso gerecht werden wie den persönlichen Bedürfnissen der Beschäftigten

(Kindererziehung, Pflege kranker Angehöriger usw.), und zwar ohne dass dazu der gesetzliche Beschäftigungsschutz überarbeitet werden müsste (Byhovskaya, 2018[93]).

Die Nutzung von KI zur Unterstützung der Arbeitsmarktfunktionen ist ebenfalls vielversprechend

Durch KI gelingt es bereits, den Prozess der Stellenvermittlung ebenso wie die Weiterbildung effizienter zu gestalten. KI kann Arbeitsuchenden und insbesondere freigesetzten Arbeitskräften helfen, passende Weiterbildungsprogramme zu finden, um die für neue, expandierende Berufe erforderlichen Kompetenzen zu erwerben. In vielen OECD-Ländern nutzen Arbeitgeber und öffentliche Arbeitsverwaltungen bereits Online-Plattformen zur Stellenbesetzung (OECD, 2018[90]). KI und andere digitale Technologien könnten künftig helfen, innovative und personalisierte Konzepte für Arbeitsuche und Stellenbesetzung zu verbessern und das Matching von Arbeitsnachfrage und Arbeitsangebot effizienter zu gestalten. Die Plattform LinkedIn nutzt KI, um Personalvermittlern zu helfen, die richtigen Kandidaten zu finden, und Kandidaten auf passende Stellen hinzuweisen. Dabei stützt sie sich auf die Profil- und Aktivitätsdaten ihrer 470 Millionen registrierten Nutzer (Wong, 2017[94]).

KI-Technologien, die große Datenmengen nutzen, können auch dazu beitragen, staatliche Stellen, Arbeitgeber und Arbeitskräfte über die lokalen Arbeitsmarktbedingungen zu informieren. Mit diesen Informationen ist es möglich, den aktuellen Kompetenzbedarf zu ermitteln und den künftigen Kompetenzbedarf vorherzusagen, Weiterbildungsressourcen zu steuern und Arbeitsuchende auf freie Stellen aufmerksam zu machen. Projekte zur Entwicklung solcher Arbeitsmarktinformationen laufen z. B. bereits in Finnland, der Tschechischen Republik und Lettland (OECD, 2018[90]).

Regeln für die Nutzung von Beschäftigtendaten festlegen

KI ist nur mit großen Datensätzen produktiv, bei der Nutzung personenbezogener Beschäftigtendaten bestehen jedoch Risiken. Dies gilt insbesondere, wenn die KI-Systeme, die die Daten analysieren, nicht transparent sind. Zur Personal- und Produktivitätsplanung wird zunehmend auf Beschäftigtendaten und Algorithmen zurückgegriffen. Politikverantwortliche und sonstige Akteure sollten daher untersuchen, wie sich Datenerfassung und -verarbeitung auf Beschäftigungsaussichten und -bedingungen auswirken. Über Anwendungen, Fingerabdrücke, Wearables und Sensoren können in Echtzeit Daten erfasst werden, aus denen sich der Standort und Arbeitsplatz der betreffenden Personen ablesen lässt. Im Kundenservice wird z. B. mit KI-Software die Freundlichkeit des Tonfalls von Mitarbeitern analysiert. Laut Aussage betroffener Mitarbeiter werden dabei allerdings keine Sprachmuster berücksichtigt, und es ist offenbar schwierig, die Auswertungsergebnisse anzufechten (UNI, 2018[95]).

In einigen Ländern wird derzeit an Vereinbarungen über Arbeitnehmerdaten und das „Recht auf Abschalten“ (*right to disconnect*) gearbeitet. Das französische Telekommunikationsunternehmen Orange France Telecom gehörte zusammen mit fünf Gewerkschaftsverbänden zu den ersten, die sich auf Verpflichtungen zum Schutz von Arbeitnehmerdaten geeinigt haben. Vereinbart wurden dabei insbesondere Regelungen für eine transparente Datennutzung, Schulungen und die Einführung neuer Geräte. Um Regelungslücken in Bezug auf Arbeitnehmerdaten zu schließen, kann es auch sinnvoll sein, Data-Governance-Organe in den Unternehmen einzurichten, Rechenschaftspflichten für die Verwendung von (personenbezogenen) Daten einzuführen sowie einen Anspruch auf Datenübertragbarkeit, Erklärung und Datenlöschung zu schaffen (UNI, 2018[95]).

