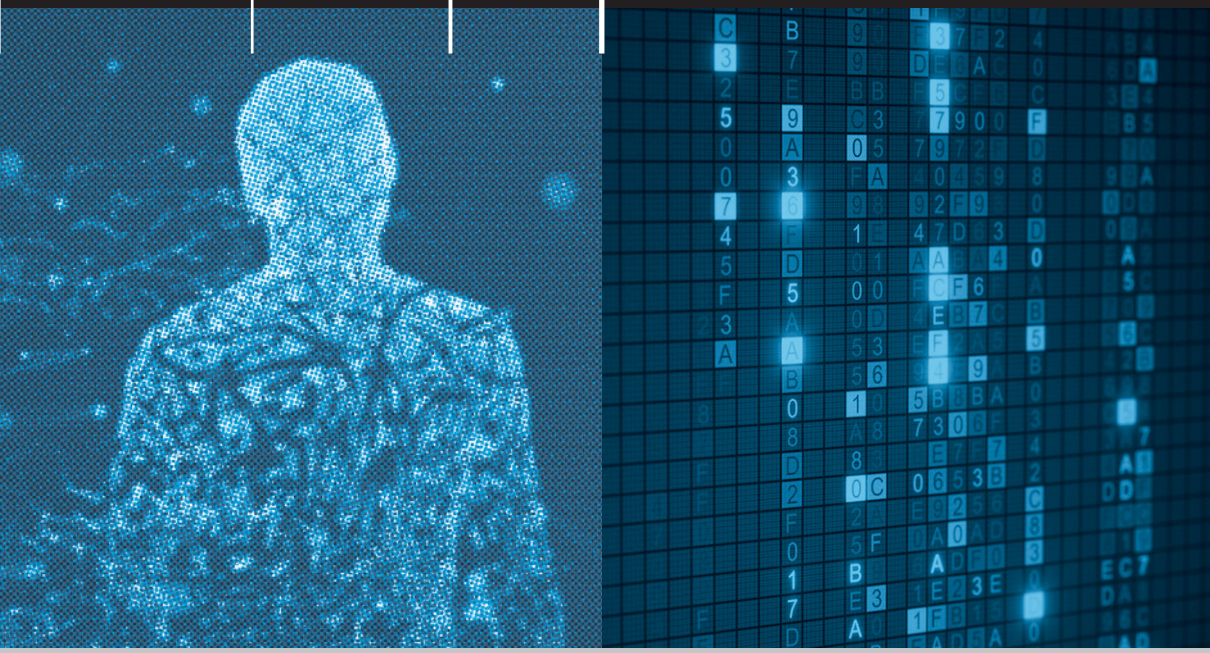

Markus Christen, Clemens Mader, Johann Čas, Tarik Abou-Chadi, Abraham Bernstein, Nadja Braun Binder, Daniele Dell'Aglio, Luca Fábián, Damian George, Anita Gohdes, Lorenz Hilty, Markus Kneer, Jaro Krieger-Lamina, Hauke Licht, Anne Scherer, Claudia Som, Pascal Sutter, Florent Thouvenin

Wenn Algorithmen für uns entscheiden: Chancen und Risiken der künstlichen Intelligenz

Brunngasse 36
CH-3011 Bern
www.ta-swiss.ch

TA-SWISS 72/2020

Markus Christen, Clemens Mader, Johann Čas, Tarik Abou-Chadi, Abraham Bernstein, Nadja Braun Binder, Daniele Dell'Aglio, Luca Fábián, Damian George, Anita Gohdes, Lorenz Hilty, Markus Kneer, Jaro Krieger-Lamina, Hauke Licht, Anne Scherer, Claudia Som, Pascal Sutter, Florent Thouvenin

Wenn Algorithmen für uns entscheiden: Chancen und Risiken der künstlichen Intelligenz

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Zitiervorschlag
Christen M., Mader C., Čas J., Abou-Chadi T., Bernstein A., Braun Binder N., Dell'Aglio D., Fábián L., George D., Gohdes A., Hilty L., Kneer M., Krieger-Lamina J., Licht H., Scherer A., Som C., Sutter P., Thouvenin F. (2020).
Wenn Algorithmen für uns entscheiden: Chancen und Risiken der künstlichen Intelligenz
In TA-SWISS Publikationsreihe (Hrsg.): TA 72/2020. Zürich: vdf

Danksagung
Das Autorenteam möchte sich hiermit herzlich bei allen Expertinnen und Experten bedanken, welche sich für das Ausfüllen der Umfrage(n) sowie für die Teilnahme an unserem Workshop Zeit genommen haben. Ohne die Mitwirkung dieser Fachpersonen wäre die Umsetzung dieser Studie nicht möglich gewesen.

Coverabbildungen:
© Links: iStock.com/DKosig
© Rechts: iStock.com/gonin

ISBN 978-3-7281-4001-2 (Printausgabe)

Download open access:
ISBN 978-3-7281-4002-9 / DOI 10.3218/4002-9

www.vdf.ethz.ch
verlag@vdf.ethz.ch

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Die digitale Transformation – die umfassende Nutzung moderner Informations- und Kommunikationstechnologie in allen Bereichen der Gesellschaft – ist ein Kernthema aktueller Technikfolgenabschätzung. Ob Arbeit, Bildung, Konsum, Medien oder Verwaltung: praktisch jeder Lebensbereich wird zunehmend durch digitale Technologien geprägt. Einer dieser Technologien – der künstlichen Intelligenz (KI) – kommt hier eine Schlüsselrolle zu. Dank Fortschritten im Bereich des maschinellen Lernens und der Rechenleistung sowie wegen der durch Internet oder Smartphones enorm gestiegenen Datenverfügbarkeit kann KI für Probleme eingesetzt werden, an denen herkömmliche Computerprogramme bislang gescheitert sind. Diese Fortschritte haben dazu geführt, dass KI-Systeme beeindruckende Erfolge bei anspruchsvollen und mehrdeutigen Aufgaben wie Bilderkennung, Übersetzung natürlicher Sprache oder Spielen erzielen konnten. KI-Systeme verbessern sich rasant und führen zu Anwendungen, die zuvor Menschen vorbehalten waren, wie z.B. das Führen von Fahrzeugen oder die Diagnose von Krankheiten. KI hat sich zu einer Basistechnologie entwickelt, die für den technologischen Fortschritt unverzichtbar erscheint.

Diese Studie wirft einen umfassenden Blick auf Chancen und Risiken dieser Basistechnologie. Der Fokus liegt auf KI-Anwendungen, die zur Unterstützung oder Automatisierung von Entscheidungsprozessen in relevanten gesellschaftlichen Bereichen eingesetzt werden sollen. Sie widmet sich fünf Anwendungsgebieten:

- **Arbeit:** Schwerpunkt sind volkswirtschaftliche Fragen sowie Auswirkungen von KI auf den individuellen Arbeitsprozess.
- **Bildung und Forschung:** Diskutiert werden die Nutzung von KI für die Personalisierung des Lernens sowie für die Förderung von Innovation.
- **Konsum:** Im Fokus liegen der Einfluss von KI auf das Verhältnis von Unternehmen zum Endkunden und wettbewerbsrechtliche Fragen.
- **Medien:** Thema ist die Rolle von KI bei Phänomenen wie *fake news* oder Filterblasen und deren Bedeutung für die politische Meinungsbildung.
- **Verwaltung und Gerichtsbarkeit:** Analysiert werden Fragen, die sich durch die Nutzung von KI bei staatlichem Handeln ergeben.

Die Untersuchungen in diesen Bereichen werden begleitet von technischen Klärungen sowie der Analyse und Beurteilung ethischer und rechtlicher Fragestellungen. In methodischer Hinsicht stützt sich die Studie auf einen Methodenmix aus Literaturstudium, Workshops und einer zweistufigen Befragung von über 300 Fachpersonen. Mittels Umfragen und Workshops wurden Chancen und Risiken von KI beurteilt. Aus diesen Arbeiten resultieren insgesamt sieben allgemeine und je zwei spezifische Empfehlungen pro untersuchten Bereich. Umgesetzt wurde die Studie von Forschenden der Digital Society Initiative der Universität Zürich, der Abteilung Technologie und Gesellschaft der Empa St. Gallen und dem Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften in Wien.

Die Autorinnen und Autoren der Studie sind dabei mit der Herausforderung konfrontiert, dass der Begriff «künstliche Intelligenz» im öffentlichen Diskurs oft diffus verwendet wird und mit anderen Technologietrends wie etwa dem Internet of Things oder der Robotik verschränkt ist. Deshalb ist es schwierig, die Rolle von KI etwa in makroökonomischen Veränderungen klar zu benennen. Um diesen Schwierigkeiten zu begegnen, wird in dieser Studie zuerst eine Einführung in KI-Technologien gegeben, wobei folgende Definitionen erarbeitet wurden:

«Künstliche Intelligenz» bezeichnet den Versuch, Verstehen und Lernen mittels eines Artefakts nachzubilden, wobei in erster Linie auf Denken bzw. Handeln fokussiert sowie ein rationales Ideal oder eine Nachbildung menschlicher Fähigkeiten angestrebt wird.

«KI-Technologie» bezeichnet einzelne, in Computer implementierbare Funktionen für die Erreichung von künstlicher Intelligenz (z.B. maschinelles Lernen).

«KI-System» bezeichnet eine strukturierte, kontextgebundene Kombination von KI-Technologien zwecks Erreichens von künstlicher Intelligenz.

«KI-Entscheidungen» sind Schlussfolgerungen von KI-Systemen mit realweltlichen Auswirkungen, die auf der Ebene des Systemdesigns, der strategischen Ebene (Entscheid über Einsatz des Systems) und der taktischen Ebene (Interaktion mit dem Bediener) von menschlichen Entscheidungen abhängig sind.

Technische Treiber der aktuellen Diskussion rund um künstliche Intelligenz sind neuere Anwendungen des maschinellen Lernens (Deep Learning), die auf neuronalen Netzen beruhen – komplexe Programme, die durch grosse Datenmengen «trainiert» werden. Rund um diese Technologie haben sich in den letzten Jahren

verschiedene Diskussionsfelder entwickelt: Bezüglich der verwendeten Trainingsdaten wird debattiert, inwieweit sich in diesen Einseitigkeiten oder Befangenheit (*bias*) verbergen. Das verwendete Verfahren (Deep Learning) ist mit dem Problem konfrontiert, dass der Lösungsweg des Algorithmus schwer nachvollziehbar ist (*black box*). Beim Design der KI-Systeme stellt sich die Frage, inwieweit dadurch bestimmte Werte und Interessen gegenüber anderen privilegiert werden. Dabei können sich schwierige Probleme ergeben, weil sich gewisse Ziele (z.B. unterschiedliche Interpretationen von Fairness) ausschliessen. All diese Aspekte führen schliesslich zur Frage, inwieweit Menschen KI-Systemen vertrauen können, die für die Unterstützung oder zum Ersatz menschlicher Entscheidungsfindung eingesetzt werden.

Nebst diesen auf die Technologie selbst fokussierten Problemfeldern werden aber auch gesellschaftliche Folgen debattiert. Dominierende Themen sind hier sozioökonomische Effekte von KI (insbesondere ein befürchteter Verlust zahlreicher Arbeitsplätze), wettbewerbsrechtliche Fragen aufgrund der Notwendigkeit grosser Datenmengen für die Erstellung bestimmter KI-Systeme sowie geostrategische Überlegungen, weil Grossmächte wie China und die USA grosse Summen in die Entwicklung von KI-Technologien investieren.

