# 5 Zukunftsvisionen

*Daniel Ullrich*

»Tut mir leid, aber ich muss Ihnen mitteilen, dass wir keine OP machen können.« Die Nachricht war erschütternd, war eine OP doch die einzige Option, die potenziell tödlich verlaufende Krankheit aufhalten zu können. Eine OP barg zwar ebenfalls Risiken, aber auf sie zu verzichten bedeutete, sich auf das reine Abwarten zu reduzieren und darauf zu hoffen, dass sich die Krankheit von alleine zurückbilden würde.

Anna-Lisa hatte die Diagnose erst vor kurzem erhalten, ihre Krankheit ist selten und verläuft schwer vorhersagbar. Aber das Urteil war nicht nur das ihres Arztes, denn dieser wiederum stützte all seine Entscheidungen auf das System »Health Guardian«, eine künstliche Intelligenz, die mit einer unüberschaubaren Anzahl von Daten gefüttert wurde und darauf basierend Handlungsempfehlungen generiert.

Und im Falle von Anna-Lisa war die Empfehlung eben, dass keine OP durchgeführt werden sollte. Anna-Lisas Mutter, die bei dem Arztgespräch zugegen war, hakte nach, ob denn kein Fehler vorliegen könnte und ob der Arzt denn der gleichen Meinung sei. Letzterer war sichtlich in einem Dilemma gefangen: Er war tatsächlich nicht unbedingt gegen eine OP und hätte im Zweifel sogar dafür argumentiert. Aber er wusste, dass die künstliche Intelligenz weit mehr Daten berücksichtigen konnte und daher seiner naturgemäß eingeschränkten Perspektive überlegen war. Und was noch schwerer wiegte: Obwohl die Ergebnisse »Handlungsempfehlungen« hießen, waren es doch eigentlich Entscheidungen und man musste als Arzt schon sehr gute Gründe vorlegen können, um sich gegen die Entscheidungen zu stellen – weit bessere Gründe als im aktuellen Fall vorlagen. So waren dem Arzt die Hände gebunden und es blieb ihm nichts Anderes übrig, als Anna-Lisa und ihre Familie zu vertrösten. Es gab ja noch die Hoffnung auf natürliche Besserung.

Künstliche Intelligenz (KI) ist eine Hype-Thema, das in periodischen Abständen medial hohe Wellen schlägt und anschließend – wie viele gehypte Themen – wieder in den Hintergrund rückt, wenn die erwarteten Durchbrüche ausbleiben.

Wenngleich KI in den vergangenen Jahren beeindruckende Ergebnisse erzielen konnte, was die Erfolge im Bereich visueller Wahrnehmung (Taigman et al., 2014) bzw. Mustererkennung (Foggia et al., 2014), Experten- und Entscheidungssysteme und in Spielsystemen wie Schach, Go (Schrittwieser et al., 2020; Silver et al., 2016) oder Computer-Strategiespielen (Vinyals et al., 2019) demonstrierten, so blieb doch

stets der Einwand der Kritiker zurück, das sei ja keine »richtige Intelligenz« gewesen (Fjelland, 2020; Crawford, 2021).

Dieser Einwand ist ebenso korrekt wie die Verwirrung und Uneinigkeit darüber, »richtige Intelligenz« zu definieren oder gar zu erkennen. Aber vielleicht ist dies auch überhaupt nicht notwendig: Wenn ein System hinreichend gute Resultate liefert, muss es dann »tatsächlich intelligent« sein? Ebenso wenig ist »tatsächliche Intelligenz« eine Voraussetzung dafür, dass Menschen einem solchen System vertrauen. Denn die Mechanismen, die bei zwischenmenschlichen Interaktionen gelten, wirken auch bei der Interaktion mit künstlichen Systemen (Costa, 2018): Manchmal reicht es, dass wir der Überzeugung sind, dass es jemand gut mit uns meint, damit wir ihm vertrauen – unabhängig davon, ob derjenige herausragende Intelligenz besitzt oder nicht. Viele Grundlagen für die Vertrauensbildung, beispielsweise freundliche Umgangsformen, ein nettes, sympathiebildendes Erscheinungsbild, können ebenso auf Maschinen und Roboter übertragen werden (Powers & Kiesler, 2006). So ist es auch kein Zufall, dass digitale Sprachassistenten in der Regel weibliche Stimmen erhalten, da diese zu messbar höheren Sympathie- und Vertrauenswerten beitragen (Ernst & Herm-Stapelberg, 2020).

Ob intelligente Systeme breiten Erfolg erzielen werden, ist maßgeblich davon abhängig, welche Resultate sie liefern. Ein schlichtes Vorspielen von Fähigkeiten ist hier nicht ausreichend, es müssen schon tatsächlich Leistungen sein, die einen Mehrwert bringen. Die aktuellen Entwicklungen zeigen, dass mehrere notwendige Kriterien mittlerweile erfüllt sind, um künstliche Intelligenz in bestimmten Bereichen zum Erfolg zu verhelfen: Neben der theoretischen Grundlage – aktuell meist künstliche neuronale Netze, die durch Training (»Machine Learning«) bessere Ergebnisse erzielen – stehen mittlerweile auch ausreichend große Beispieldatenbanken zur Verfügung und Rechenleistung, um aus diesen Beispielen durch wiederholtes automatisiertes Training und einer Vielzahl unterschiedlicher Netztopologien die jeweils besten Ergebnisse auszuwählen (Ongsulee, 2017; Gupta et al., 2018). Hier findet im Prinzip digitale Evolution statt, die als Resultat für eine Fragestellung das jeweils beste künstliche System ausbildet.