Den Wandel begleiten

Die Politik muss den Übergang zu einer KI-basierten Wirtschaft begleiten, insbesondere mit Maßnahmen im Bereich des Sozialschutzes

Wenn der organisatorische Wandel nicht mit der technologischen Entwicklung Schritt hält, kann es an den Arbeitsmärkten zu Verwerfungen kommen (OECD, 2018[13]). Dass langfristig Grund zu Optimismus besteht, heißt nicht, dass der Übergang zu einer zunehmend KI-basierten Wirtschaft reibungslos verlaufen wird: Einige Wirtschaftszweige werden wahrscheinlich wachsen, andere werden schrumpfen. Manche Arbeitsplätze werden verschwinden, andere werden entstehen. Daher wird die wichtigste beschäftigungspolitische Herausforderung in Bezug auf KI darin bestehen, diese Veränderungen unterstützend zu begleiten. Dabei helfen können soziale Sicherheitsnetze, Krankenversicherungsregelungen, eine progressive Besteuerung von Arbeit und Kapital sowie Bildungsmaßnahmen. OECD-Analysen zeigen, dass zudem Maßnahmen in anderen Bereichen, z. B. im Wettbewerbsrecht, berücksichtigt werden müssen, die sich auf Marktkonzentration, Marktmacht und Einkommensverteilung auswirken können (OECD, 2020[61]).

Zur Nutzung von KI erforderliche Kompetenzen

Mit dem Wandel der Arbeitswelt verändern sich auch die Kompetenzanforderungen

Mit dem Wandel der Arbeitswelt verändern sich auch die Kompetenzanforderungen, die an die Beschäftigten gerichtet werden (OECD, 2017[96]; Acemoglu, D. und P. Restrepo, 2018[97]; Brynjolfsson, E. und T. Mitchell, 2017[98]). In diesem Unterabschnitt wird erörtert, welche Auswirkungen künstliche Intelligenz auf die von den Arbeitskräften benötigten Kompetenzen haben könnte. Auch hier verändert sich die Situation rasch, und die evidenzbasierte Analyse steht noch ganz am Anfang. In der Bildungspolitik dürften Anpassungen erforderlich sein, um die Möglichkeiten für lebenslanges Lernen, Fort- und Weiterbildung und Kompetenzentwicklung zu verbessern. Wie andere Technologien auch dürfte KI voraussichtlich eine zusätzliche Nachfrage nach drei Arten von Kompetenzen erzeugen: Erstens werden **Fachkompetenzen** zur Programmierung und Entwicklung von KI-Anwendungen benötigt werden. Dazu könnten Kompetenzen für KI-bezogene Grundlagenforschung, Technik und Anwendungen sowie Datenwissenschaft und rechnergestütztes Denken gehören. Zweitens werden **allgemeine Kompetenzen** zur Nutzung von KI benötigt werden, z. B. um gemischte Teams aus Menschen und KI im Fertigungsbereich und in der Qualitätskontrolle zu ermöglichen. Drittens wird KI **komplementäre Kompetenzen** erfordern, so z. B. kritisches Denken, Kreativität, Innovationsgeist und unternehmerische Initiative sowie Einfühlungsvermögen (Vereinigte Staaten, 2016[85]; OECD, 2017[20]).

Es bedarf Initiativen zum Aufbau und zur Weiterentwicklung von KI-Kompetenzen, um Kompetenzengpässen entgegenzuwirken

Es wird erwartet, dass die Kompetenzengpässe im KI-Bereich zunehmen und infolge der wachsenden Nachfrage nach Spezialisten in Bereichen wie maschinellem Lernen immer deutlicher zutage treten werden. KMU, Hochschulen und öffentliche Forschungszentren konkurrieren bereits mit den großen marktführenden Unternehmen um Talente. Im öffentlichen, im privaten und im Hochschulsektor werden Initiativen zum Aufbau und zur Weiterentwicklung von KI-Kompetenzen gestartet. Zum Beispiel hat die Regierung von Singapur an der Singapore Management University ein fünfjähriges Forschungsprogramm zur Governance von KI und zur Datennutzung eingerichtet. Das dortige Centre for AI &

Data Governance betreibt industrierelevante Forschung zu den Themen KI und Industrie, Gesellschaft und Vermarktung. Das Massachusetts Institute of Technology (MIT) hat 1 Mrd. USD für die Gründung des Schwarzman College of Computing bereitgestellt. Ziel ist es, Studierende und Forscher aller Fachrichtungen Mittel an die Hand zu geben, um ihre jeweiligen Fachrichtungen mithilfe von Informatik und KI voranzubringen und so umgekehrt auch die KI voranzubringen.

Die Kompetenzengpässe im KI-Bereich haben einige Länder auch dazu veranlasst, die Zuwanderungsverfahren für hochqualifizierte Fachkräfte zu vereinfachen. So hat das Vereinigte Königreich beispielsweise die Zahl der Tier-1-Visa (für Hochqualifizierte) auf 2 000 pro Jahr verdoppelt und das Verfahren zur Erlangung einer Arbeitsgenehmigung für Spitzenstudenten und -forscher beschleunigt (Vereinigtes Königreich, 2017[99]). Kanada hat die Bearbeitungsfrist für Visumanträge von Hochqualifizierten auf zwei Wochen verkürzt und Visumbefreiungen für kurzfristige Forschungsaufträge eingeführt. Dies war Teil der Global Skills Strategy 2017, um hochqualifizierte Arbeitskräfte und Forscher aus dem Ausland anzuziehen (Kanada, 2017[100]).