Diese Breite an Themen spiegelt sich in der internationalen Debatte zu KI wider, die in der vorliegenden Studie dargestellt wird. Die Darstellung belegt das enorme Interesse an einem besseren Verständnis der Chancen und Risiken von KI. Ein wichtiger Treiber sind hierbei **ethische Fragen**. Unter anderem weckt die zunehmende Nutzung von KI in sensitiven Bereichen Befürchtungen bezüglich mangelnder Transparenz, der Verfestigung von diskriminierenden Praktiken und einer unklaren Zuschreibung von Verantwortung. Die künstliche Intelligenz steht damit beispielhaft für die Befürchtung eines menschlichen Kontrollverlustes als Folge des sich beschleunigenden technischen Fortschritts. Ergänzt wird die Erörterung ethischer Fragen mit einer umfassenden Darstellung der **rechtlichen Fragen**, welche die Nutzung von KI-Systemen z.B. durch Unternehmen aufwirft. Besprochen werden dabei das Haftungsrecht, das Immaterialgüterrecht, das Datenschutzrecht und das verfassungsrechtliche Diskriminierungsverbot.

Die Untersuchungen in den fünf Themenbereichen fokussieren auf aktuelle oder in naher Zukunft erwartbare KI-Anwendungen; Spekulationen über «superintelligente KI» – also Systeme mit Fähigkeiten, die den Menschen generell und nicht nur bei spezifischen Aufgaben weit übertreffen – waren nicht Teil der Arbeit. Die

Analysen sind wie folgt strukturiert: In einem ersten Schritt wird die aktuelle Literatur dargestellt, um einen Überblick über die zu erwartenden Einsatzformen und Effekte von KI zu erlangen. Dies beinhaltet auch eine Darstellung der wichtigsten in der internationalen Debatte vorgebrachten Vorschläge zur Minimierung von Risiken und der Förderung positiver Anwendungen von KI. In einem nächsten Schritt sind in einer zweistufigen Umfrage unter Expertinnen und Experten – gemeint sind sowohl technische als auch bereichsspezifische Fachpersonen – Beurteilungen zu KI-Anwendungsformen, zu Chancen und Risiken sowie zu möglichen Massnahmen eingeholt worden. Die Ergebnisse der Umfrage sind schliesslich in zwei Workshops verdichtet worden und flossen in die hier vorliegenden Empfehlungen ein.

Im Themenfeld **Arbeitswelt** geht es zunächst um die weitverbreitete Befürchtung, dass KI-Systeme menschliche Arbeitskraft in grossem Umfang obsolet machen könnten, was enorme Auswirkungen auf die Gesellschaft haben würde. Eine Analyse der Literatur zeigt aber die Schwierigkeit, hierzu genaue Prognosen zu machen. Der Rückgriff auf historische Erfahrungen mit früheren Automatisierungsschüben ist dabei ebenfalls problembehaftet, weil unklar ist, ob eine Vergleichbarkeit gegeben ist. Die Analysen fokussieren insbesondere auf Massnahmen, die einer Polarisierung am Arbeitsmarkt (zunehmende Spezialisierung versus prekäre Arbeitsbedingungen), der damit verknüpften Polarisierung der Löhne sowie einer Abnahme des Beschäftigungsvolumens entgegenwirken sollen. Hinsichtlich der Effekte von KI auf die Gestaltung der Arbeit selbst werden die Selektion von Arbeitskräften, deren Überwachung und Karriereplanung und indirekte Effekte von KI auf die Gestaltung der Arbeit thematisiert. Unter anderem rückt dabei die Bedeutung der Weiterbildung als Massnahme für die positive Gestaltung des digitalen Wandels in das Blickfeld.

Im Themenfeld **Bildung und Forschung** liegt der Schwerpunkt bei der Bildung auf der obligatorischen Schulstufe. Untersucht worden sind KI-Anwendungen zur Unterstützung der Administration, der Lernenden und der Lehrenden. Die diskutierten Massnahmen umfassen unter anderem die Darlegung von KI-Kompetenzen, die in den Schulen vermittelt werden sollten, sowie den Schutz von Daten aus dem Bildungswesen, die Voraussetzungen für die Nutzung von KI sind. Im Bereich Forschung sind beispielhaft Anwendungen von KI zur Innovationsförderung analysiert worden. Die diskutierten Massnahmen betreffen unter anderem die Frage, wie interdisziplinäre Forschung zu KI gefördert werden kann.

Im Themenfeld **Konsum** werden insbesondere KI-Anwendungen für Personalisierung (z.B. von Angeboten und Preisen), Empfehlungssysteme und digitale Assistenten untersucht. Nebst den Vorteilen für Konsumentinnen und Konsumenten sowie Unternehmen kommen die Themen «Erkennbarkeit der KI» sowie «KI als *Blackbox* für den Konsumenten» und daraus resultierende Fragen zum Schutz der Privatsphäre und des Vertrauens in KI zur Sprache. Wettbewerbsrechtliche Überlegungen betreffen Datenmonopole und Netzwerkeffekte. Die diskutierten Massnahmen betreffen insbesondere die Förderung des Kundenvertrauens, die Sicherung der Privatsphäre und die Verhinderung von Oligopolen.

Im Themenfeld **Medien** wird der Einsatz von KI im Kontext eines sich verändernden Informationsverhaltens untersucht, zumal sich die Art des Medienkonsums und die Rolle traditioneller Medien wie Presse und TV in den letzten Jahren stark gewandelt haben. Im Fokus liegen hier zum einen sogenannte Filterblasen und Echokammern, welche zu einer inhaltlichen Uniformierung des Medienkonsums von Einzelnen oder ganzen Gruppen führen können. Zum anderen wird das Thema *fake news* ins Zentrum gerückt – auch weil KI-Technologien neue Möglichkeiten zur Fälschung von Audio- und Videoinhalten eröffnet haben. Die diskutierten Massnahmen fokussieren auf den Schutz des öffentlichen Diskurses vor ungewollten oder bewussten manipulativen Einflüssen durch KI-Technologien.

Im Themenfeld **Verwaltung und Gerichtsbarkeit** schliesslich wird der staatliche KI-Einsatz untersucht; insbesondere im Kontext von hoheitlichem Handeln – also beispielsweise der KI-gestützte Erlass von Verfügungen oder die vorausschauende Polizeiarbeit (*predictive policing*). Im Zentrum der Analyse steht dabei die Frage, wie Anforderungen an staatliches Handeln – z.B. die Begründungspflicht, das Diskriminierungsverbot oder der Anspruch auf rechtliches Gehör – durch KI-Einsatz tangiert werden und mit welchen Massnahmen sichergestellt werden kann, dass sich diese Anforderungen weiterhin erfüllen lassen. Als Risikopunkte werden unter anderem eine Gefährdung der Unschuldsvermutung, intransparente Verfahren sowie Maschinenhörigkeit identifiziert.

Aus diesen Analysen gehen sieben generelle und zehn bereichsspezifische Empfehlungen zuhanden des Gesetzgebers, der Bundesbehörden sowie anderer Stakeholder (z.B. Unternehmen oder Öffentlichkeit) hervor. Diese beschränken sich nicht auf regulatorische Aspekte. Sie benennen auch Diskussionsbedarf und verweisen auf die Förderung technischer Lösungen für die Unterstützung von bestehendem Recht. Schliesslich beinhalten sie Massnahmen, wie einzelne Akteure befähigt werden können, ihre Aufgaben angesichts der Herausforderungen von KI

besser wahrzunehmen. Zu jeder Empfehlung wird eine kurze Erläuterung gegeben und der Adressat bzw. die Adressatin spezifiziert. Die Empfehlungen orientieren sich dabei an folgenden Leitideen:

- **Fokussierte Regulierung:** Obgleich moderne KI-Systeme gewisse Charakteristika wie Bedarf an grossen Datenmengen und fehlende Erklärbarkeit oft teilen, hängt die Art der Risiken und die Möglichkeit von deren Beeinflussung entscheidend von den jeweiligen KI-Anwendungen ab. Entsprechend sollten Notwendigkeit und Ausgestaltung von Regulierungen bereichsspezifisch geprüft werden; allgemeine Regulierungsansätze wie beispielsweise eine Regelung im Datenschutzrecht sind unzureichend und oft auch ungeeignet, um KI-Risiken sachgerecht in den Griff zu bekommen.
- **Regelmässiges und integratives Monitoring:** Die Forschung im Bereich KI ist sehr dynamisch; es ist durchaus denkbar, dass einige der heute hochgelobten Anwendungen sich als Fehlschlag erweisen, während neue Ideen heute noch ungeahnte Folgefragen stellen werden. Entsprechend ist die Entwicklung im Bereich KI aufmerksam zu beobachten und der Gesetzgeber ist angehalten, in regelmässigen Abständen den Regulierungsbedarf zu prüfen – dies unter Einbezug internationaler Entwicklungen (insbesondere in der EU) sowie durch Mitgestaltung eines öffentlichen Dialogs.
- **Förderung von Befähigungen:** Eine zentrale Leitidee für den Umgang mit KI ist die kontextspezifische Umsetzung von Massnahmen, die eine optimale Befähigung der Nutzerinnen und Nutzern bzw. der Betroffenen ermöglicht, sodass die Chancen von KI maximiert und deren Risiken kontrolliert werden. Dazu gehört die Einsicht, dass die staatliche Nutzung von KI bei hoheitlichem Handeln höheren Anforderungen unterliegen soll und dass Transparenz über die Massgabe des Datenschutzrechtes hinaus die Erkennung von Fehlentwicklungen erlauben soll. Institutionen der Zivilgesellschaft bzw. die Regulierungsinstanzen sollen zudem befähigt werden, private KI-Zertifizierungen zu prüfen. Fachleute, welche KI-Systeme entwickeln, implementieren oder über deren Einsatz entscheiden, sollen sich schliesslich Kenntnisse über ethische, rechtliche und soziale Aspekte der Nutzung von KI aneignen.

Ausgehend von diesen Leitideen werden in dieser Studie folgende sieben allgemeinen Empfehlungen gemacht (die Reihenfolge reflektiert keine Priorisierung):

1. Der Gesetzgeber soll im Bereich der KI eine technologieneutrale und bereichsspezifische Herangehensweise verfolgen: Anstelle eines allgemeinen «KI-Gesetzes» sollen bereichsbezogen konkrete Probleme und Fehlentwicklungen in regelmässigen Abständen identifiziert, evaluiert und gegebenenfalls mittels geeigneter Rechtsnormen gelöst werden. Dabei sollen insbesondere die Entwicklungen in der Europäischen Union Beachtung finden.

2. Der Gesetzgeber soll den Einsatz von KI-Systemen nicht als datenschutzrechtliches Problem auffassen. Zwar greift der Datenschutz, wenn Personendaten zur Entwicklung und Anwendung von KI genutzt werden. Risiken, die durch die Nutzung von Sachdaten entstehen, oder Diskriminierung als Folge der Datenbearbeitung brauchen aber neue bzw. andere Ansätze, die entworfen und weiterentwickelt werden sollten.

3. Gesetz- und Verordnungsgeber sollen sicherstellen, dass für staatliche Akteure (z.B. Gerichte, Polizei und Verwaltung) höhere Anforderungen an die Nutzung von KI gelten als für Private, wenn hoheitliche KI-Nutzung Menschen in relevanter Weise betrifft: In diesem Fall muss stets gewährleistet sein, dass die Betroffenen die Rechtmässigkeit des staatlichen Handelns beurteilen können.

4. Setzen Private (Unternehmen und andere Organisationen) KI für Entscheidungen ein, die Menschen in relevanter Weise betreffen, sollen sie neben der aktiven Information über den Umstand des KI-Einsatzes auch eine Transparenz auf Nachfrage sicherstellen. Über die Massgabe des Datenschutzrechtes hinaus sollen beispielsweise Institutionen der Zivilgesellschaft auf Nachfrage alle Informationen erhalten, die eine Einschätzung möglicher Fehlentwicklungen erlauben. Der Schutz der Geschäftsgeheimnisse der Unternehmen und anderen Organisationen ist dabei in angemessener Weise zu gewährleisten.