Ein fundamentaler Nachteil an diesem Verfahren ist es, dass die Herausbildung der »Intelligenz« für Menschen nur auf theoretischer Ebene nachvollziehbar ist und das fertige System den menschlichen Verständnishorizont übersteigt (Rudin, 2019). Dies kann leicht dadurch gezeigt werden, dass ein solches System in der Regel nicht repariert werden kann: Es ist einem Menschen nicht möglich, bestimmte Fehlentscheidungen aus solch einem System zu entfernen, ohne zahlreiche andere Stellen zu beschädigen. Dies liegt daran, dass zwar die grundlegenden Mechanismen verstanden sind, aber die konkrete Ausgestaltung so komplex ist, dass sie Menschen schlicht überfordert. Hier hilft dann häufig nur ein kompletter Neuanfang, d. h. ein Training eines neuen Systems mit geänderten Startparametern, das am Ende hoffentlich den fraglichen Fehler nicht mehr aufweist.

Hieraus erwächst auch das im Eingangsbeispiel der KI im OP-Saal angedeutete Problem: Entscheidungen sind prinzipiell intransparent, sie werden auf Basis des Entscheidungssystems getroffen, das auf dem Beispielmaterial (aus der Vergangenheit) basiert und mit neuen Beispielen Vorhersagen in die Zukunft erstellen soll. Welche Variablen genau berücksichtigt und wie diese gewichtet werden und in

Beziehung miteinander wechselwirken, bleibt dem Nutzer (und auch dem Entwickler, wenn es um ein tieferes Verständnis und nicht eine zahlenmäßige Auflistung geht) verborgen (Kim & Routledge, 2021). Ging es im Falle der Entscheidung pro oder contra OP tatsächlich um die Effektivität des Eingriffs? Oder wurden noch weitere Variablen wie Erfolgswahrscheinlichkeit, Kosten-Nutzen-Abwägung, Budget des Gesundheitsapparats und Ressourcen-Auslastung berücksichtigt? Wie steht es mit der Entscheidung, wenn es andere Patienten gibt, für die es vielversprechendere Eingriffe gäbe, für die dann die Ressourcen fehlen? Man muss sich keine dystopische Gesellschaft ausmalen, um sich vorzustellen, dass eine solche künstliche Intelligenz auch die normativen Zwänge der Auftraggeber berücksichtigen wird, insbesondere in Gesellschaften, in denen Ressourcen im Gesundheitssystem nicht grenzenlos zur Verfügung stehen.

Transparenz und Erklärbarkeit künstlicher Systeme ist ein neues Forschungsfeld, in dem Ansätze gesucht werden, den intransparenten Entscheidungen künstlicher Intelligenz zu begegnen (Larsson & Heintz, 2020). Fraglich bleibt jedoch, ob diese Bemühungen um Transparenz mit den Fortschritten im Bereich der künstlichen Entscheidungsfindung Schritt halten können, da letztere naturgemäß ein weit größeres wirtschaftliches Potenzial bergen.

## 5.1 Kapitelausblick

Was bedeutet es für unsere Gesellschaft, wenn sich aktuelle Trends und Technologien weiter fortsetzen? Welche Konsequenzen ergeben sich aus psychologischer Perspektive für das Erleben der Menschen? Welche moralischen Überlegungen spielen hier eine Rolle und wo gilt es, Entscheidungen zu treffen? Beispielhaft werden in diesem Abschlusskapitel drei Bereiche zentraler gesellschaftlicher und psychologischer Relevanz näher beleuchtet: (1) Soziale Normen im Kontext digitaler Systeme, (2) Überwachung und Social Scoring sowie (3) Künstliche Intelligenz als Entscheidungshilfe und Entscheidungsinstanz.

## 5.2 Soziale Normen

Warum sollte man überhaupt soziale Normen im Kontext von digitalen Technologien betrachten? Was ist das Besondere am digitalen Raum?

Der digitale Raum unterscheidet sich in einigen Aspekten fundamental vom nicht-digitalen Raum. Dadurch gelten in ihm andere Gesetzmäßigkeiten, was sich wiederum auf die Herausbildung, Veränderung und Durchsetzung bestimmter so-

zialer Normen auswirkt. Beispiele für solche Unterschiede werden im Folgenden dargestellt.

**Charakteristika der sozialen Interaktion im digitalen Raum**

**Gefühlte Anonymität. »Im Internet weiß niemand, dass du in Wahrheit ein Hund bist.«**

Aus der Tatsache, dass andere Interaktionspartner häufig als Avatar auftreten UND daraus, dass man selbst nicht weiß, wer der andere genau ist, entsteht die Illusion, man selbst wäre ebenfalls komplett anonym. Genaugenommen sind aber nur Nutzer einander anonym, technisch gesehen können Nutzer identifiziert werden. Aber die Pseudoanonymität reicht aus, sich »sicher« zu fühlen und führt dazu, dass mitunter handlungsregulierende Hemmnisse wegfallen und Nutzer sich nicht an soziale Regeln gebunden fühlen, so wie Vermummte auf einer Demonstration (Suler, 2004; Macdonald, 2020). Diese »Freiheit« nutzen nicht alle Nutzer aus, aber ein signifikanter Teil eben schon.

**Distanz zu den Interaktionspartnern**

Interaktion besteht häufig nur aus Schreiben und Lesen von Texten. Alle sozialen Cues (menschliche Charakteristika; Aussehen, Stimme, körperliche Präsenz) fehlen und man kann somit leicht vergessen, dass man nicht mit Texten interagiert, sondern mit Menschen, die als solche ihrerseits Motive, ein eigenes Wertsystem sowie Gefühle und Emotionen besitzen, die man durch das eigene Handeln verletzen kann. Diese Verletzungen können aber durch das digitale Medium nicht übermittelt werden. Man bekommt nicht mit, wenn man sein Gegenüber verletzt und empathische Mechanismen, die die Konsequenzen eigenen Handelns aufzeigen könnten, sind nicht verfügbar (Carrier et al., 2015). Auf der anderen Seite wird niemand kommunizieren, dass er durch eine Äußerung getroffen wurde, um nicht den Verdacht der Verletzlichkeit aufkommen zu lassen. Hierdurch wird negativem Verhalten weiter der Boden geebnet, da die negativen Konsequenzen nur noch im Verborgenen stattfinden – der Aggressor weiß nicht, dass er sein Gegenüber verletzt hat und der Angegriffene verschweigt es.