Zur Nutzung von KI erforderliche allgemeine Kompetenzen

Alle OECD-Länder führen Kompetenzbeurteilungen durch und stellen Prognosen zum kurz-, mittel- oder langfristigen Kompetenzbedarf an. In Finnland wurde ein Programm für künstliche Intelligenz vorgeschlagen, das ein Kompetenzkonto bzw. Gutscheinsystem für lebenslanges Lernen umfasst, das die Nachfrage nach allgemeiner und beruflicher Bildung erhöhen soll (Finnland, 2017[101]). Das Vereinigte Königreich engagiert sich für eine vielfältige Erwerbsbevölkerung mit KI-Fachkompetenzen und investiert rd. 406 Mio. GBP (530 Mio. USD) in die Kompetenzentwicklung. Der Schwerpunkt liegt dabei auf Naturwissenschaften, Technologie, Ingenieurwesen und Mathematik sowie auf der Ausbildung von Informatiklehrern (Vereinigtes Königreich, 2017[99]).

Zunehmend wichtig im Arbeitsleben wird die „branchenspezifische Zweisprachigkeit“: Es reicht dann nicht mehr, sich nur im eigenen Fachbereich auszukennen, z. B. in Wirtschaft, Biologie oder Recht, erforderlich sind zudem Kompetenzen in KI-Techniken wie maschinellem Lernen. Daher kündigte das MIT im Oktober 2018 die bedeutendste Änderung seiner Organisationsstruktur seit fünfzig Jahren an – die Schaffung eines neuen Informatikinstituts, das von den Ingenieurwissenschaften unabhängig und mit allen anderen Fachbereichen verflochten sein wird. Dort sollen solche „zweisprachigen“ Studierenden ausgebildet werden, die KI und ML auf die Herausforderungen ihrer jeweiligen Fachrichtung anwenden können. Die Art und Weise, wie Informatik am MIT gelehrt wird, wandelt sich damit grundlegend. Das MIT hat 1 Mrd. USD für die Gründung dieser neuen Fakultät zur Verfügung gestellt (MIT, 2018[102]).

Komplementäre Kompetenzen

Soziale Kompetenzen („Soft Skills“) werden immer wichtiger. Nach dem aktuellen Stand der Forschung geht es dabei insbesondere um menschliches Urteilsvermögen, Analysefähigkeiten und zwischenmenschliche Kommunikation (Agrawal, A., J. Gans und A. Goldfarb, 2018[103]; Deming, 2017[104]; Trajtenberg, 2018[105]). Die OECD wird ihre internationale Schulleistungsstudie PISA 2021 um ein Modul ergänzen, mit dem kreative und kritische Denkfähigkeiten getestet werden sollen. Die Ergebnisse sollen einen Vergleich des Kreativitätspotenzials der verschiedenen Länder ermöglichen, der dann als Grundlage für von der Politik und den Sozialpartnern zu ergreifenden Maßnahmen dienen kann.

Messung

Die Umsetzung einer menschenzentrierten und vertrauenswürdigen KI hängt vom Kontext ab. Ein wichtiger Teil des Engagements der Politik für eine menschenzentrierte KI wird allerdings darin bestehen, Ziele und Messgrößen zur Bewertung der Leistung von KI-Systemen festzulegen, insbesondere im Hinblick auf die Aspekte Genauigkeit, Effizienz, gesellschaftlicher Wohlstand, Fairness und Robustheit.

Literaturverzeichnis

Abrams, M. et al. (2017), *Artificial Intelligence, Ethics and Enhanced Data Stewardship*, The Information Accountability Foundation, Plano, Texas, http://informationaccountability.org/wp-content/uploads/Artificial-Intelligence-Ethics-and-Enhanced-Data-Stewardship.pdf. [16]

Acemoglu, D. und P. Restrepo (2018), "Artificial Intelligence, Automation and Work", *NBER Working Paper* No. 24196. [97]

Agrawal, A., J. Gans und A. Goldfarb (2018), "Economic policy for artificial intelligence", *NBER Working Paper*, No. 24690, http://dx.doi.org/10.3386/w24690. [49]

Agrawal, A., J. Gans und A. Goldfarb (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business School Press, Brighton, MA. [103]

Autor, D. und A. Salomons (2018), "Is automation labor-displacing? Productivity growth, employment, and the labor share", *NBER Working Paper*, No. 24871, http://dx.doi.org/10.3386/w24871. [71]

Bajari, P. et al. (2018), "The impact of big data on firm performance: An empirical investigation", *NBER Working Paper*, No. 24334, http://dx.doi.org/10.3386/w24334. [63]

Barocas, S. und A. Selbst (2016), "Big Data's Disparate Impact", *California Law Review*, Vol. 104, S. 671-729, http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf. [29]

Berk, R. und J. Hyatt (2015), "Machine Learning Forecasts of Risk to Inform Sentencing Decisions", *Federal Sentencing Reporter*, Vol. 27/4, S. 222-228, http://dx.doi.org/10.1525/fsr.2015.27.4.222. [23]