5. Organisationen (beispielsweise im Bereich Konsumentenschutz) sollen durch staatliche Unterstützung besser befähigt werden, private KI-Zertifizierungen zu prüfen. Private Initiativen für KI-Zertifizierung und die Vergabe entsprechender Labels sind zu begrüssen. Die Nutzung von KI-Systemen soll aber nicht generell von einer Marktzulassung abhängig gemacht werden.

6. Hochschulen und weitere Bildungsinstitutionen, welche KI-Fachleute ausbilden, sollen auch nicht technische Kompetenzen fördern: Fachleute, welche

KI-Systeme entwickeln, implementieren oder über deren Einsatz entscheiden, sollen sich Kenntnisse über rechtliche, ethische und soziale Aspekte der Nutzung von KI aneignen.

7. Bund, Hochschulen, Unternehmen und zivilgesellschaftliche Organisationen sollen gemeinsam den gesellschaftlichen Dialog über Chancen und Risiken der KI fördern. In Bereichen mit unklarer Risikolage muss dabei auch die Forschung zur Erkennung solcher Risiken intensiviert werden, was durch entsprechende Massnahmen der Hochschulen und Institutionen der Drittmittelförderung unterstützt werden soll.

Nebst diesen sieben allgemeinen Empfehlungen werden für jeden untersuchten Bereich je zwei spezifische Empfehlungen formuliert. Die Empfehlungen im Bereich **Arbeit** bringen dabei zum Ausdruck, dass im makroökonomischen Bereich konkrete Effekte von KI schwer von anderen Faktoren des digitalen Wandels abgrenzbar sind, diese aber insgesamt den Bedarf an einer Debatte über Anpassungsprozesse erhöhen. Die Nutzung von KI-Systemen im Arbeitsprozess selbst soll wiederum so gestaltet werden, dass die Durchsetzung geltender Rechte (z.B. Mitspracherechte von Mitarbeitenden) nicht behindert wird. Die entsprechenden Empfehlungen lauten:

- Der Bund soll mögliche makroökonomische Auswirkungen der digitalen Transformation im Allgemeinen und von KI im Speziellen verstärkt zum Anlass nehmen, gesellschaftliche Debatten über Anpassungsprozesse anzustossen: Dies betrifft insbesondere die Bereiche Arbeitszeitverkürzung, Flexibilisierung der Arbeit hinsichtlich Zeit, Ort und Mittel, ökonomische Polarisierung und Weiterbildung.
- Der Gesetz- und Verordnungsgeber soll das Mitspracherecht der Mitarbeitenden sicherstellen, wenn Unternehmen KI-Systeme für deren Überwachung und Kontrolle einsetzen: Die Arbeitsinspektorate müssen angemessen ausgestattet werden, um die Einhaltung der gesetzlichen Bestimmungen effektiv kontrollieren zu können.

Die Empfehlungen im Bereich **Bildung** (jene zur Forschung benennen vorab Forschungslücken; siehe unten) bringen einerseits zum Ausdruck, dass Daten von Schülerinnen und Schülern für die Personalisierung des Lernens besonders schützenswert sind. Andererseits muss sichergestellt werden, dass die positiven Nutzungen von KI in der Bildung vorangetrieben und entsprechende Erfahrungen ausgetauscht werden. Die entsprechenden Empfehlungen lauten:

- Kantonale Gesetz- und Verordnungsgeber und die Erziehungsdirektoren sollen Leitlinien formulieren, wie mit Daten über Leistung und Verhalten von Lernenden und den daraus mittels KI-Systemen gewonnenen Schlüssen umgegangen werden soll: Insbesondere ist zu prüfen, ob und welche über den aktuellen Datenschutz hinausgehenden Mechanismen zu schaffen sind, um Lernende vor negativen Folgen der Nutzung und Bekanntgabe ihrer Lern- und Leistungsdaten an Dritte zu schützen.
- Die Bildungsinstitutionen und insbesondere die pädagogischen Hochschulen sollen untersuchen, welche spezifischen Kompetenzen vermittelt werden müssen, um ein allgemeines Verständnis von Fähigkeiten und Grenzen von KI-Systemen zu erhalten: Entsprechende Erkenntnisse sollen in Lehrmittel einfliessen und unter Nutzung bestehender Plattformen für Lehrkräfte und Lernende verfügbar gemacht werden.

Die Empfehlungen im Bereich **Konsum** zielen in erster Linie darauf, wie das geltende Datenschutzrecht ergänzt werden kann, sodass Konsumentinnen und Konsumenten transparent über die Nutzung von KI informiert werden und dass der Wettbewerb auch bei datenintensiven KI-Anwendungen erhalten bleibt. Die entsprechenden Empfehlungen lauten:

- Unternehmen, welche KI-Systeme im Konsumbereich nutzen und dafür Personendaten erheben, sollen die Transparenz des KI-Einsatzes und die sonstigen Anforderungen an den Datenschutz möglichst einfach vermitteln. Entsprechende Forschung und *best practices* sind zu fördern.
- Der Gesetzgeber soll prüfen, wie Datenportabilität im Bereich von KI-Systemen umgesetzt werden kann, insbesondere um Konsumentinnen und Konsumenten den Wechsel zu einem anderen Anbieter zu erleichtern.

Die Empfehlungen im Bereich **Medien** zielen darauf ab, das Bewusstsein des Effekts von Personalisierung beim Einzelnen zu schärfen und eine gesellschaftliche Debatte zu fördern über grundlegende Fragen wie jene nach der Meinungsfreiheit, dem Umgang mit *fake news* sowie der Rolle des Staates beim Schutz des demokratischen Meinungsbildungsprozesses vor Kampagnen illegitimer Akteure (z.B. Drittstaaten). Die entsprechenden Empfehlungen lauten:

- Die Betreiber von Medienplattformen sollen ihren Nutzerinnen und Nutzern auf einfache Weise erkennbar machen, wie die Personalisierung von Medieninhalten mittels KI die Auswahl angezeigter Inhalte beeinflusst.

- Der Bund soll in Zusammenarbeit mit Medienunternehmen und zivilgesellschaftlichen Akteuren die gesellschaftliche Diskussion über den Umgang mit *fake news*, Filterblasen und Echokammern intensivieren. Sicherheitsbehörden (Polizei, Nachrichtendienst und Armee) sollen unter der Voraussetzung der parlamentarischen Kontrolle Fähigkeiten entwickeln, systematische *fake-news*-Kampagnen mit dem Ziel politischer Manipulation rascher zu identifizieren und die Öffentlichkeit entsprechend zu informieren.

Die Empfehlungen im Bereich **Verwaltung** schliesslich konkretisieren die besonderen Anforderungen der staatlichen KI-Nutzung bei hoheitlichem Handeln. Die entsprechenden Empfehlungen lauten:

- Die öffentliche Verwaltung soll Kriterien definieren, anhand derer bestimmt werden kann, wie eine verantwortliche staatliche KI-Nutzung konkret umgesetzt werden kann.
- Die öffentliche Verwaltung soll sicherstellen, dass Daten, welche für die staatliche KI-Nutzung genutzt werden, eine ausreichende Qualität haben.

Der Bericht schliesst mit einer Auflistung von **Forschungslücken**, deren Schliessung einen sachgerechten Umgang mit KI und entsprechende Innovationen fördern können. Dies betrifft nicht nur technische Aspekte wie die Verbesserung des Verständnisses neuer Formen des maschinellen Lernens (erklärbare KI). Auch Forschung ist nötig, beispielsweise in den Rechtswissenschaften (z.B. regulatorischer Umgang mit autonomen Systemen), der Psychologie (z.B. Bedingungen für Vertrauen in KI) oder den Humanwissenschaften (z.B. was bedeutet es, ein KI-System zu kontrollieren). Um das Potenzial von KI-Systemen in der Forschung und Innovation zu nutzen, wird Forschungseinrichtungen empfohlen, KI-Zentren einzurichten, um die genannten Forschungslücken besser angehen zu können.

Die Studie soll, so ist abschliessend festzuhalten, einen Beitrag zur «Entmystifizierung» der KI-Debatte weg von übertriebenen Ängsten und Erwartungen hin zu einer Analyse konkreter Chancen und Risiken einer grundsätzlich vielversprechenden Technologie liefern. Das Autorenteam hofft, dass die umfangreiche Darstellung zahlreicher aktueller Entwicklungen im Bereich KI den Leserinnen und Lesern wichtige Einsichten und Denkanstösse gibt.

Summary

Digital transformation – the comprehensive use of modern information and communication technology across all areas of modern society – forms a central theme in technology assessment today. Practically every area of life, whether it be work, education, consumption, media or administration, is being shaped by digital technologies. And of these technologies, there is one – artificial intelligence (AI) – that plays a particularly key role. Thanks to the advances made in machine learning and computing performance, as well as the enormous increase in the availability of data through the internet and smartphones, AI can be used to solve problems at which traditional computer programmes have so far failed. This progress has led to impressive success with AI systems in tackling highly sophisticated and complex tasks such as image recognition, natural language translation and gaming. AI systems are rapidly improving, with applications being developed in domains previously exclusive to humans, such as driving vehicles or diagnosing diseases. As a result, AI has evolved into a fundamental technology without which any new technological advances hardly seem possible.

This study takes a comprehensive look at the opportunities and risks of this fundamental technology, focussing on AI applications designed to assist or automate decision-making processes in relevant areas of society. The study is divided into five areas of application:

- **Work:** here focus is given to economic issues and the impact of AI on the individual work process.
- **Education and research:** this section discusses the use of AI for the personalisation of learning and the advancement of innovation.
- **Consumption:** in this section, the study focusses on the influence of AI on the relationship between companies and end customers, and competition law issues.
- **Media:** here the topic is the role of AI in phenomena such as fake news or filter bubbles and their significance in the formation of political opinion.
- **Administration and jurisdiction:** this section contains an analysis of the questions arising from the use of AI in governmental actions.

The research conducted in these areas is complemented by technical explanations as well as analyses and assessments of ethical and legal issues. From a methodological point of view, the study is based on a combination of methods consisting of literature studies, workshops and a two-stage survey involving over three hundred experts. The opportunities and risks of AI were assessed by means of surveys and workshops. This led to a total of seven general and two specific recommendations for each subject area. The study was conducted by researchers from the Digital Society Initiative of the University of Zurich, the Technology and Society Lab of Empa St. Gallen and the Institute of Technology Assessment of the Austrian Academy of Sciences in Vienna.

One challenge the study faced was the fact that the term ‘artificial intelligence’ is often used in rather vague terms in public discourse and is easily muddled with other technological trends such as the ‘Internet of Things’ or robotics. This makes it difficult to clearly identify the role of AI in macroeconomic changes, for example. To address this difficulty, the study first provides an introduction to AI technologies, using definitions that were drawn up for this purpose:

‘Artificial intelligence’ refers to the attempt to replicate understanding and learning by means of an artefact. The main focus lies on thought and/or action, and the pursuit to achieve a rational ideal or the replication of human abilities.

‘AI technology’ refers to individual functions that can be implemented in computers to achieve artificial intelligence (e.g. machine learning).

‘AI system’ means a structured, context-bound combination of AI technologies for the purpose of achieving artificial intelligence.

‘AI decisions’ are conclusions made by AI systems with real-world consequences that depend on human decisions at the level of system design, at the strategic level (the decision to deploy the system) and at the tactical level (interaction with the operator).