**Avatar statt Authentizität**

Aktionen im digitalen Raum sind an einen Avatar gekoppelt, den man im Zweifel gegen einen neuen eintauschen kann. Ein solcher Neustart ist im nicht-digitalen Kontext nur sehr schwer möglich. Im digitalen Raum hingegen ist ein Neu-Erstellen eines anderen Accounts schnell erledigt, womit man mit einer weißen Weste neu starten kann (Interaktionen im anonymen bzw. pseudonymen Raum vorausgesetzt). Selbst wenn der Avatar nicht leicht gewechselt werden kann, hat der Nutzer eine viel größere Kontrolle darüber, welche Informationen über sich preisgegeben werden. Insbesondere unwillkürliche Aspekte der Kommunikation (Mimik, affektive Reaktionen, Stimmfarbe) sind im digitalen Raum stark reduziert (Suler, 2004).

**Digital-exklusive Mechanismen**

Im digitalen Raum existieren Interaktionsmechaniken, die im nicht-digitalen Raum unbekannt und mitunter unmöglich sind. Als Beispiel sei *Ghost-Banning* genannt. Dies ist eine Technik, die beispielsweise gegen sogenannte Trolle, also Störenfriede, die Genugtuung daraus ziehen, andere Nutzer mit polarisierenden Äußerungen zu provozieren, eingesetzt wird. Wird ein Troll schlicht gebannt (gelöscht), legt dieser sich einen neuen Account zu und startet von neuem mit der Erstellung provokativer Postings, womit das eigentliche Problem nicht gelöst wird. Ein Versuch, dieses Problem nachhaltiger zu lösen ist Ghost-Banning, bei dem der Troll für alle anderen Nutzer ausgeblendet wird (so als wäre er gelöscht), aber für ihn selbst wird alles so dargestellt, wie er es selbst erwartet. Er hat zunächst keine Möglichkeit, sein eigenes Ghost-Banning festzustellen (er müsste die Interaktion aus der Sicht eines anderen Accounts betrachten) und kann sich höchstens über die ausbleibenden Reaktionen auf seine Provokationen wundern. Würde man diese Technik auf den nicht-digitalen Raum übertragen, so käme sie einer Tarnkappe gleich, die man einem Störenfried in der realen Welt aufsetzen könnte, ohne dass dieser dies mitbekommen würde. Was in der Realität pure Fiktion ist, ist im digitalen Raum Alltag: Jeder Nutzer erhält seine individuelle Sicht auf die (digitale) Welt – und wo die Unterschiede liegen, wird im Zweifel nicht kommuniziert.

Bereits heute ist es so, dass durch die allgegenwärtige digitale Nutzung entsprechende digitale Normen immer mehr an Gewicht gewinnen. Durch die Eigenheiten im digitalen Raum bilden sich hierbei Normen heraus, die durch die digitalen Regeln beeinflusst werden.

### 5.2.1 Ein mögliches Zukunftsszenario

Normen werden implizit gelernt und eingehalten und Normen aus der nicht-digitalen Welt beeinflussen solche aus der digitalen Welt und umgekehrt (Diefenbach & Ullrich, 2019).

Wenn digitale Technologien unser gesamtes Leben durchdringen und immer größeren Raum einnehmen, steigt auch der Einfluss sozialer Normen aus der digitalen Welt, da wir ihnen in immer größerem Maße ausgesetzt sind. Dies kann letztendlich dazu führen, dass diese Normen irgendwann gegenüber traditionellen Normen, die in der nicht-digitalen Welt entstanden sind, dominieren. Unsere Verhaltensweisen werden dann primär durch Technologie bestimmt und den Regeln, die dort bewusst oder unbewusst aufgestellt wurden.

Konkret könnte dies in einem ruppigeren Umgang miteinander resultieren, bei dem wenig Rücksicht auf die gegenseitige Gefühlswelt genommen wird. Ein Seiteneffekt könnte darüber hinaus in der Ausbildung von Vermeidungsstrategien gegen direkte, nicht-digitale Interaktion liegen: Bereits heute gibt es einen wahrnehmbaren Trend jüngerer Menschen, direkte Interaktion wie zum Beispiel direkte Gespräche oder auch Telefonate zu vermeiden. Stattdessen wird auf technik-medi-

ierte Kommunikation ausgewichen, wo immer dies möglich ist (▶ Kap. 2). Dies könnte auch darin begründet sein, dass die gewohnten Normen aus der digitalen Welt in der nicht-digitalen Welt zu irritierenden Erlebnissen führen kann. Und statt sich mit der erweiterten Kommunikationsmodalität sowie eigenen empathischen Reaktionen auseinanderzusetzen, wird stattdessen die nicht-digitale Situation gemieden. Als Folge werden empathische Fähigkeiten noch seltener genutzt und entsprechend weniger ausgebildet.