Borges, G. (2017), *Liability for Machine-Made Decisions: Gaps and Potential Solutions*, Präsentation bei der OECD-Konferenz "AI: Intelligent Machines, Smart Policies", Paris, 26.-27. Oktober, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-borges.pdf. [44]

Brundage, M. et al. (2018), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Centre for a New American Security, Electronic Frontier Foundation and Open AI, arXiv:1802.07228, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf. [37]

Brynjolfsson, E. und T. Mitchell (2017), "What can machine learning do? Workforce implications", *Science*, Vol. 358/6370, S. 1530-1534, http://dx.doi.org/10.1126/science.aap8062. [98]

Brynjolfsson, E., D. Rock und C. Syverson (2017), "Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics", *NBER Working Paper*, No. 24001, http://dx.doi.org/10.3386/w24001. [50]

Burgess, M. (2016), "Holding AI to account: Will algorithms ever be free of bias if they are created by humans?", *WIRED,* 11. Januar, https://www.wired.co.uk/article/creating-transparent-ai-algorithms-machine-learning. [31]

Byhovskaya, A. (2018), *Overview of the national strategies on work 4.0: a coherent analysis of the role of the social partners*, Europäischer Wirtschafts- und Sozialausschuss, Brüssel, https://www.eesc.europa.eu/sites/default/files/files/qe-02-18-923-en-n.pdf. [93]

Cellarius, M. (2017), *Artificial Intelligence and the Right to Informational Self-determination*, The Forum Network, OECD, Paris, https://www.oecd-forum.org/users/75927-mathias-cellarius/posts/28608-artificial-intelligence-and-the-right-to-informational-self-determination. [9]

Chouldechova, A. (2016), "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", arXiv:1610.07524, https://arxiv.org/abs/1610.07524. [24]

Citron, D. und F. Pasquale (2014), "The Scored Society: Due Process for Automated Predictions", *Washington Law Review*, Vol. 89, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376209. [36]

Cockburn, I., R. Henderson und S. Stern (2018), "The impact of artificial intelligence on innovation"*, NBER Working Paper*, No. 24449, http://dx.doi.org/10.3386/w24449. [51]

Crawford, K. (2016), "Artificial Intelligence's White Guy Problem", *The New York Times,* 26. Juni, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=0. [30]

Daugherty, P. und H. Wilson (2018), *Human Machine: Reimagining Work in the Age of AI*, Harvard Business Review Press, Cambridge, MA. [74]

Deloitte (2017), *HR Technology Disruptions for 2018: Productivity, Design and Intelligence Reign*, Deloitte, http://marketing.bersin.com/rs/976-LMP-699/images/HRTechDisruptions2018-Report-100517.pdf. [92]

Deming, D. (2017), "The Growing Importance of Social Skills in the Labor Market", *The Quarterly Journal of Economics*, Vol. 132/4, S. 1593-1640, http://dx.doi.org/10.1093/qje/qjx022. [104]

Deutschland (2018), "Eckpunkte der Bundesregierung für eine Strategie Künstliche Intelligenz", Gemeinsame Pressemitteilung der Bundesregierung und des BMWi, 18. Juli, Bundesministerium für Wirtschaft und Energie, https://www.bmwi.de/Redaktion/EN/Pressemitteilungen/2018/20180718-key-points-for-federal-government-strategy-on-artificial-intelligence.html. [69]

Dormehl, L. (2018), "Meet the British whiz kid who fights for justice with robo-lawyer sidekick", *Digital Trends,* 25. März, https://www.digitaltrends.com/cool-tech/robot-lawyer-free-acess-justice/. [82]

Doshi-Velez, F. et al. (2017), "Accountability of AI under the law: The role of explanation", *arXiv* arXiv:1711.01134v2, https://arxiv.org/pdf/1711.01134v2.pdf. [28]

Dowlin, N. (2016), "CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy", *MSR-TR-2016-3* Microsoft Research, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/CryptonetsTechReport.pdf. [60]

Dressel, J. und H. Farid (2018), "The accuracy, fairness and limits of predicting recidivism", *Science Advances*, Vol. 4/1, http://advances.sciencemag.org/content/4/1/eaao5580. [33]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264284395-en. [79]

EPO (2018), *Patenting Artificial Intelligence – Conference Summary*, Europäisches Patentamt, München, 30. Mai, http://documents.epo.org/projects/babylon/acad.nsf/0/D9F20464038C0753C125829E0031B814/$FILE/summary_conference_artificial_intelligence_en.pdf. [67]

EWSA (2017), *Künstliche Intelligenz – die Auswirkungen der künstlichen Intelligenz auf den (digitalen) Binnenmarkt sowie Produktion, Verbrauch, Beschäftigung und Gesellschaft*, Europäischer Wirtschafts- und Sozialausschuss, Brüssel, https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence. [45]

Finnland (2017), "Artificial intelligence programme" Webseite, Ministry of Economic Affairs and Employment, https://tem.fi/en/artificial-intelligence-programme. [101]