The technologies driving the current discussion around artificial intelligence are the most recent machine learning applications (deep learning) based on neural networks, i.e. complex programmes that are ‘trained’ thanks to large amounts of data. Various subjects of discussion have emerged in recent years around this technology: with regard to the training data used, there is debate about how much subjectivity or bias this data conceals. The problem that the method applied (deep learning) faces is that the solution path of the algorithm is difficult to follow (black box). In designing AI systems, the question arises as to what extent certain values

and interests are being favoured over others. This can lead to serious problems as certain aims (e.g. different interpretations of fairness) contradict each other. All these aspects ultimately lead to the question as to how far people can trust AI systems that are used to assist or replace human decision making.

However, the discussions focus not only on the dilemmas raised by the technology itself, but also its consequences on society. The dominant topics here are the socio-economic effects of AI (particularly the fear of numerous job losses), competition law issues due to the need for large amounts of data for the creation of certain AI systems, and geostrategic considerations since major powers such as China and the USA are investing large sums in the development of AI technologies.

This wide range of issues is reflected in the international debate on AI presented in this report, demonstrating the huge interest in gaining a better understanding of the opportunities and risks of AI. One important topic driving this interest is the subject of **ethical issues**. Among other things, the rise in the use of AI in sensitive areas fuels fears regarding the lack of transparency, the consolidation of discriminatory practices, and the confusion surrounding responsibility. In other words, artificial intelligence reflects the fear of losing human control as a result of rapidly accelerating technological advances. This introduction to the ethical aspects of AI is supplemented by a comprehensive presentation of the **legal issues** posed by the use of AI systems – by companies, for example. The legal aspects discussed include liability law, intellectual property law, data protection law and the constitutional ban on discrimination.

The research in the five subject areas focuses on current or soon-to-be-expected AI applications; speculations about superintelligent AI systems – i.e. systems with capabilities that far exceed those of human beings on a general basis and not only in specific tasks – did not form part of the work. The analyses all follow the same pattern: firstly, current literature on the topic is presented in order to provide an overview of the envisaged applications and effects of AI. This also includes key proposals emerging from international debate on minimising risks and promoting beneficial AI applications. The second stage involves a two-part survey held with experts – both technical experts and subject area specialists – to obtain their assessments of different forms of AI applications, opportunities and risks, and potential measures. The results were consolidated over the course of two workshops and incorporated into the recommendations presented in this report.

With regard to the subject area **working world**, the report first looks at the common fear that AI systems could render human manpower redundant on a massive

scale which would have huge consequences on society. However, an analysis of the literature shows that it is difficult to make accurate predictions on this subject. And resorting to a review of historical experiences of previous booms in automation is also problematic as it is unclear whether reasonable comparisons can actually be made. The analyses then focus on measures designed to counteract the issues of polarisation in the labour market (increasing specialisation versus precarious working conditions), the polarisation of wages that this brings and the decline in employment volume. With regard to the effects of AI on the organisation of work itself, the study looks at employee selection, employee monitoring and career planning, and the indirect effects of AI on the organisation of work in the future. Among other things, attention is drawn to the importance of further education as a way of shaping digital transformation for the common good.

In the subject area **education and research**, the focus in terms of education lies on compulsory schooling. AI applications that support administration, teachers and learners were examined. The measures discussed include identifying the AI skills to be taught in schools as well as the protection of data in the education system that is requisite to the use of AI. In the area of research, possible AI applications to promote innovation were analysed. The measures discussed include the question of how to promote interdisciplinary research on AI.

In the subject area **consumption**, AI applications for personalisation (e.g. of offers and prices), recommendation systems and digital assistants were examined. Besides the advantages for consumers and companies, the topics ‘recognisability of AI’ and ‘AI as a black box for consumers’ and the resulting questions on the protection of privacy and trust in AI are addressed. Considerations regarding competition law involve data monopolies and network effects. The measures discussed in the report mainly concern the promotion of customer confidence, privacy protection and the prevention of oligopolies.

Under the subject area **media**, the use of AI is examined in the context of today’s constantly evolving information behaviour, especially since the nature of media consumption and the role of traditional media such as the press and TV have changed dramatically in recent years. The focus here is on what are known as filter bubbles and echo chambers, which can lead to a uniformity of the media content consumed by individuals or entire groups. In addition, attention is given to the subject of fake news – not least because AI technologies have opened up new possibilities for falsifying audio and video content. The measures discussed focus on

the protection of public discourse from unintentional or deliberate manipulative influences by AI technologies.

Under the subject area **administration and jurisdiction,** the government's use of AI is examined, in particular in the context of sovereign action – such as issuing decrees on an AI-assisted basis, or predictive policing. The analysis concentrates on the question of how requirements for government action – e.g. the obligation to give reasons, the principle of non-discrimination or the right to a legal hearing – are affected by the use of AI and what measures can be taken to ensure that these requirements will continue to be met. The risk points identified include the danger to the presumption of innocence, non-transparent procedures and servility to machines.

These analyses resulted in seven general and ten subject area specific recommendations for the attention of legislators, the federal authorities and other stakeholders (e.g. companies or the public). The recommendations are not limited to regulatory aspects – they also specify the need for discussion and underline the importance of developing technical solutions to support existing legislation. Finally, they include measures to enable individual actors to perform their duties better in light of the challenges posed by AI. With every recommendation comes a short explanation and a reference to whom the recommendation is directed. The recommendations are based on the following guiding principles:

- **Focussed regulation:** although modern AI systems often share certain characteristics such as the need for large amounts of data and a lack of explainability, the nature of their associated risks and the potential to influence these risks depend on the AI applications in question. Consequently, the necessity for regulations and the formulation of regulations should be examined in relation to each subject area; general regulatory strategies such as incorporating one single regulation in the data protection law are insufficient and often inappropriate for managing AI risks properly.
- **Regular and integrative monitoring:** research in the field of AI is highly dynamic and it is quite conceivable that some of the applications that are highly acclaimed today may prove to be a failure, while new ideas may lead on to questions quite literally unimaginable to us today. Consequently, developments in the field of AI must be closely monitored and legislators are called upon to review the need for new regulatory measures at regular intervals – after taking international developments (especially in the EU) into account and facilitating public discourse.

- **Building proficiency:** a key guiding principle in managing AI is to implement measures on a context-specific basis that provide users and stakeholders with optimal levels of proficiency in order to maximise the opportunities and control the risks of AI. Included here is the insight that the government's use of AI for sovereign acts should be subject to higher regulatory requirements and that transparency beyond the scope of the data protection law should enable undesirable developments to be identified. Civil society institutions and regulatory bodies should also be granted the authority to check private AI certifications. Finally, experts who develop, implement or have decision-making powers over the use of AI systems should receive training in the ethical, legal and social aspects of the use of AI.

Based on these guiding principles, the report makes the following seven general recommendations (in random order):

1. Legislators should adopt a technology-neutral approach to AI, specific to each subject area: instead of a general 'AI law', concrete problems and undesirable developments based on subject areas should be identified, assessed and accordingly solved by introducing suitable legal measures at regular intervals. Developments in the European Union should be taken into account during this process.
2. Legislators should not regard the use of AI systems as a data protection problem, even though data protection law does apply when personal data is used for the development and application of AI. However, the risks arising from the use of factual data or the discrimination that results from data processing call for new or different approaches to be drawn up and further developed.
3. Legislators and regulators should ensure that government players (e.g. courts, police and administration) are subject to higher regulatory requirements regarding the use of AI than the private sector if the sovereign use of AI affects people in a relevant way: should this be the case, measures must be in place to ensure that those affected are able to assess the legality of such action taken by the government.
4. If actors from the private sector (companies and other organisations) use AI for decisions that have a relevant effect on people, they should not only actively communicate the circumstances surrounding their use of AI but also provide transparency on request. For example: civil society institutions should, when they so request, be provided with all the information – beyond what is

stipulated in data protection law – that would allow them to assess any possible undesirable developments. At the same time, the protection of business secrets held by companies and other organisations must be appropriately guaranteed.

5. Organisations (in consumer protection, for example) should, with support from the government, be better enabled to check private AI certifications. Any private initiatives on the subject of AI certification and the awarding of corresponding labels are to be welcomed. However, the use of AI systems should not generally be made dependent on market approval.

6. Universities and other educational institutions that train AI experts should also teach skills unrelated to the technology, i.e. experts who develop, implement or decide on the use of AI systems should receive training in the legal, ethical and social aspects of the use of AI.

7. The federal government, universities, companies and civil society organisations should collaborate to promote social dialogue on the opportunities and risks of AI. In areas where the risks are unclear, research to identify such risks must also be stepped up with the help of appropriate measures from universities and third-party funding institutions.

In addition to these seven general recommendations, two specific recommendations were also formulated for each subject area examined. The recommendations made for the subject area **working world** express the fact that, in the macroeconomic field, the concrete effects caused by AI are difficult to distinguish from other factors resulting from the current digital transformation; overall, however, these effects increase the need for a debate on developing processes that will facilitate adaptation. In turn, the use of AI systems in the work process itself should be organised in such a way that workers' current rights (e.g. the right of employees to have a say) are not endangered. The corresponding recommendations are as follows:

- The Federal Government should take the potential macroeconomic effects of the digital transformation in general and of AI in particular as an opportunity to initiate social debate on processes to facilitate adaptation: this would include the reduction of working hours, flexibilisation of work in terms of time, place and means, economic polarisation and further training.

- Legislators and regulators should ensure that employees have a say when companies use AI systems to monitor and control them: labour inspectorates must be adequately equipped to effectively monitor compliance with the law.

Recommendations in the area of **education** (the recommendations on research primarily identify gaps in research: see below) state, firstly, that learners' data for the purpose of personalising the learning process is particularly worth protecting. Secondly, it should be ensured that the beneficial applications of AI in education are promoted and corresponding experiences shared. The corresponding recommendations are as follows:

- Cantonal legislators and regulators as well as ministers of education should formulate guidelines on how to deal with data gathered on the performance and behaviour of learners and the conclusions drawn from the data obtained from AI systems: in particular, assessments should be made on whether and/or which mechanisms should be created beyond current data protection regulations to protect learners from any negative consequences of the use of their learning and performance data and its disclosure to third parties.
- Educational institutions, and teacher training colleges in particular, should examine what specific skills need to be taught in order to gain a general understanding of the capabilities and limitations of AI systems: corresponding findings should be incorporated into teaching material and made available to teachers and learners via existing platforms.

The recommendations in the area of **consumption** are primarily aimed at seeing how the current data protection law can be supplemented so that consumers receive transparent information about the use of AI and at ensuring that competition is maintained even with data-intensive AI applications. The corresponding recommendations are as follows:

- Companies that use AI systems in the consumer sector and collect personal data for this purpose should communicate the transparency of their use of AI and their compliance with other data protection requirements in the simplest way possible: corresponding research and best practices should be promoted.
- Legislators should examine how data portability can be implemented with regard to AI systems, in order to make it easier for consumers to switch to another provider.

The recommendations in the subject area of **media** are aimed at raising awareness of the effect of personalisation on each individual and advocating social debate on fundamental issues such as freedom of opinion, dealing with fake news and the role of the government in protecting the democratic opinion-forming process from campaigns by illegitimate actors (e.g. third countries). The corresponding recommendations are as follows:

- The operators of media platforms should make it easy for their users to see how the personalisation of media content via AI has an influence on the content chosen for display.
- In collaboration with media companies and civil society actors, the federal government should boost public debate on how to deal with fake news, filter bubbles and echo chambers. Security authorities (the police, intelligence service and army) should – subject to parliamentary control – develop the capabilities required to identify more rapidly any systematic fake news campaigns that have the aim of political manipulation, and inform the public accordingly.

Finally, the recommendations in the area of **administration** specify the special requirements of the government's use of AI in sovereign acts. The corresponding recommendations are as follows:

- Public administration bodies should define the criteria needed to determine how a responsible use of AI by the government can be implemented in practice.
- Public administration bodies should ensure that data used in AI by the government is of sufficient quality.