Bei diesen Prognosen muss berücksichtigt werden, dass die Summe unserer aktuellen Normen immer auch auf schon bestehenden Normen basiert, die durch Einhaltung repliziert und selbst verstärkt werden. Das heißt gleichzeitig auch, dass das über Jahrhunderte angeeignete Normen-Repertoire aus der nicht-digitalen Welt unser Verhalten weiterhin prägt und dafür sorgt, dass die Dominanz von Normen aus der digitalen Welt nicht noch weiter vorangeschritten ist als aktuell der Fall. Eine fiktive Gesellschaft, die bei »Null« starten würde, wäre vermutlich noch stärker durch Normen aus der digitalen Welt geprägt, da ihr Einfluss eben stärker ist, als wir es aktuell am Resultat erkennen können, welches stark zugunsten konservativer Normen verzerrt ist. Diesen Gedanken folgend wird aber auch jede existierende Gesellschaft im Laufe der Zeit immer stärker von digitalen Normen beeinflusst werden, schon aus dem Grund heraus, dass ältere Menschen, tendenziell Vertreter konservativer Normen, sterben und nachkommende Menschen, stärker geprägt von Normen aus der digitalen Welt, diese ersetzen.

## 5.3 Überwachung und Social Scoring

Als das Internet entstand, galt es zunächst, eine ausfallsichere Kommunikationsinfrastruktur zu schaffen, die auch dann noch funktionierte, wenn Teile davon wegbrachen (Leiner et al., 2009). Die Ideen, zielgerichtet soziale Netzwerke zu schaffen, Nutzer auf ihren Pfaden durch das digitale Netz zu tracken, Profile zu erstellen und zielgruppenbezogene Werbung zu präsentieren, kamen erst später. So war die erste Zeit des Internets vor allem durch Freiheit geprägt: Freiheit im Handeln der Nutzer und Freiheit von Kontrolle. Diese Zeit wird von manchem Nutzer auch als die goldene Zeit des Internets bezeichnet, von anderen als Wild-West-Zeit ohne Regeln (Palacios, 2019).

Mit steigender Popularität des Internets wurden aber auch die Potenziale erkannt, die große Nutzergruppen bildeten. Allen voran das Anzeigen von Werbung und das Anbieten zahlreicher Handelsplätze, welche eine immer größere Konkurrenz zu ihren nicht-digitalen Vertretern darstellen (Taylor, 2021).

Darüber hinaus kam aber auch dem Verbreiten von Nachrichten und Informationen eine immer größere Rolle zu, verstärkt durch die Tatsache, dass immer mehr Menschen ihre Informationen aus dem Netz beziehen und die Sender von Informationen dadurch eine stetig wachsende Reichweite erhielten (Beisch & Schäfer, 2020). Eine natürliche Folgefrage war, wie man den Einfluss auf die Nutzer maxi-

mieren konnte und wie eine Informationshoheit herzustellen war: Wer bestimmt, welche von zwei Informationen »richtig« ist, wenn diese sich inhaltlich widersprechen.

Somit dauerte es nicht lange, bis verschiedenste Interessengruppen das weltweite Netz und seine Nutzer für sich entdeckten und in ihrem Sinne versuchten, Einfluss zu nehmen: Politik, Nachrichtenportale, die Werbeindustrie, Anbieter von Konsumprodukten, Aktivisten und individuelle Meinungsträger sowie »Influencer« (Moffett & Santos, 2014). Der Einfluss der verschiedenen Gruppen nahm mehr und mehr zu, gespeist durch Monopolisierung, Lobby-Arbeit und Regulierungen von Seiten der Staaten als auch der Plattformanbieter.

Das Internet ist die Anti-These zur klassischen demokratischen Gesellschaft, in der insbesondere die Informationshoheit in der Kontrolle einiger weniger Personen bzw. beim Staat lag. Im Internet hingegen ist jeder Sender und Empfänger – und damit jeder potenziell ein Konkurrent zu den großen etablierten Medien. Und jeder kann potenziell an der Meinungsbildung mitwirken (Bakshy et al.,2012; Burbach et al., 2020). Ein Zustand, den (alte) Medien und Politik nicht unbedingt wünschenswert finden, da er unkontrollierbare Effekte birgt.

Damit einher gehen Versuche der Überwachung und Informationskontrolle wie etwa Upload-Filter oder Sabotage von Verschlüsselungstechnologien – meist argumentativ durchgesetzt mit populären Zielen wie Strafverfolgung und mit einer kleinen Zahl von Straftätern im Fokus (z. B. Kinder-Pornografie, illegale Schwarzmärkte). Die negativen Auswirkungen betreffen aber alle Nutzer gleichermaßen und das Missbrauchspotenzial ist naturgemäß groß.

### 5.3.1 Ein mögliches Zukunftsszenario

Mit der fortschreitenden Digitalisierung und Nutzung digitaler Technologien steigt auch das Potenzial zur Überwachung: Nutzer hinterlassen bei jeder Aktion im Netz Spuren, die sie zu gläsernen Nutzern machen können, sofern entsprechende Gesetze dies ermöglichen.

Auf Seite der Nutzer erzeugt das Bewusstsein, überwacht zu werden, Stress und führt zu angepasstem Verhalten – ein Symptom das auch als »*Chilling Effekt*« bekannt ist (Büchi et al., 2022). Hierbei reicht bereits das Gefühl aus, man könne überwacht werden, ob die eigenen Handlungen dann auch tatsächlich überwacht werden, ist nicht ausschlaggebend. Dieser Chilling Effekt kann natürlich gezielt genutzt werden, um das Verhalten der Nutzer in gewünschte Bahnen zu lenken. Und weil nicht tatsächlich jeder überwacht werden muss, ist die Methode zudem kosteneffektiv.

Auf der anderen Seite ist es mit der fortschreitenden Entwicklung künstlicher Intelligenzen immer weniger relevant, kosteneffektiv zu sein: Wo früher noch tatsächliche Menschen für die Überwachung eingesetzt wurden und Vergehen erkennen mussten, bedient man sich mehr und mehr Algorithmen, die dann auch tatsächlich jede Aktion überwachen können. Solche Algorithmen kommen beispielsweise auf sozialen Plattformen zum Einsatz und erkennen automatisiert

Copyright-Verstöße, (Kinder-)Pornografie oder bestimmte Schlüsselwörter, die auf den Plattformen tabu sind.