Flanagan, M., D. Howe und H. Nissenbaum (2008), "Embodying values in technology: Theory and practice", in van den Hoven, Jeroen und J. Weckert (Hrsg.), S. 322-353, http://dx.doi.org/10.1017/cbo9780511498725.017. [15]

Freeman, R. (2017), *Evolution or Revolution? The Future of Regulation and Liability for AI*, Präsentation bei der OECD-Konferenz "AI: Intelligent Machines, Smart Policies", Paris, 26.-27. Oktober, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-freeman.pdf. [40]

Frey, C. und M. Osborne (2017), "The future of employment: How susceptible are Jobs to computerisation?", *Technological Forecasting and Social Change*, Vol. 114, S. 254-280, http://dx.doi.org/10.1016/j.techfore.2016.08.019. [83]

Gartner (2017), "Gartner says by 2020, artificial intelligence will create more jobs than it eliminates", Pressemitteilung, 13. Dezember, Gartner, https://www.gartner.com/en/newsroom/press-releases/2017-12-13-gartner-says-by-2020-artificial-intelligence-will-create-more-jobs-than-it-eliminates. [88]

Golson, J. (2016), "Google's self-driving cars rack up 3 million simulated miles every day", *The Verge,* 1. Februar, https://www.theverge.com/2016/2/1/10892020/google-self-driving-simulator-3-million-miles. [41]

Goodfellow, I., J. Shlens und C. Szegedy (2015), "Explaining and harnessing adversarial examples", arXiv:1412.6572, https://arxiv.org/pdf/1412.6572.pdf. [38]

Goos, M., A. Manning und A. Salomons (2014), "Explaining Job Polarization: Routine-Biased Technological Change and Offshoring", *American Economic Review*, Vol. 104/8, S. 2509-2526, http://dx.doi.org/10.1257/aer.104.8.2509. [78]

Graetz, G. und G. Michaels (2018), "Robots at Work", *Review of Economics and Statistics*, Vol. 100/5, S. 753-768, http://dx.doi.org/10.1162/rest_a_00754. [76]

Harkous, H. (2018), "Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning", *arXiv* arXiv:1802.02561v2, https://arxiv.org/pdf/1802.02561.pdf. [14]

Heiner, D. und C. Nguyen (2018), "Amplify Human Ingenuity with Intelligent Technology", *Shaping Human-Centered Artificial Intelligence, A.Ideas Series*, The Forum Network, OECD, Paris, https://www.oecd-forum.org/users/86008-david-heiner-and-carolyn-nguyen/posts/30653-shaping-human-centered-artificial-intelligence. [6]

Helgason, S. (1997), *Towards Performance-Based Accountability: Issues for Discussion*, Public Management Service, OECD, Paris, http://www.oecd.org/governance/budgeting/1902720.pdf. [46]

Indien (2018), *National Strategy for Artificial Intelligence #AIforall*, Discussion Paper, NITI Aayog, Juni, http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf. [5]

Ingels, H. (2017), *Artificial Intelligence and EU Product Liability Law*, Präsentation bei der OECD-Konferenz "AI: Intelligent Machines, Smart Policies", Paris, 26.-27. Oktober, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-ingels.pdf. [43]

ITF (2017), "Driverless Trucks: New Report Maps Out Global Action on Driver Jobs and Legal Issues", Weltverkehrsforum, Paris, https://www.itf-oecd.org/driverless-trucks-new-report-maps-out-global-action-driver-jobs-and-legal-issues. [81]

Jain, S. (2017), "NanoNets: How to use Deep Learning when you have Limited Data, Part 2: Building Object Detection Models with Almost no Hardware" Medium, 30. Januar, https://medium.com/nanonets/nanonets-how-to-use-deep-learning-when-you-have-limited-data-f68c0b512cab. [58]

Kanada (2017), "Government of Canada launches the Global Skills Strategy", Pressemitteilung, Immigration, Refugees and Citizenship Canada, 12. Juni, https://www.canada.ca/en/immigration-refugees-citizenship/news/2017/06/government_of_canadalaunchestheglobalskillsstrategy.html. [100]

Kasparov, G. (2018), *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*, Public Affairs, New York. [91]

Kendall, A. (2017), "Deep Learning Is Not Good Enough, We Need Bayesian Deep Learning for Safe AI", Alex Kendall blog, 23. Mai, https://alexgkendall.com/computer_vision/bayesian_deep_learning_for_safe_ai/. [59]

Knight, W. (2017), "The Financial World Wants to Open AI's Black Boxes", *MIT Technology Review*, 13. April, https://www.technologyreview.com/s/604122/the-financial-world-wants-to-open-ais-black-boxes/. [32]

Kosack, S. und A. Fung (2014), "Does Transparency Improve Governance", *Annual Review of Political Science*, Vol. 17, S. 65-87, https://www.annualreviews.org/doi/pdf/10.1146/annurev-polisci-032210-144356. [26]

Kosinski, M., D. Stillwell und T. Graepel (2013), "Private traits and attributes are predictable from digital records of human behavior", *PNAS,* 11. März, http://www.pnas.org/content/pnas/early/2013/03/06/1218772110.full.pdf. [2]