The report finishes with a list of **research gaps** that need to be filled in order to facilitate the correct use of AI and to encourage the development of corresponding innovations. This not only includes technical aspects such as improving people's understanding of the new forms of machine learning (explainable AI): research is also needed in the legal sciences (e.g. the regulatory approach to autonomous systems), in psychology (e.g. what conditions do people need to trust AI), and in the human sciences (e.g. what does it mean to control an AI system?). In order to leverage the potential of AI systems in research and innovation, and to be better equipped to respond to the research gaps mentioned above, research institutions are advised to set up dedicated AI centres.

In conclusion, the report aims to help ‘demystify’ the AI debate by putting exaggerated fears and expectations into perspective with an objective analysis of the concrete opportunities and risks of a fundamentally promising technology. The team of authors hopes that this comprehensive presentation of the numerous ongoing developments in the field of AI will provide readers with important insights and food for thought.

Résumé

La transformation numérique, c'est-à-dire l'utilisation extensive des technologies d'information et de communication modernes dans tous les domaines de la société, est un des thèmes centraux de l'évaluation des choix technologiques aujourd'hui. Travail, formation, consommation, médias ou administration : quasiment tous les domaines de la vie courante sont impactés par les technologies numériques. L'une de ces technologies – l'intelligence artificielle (IA) – joue un rôle prépondérant à cet égard. Grâce aux progrès en matière d'apprentissage automatique et de puissance de calcul, et en raison de l'explosion de la quantité disponible de données générées par Internet et par les *smartphones*, l'IA peut être utilisée pour résoudre des problèmes sur lesquels les programmes informatiques conventionnels avaient buté jusqu'ici. Ces progrès ont entraîné le succès impressionnant des systèmes d'IA pour certaines tâches exigeantes et complexes comme la reconnaissance d'image, la traduction en langue naturelle ou les jeux. Les systèmes d'IA s'améliorent rapidement et ont aujourd'hui des applications jusqu'ici réservées aux êtres humains, comme par exemple la conduite de véhicules ou le diagnostic de maladies. L'IA est devenue une technologie fondamentale qui paraît incontournable en matière de progrès technologique.

La présente étude donne un aperçu complet des chances et des risques liés à cette technologie fondamentale. Elle met l'accent sur les applications de l'IA qui doivent être utilisées pour l'assistance ou l'automatisation de processus décisionnels dans les domaines sociétaux pertinents, c'est-à-dire dans les cinq domaines suivants :

- **Travail :** l'étude porte ici essentiellement sur les questions économiques et les répercussions de l'IA sur les processus de travail individuels.
- **Formation et recherche :** le recours à l'IA en termes de personnalisation de l'apprentissage et de promotion de l'innovation est ici au cœur du débat.
- **Consommation :** l'influence de l'IA sur la relation entre les entreprises et les clients finaux et sur les questions en matière de droit de la concurrence est examinée ici.

- **Médias :** l'accent est mis ici sur le rôle de l'IA dans les phénomènes de type *fake news* ou bulles de filtres et leur importance pour la formation des opinions politiques.
- **Administration et juridiction :** l'analyse porte ici sur les questions soulevées par l'utilisation de l'IA dans le cadre de l'action publique.

Des précisions d'ordre technique, ainsi que l'analyse et l'évaluation des questions éthiques et juridiques complètent les recherches dans ces cinq domaines. D'un point de vue méthodologique, l'étude s'appuie sur une combinaison de méthodes comprenant une analyse de littérature, des ateliers et un sondage en deux étapes mené auprès de plus de trois cents spécialistes. Les chances et les risques liés à l'IA ont été évalués au moyen d'enquêtes et d'ateliers. Ces travaux ont abouti à un ensemble de sept recommandations générales et de deux recommandations spécifiques pour chaque domaine examiné. L'étude a été réalisée par des chercheurs de la Digital Society Initiative de l'université de Zurich, du Technology and Society Lab de l'Empa à Saint-Gall et de l'Institut pour l'évaluation des choix technologiques de l'Académie autrichienne des sciences (Österreichische Akademie der Wissenschaften) à Vienne.

L'étude est confrontée à l'utilisation souvent floue du terme « intelligence artificielle » dans le discours public et à son imbrication avec d'autres tendances technologiques telles que l'Internet des objets ou la robotique. Par conséquent, il est parfois difficile d'identifier clairement le rôle de l'IA, notamment dans les changements macroéconomiques. Afin de remédier à ces difficultés, la présente étude fournit tout d'abord une introduction aux technologies de l'IA où les définitions suivantes ont été retenues :

L'« intelligence artificielle » désigne la tentative de reproduire la compréhension et l'apprentissage au moyen d'un artefact en se focalisant principalement sur la pensée ou l'action et en poursuivant un idéal rationnel ou une reproduction des aptitudes humaines.

La « technologie de l'IA » désigne des fonctions individuelles qui peuvent être implémentées dans les ordinateurs pour obtenir de l'intelligence artificielle (p. ex. apprentissage automatique).

Le « système d'IA » désigne une combinaison structurée et contextuelle de technologies de l'IA afin d'obtenir de l'intelligence artificielle.

Les « décisions d'IA » sont les conclusions des systèmes d'IA ayant des effets réels qui dépendent de décisions humaines au niveau de la conception du système, de la stratégie (décision sur le recours à un système) et de la tactique (interaction avec l'opératrice ou l'opérateur).

Les applications plus récentes de l'apprentissage automatique qui s'appuient sur des réseaux neuronaux (*deep learning*) sont les instigateurs technologiques de la discussion actuelle sur l'intelligence artificielle ; il s'agit de programmes complexes qui peuvent être « entraînés » grâce à de grandes quantités de données. Différents axes de discussion ont émergé autour de cette technologie ces dernières années, comme de savoir si les données utilisées pour l'« entraînement » masquent une certaine partialité ou de la subjectivité (*bias*), et dans quelle mesure. La procédure utilisée (*deep learning*) est confrontée au fait que la solution proposée par l'algorithme est difficile à comprendre (*black box*). Il s'agit de déterminer dans quelle mesure certaines valeurs et intérêts sont privilégiés par rapport à d'autres lors de la conception des systèmes d'IA. Des problèmes complexes peuvent se poser lorsque des objectifs s'annulent mutuellement (p. ex. différentes interprétations de l'équité *‹fairness›*). Tous ces éléments mènent à la question ultime de savoir dans quelle mesure il est possible de faire confiance aux systèmes d'IA lorsqu'ils sont utilisés pour assister ou remplacer le processus décisionnel humain.

Toutefois, en plus de ces problématiques au sujet de la technologie elle-même, les conséquences d'ordre sociétal font également l'objet de discussions. Les thèmes dominants à cet égard sont les répercussions socioéconomiques de l'IA (en particulier la perte redoutée de nombreuses places de travail), les questions en matière de droit de la concurrence en raison des énormes quantités de données nécessaires à l'élaboration de certains systèmes d'IA, et des réflexions d'ordre géostratégique car les grandes puissances comme la Chine et les États-Unis investissent des sommes importantes dans le développement des technologies de l'IA.

Le vaste éventail de thèmes présenté dans ce rapport se reflète dans le débat sur l'IA au niveau international et montre qu'il y a un énorme intérêt à mieux comprendre les chances et les risques liés à l'IA. Un moteur important de cet intérêt est la **question éthique**. Dans des domaines sensibles notamment, le recours croissant à l'IA éveille des craintes quant au manque de transparence, au renforcement de pratiques discriminantes et à l'affaiblissement de la responsabilité. L'intelligence artificielle incarne ainsi de manière exemplaire la peur d'une perte de

contrôle humain due à l'accélération du progrès technique. Cette introduction aux questions éthiques est complétée par une présentation détaillée des **questions juridiques** que pose l'utilisation de l'IA, notamment par les entreprises. La responsabilité civile, le droit de la propriété intellectuelle, le droit de la protection des données et l'interdiction de discrimination inscrite dans la Constitution y sont abordés.

Les recherches menées dans les cinq domaines indiqués se concentrent sur les applications actuelles ou envisageables dans un avenir proche ; les spéculations concernant une « IA superintelligente » – c'est-à-dire des systèmes dont les capacités dépassent de loin celles des êtres humains de manière générale et non plus seulement pour des tâches précises – n'ont pas été abordées dans le cadre de cette étude. Les analyses suivent toutes le même schéma : elles présentent dans un premier temps la littérature spécialisée actuelle afin de donner une vue d'ensemble des formes d'utilisation et des effets prévisibles de l'IA. Ceci comprend également une présentation des propositions les plus importantes amenées à un niveau international pour minimiser les risques et promouvoir les utilisations positives de l'IA. Dans un deuxième temps, une enquête en deux étapes a été menée auprès d'expertes et d'experts – spécialistes sur le plan technique ou sectoriel – pour obtenir des évaluations des différentes formes d'application de l'IA, des chances et risques qui y sont liés ainsi que des mesures potentielles à prendre. Ces résultats ont été consolidés dans le cadre de deux ateliers et intégrés aux recommandations ci-dessous.

Dans le domaine du **monde du travail**, il s'agit avant tout d'étudier la crainte largement répandue que les systèmes d'IA sont susceptibles de rendre le travail humain en grande partie obsolète, ce qui aurait d'énormes conséquences sur la société. L'analyse de la littérature indique toutefois qu'il est difficile de faire des pronostics précis en la matière. La référence à l'expérience historique et aux avancées de l'automatisation qui ont déjà eu lieu est également problématique ici car la comparabilité n'est pas établie. Les analyses portent par conséquent surtout sur des mesures censées contrer la polarisation du marché du travail (spécialisation croissante versus précarisation croissante des conditions de travail), ainsi que la polarisation des salaires qui en découle ou la réduction du taux d'occupation. En ce qui concerne l'impact de l'IA sur l'organisation du travail proprement dit, la sélection, la surveillance et la gestion de carrière de la main d'œuvre ont été examinées, de même que les effets indirects que l'IA a sur cette organisation. Entre autres constatations, l'importance de la formation continue en tant que mesure permettant de façonner de manière positive le virage numérique a été établie.

Dans le domaine de la **formation et** de la **recherche**, l'accent est mis sur la formation à l'école obligatoire. Les applications de l'IA visant à apporter un soutien à l'administration, aux apprenantes, aux apprenants et au corps enseignant ont été examinées. Les mesures envisagées comprennent notamment la définition des compétences en matière d'IA qui devraient être enseignées à l'école, ainsi que la protection des données issues des institutions de formation comme condition préalable à l'utilisation de l'IA. Dans le domaine de la recherche, des applications de l'IA en matière de promotion de l'innovation ont été analysées. Les mesures envisagées portent entre autres sur la question de savoir comment la recherche interdisciplinaire en matière d'IA peut être encouragée.

Dans le domaine de la **consommation**, les applications de l'IA pour la personnalisation (p. ex. des offres et des prix), les systèmes de recommandation et les assistants numériques ont été examinés. Outre les avantages pour les consommatrices, les consommateurs et les entreprises, les thèmes « reconnaissabilité de l'IA » et « l'IA comme *black box* pour les consommateurs » ainsi que les questions qui en découlent en termes de protection de la sphère privée et de confiance dans l'IA ont été abordées. Les réflexions en matière de droit de la concurrence concernent le monopole des données et les effets de réseau. Les mesures envisagées portent en particulier sur la promotion de la confiance des clients, la protection de la sphère privée et la prévention des oligopoles.