Dennoch waren die Maßnahmen bislang eher auf sanfte Beeinflussung der Nutzer angelegt, denn selbst die Verbannung von einer Plattform hatte für Nutzer in der Regel keine wirklich gravierenden Folgen.

Mit der Einführung des *Social Scorings* hat sich dies grundsätzlich geändert. Social Scoring hebt den Überwachungsaspekt auf eine neue Ebene und macht aus der impliziten, beiläufigen Beeinflussung eine explizite, zielgerichtete: mit dem Einsatz von Social Scoring – Bürger bekommen Punkte für gewünschte Verhaltensweisen und Abzug für unerwünschte Verhaltensweisen – wird explizit Verhalten vorgegeben, das normativ für wünschenswert erachtet wird (z. B. Hoffrage & Marewski, 2020; Kostka, 2019). Verstöße gegen gewünschte Verhaltensweisen haben konkrete und fühlbare Konsequenzen für die Nutzer, etwa wenn es darum geht, eine Wohnung zu finden und die Interessenten nach Social Score sortiert werden.

Weil gleichzeitig Kritik an diesem System naturgemäß als nicht sozial wünschenswerte Handlung klassifiziert werden wird, muss ein solches System in einer selbstverstärkenden Spirale münden, in der Regeln immer extremer und umfassender werden, bis alle Bereiche menschlichen Verhaltens abgedeckt sind. Ein Entziehen von solch einem System wird nahezu unmöglich werden, sobald kritische Funktionalitäten (Reisefreiheit, Bezahlfunktionen, Priorisierung bei der Wohnungssuche, bei Jobs, Einstellungskriterium analog zu polizeilichem Führungszeugnis) an den Social Score gekoppelt sind.

## 5.4 KI als Entscheidungshilfe und Entscheidungsinstanz

Künstliche Intelligenz dient schon heute als Unterstützung bei komplexen Entscheidungen. Sie wird im Versicherungskontext eingesetzt, beispielsweise um zu prüfen, ob im Rahmen konkreter Versicherungspolicen und Randfaktoren ein Versicherungsfall vorliegt (Eling et al., 2021; Riikkinen et al., 2018). Im Bereich der Medizin hilft künstliche Intelligenz bei der Diagnostik und Verfahren der Mustererkennung bei der Auswertung bildgebender Verfahren (Kermany et al., 2018; Esteva et al., 2017). In Personalabteilungen kann künstliche Intelligenz dabei helfen, den passendsten Kandidaten für eine ausgeschriebene Stelle zu identifizieren (Upadhyay & Khandelwal, 2018; Nawaz & Mary, 2019)).

Begrenzt werden die Möglichkeiten künstlicher Intelligenz im Wesentlichen durch drei Faktoren (hierbei wird insbesondere Bezug genommen auf das aktuell vorherrschende Verfahren des Machine Learnings):

- die Spezifikation der Methode, des Algorithmus bzw. der Netztopologie,
- der Anzahl zur Verfügung stehender Datensätze, die mögliche Eingangsdaten mit gewünschten Ausgangsdaten verknüpfen (zum Beispiel eine große Sammlung von verschiedenen Tierbildern, jeweils mit Hinweis darauf, welches Tier abgebildet ist),
- der zur Verfügung stehenden Rechenleistung für das Training der KI.

Aktuell stehen je nach Anwendungsdomäne alle drei Faktoren in ausreichend guter Qualität bzw. Quantität zur Verfügung, um künstliche Intelligenzen zu generieren, die Ergebnisse liefern, die häufig denen von Menschen ebenbürtig oder überlegen sind. Insbesondere für den zweiten Faktor – die Datensätze, die Eingangsmuster mit den gewünschten Ergebnissen verknüpfen – werden quasi beiläufig Daten generiert, denn es sind die Daten, Aktivitäten, Bilder, Verhaltensweisen der Nutzer, die gespeichert werden und nach entsprechender Aufbereitung als Trainingsdatensätze zur Verfügung stehen. Die Datenlage wird also täglich besser – zumindest in Bezug auf die Daten, die Nutzer generieren können und zumindest für diejenigen, die sie speichern oder auswerten können.

### 5.4.1 Ein mögliches Zukunftsszenario

Sobald KI-Methoden in der Lage sind, menschliche Arbeitskraft oder Fähigkeiten in ebenbürtiger Qualität zu ersetzen, stellt sich die Frage kaum noch, ob diese KI-Methoden zur Anwendung kommen werden: Wer sie nicht einsetzt, hat schnell einen Wettbewerbsnachteil oder verschenkt Potenzial und riskiert, vom Markt zu verschwinden. Mit fortschreitender Entwicklung von Methoden und Datensammlungen wird KI als Entscheidungshilfe in immer weiteren Feldern Einzug halten, z. B. Rechtsprechung (Sourdin, 2021; Vermeys, 2021), Partnerwahl (Agudo & Matute, 2021; Scavarelli, 2018) und vielen weiteren.

Es ist stark anzunehmen, dass KI-Methoden insgesamt immer stärker etabliert werden und die Akteure nicht vorher abklären, ob ihr Einsatz denn redlich wäre. Denn mit ihrem Einsatz ergeben sich viele Fragen: Gibt es Tabus oder darf KI in alle Domänen der menschlichen Gesellschaft durchdringen? Was, wenn KI Empfehlungen liefert, die politisch nicht korrekt und damit nicht erwünscht sind? Wie kann sichergestellt werden, dass die Trainingsdaten »neutral« sind, so dass sich kein Bias auf die trainierte KI überträgt? Gibt es ein Anrecht darauf, zu verstehen, auf welcher Basis eine KI konkret entscheidet und könnte man einem solchen Anrecht überhaupt gerecht werden, wenn die KI doch prinzipiell immer ein Stück weit eine Black Box bleibt?