Kurakin, A., I. Goodfellow und S. Bengio (2017), "Adversarial examples in the physical world", *arXiv* arXiv:1607.02533v4, https://arxiv.org/abs/1607.02533. [39]

Lakhani, P. und B. Sundaram (2017), "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks", *Radiology*, Vol. 284/2, S. 574-582, http://dx.doi.org/10.1148/radiol.2017162326. [75]

Matheson, R. (2018), *Artificial intelligence model "learns" from patient data to make cancer treatment less toxic*, 9. August, http://news.mit.edu/2018/artificial-intelligence-model-learns-patient-data-cancer-treatment-less-toxic-0810. [106]

MGI (2017), *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation*, McKinsey Global Institute, New York, https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages. [87]

Michaels, G., A. Natraj und J. Van Reenen (2014), "Has ICT Polarized Skill Demand? Evidence from Eleven Countries over Twenty-Five Years", *Review of Economics and Statistics*, Vol. 96/1, S. 60-77, http://dx.doi.org/10.1162/rest_a_00366. [77]

Mims, C. (2010), "AI That Picks Stocks Better Than the Pros", *MIT Technology Review,* 10. Juni, https://www.technologyreview.com/s/419341/ai-that-picks-stocks-better-than-the-pros/. [84]

MIT (2018), "Cybersecurity's insidious new threat: workforce stress", *MIT Technology Review* 7. August, https://www.technologyreview.com/s/611727/cybersecuritys-insidious-new-threat-workforce-stress/. [102]

Mousave, S., M. Schukat und E. Howley (2018), "Deep Reinforcement Learning: An Overview", arXiv:1806.08894v1, https://arxiv.org/abs/1806.08894. [56]

Narayanan, A. (2018), "Tutorial: 21 fairness definitions and their politics", https://www.youtube.com/watch?v=jIXIuYdnyyk. [17]

Nedelkoska, L. und G. Quintini (2018), "Automation, skills use and training", *OECD Social, Employment and Migration Working Papers*, No. 202, OECD Publishing, Paris, https://dx.doi.org/10.1787/2e2f4eea-en. [89]

Neppel, C. (2017), *AI: Intelligent Machines, Smart Policies*, Präsentation bei der OECD-Konferenz "AI: Intelligent Machines, Smart Policies", Paris, 26.-27. Oktober, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-neppel.pdf. [55]

OECD (erscheint demnächst), *Enhanced Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-Use across Societies*, OECD Publishing, Paris. [54]

OECD (2020), *Going Digital: Den digitalen Wandel gestalten, das Leben verbessern*, OECD Publishing, Paris, https://doi.org/10.1787/e78eb379-de. [61]

OECD (2019), *An Introduction to Online Platforms and Their Role in the Digital Transformation*, OECD Publishing, Paris, https://dx.doi.org/10.1787/53e5f593-en. [62]

OECD (2019), *Empfehlung des Rats zu künstlicher Intelligenz*, OECD, Paris, http://www.oecd.org/berlin/presse/Empfehlung-des-Rats-zu-kuenstlicher-Intelligenz.pdf. [35]

OECD (2019), *Scoping Principles to Foster Trust in and Adoption of AI – Proposal by the Expert Group on Artificial Intelligence at the OECD (AIGO)*, OECD, Paris, http://oe.cd/ai. [34]

OECD (2018), "AI: Intelligent machines, smart policies: Conference summary", *OECD Digital Economy Papers*, No. 270, OECD Publishing, Paris, http://dx.doi.org/10.1787/f1a650d9-en. [13]

OECD (2018), *Job Creation and Local Economic Development 2018: Preparing for the Future of Work*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264305342-en. [90]

OECD (2018), *OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption*, OECD Publishing, Paris, https://dx.doi.org/10.1787/sti_in_outlook-2018-en. [52]

OECD (2018), "Perspectives on innovation policies in the digital age", in *OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption*, OECD Publishing, Paris, https://dx.doi.org/10.1787/sti_in_outlook-2018-8-en. [48]

OECD (2017), *Algorithms and Collusion: Competition Policy in the Digital Age*, OECD Publishing, Paris, http://www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.html. [66]

OECD (2017), *Getting Skills Right: Skills for Jobs Indicators*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264277878-en. [96]

OECD (2017), *OECD Digital Economy Outlook 2017*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264276284-en. [20]

OECD (2017), *The Next Production Revolution: Implications for Governments and Business*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264271036-en. [68]

OECD (2016), "Big Data: Bringing Competition Policy to the Digital Era – Executive Summary", OECD-Dokument DAF/COMP/M(2016)2/ANN4/FINAL, Direktion Finanz- und Unternehmensfragen (DAF), Wettbewerbsaussschuss, OECD, Paris, https://one.oecd.org/document/DAF/COMP/M(2016)2/ANN4/FINAL/en/pdf. [64]

OECD (2013), *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, OECD, Paris, http://www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf. [12]

OECD (2011), *OECD-Leitsätze für multinationale Unternehmen*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264122352-de. [8]