Dans le domaine des **médias**, le recours à l'IA dans un contexte où les comportements en matière d'information ont beaucoup évolué a été examiné d'autant plus près que la nature de la consommation des médias et le rôle des médias traditionnels comme la presse et la télévision ont considérablement changé au cours des dernières années. L'accent est mis ici sur ce qu'on appelle les bulles de filtres et les chambres d'écho ; celles-ci peuvent mener à une uniformisation de la consommation des contenus médias de la part d'individus ou de groupes entiers. Par ailleurs, le thème des *fake news* a pris de l'importance, notamment parce que les technologies de l'IA offrent de nouvelles possibilités pour falsifier des contenus audio et vidéo. Les mesures envisagées portent en particulier sur la protection du discours public contre l'influence manipulatrice non désirée ou délibérée des technologies de l'IA.

Enfin, dans le domaine de l'**administration et** de la **juridiction**, c'est le recours à l'IA par l'État qui est examiné, sous l'angle spécifique des actes d'autorité, notamment la publication de décisions avec l'aide de l'IA ou la prévision policière (*pre-*

dictive policing). L'analyse porte essentiellement sur la question de savoir comment le recours à l'IA peut impacter les exigences concernant l'action publique – p. ex. l'obligation de motivation, l'interdiction de discrimination ou le droit à être entendu – et quelles mesures peuvent être prises pour garantir que ces exigences continuent d'être remplies. Les éléments à risque qui ont été identifiés sont, entre autres, la mise en danger de la présomption d'innocence, les procédures opaques et la question de la soumission aux machines.

Ces analyses ont donné lieu à sept recommandations générales et dix recommandations spécifiques à l'intention du législateur, des autorités fédérales et d'autres acteurs concernés (p. ex. les entreprises ou le grand public). Elles ne se limitent pas aux aspects réglementaires. Elles identifient aussi les besoins en termes de débat et se réfèrent à la promotion de solutions techniques pour soutenir la législation existante. Enfin, elles comprennent des mesures visant à permettre aux différents acteurs de mieux s'acquitter de leurs tâches face aux défis posés par l'IA. Chaque recommandation est accompagnée d'une brève explication et d'une précision concernant la ou le destinataire. Les recommandations sont fondées sur les idées directrices suivantes :

- **Réglementation ciblée :** Bien que les systèmes d'IA modernes partagent souvent certaines caractéristiques, comme le fait que de grandes quantités de données sont nécessaires et que les explications sont lacunaires, la nature des risques et la possibilité de les influencer dépendent essentiellement des applications d'IA en question. Par conséquent, la nécessité d'une réglementation et sa conception devraient être examinées sur une base sectorielle ; les approches réglementaires d'ordre général, telles que la réglementation en matière de protection des données, sont insuffisantes et souvent aussi inadaptées pour gérer correctement les risques de l'IA.
- **Monitoring régulier et intégré :** La recherche dans le domaine de l'IA est très dynamique et il est tout à fait concevable que certaines des applications très appréciées aujourd'hui se révèlent être un échec, tandis que de nouvelles idées donnent lieu à des hypothèses encore insoupçonnées à ce jour. Par conséquent, les développements dans le domaine de l'IA doivent être suivis avec attention et le législateur doit être encouragé à réexaminer la nécessité d'une réglementation à intervalles réguliers en tenant compte de l'évolution au niveau international (notamment dans l'UE) et en contribuant à créer un dialogue public.

- **Promotion des qualifications :** L'une des idées directrices centrales pour aborder l'IA est la mise en œuvre, en fonction du contexte, de mesures qui permettent une qualification optimale des utilisatrices, utilisateurs et personnes concernées, de manière à maximiser les chances offertes par l'IA et à maîtriser les risques qu'elle présente. Cela inclut l'idée que le recours à l'IA par l'État doit être soumis à des exigences plus strictes dans le cadre des actes d'autorité, et qu'une transparence allant au-delà des dispositions de la loi sur la protection des données est nécessaire pour permettre de détecter une évolution indésirable. Les institutions de la société civile et les instances de réglementation doivent également être habilitées à vérifier les certifications privées d'IA. Enfin, les spécialistes qui élaborent, mettent en œuvre ou décident de l'utilisation de systèmes d'IA doivent acquérir des connaissances sur les aspects éthiques, juridiques et sociaux dans le cadre du recours à l'IA.

Sur la base de ces idées directrices, les sept recommandations générales suivantes sont formulées dans ce rapport (l'ordre donné ne reflète pas le degré de priorité) :

1. Le législateur doit adopter une approche à la fois neutre sur un plan technologique et spécifique à chaque domaine en matière d'IA : au lieu d'une « loi générale sur l'IA », il faudrait, en fonction des secteurs, identifier, évaluer et, si nécessaire, résoudre à intervalles réguliers les problèmes concrets et les développements indésirables au moyen de normes juridiques appropriées. En particulier, il convient également de tenir compte de l'évolution de la situation dans l'Union européenne.

2. Le législateur ne doit pas considérer l'utilisation des systèmes d'IA comme un problème de protection des données. La protection des données s'applique lorsque des données à caractère personnel sont utilisées pour le développement et l'application de l'IA. Toutefois, les risques découlant de l'utilisation de données factuelles ou la discrimination résultant du traitement des données exigent la conception et le développement d'approches nouvelles ou différentes.

3. Le législateur doit veiller à ce que les pouvoirs publics (p. ex. les tribunaux, la police et l'administration) soient soumis à des exigences plus strictes que le secteur privé lorsque l'IA est utilisée pour des actes d'autorité et qu'elle affecte les personnes d'une manière notable : dans ce cas, il faut toujours veiller à ce que les personnes concernées soient à même d'évaluer la légalité de l'action publique.

4. Si des acteurs du secteur privé (entreprises et autres organisations) utilisent l'IA pour prendre des décisions qui affectent les personnes de manière importante ils doivent non seulement assurer une communication active sur les circonstances du recours à l'IA, mais aussi faire preuve de transparence sur demande. Au-delà des dispositions de la loi sur la protection des données, les institutions de la société civile, par exemple, doivent avoir accès sur demande à toutes les informations permettant d'évaluer d'éventuels développements indésirables. À cet égard, la protection des secrets commerciaux des sociétés et des autres organisations doit être garantie de manière adéquate.

5. Les organisations (p. ex. dans le domaine de la protection des consommateurs) doivent être mieux à même de tester les certifications privées d'IA avec le soutien des pouvoirs publics. Les initiatives privées en faveur de certifications d'IA et d'attribution de labels appropriés doivent être saluées. Toutefois, l'utilisation de systèmes d'IA ne doit pas systématiquement dépendre de l'approbation du marché.

6. Les hautes écoles et autres institutions de formation qui forment des spécialistes de l'IA doivent également promouvoir les compétences non techniques : les spécialistes qui élaborent, mettent en œuvre ou décident du recours aux systèmes d'IA doivent acquérir des connaissances sur les aspects juridiques, éthiques et sociaux de l'utilisation de l'IA.

7. La Confédération, les hautes écoles, les entreprises et les organisations de la société civile doivent collaborer pour promouvoir le dialogue sociétal sur les chances et les risques de l'IA. Dans les domaines où le niveau de risque est incertain, la recherche visant à identifier ces risques doit également être intensifiée et soutenue par des mesures appropriées de la part des hautes écoles et des institutions de financement par des tiers.

En plus de ces sept recommandations générales, deux recommandations spécifiques ont été formulées pour chaque domaine examiné. Les recommandations dans le domaine du **monde du travail** expriment le fait que les effets concrets de l'IA sont difficiles à distinguer des autres facteurs du virage numérique dans le domaine macroéconomique, mais qu'ils augmentent dans l'ensemble le besoin d'ouvrir le débat sur les processus d'adaptation. En revanche, l'utilisation des systèmes d'IA dans le processus de travail lui-même doit être conçue de manière à ne pas entraver l'application des droits existants (p. ex. le droit à la participation des employées et employés). Les recommandations correspondantes sont les suivantes :

- La Confédération doit profiter davantage des effets macroéconomiques potentiels de la transformation numérique en général, et de l'IA en particulier, pour lancer des débats de société sur les processus d'adaptation : cela concerne en particulier le raccourcissement du temps de travail, l'assouplissement du travail en termes de temps, de lieu et de ressources, la polarisation économique et la formation continue.
- Le législateur doit veiller à ce que les employées et employés aient un droit de regard lorsque les entreprises utilisent des systèmes d'IA pour les surveiller et les contrôler : les organismes de l'inspection du travail doivent être équipés de manière adéquate pour contrôler efficacement le respect des dispositions légales.

Les recommandations dans le domaine de la **formation** (celles concernant la recherche désignent les lacunes en matière de recherche qui se profilent, voir ci-dessous) soulignent d'une part que les données des élèves destinées à la personnalisation de l'apprentissage doivent tout particulièrement être protégées ; et que, d'autre part, il faut veiller à ce que les utilisations positives de l'IA en matière de formation soient encouragées et que les expériences correspondantes soient échangées. Les recommandations correspondantes sont les suivantes :

- Le législateur cantonal et les directions de l'instruction publique doivent formuler des directives relatives à la manière de gérer les données sur les résultats et le comportement des apprenantes et apprenants, ainsi que les conclusions qui ont été tirées de ces données grâce à l'utilisation de systèmes d'IA. En particulier, il faut vérifier s'il convient de créer des mécanismes plus stricts que les dispositions actuelles de protection des données afin de protéger les apprenantes et apprenants des effets négatifs de l'utilisation et de la divulgation de leurs données à des tiers.
- Les institutions de formation, et en particulier les hautes écoles pédagogiques, doivent analyser quelles sont les compétences spécifiques qu'il faut enseigner pour acquérir une compréhension générale des capacités et des limites des systèmes d'IA : ces résultats doivent être incorporés dans le matériel pédagogique et mis à la disposition du corps enseignant, des apprenantes et des apprenants au moyen des plateformes existantes.

Les recommandations dans le domaine de la **consommation** visent principalement à compléter la législation actuelle sur la protection des données afin que les consommatrices et consommateurs soient informés de manière transparente sur

l'utilisation de l'IA et que la concurrence soit maintenue même avec des applications d'IA nécessitant de grandes quantités de données. Les recommandations correspondantes sont les suivantes :

- Les entreprises qui utilisent des systèmes d'IA dans le secteur de la consommation et collectent des données à caractère personnel à cette fin doivent fournir une information aussi simple que possible sur la transparence du recours à l'IA et sur les autres exigences en matière de protection des données : la recherche dans ce domaine et les meilleures pratiques (*best practice*s) doivent être encouragées.
- Le législateur doit examiner comment la portabilité des données peut être mise en œuvre dans le domaine des systèmes d'IA, en particulier pour faciliter le passage des consommatrices et consommateurs à un autre fournisseur.

Les recommandations dans le domaine des **médias** visent à sensibiliser l'opinion à l'impact sur l'individu de la personnalisation et à promouvoir un débat sociétal sur des questions fondamentales comme la liberté d'expression, le traitement des *fake news* et le rôle des pouvoirs publics en matière de protection du processus démocratique de formation de l'opinion contre les campagnes menées par des acteurs illégitimes (p. ex. pays tiers). Les recommandations correspondantes sont les suivantes :

- Les opérateurs de plateformes médiatiques doivent permettre à leurs utilisatrices et utilisateurs de reconnaître facilement comment la personnalisation du contenu des médias par le biais de l'IA influence le choix des contenus affichés.
- En collaboration avec les médias et les acteurs de la société civile, la Confédération doit intensifier le débat public sur la manière de traiter les *fake news*, les bulles de filtres et les chambres d'écho. Les autorités chargées de la sécurité (police, services de renseignement et armée) doivent, sous contrôle parlementaire, développer la capacité de détecter plus rapidement les campagnes de *fake news* systématiques dont le but est la manipulation politique ; elles doivent informer le grand public en conséquence.