Fest steht (nach aktuellem Stand der Technik), dass uns die KI weder fehlerfreie Entscheidungsinstanzen noch transparente Begründungen für ihre Entscheidung bieten kann – man könnte sagen, hier steht sie ihren menschlichen Pendants in nichts nach (im negativen Sinne). Auf der anderen Seite müssen diese Mängel nicht notwendigerweise dazu führen, auf sie zu verzichten – denn auch die Vorteile, die sie verspricht, sind nicht zu ignorieren.

Wesentlich wird also sein, wie die Menschen dazu stehen, wenn KI wichtige Entscheidungen in der Gesellschaft zu treffen hat. Wäre es beispielsweise wünschenswert, wenn – anstelle von Politikern – eine KI über die Zukunft entscheiden würde, die Zugriff auf Ihre Daten hat und sich so für Ihre Interessen einsetzen könnte? Diese Frage wurde in einer Studie gestellt, mit insgesamt hohen Zustimmungswerten zugunsten der KI: Im europäischen Raum liegt die Zustimmungsrate bei durchschnittlich 51 %, besonders deutlich ist der Zuspruch für die KI in Spanien (66 %), Italien (59 %) und Estland (56 %). In China sind es sogar 75 % die KI als Gestalter politischer Entscheidungen befürworten, wohingegen in den USA nur 40 % derartige Entscheidungen an KI delegieren wollen (Jonsson & de Tena, 2021).

## 5.5 Ausblick

Der Einsatz von KI und Digitalisierung in vielen Bereichen des Arbeits- und Privatlebens wird in Zukunft weiter zunehmen und birgt insgesamt große Potenziale. Unliebsame Aufgaben können an die Technik delegiert werden; KI kann Aufgaben übernehmen, die den Menschen überfordern oder langweilen – und andersherum. Was wir allerdings gleichermaßen im Blick haben müssen, sind die großen gesellschaftlichen Umwälzungen, die der Einsatz von KI mit sich bringen kann. Ein System, das auf Angebot und Nachfrage von Arbeitsleistung beruht, kann schwer existieren, wenn das System mit künstlichen Agenten überschwemmt wird, die mit Menschen in Konkurrenz treten. Neue gesellschaftliche Ideen des Zusammenlebens sind gefragt – insbesondere solche, die nicht auf bereits gescheiterten Gesellschaftsmodellen beruhen. Zwar bleibt bis zum großen Durchbruch der künstlichen Agenten noch etwas Zeit, doch niemand weiß wie viel genau. Der Durchbruch der KI nähert sich nicht graduell mit gleichbleibendem Tempo, sondern wird uns vermutlich mit einem Schub erreichen, so dass es zu diesem Zeitpunkt bereits einen Aktionsplan braucht, welchen Raum wir KI in der Gesellschaft zugestehen wollen. Andernfalls lässt sich nur noch reagieren statt agieren. Die Gestaltung der Zukunft ist dann nur noch Reaktion auf die neue faktische Realität.

Diese Überlegungen zeigen: Die unschuldige goldene Zeit von KI und Digitalisierung ist vorbei. Deren Effekte und Nebeneffekte auf unsere Gesellschaft einfach hinzunehmen, weil Entwickler und Designer sich darüber keine Gedanken gemacht hatten bzw. es an Erfahrungswerten mangelte, ist nicht akzeptabel. Wie auch in der physischen Welt, wird unser Verhalten im digitalen Raum durch Designentscheidungen beeinflusst (Diefenbach & Ullrich, 2019). Es gibt auch hier kein neutrales Design.

Es braucht bewusste Überlegungen dazu, wie bestimmte Features der Technik sich auf soziale Dynamiken auswirken, um gewünschtes, prosoziales Verhalten zu fördern und asoziales Verhalten zu verringern. Hier tatsächlich funktionierende Lösungen zu entwickeln, ist und bleibt eine große Herausforderung. Selbst wenn bewusste Überlegungen stattfinden, können die gewählten Ansätze zur Förderung

prosozialen Verhaltens, auch wieder nicht gewünschte Seiteneffekte mit sich bringen. Wer beispielsweise asoziales Verhalten verhindern will, indem er Nutzer komplett gläsern macht, hat ein Problem gegen ein anderes getauscht. Die Entwicklung guter Lösungen, die moralisch und gesellschaftlich akzeptabel sind, ist somit eine der aktuellen Kernaufgaben im Feld von KI und Digitalisierung. Ähnliches gilt im Bereich der Überwachung/Social Scoring. Nicht alles, was technisch machbar ist, ist moralisch vertretbar. Negative Auswirkungen von Social Scores müssen bereits im Vorfeld erforscht werden, um keine Faktenlage zu schaffen, aus der man sich später kaum befreien kann.

### 5.5.1 Das grundsätzliche Problem mit Vorhersagen

Das Bemühen um Vorhersagen, wie die Technik sich weiter entwickeln wird, welche Effekte wir auf unsere Gesellschaft erwarten können, und wie wir diesen Einflüssen mit Voraussicht begegnen und eine gute Zukunft gestalten können, ist wünschenswert und löblich – konfrontiert uns aber auch immer wieder mit grundsätzlichen Problemen von Vorhersagen.