OHCHR (2011), *Guiding Principles on Business and Human Rights*, Hohes Kommissariat der Vereinten Nationen für Menschenrechte (OHCHR), https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf. [7]

O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Broadway Books, New York. [25]

OpenAI (2018), "AI and compute", OpenAI blog, 16. Mai, https://blog.openai.com/ai-and-compute/. [53]

Pan, S. und Q. Yang (2010), "A Survey on Transfer Learning", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22/10, S. 1345-1359, http://dx.doi.org/10.1109/TKDE.2009.191. [57]

Paper, I. (ed.) (2018), "Artificial intelligence and privacy", Issues Paper, Juni, Office of the Victorian Information Commissioner, https://ovic.vic.gov.au/wp-content/uploads/2018/08/AI-Issues-Paper-V1.1.pdf. [11]

Patki, N., R. Wedge und K. Veeramachaneni (2016), "The Synthetic Data Vault", in *IEEEE Proceedings – 3rd IEEE International Conference on Data Science and Advanced Analytics (DSAA 2016)*, S. 399-410, http://dx.doi.org/10.1109/dsaa.2016.49. [107]

Privacy International and ARTICLE 19 (2018), "Privacy and Freedom of Expression in the Age of Artificial Intelligence" Scoping Paper, https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf. [10]

Purdy, M. und P. Daugherty (2016), "Artificial Intelligence Poised to Double Annual Economic Growth Rate in 12 Developed Economies and Boost Labor Productivity by up to 40 Percent by 2035, According to New Research by Accenture", Pressemitteilung, Accenture, 28. September, http://www.accenture.com/news/artificial-intelligence-poised-to-double-annual-economic-growth-rate-in-12-developed-economies-and-boost-labor-productivity-by-up-to-40-percent-by-2035-according-to-new-research-by-accenture.htm. [72]

Selbst, A. (2017), "Disparate impact in big data policing", *Georgia Law Review*, Vol. 52/109, S. 109-195, http://dx.doi.org/10.2139/ssrn.2819182. [21]

Simonite, T. (2018), "Probing the dark side of Google's ad-targeting system", *MIT Technology Review,* 6. Juli, https://www.technologyreview.com/s/539021/probing-the-dark-side-of-googles-ad-targeting-system/. [19]

Slusallek, P. (2018), *Artificial Intelligence and Digital Reality: Do We Need a CERN for AI?*, The Forum Network, OECD, Paris, https://www.oecd-forum.org/channels/722-digitalisation/posts/28452-artificial-intelligence-and-digital-reality-do-we-need-a-cern-for-ai. [42]

Smith, M. und S. Neupane (2018), *Artificial intelligence and human development: toward a research agenda*, International Development Research Centre, Ottawa, https://idl-bnc-idrc.dspacedirect.org/handle/10625/56949. [4]

Stewart, J. (2018), "As Uber Gives up on Self-Driving Trucks, Another Startup Jumps In", *WIRED*, 8. Juli, https://www.wired.com/story/kodiak-self-driving-semi-trucks/. [80]

Talbot, D. et al. (2017), "Charting a Roadmap to Ensure Artificial Intelligence (AI) Benefits All" Medium, 30. November, https://medium.com/berkman-klein-center/charting-a-roadmap-to-ensure-artificial-intelligence-ai-benefits-all-e322f23f8b59. [3]

Trajtenberg, M. (2018), "AI as the next GPT: A political-economy perspective"*, NBER Working Paper*, No. 24245, http://dx.doi.org/10.3386/w24245. [105]

UNI (2018), *Die 10 wichtigsten Grundsätze für Arbeitnehmerdatenschutz und -sicherheit*, UNI Global Union, http://www.thefutureworldofwork.org/media/35483/uni-global-union_-arbeitnehmerdatenschutz-und-sicherheit.pdf. [95]

Varian, H. (2018), "Artificial intelligence, economics and industrial organization", *NBER Working Paper*, Vol. 24839, http://dx.doi.org/10.3386/w24839. [65]

Vereinigtes Königreich (2017), *UK Digital Strategy*, Government of the United Kingdom, Department for Digital, Culture, Media & Sport, https://www.gov.uk/government/publications/uk-digital-strategy/uk-digital-strategy. [70]

Vereinigtes Königreich (2017), *UK Industrial Strategy: A Leading Destination to Invest and Grow*, Great Britain & Northern Ireland, Department for Business, Energy & Industrial Strategy, http://https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/668161/uk-industrial-strategy-international-brochure.pdf. [99]

Vereinigte Staaten (2016), *Artificial Intelligence, Automation and the Economy*, Executive Office of the President, Government of the United States, https://www.whitehouse.gov/sites/whitehouse.gov/files/images/EMBARGOED AI Economy Report.pdf. [85]

Wachter, S., B. Mittelstadt und C. Russell (2017), "Counterfactual explanations without opening the black box: Automated decisions and the GDPR", arXiv:1711.00399, https://arxiv.org/pdf/1711.00399.pdf. [27]