Enfin, les recommandations dans le domaine de l'**administration** précisent quelles sont les exigences spécifiques à l'utilisation par l'État de l'IA dans le cadre des actes d'autorité. Les recommandations correspondantes sont les suivantes :

- L'administration publique doit définir les critères qui permettent de déterminer comment une utilisation responsable de l'IA par les pouvoirs publics peut être mise en œuvre de manière concrète.
- L'administration publique doit s'assurer que les données utilisées en cas d'utilisation de l'IA par les pouvoirs publics sont de qualité suffisante.

Le rapport se termine par une liste de **lacunes en matière de recherche** dont le comblement peut promouvoir une utilisation appropriée de l'IA et encourager les innovations correspondantes. Cela ne s'applique pas seulement aux aspects techniques comme une meilleure compréhension des nouvelles formes d'apprentissage automatique (IA explicable), mais aussi aux recherches qui sont également nécessaires, notamment en droit (p. ex. sur le traitement réglementaire des systèmes autonomes), en psychologie (p. ex. sur les conditions de confiance dans l'IA) ou en sciences humaines (p. ex. que signifie contrôler un système d'IA). Afin d'exploiter le potentiel des systèmes d'IA dans la recherche et l'innovation, il est recommandé aux institutions de recherche de créer des centres d'IA pour mieux combler les lacunes en matière de recherche mentionnées ci-dessus.

Enfin, il faut souligner que le rapport vise à contribuer à « démystifier » le débat sur l'IA en s'éloignant des craintes et des attentes exagérées pour analyser les chances et les risques concrets d'une technologie fondamentalement prometteuse. L'équipe des auteurs espère que la présentation détaillée des nombreux développements actuels dans le domaine de l'IA fournira aux lectrices et lecteurs des connaissances et des pistes de réflexion importantes.

Sintesi

La trasformazione digitale, ossia l'uso esteso delle moderne tecnologie d'informazione e di comunicazione in tutti gli ambiti della società, è un tema centrale nell'odierna valutazione delle conseguenze tecnologiche. In un panorama in cui le tecnologie digitali esercitano un'influenza crescente ormai su quasi tutti gli aspetti della vita – lavoro, istruzione, consumo, media, amministrazione – a una tecnologia in particolare viene riconosciuto un ruolo chiave: l'intelligenza artificiale (IA). Grazie ai progressi nel campo dell'apprendimento automatico e della potenza di calcolo, e in seguito all'enorme aumento della disponibilità di dati generato da Internet e dagli smartphone, è ora possibile ricorrere all'IA per risolvere problemi in cui sinora i programmi informatici convenzionali avevano fallito. Questi progressi hanno consentito ai sistemi di IA di ottenere risultati impressionanti in compiti impegnativi e molto complessi, come il riconoscimento delle immagini, la traduzione della lingua naturale e i giochi. I sistemi di IA stanno compiendo passi da gigante e si stanno affermando in applicazioni un tempo riservate agli esseri umani, come la conduzione dei veicoli e la diagnosi delle malattie. L'IA pare ormai diventata una tecnologia di base indispensabile al progresso tecnologico.

Il presente studio offre una prospettiva ad ampio raggio sulle opportunità e i rischi di questa tecnologia di base. L'attenzione si concentra sulle applicazioni dell'IA finalizzate a sostenere o automatizzare i processi decisionali in cinque aree sociali rilevanti considerate nella presente analisi:

- **lavoro:** ci si concentra su questioni politico-economiche e sugli effetti dell'IA sul processo lavorativo individuale
- **istruzione e ricerca:** si discute dell'uso dell'IA per personalizzare l'apprendimento e promuovere l'innovazione
- **consumo:** l'attenzione viene posta sull'influenza che l'IA esercita sul rapporto tra aziende e clienti finali e su questioni relative al diritto della concorrenza
- **media:** il focus riguarda il ruolo dell'IA in fenomeni come le *fake news* o le bolle di filtraggio (*filter bubble*) e sull'importanza di queste ultime nella formazione dell'opinione politica

- **amministrazione e giurisprudenza:** vengono analizzate le problematiche generate dall'uso dell'IA nell'intervento statale.

Le indagini in questi ambiti sono accompagnate da chiarimenti tecnici nonché dall'analisi e dalla valutazione di questioni etiche e giuridiche. Sul piano della metodologia, lo studio si fonda su un mix di metodi tratti da studi della letteratura in materia, affiancati da un workshop e da un sondaggio in due fasi che ha coinvolto più di trecento esperte/-i. Le opportunità e i rischi dell'IA sono stati vagliati utilizzando sondaggi e workshop. Da queste attività sono emerse nel complesso sette raccomandazioni generali e due raccomandazioni specifiche per ogni area esaminata. Lo studio è stato realizzato da ricercatrici e ricercatori della Digital Society Initiative/Università di Zurigo, del Technology and Society Lab/Empa di San Gallo e dell'Institute of Technology Assessment (ITA)/Accademia austriaca delle scienze di Vienna.

Lo studio ha dovuto affrontare la difficoltà dovuta all'uso poco chiaro del termine «intelligenza artificiale» nel discorso pubblico, spesso sovrapposto ad altri trend tecnologici come l'Internet of Things e la robotica. Risulta quindi complesso identificare chiaramente il ruolo dell'IA ad es. nei cambiamenti macroeconomici. Per far fronte a queste difficoltà, lo studio esordisce con una breve introduzione sulle tecnologie IA partendo dalle seguenti definizioni:

«Intelligenza artificiale»: descrive il tentativo di riprodurre la comprensione e l'apprendimento per mezzo di un artefatto; si concentra principalmente su pensiero e azione, cercando di ottenere un modello di comportamento razionale o di replicare abilità umane.

«Tecnologia di IA»: si riferisce alle singole funzioni che possono essere implementate nei computer per produrre un'intelligenza artificiale (ad esempio l'apprendimento automatico).

«Sistema di IA»: descrive una combinazione strutturata e legata al contesto di tecnologie di IA con lo scopo di produrre un'intelligenza artificiale.

«Decisioni di IA»: sono deduzioni dei sistemi di IA con effetti sul mondo reale che dipendono dalle decisioni umane a livello di progettazione del sistema, sul piano strategico (decisione sull'uso del sistema) e tattico (interazione con l'operatore).

A incentivare l'attuale discussione sull'IA sono, a livello tecnologico, le più recenti applicazioni dell'apprendimento automatico (*deep learning*) basate sulle reti neurali, ossia programmi complessi «addestrati» per mezzo di grandi quantità di dati.

Negli ultimi anni si sono sviluppati vari campi di discussione attorno a questa tecnologia. Per quanto concerne i dati utilizzati per l'addestramento, la questione verte sul livello di parzialità o distorsione (*bias*) che questi possono celare. Il metodo utilizzato (*deep learning*) presenta il problema che i risultati prodotti dagli algoritmi sono difficili da spiegare (*black box*). Nella progettazione dei sistemi di IA ci si chiede in che misura determinati valori e interessi vengano privilegiati rispetto ad altri. Ciò può far emergere problemi spinosi, dal momento che determinati obiettivi (ad es. divergenti interpretazioni di *fairness*) si escludono a vicenda. Tutti questi aspetti fanno sì che ci si chieda in che misura le persone possano fidarsi dei sistemi di IA utilizzati per supportare o sostituire il processo decisionale umano.

Oltre a queste aree problematiche che si concentrano sulla tecnologia in sé, il dibattito verte sulle conseguenze sociali. Tra i principali argomenti di discussione figurano gli effetti socio-economici dell'IA (in particolare il timore di un taglio drastico dei posti di lavoro), le questioni legate al diritto della concorrenza indotte dalla grande quantità di dati necessari a creare determinati sistemi di IA e considerazioni geostrategiche generate dai copiosi investimenti nello sviluppo di tecnologie di IA da parte di superpotenze come la Cina e gli Stati Uniti.

Questo ampio ventaglio di tematiche si riflette nella disputa internazionale sull'IA esposta nello studio. La trattazione evidenzia l'enorme interesse verso una maggiore comprensione delle opportunità e dei rischi dell'IA. A nutrire questo interesse sono soprattutto **questioni etiche**. Il crescente ricorso all'IA in settori sensibili suscita ad es. timori legati alla mancanza di trasparenza, al consolidamento di pratiche discriminatorie e alla confusione delle responsabilità. In questo l'IA è emblematica della paura di perdita del controllo umano dovuta all'accelerazione del progresso tecnologico. L'introduzione alle questioni etiche è integrata da un'ampia presentazione delle **questioni giuridiche** generate dall'uso dei sistemi di IA, ad es. da parte delle aziende. In questo contesto vengono discusse problematiche relative alla regolamentazione della responsabilità, al diritto dei beni immateriali, alla protezione dei dati e al divieto costituzionale di discriminazione.

Le indagini nelle cinque aree tematiche si concentrano sulle applicazioni di IA esistenti o previste per il prossimo futuro. Le speculazioni sull'«IA super intelligente» – ossia su sistemi dotati di abilità di gran lunga superiori a quelle umane non nello specifico, ma in generale **–** non rientrano nell'indagine. Le analisi seguono tutte il medesimo schema: in una prima fase riassumono la letteratura attuale per fornire una panoramica delle forme di utilizzo e degli effetti prevedibili dell'IA. Ciò include anche la presentazione delle principali proposte emerse a livello internazionale per

il contenimento dei rischi e la promozione delle applicazioni positive dell'IA. In una seconda fase, per mezzo di un sondaggio in due fasi condotto tra esperte/-i, ovvero persone con una formazione tecnica o settoriale specifica, si raccolgono valutazioni delle forme di utilizzo dell'IA, delle opportunità e dei rischi e, infine, dei possibili provvedimenti da adottare. Questi ultimi sono stati condensati nell'ambito di due seminari per poi confluire nelle raccomandazioni finali.

Nell'area tematica dedicata al mondo del **lavoro**, lo studio affronta innanzitutto il diffuso timore che i sistemi di IA possano rendere obsoleta gran parte della forza lavoro umana, con enormi ripercussioni sulla società. Dall'analisi della letteratura in materia emerge però la difficoltà di formulare previsioni precise. Risulta ugualmente problematico rifarsi alle esperienze storiche (precedenti balzi in avanti nell'automazione) perché non è chiaro se esista una base di comparabilità. Le analisi si concentrano dunque soprattutto sulle misure che dovrebbero contrastare la polarizzazione del mercato del lavoro (crescente specializzazione a fronte di condizioni di lavoro precarie), la divaricazione dei salari a ciò associata e la diminuzione del volume occupazionale. Per quanto riguarda gli effetti dell'IA sull'organizzazione del lavoro stesso, lo studio tratta la selezione della forza lavoro, il monitoraggio e la pianificazione della carriera dei lavoratori nonché gli effetti indiretti. Mette in evidenza, tra le altre cose, l'importanza del perfezionamento come strumento per la strutturazione positiva della svolta digitale.

Nel campo dell'**istruzione e** della **ricerca**, l'attenzione è rivolta innanzitutto alla scuola dell'obbligo. Sono state esaminate le applicazioni di IA a supporto dell'amministrazione, delle/-i discenti e delle/-gli insegnanti. Le misure discusse includono, tra l'altro, la presentazione delle competenze in IA che dovrebbero essere apprese a scuola e la protezione dei dati nel sistema educativo, entrambe prerequisiti per l'uso dell'IA. Nel campo della ricerca lo studio analizza ad es. applicazioni di IA finalizzate a promuovere l'innovazione. Le misure discusse si riferiscono anche alle modalità di finanziamento della ricerca interdisciplinare sull'IA.