Die Vorhersage der Zukunft basiert notwendigerweise auf falschen Prämissen: In der Regel werden aktuelle Entwicklungen analysiert und es wird versucht, diese in die Zukunft zu projizieren und ihre Wechselwirkungen mit anderen Entwicklungen zu antizipieren. Daraus wird eine Vorhersage gebildet, die möglichst gut mit der tatsächlich in der Zukunft stattfindenden Entwicklung übereinstimmen soll. Ein Grundproblem hierbei ist, dass sogenannte disruptive, nicht vorhergesehene (engl. »to disrupt« = unterbrechen/stören) Technologien, Ereignisse oder Erkenntnisse nicht berücksichtigt werden. So steht der Begriff disruptive Technologien für Innovationen, die etablierte Produkte oder Dienstleistungen ersetzen oder verdrängen, und die Erfolgsserie bislang vorherrschender Ansätze unterbrechen (Danneels, 2004). Ein Beispiel wäre das Internet, das viele neue Geschäftsfelder eröffnet hat, gleichzeitig aber für viele bis dahin erfolgreiche Geschäftsmodelle einen Einbruch brachte, den so wahrscheinlich ein paar Jahre zuvor niemand prognostiziert hätte.

Als man Menschen in den 1950er Jahren fragte, wie diese sich das Jahr 2000 vorstellten, malten diese sich aus, dass Menschen in der Zukunft sich vermutlich mit fliegenden Autos, angetrieben von miniaturisierten Atomkraftantrieben, fortbewegen würden (Paul, 1955). Hier wurden schlicht zwei derzeit erfolgreiche existierende Technologien, das Auto und die Kernkraft, als Basis genommen und in die Zukunft projiziert. Die Gefahren und auch die technischen Grenzen der Kernkraft konnten nicht antizipiert werden.

Hätten die Menschen der Vergangenheit eine bessere Vorhersage treffen können, wenn sie sich intensiver mit der Kernkraft beschäftigt hätten? Möglicherweise. Aber selbst, wenn eine Fehleinschätzung berücksichtigt und korrigiert wird, lauern noch unzählige weitere.

In der zweiten Hälfte des letzten Jahrhunderts veröffentlichten Forscher des Massachusetts Institute of Technology (MIT) eine Studie zur Zukunft der Weltwirtschaft (Meadows et al., 1972). Die Kernfrage war hierbei, wann das aktuelle auf exponentiellem Wachstum basierende Wirtschaftssystem aufgrund seiner inhären-

ten Mängel notwendigerweise kollabieren müsste. In die Vorhersage mit einbezogen wurden zahlreiche Parameter wie Bevölkerungsentwicklung und -dichte, Überalterung der Gesellschaft, Warenverkehr, Staatshaushalt und -verschuldung und viele mehr. Der Zeitpunkt des Zusammenbruchs würde gemäß der Modellberechnung innerhalb der nächsten 100 Jahre liegen, also bis einschließlich des Jahres 2070. Dann sollten die Parameter in ihrer Gesamtheit ein Zusammenbrechen unvermeidlich machen und zu einem unkontrollierten Zusammenbrechen der Industrie und zu einem starken Bevölkerungsrückgang führen. Nicht berücksichtigt wurde aber der Zerfall der Sowjetunion, das rasche Aufsteigen Chinas zur Weltmacht und die Bedeutung des Klimawandels für den Planeten. Allesamt Entwicklungen, die jeweils für sich alleine genommen zahlreiche Umwälzungen in vielen Staaten der Welt produzieren könnten. Diese Entwicklungen waren aber in ihrer Stärke nicht absehbar, als die Berechnungen erstellt wurden, ihr Einfluss konnte also im Vorhersagemodell nicht angemessen berücksichtigt werden. Aus diesem Grund musste das Vorhersagemodell in der Zwischenzeit mit neuen Parametern aktualisiert werden (Herrington, 2021) – wobei zweifelhaft bleibt, ob es dieses Mal tatsächlich gelungen ist, alle relevanten Parameter zu berücksichtigen.

Vorhersagen sind daher in erster Linie Gedankenexperimente und erlauben keine perfekten Kenntnisse darüber, was tatsächlich passieren wird. Dies soll die Bedeutung derartiger Gedankenexperimente aber nicht schmälern. Auch nicht-perfekte Gedankenexperimente sind immer noch besser, als sich gar keine Gedanken zu machen. Sie können aufzeigen, was passieren könnte und sind damit auch Hinweisgeber auf mögliche Handlungsalternativen – damit wir eben nicht bloße Passagiere sind, die von der Zukunft überrollt werden, sondern diese aktiv mitgestalten können.

So enthält beispielsweise die Grundidee des Social Scoring bereits so viel an negativem Potenzial, dass die konkrete Ausgestaltung nur dem Versuch gleichkommen kann, verschiedene negative Szenarien in eine Rangreihe zu bringen. Hier wäre dann auch die Hoffnung auf ein disruptives Ereignis, das den Social Score obsolet machen würde, überflüssig, wenn wir schon heute durch unser Handeln die Etablierung eines solchen Konzepts verhindern können.

## Literatur

Agudo, U., & Matute, H. (2021). The influence of algorithms on political and dating decisions. *PLOS ONE, 16*(4), e0249454.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. *Proceedings of the 21st International Conference on World Wide Web – WWW ›12*, 519–528.

Beisch, N., & Schäfer, C. (2020). Ergebnisse der ARD/ZDF-Onlinestudie 2020. Internetnutzung mit großer Dynamik: Medien, Kommunikation, Social Media. *Media Perspektiven, 9*, 462–481.

Büchi, M., Festic, N., & Latzer, M. (2022). The Chilling Effects of Digital Dataveillance: A Theoretical Model and an Empirical Research Agenda. Big Data & Society, 9(1), 205395172110653.

Burbach, L., Halbach, P., Ziefle, M., & Calero Valdez, A. (2020). Opinion formation on the internet: The influence of personality, network structure, and content on sharing messages online. *Frontiers in Artificial Intelligence, 3*, 45.