Wachter, S., B. Mittelstadt und L. Floridi (2017), "Transparent, explainable and accountable AI for robotics", *Science Robotics,*, Vol. 2/6, 31. Mai, http://robotics.sciencemag.org/content/2/6/eaan6080. [47]

Weinberger, D. (2018), "Optimization over explanation – Maximizing the benefits of machine learning without sacrificing its intelligence", Medium, 28. Januar, https://medium.com/@dweinberger/optimization-over-explanation-maximizing-the-benefits-we-want-from-machine-learning-without-347ccd9f3a66. [1]

Weinberger, D. (2018), "Playing with AI Fairness", Google PAIR, 17. September, https://pair-code.github.io/what-if-tool/ai-fairness.html. [22]

Winick, E. (2018), "Every study we could find on what automation will do to jobs, in one chart", *MIT Technology Review,* 25. Januar, https://www.technologyreview.com/s/610005/every-study-we-could-find-on-what-automation-will-do-to-jobs-in-one-chart/. [86]

Wong, Q. (2017), "At LinkedIn, artificial intelligence is like 'oxygen'", *Mercury News,* 1. Juni, http://www.mercurynews.com/2017/01/06/at-linkedin-artificial-intelligence-is-like-oxygen. [94]

Yona, G. (2017), "A Gentle Introduction to the Discussion on Algorithmic Fairness", Towards Data Science", 5. Oktober, https://towardsdatascience.com/a-gentle-introduction-to-the-discussion-on-algorithmic-fairness-740bbb469b6. [18]

Zeng, M. (2018), "Alibaba and the future of business", *Harvard Business Review,* September-Oktober, https://hbr.org/2018/09/alibaba-and-the-future-of-business. [73]

Anmerkungen

[1] Weitere Informationen unter: https://www.microsoft.com/en-us/ai/ai-for-good.

[2] Vgl. https://deepmind.com/applied/deepmind-ethics-society/.

[3] Vgl. https://www.blog.google/technology/ai/ai-principles/.

[4] Vgl. https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

[5] Als Beispiele sind die Internationale Arbeitsorganisation, die OECD-Leitsätze für multinationale Unternehmen oder die Leitprinzipien für Wirtschaft und Menschenrechte der Vereinten Nationen zu nennen.

[6] Wegen weiterer Studien zu diesem Thema vgl. https://www.dudumimran.com/2018/05/speaking-about-ai-and-cyber-security-at-the-oecd-forum-2018.html und https://maliciousaireport.com/.

[7] Pierre Chalançon, Vorsitzender der Taskforce Verbraucherpolitik des Beratenden Ausschusses der Wirtschaft bei der OECD (BIAC) und Vice President Regulatory Affairs, Vorwerk & Co KG, Vertretung bei der Europäischen Union – *Science-Fiction is not a Sound Basis for Legislation.*

[8] Diese Technik wird u. a. eingesetzt, um autonome Fahrzeuge in der Ausführung komplexer Manöver zu trainieren, das AlphaGo-Programm zu trainieren und Krebspatienten zu behandeln (z. B. indem die kleinste Dosierung und Verabreichungshäufigkeit bestimmt wird, die zur Reduzierung von Hirntumoren noch wirksam ist (Matheson, 2018[106])).

[9] Aus den Ergebnissen einer neueren Studie geht hervor, dass reale Daten vielfach erfolgreich durch synthetische Daten ersetzt werden können, was für Wissenschaftler gerade im Fall datenschutzrechtlicher Beschränkungen sehr hilfreich sein kann (Patki, N., R. Wedge und K. Veeramachaneni, 2016[107]). Die Autoren zeigen, dass sich die Ergebnisse, die mit synthetischen Daten erzeugt wurden, in 70 % der Fälle nicht signifikant von den Ergebnissen unterschieden, die mit realen Daten erzielt wurden.

[10] Lösungen, die z. B. eine vollhomomorphe Verschlüsselung mit neuronalen Netzen kombinieren, wurden in dieser Hinsicht erfolgreich getestet und eingesetzt (Dowlin, 2016[60]).

[11] Vgl. https://www.wipo.int/about-ip/en/artificial_intelligence/ und https://www.uspto.gov/about-us/events/artificial-intelligence-intellectual-property-policy-considerations.

[12] Vgl. https://www.ibm.com/watson/stories/creditmutuel/.

[13] Alibaba beschäftigt z. B. keine Zeitarbeiter mehr, um an Tagen, an denen Hochbetrieb herrscht oder Sonderaktionen angeboten werden, Kundenanfragen zu bearbeiten. An dem Tag, an dem Alibaba 2017 den höchsten Umsatz erzielte, bearbeitete der Chatbot des Unternehmens mehr als 95 % der Kundenanfragen und antwortete rd. 3,5 Millionen Kunden (Zeng, 2018[73]). Wenn Chatbots immer mehr Kundendienstfunktionen übernehmen, verändern sich die Aufgaben der menschlichen Kundenberater, die sich dann auf komplexere oder individuellere Anfragen konzentrieren müssen.