Carrier, L. M., Spradlin, A., Bunce, J. P., & Rosen, L. D. (2015). Virtual empathy: Positive and negative impacts of going online upon empathy in young adults. *Computers in Human Behavior, 52*, 39–48.

Costa, P. (2018). Conversing with personal digital assistants: On gender and artificial intelligence. *Journal of Science and Technology of the Arts,10*(3), 59–72.

Crawford, K. (2021, June 6). Microsoft's Kate Crawford: ›AI is neither artificial nor intelligent‹ (Z. Corbyn, Interviewer) [Interview]. https://www.theguardian.com/technology/2021/jun/06/microsofts-kate-crawford-ai-is-neither-artificial-nor-intelligent [15.09.2022]

Danneels, E. (2004). Disruptive technology reconsidered: A critique and research agenda. *Journal of Product Innovation Management, 21*(4), 246–258.

Diefenbach, S., & Ullrich, D. (2019). Disrespectful technologies: Social norm conflicts in digital worlds. In T. Z. Ahram & C. Falcão (Hrsg.), *Advances in usability, user experience and assistive technology* (Vol. 794, S. 44–56). Basel: Springer International Publishing.

Eling, M., Nuessle, D., & Staubli, J. (2021). The impact of artificial intelligence along the insurance value chain and on the insurability of risks. *The Geneva Papers on Risk and Insurance – Issues and Practice, 47*, 205–241.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118.

Ernst, C.-P. H., & Herm-Stapelberg, N. (2020). The impact of gender stereotyping on the perceived likability of virtual assistants. 7. https://aisel.aisnet.org/amcis2020/cognitive_in_is/cognitive_in_is/4 [15.09.2022]

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications, 7*(1), 10.

Foggia, P., Percannella, G., & Vento, M. (2014). Graph matching and learning in pattern recognition in the last 10 years. *International Journal of Pattern Recognition and Artificial Intelligence, 28*(01), 1450001.

Gupta, S., Kar, A. K., Baabdullah, A., & Al-Khowaiter, W. A. A. (2018). Big data with cognitive computing: A review for the future. *International Journal of Information Management, 42*, 78–89.

Herrington, G. (2021). Update to limits to growth: Comparing the World3 model with empirical data. *Journal of Industrial Ecology, 25*(3), 614–626.

Hoffrage, U., & Marewski, J. N. (2020). Social Scoring als Mensch-System-Interaktion. In O. Everling (Hrsg.), *Social Credit Rating* (S. 305–329). Wiesbaden: Springer Fachmedien.

Jonsson, O., & de Tena, C. L. (2021). *European tech insights 2021. Part II embracing and govering technological disruption.* Center for Governance of Change. https://docs.ie.edu/cgc/IE-CGC-European-Tech-Insights-2021-%28Part-II%29.pdf [15.09.2022]

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., … Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell, 172*(5), 1122–1131.e9.

Kim, T. W., & Routledge, B. R. (2021). Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. business ethics quarterly, 1–28.

Kostka, G. (2019). China's social credit systems and public opinion: Explaining high levels of approval. *New Media & Society, 21*(7), 1565–1593.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review, 9*(2).

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (2009). A brief history of the internet. *ACM SIGCOMM Computer Communication Review, 39*(5), 22–31.

Macdonald, C. (2020). Avatars, disconnecting agents: Exploring the nuances of the avatar effect in online discourse. *Open Science Journal, 5*(2).

Meadows, D. H., Meadows, D. L., Randers, J., & Behrens, W. W. (1972). *The Limits to growth: A report for the Club of Rome's project on the predicament of mankind.* New York: Universe Books.

Moffett, S., & Santos, J. (2014). Social media as an influencer of public policy, cultural engagement, societal change and human impact. *Proceedings of the European Conference on Social Media: ECSM 2014*, 312–319.

Nawaz, N., & Mary, A. (2019). Artificial intelligence chatbots are new recruiters. *International Journal of Advanced Computer Science and Applications, 10*(9).

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 1–6.

Palacios, A. (2019, November 10). The internet's »wild west« era: A love letter to the early 00's internet. Alejandro Palacios. https://medium.com/@alejandropalacios_98575/the-internets-wild-west-era-a-love-letter-to-the-early-00-s-internet-3075722f79ae [15.09.2022]

Paul, F. R. (1955). Tandem wheel, gyroscopic, atomic-powered flying car.

Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. *Proceeding of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction – HRI ›06*, 218.

Riikkinen, M., Saarijärvi, H., Sarlin, P., & Lähteenmäki, I. (2018). Using artificial intelligence to create value in insurance. *International Journal of Bank Marketing, 36*(6), 1145–1168.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Scavarelli, C. M. (2018). The future of dating (No. *6*) [Song]. https://soundcloud.com/user-145965453 [22.09.2022]

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature, 588*(7839), 604–609.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489.

Sourdin, T. (2021). Judges, technology and artificial intelligence: The artificial judge. Edward Elgar Publishing.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321–326.

Taylor, K. (2021, November 1). One statistic shows how much Amazon could dominate the future of retail. Business Insider. https://www.businessinsider.com/retail-apocalypse-amazon-accounts-for-half-of-all-retail-growth-2017-11 [15.09.2022]

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1701–1708.

Upadhyay, A. K., & Khandelwal, K. (2018). Applying artificial intelligence: Implications for recruitment. *Strategic HR Review, 17(5)*, 255–258.

Vermeys, N. (2021). The computer as the court: How will artificial intelligence affect judicial processes? In X. Kramer, A. Biard, J. Hoevenaars, E. Themeli (Hrsg.), *New pathways to civil justice in Europe.* Basel: Springer.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., & Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature, 575*(7782), 350–354.