Nel campo tematico del **consumo** lo studio esamina in particolare le applicazioni di IA per la personalizzazione (ad esempio di offerte e prezzi), i sistemi di raccomandazione e gli assistenti digitali. Oltre ai vantaggi per le/i consumatrici/-tori e le aziende, affronta le questioni della «riconoscibilità dell'IA» e dell'«IA come *black box* per i consumatori» con i conseguenti dubbi sulla protezione della privacy e sulla fiducia nell'IA. Le considerazioni sul diritto della concorrenza riguardano il

monopolio dei dati e gli effetti del network. Le misure discusse vertono in particolare sulla promozione della fiducia dei clienti, la tutela della privacy e la prevenzione di oligopoli.

Nell'area tematica dedicata ai **media** lo studio esamina l'uso dell'IA nel contesto di un comportamento informativo mutevole, visto che negli ultimi anni la modalità di consumo dei media e il ruolo di canali tradizionali come la stampa e la TV sono drasticamente cambiati. L'attenzione in questo ambito si concentra da un lato sulle cosiddette bolle di filtraggio e camere d'eco, che possono portare all'uniformazione contenutistica nel consumo mediatico dei singoli o di interi gruppi. Dall'altro lato si è esaminato il tema delle *fake news*, anche perché le tecnologie di IA hanno schiuso nuove possibilità di falsificazione dei contenuti audio e video. Le misure discusse si concentrano sulla tutela del discorso pubblico da influssi manipolativi volontari o involontari determinati dalle tecnologie di IA.

Infine, nell'area tematica di **amministrazione e giurisprudenza** viene analizzato l'uso dell'IA da parte dello stato, soprattutto nell'esercizio della sua sovranità – ad esempio nell'emissione di provvedimenti basati sull'IA o nel lavoro della polizia predittiva (*predictive policing*). Al centro dell'analisi si pone la questione di come requisiti imprescindibili dell'intervento statale – ad es. l'obbligo di motivazione, il divieto di discriminazione o il diritto al contraddittorio – vengano influenzati dall'uso dell'IA e quali misure si possano adottare per continuare a garantirne il soddisfacimento. Tra i punti critici identificati figurano la messa a repentaglio della presunzione di innocenza, procedure non trasparenti e la sottomissione alla tecnologia.

Da queste analisi sono derivate sette raccomandazioni generali e dieci raccomandazioni specifiche per area destinate al legislatore, alle autorità federali e ad altre parti interessate (ad es. aziende e pubblico in generale). Le raccomandazioni non si limitano agli aspetti normativi. Riguardano anche l'esigenza di discutere e promuovere soluzioni tecniche a sostegno della legislazione esistente. Infine includono misure su come consentire ai singoli attori di svolgere meglio i propri compiti facendo fronte anche alle sfide dell'IA. Di ciascuna raccomandazione viene fornita una breve spiegazione integrata dai destinatari cui è rivolta. Le raccomandazioni si ispirano ai seguenti criteri di base:

- **regolamentazione mirata:** sebbene i moderni sistemi di IA condividano spesso determinate caratteristiche, come la necessità di grandi quantità di dati e la mancanza di spiegabilità, la natura dei rischi e la possibilità di influenzarli dipendono in modo determinate dalle applicazioni di IA che detti sistemi utiliz-

zano. Di conseguenza la necessità di regolamentarli (e, nel caso, in che termini) va esaminata nello specifico; approcci regolamentativi generici, come ad esempio il disciplinamento ai fini della protezione dei dati, sono insufficienti e spesso anche inadatti alla corretta gestione dei rischi legati all'IA

- **monitoraggio regolare e integrativo:** la ricerca nel campo dell'IA è molto dinamica ed è senz'altro ipotizzabile che alcune applicazioni oggi molto apprezzate si rivelino un buco nell'acqua, mentre nuove idee mettano in luce problemi inattesi. Di conseguenza gli sviluppi dell'IA vanno monitorati con attenzione e il legislatore è tenuto a rivedere periodicamente l'esigenza di regolamentazione, tenendo conto delle evoluzioni internazionali (specialmente nell'UE) e partecipando al pubblico scambio di opinioni
- **promozione di competenze:** un principio guida centrale nell'affrontare l'IA è l'implementazione specifica di misure che consentano di abilitare al meglio utenti e interessati, così da massimizzare le opportunità dell'IA e al contempo contenerne i rischi. Ciò comprende la convinzione che l'uso dell'IA nell'esercizio della sovranità statale debba essere soggetto a requisiti più rigorosi e che la trasparenza debba consentire di individuare sviluppi indesiderati, più di quanto già previsto dalla legge sulla privacy. Inoltre le istituzioni della società civile e le istanze regolatrici dovrebbero essere autorizzate a verificare le certificazioni IA private. Le esperte e gli esperti che sviluppano e implementano sistemi di IA o ne decidono l'uso dovrebbero infine acquisire conoscenze sugli aspetti etici, legali e sociali dell'utilizzo dell'IA.

Sulla base di questi principi guida, lo studio formula le seguenti sette raccomandazioni generali (l'ordine non esprime una priorità):

1. nel campo dell'IA il legislatore deve perseguire un approccio neutro sul piano tecnologico e settoriale: invece di promulgare una generica «legge sull'intelligenza artificiale», sarebbe opportuno individuare periodicamente, valutare e, se del caso, risolvere con interventi normativi problemi e sviluppi scorretti campo per campo, tenendo conto anche delle evoluzioni in seno all'Unione Europea
2. il legislatore non deve interpretare l'uso dei sistemi di IA come un problema di protezione dei dati. È vero che entra in gioco la protezione dei dati, laddove vengono utilizzati dati personali per lo sviluppo e l'applicazione dell'IA; tuttavia i rischi derivanti dall'uso di dati materiali o da discriminazione in conseguenza

al trattamento dei dati richiedono approcci nuovi o diversi che vanno ideati e perfezionati

3. le autorità legislative e regolamentative devono garantire che i soggetti statali (ad es. tribunali, polizia e amministrazione), laddove utilizzo dell'IA nell'esercizio della sovranità abbia ripercussioni di rilievo sulle persone, siano tenuti a soddisfare requisiti più rigorosi rispetto ai privati: in questi casi le persone interessate devono sempre essere messe nelle condizioni di valutare la liceità dell'intervento statale
4. se dei privati (aziende e altre organizzazioni) utilizzano l'IA per decisioni che esercitano un'influenza rilevante sulle persone, oltre a fornire spontaneamente informazioni sulle circostanze di utilizzo dell'IA, su richiesta devono garantire anche la trasparenza. In aggiunta a quanto già previsto dalla legge sulla protezione dei dati, ad es. le istituzioni della società civile devono poter ricevere a richiesta tutte le informazioni atte a stimare potenziali sviluppi indesiderati. Ciò detto, va contemporaneamente garantita la protezione adeguata dei segreti commerciali di aziende e altre organizzazioni
5. le organizzazioni (ad es. nel settore della tutela dei consumatori) devono essere messe in grado, grazie al sostegno statale, di controllare meglio le certificazioni private di IA. Ben vengano le iniziative private per la certificazione di IA e l'attribuzione dei relativi label; in generale però l'uso dei sistemi di IA non deve essere subordinato all'autorizzazione all'immissione sul mercato
6. università e altri istituti di istruzione per la formazione di specialiste/-i in IA devono fornire anche competenze non tecniche: le esperte e gli esperti che sviluppano e implementano sistemi di IA o ne decidono l'uso dovrebbero acquisire conoscenze sugli aspetti etici, legali e sociali dell'impiego dell'IA
7. la Confederazione, le università, le aziende e le organizzazioni della società civile devono promuovere congiuntamente il dialogo sociale sulle opportunità e sui rischi dell'IA. In settori caratterizzati da una situazione di rischio poco chiara, occorre intensificare anche la ricerca per identificare i pericoli, da supportare con misure appropriate da parte di università e istituti di finanziamento terzi.

Oltre a queste sette raccomandazioni generali, ne sono state formulate due specifiche per ciascuna delle aree esaminate. Le raccomandazioni relative all'area **lavoro** evidenziano come, sebbene nell'ambito della macroeconomia sia difficile

distinguere gli effetti concreti dell'IA da altri fattori di cambiamento digitale, tali effetti incrementino nel complesso la necessità di un dibattito sui processi di adattamento. D'altro canto l'uso dei sistemi di IA nel processo di lavoro stesso dovrebbe essere studiato in modo da non ostacolare il rispetto dei diritti vigenti (ad es. il diritto delle/dei collaboratrici/-tori di esprimere la propria opinione). Le relative raccomandazioni sono:

- la Confederazione deve cogliere l'occasione dei possibili impatti macroeconomici della trasformazione digitale in generale e dell'IA in particolare per avviare dibattiti sociali sui processi di adattamento. Ciò vale in particolare per quanto concerne la riduzione dell'orario di lavoro, la flessibilizzazione del lavoro dal punto di vista di orario, luogo e mezzi, la polarizzazione economica e la formazione continua
- le autorità legislative e regolamentative devono garantire al personale dipendente di poter esprimere la propria opinione nei casi in cui le aziende utilizzino sistemi di IA per monitorarlo e controllarlo. Gli ispettorati del lavoro devono essere attrezzati in modo adeguato per poter sorvegliare con efficacia il rispetto delle disposizioni di legge.

Le raccomandazioni nel campo dell'**istruzione** (quelle relative alla ricerca menzionano in apertura determinate lacune; vedere sotto) evidenziano da un lato che vanno particolarmente tutelati i dati di studentesse e studenti utilizzati per personalizzare l'apprendimento, e dall'altro che bisogna garantire che l'istruzione promuova gli usi positivi dell'IA e lo scambio di esperienze in merito. Segue il testo delle raccomandazioni:

- le autorità legislative e regolamentative cantonali e i direttori dell'educazione devono formulare linee guida sulle modalità di gestione dei dati sulle prestazioni e la condotta di studentesse e studenti nonché sulle conclusioni che se ne traggono con i sistemi di IA. In particolare va verificato se e quali meccanismi debbano essere messi in atto, oltre all'attuale sistema di protezione dei dati, per tutelare studentesse e studenti dalle conseguenze negative dell'uso e della divulgazione a terzi dei loro dati di apprendimento e rendimento
- gli istituti di formazione e in particolare le scuole superiori a indirizzo pedagogico devono esaminare quali competenze specifiche fornire per favorire la comprensione generale delle possibilità e dei limiti dei sistemi di IA: le relative nozioni vanno integrate nei materiali didattici e rese disponibili a insegnanti e studentesse/studenti attraverso le piattaforme esistenti.

Le raccomandazioni in materia di **consumo** mirano principalmente a integrare l'attuale normativa sulla protezione dei dati per informare con la massima trasparenza le consumatrici e i consumatori sull'uso dell'IA e preservare la concorrenza anche nelle applicazioni di IA ad alta intensità di dati. Segue il testo delle raccomandazioni:

- le aziende che utilizzano i sistemi di IA nel settore del **consumo** e raccolgono dati personali a tale scopo devono comunicare nel modo più semplice possibile la trasparenza nell'utilizzo dell'IA e gli altri requisiti ai fini della protezione dei dati. Vanno promosse la ricerca e le best practice appropriate
- il legislatore deve indagare su come implementare la portabilità dei dati nel campo dei sistemi di IA, soprattutto per agevolare alle consumatrici e ai consumatori il passaggio a un altro fornitore.

Le raccomandazioni nel settore dei **media** mirano ad aumentare la consapevolezza dell'effetto che la personalizzazione ha sull'individuo e a promuovere un dibattito sociale su questioni fondamentali come la libertà di espressione, la gestione delle *fake news* e il ruolo dello stato nella tutela da campagne illecite (ad esempio ad opera di paesi terzi) del processo democratico di formazione dell'opinione pubblica. Segue il testo delle raccomandazioni:

- i gestori di piattaforme mediali devono rendere facilmente riconoscibile alla propria utenza in che modo la personalizzazione dei contenuti mediali effettuata dall'IA influenzi la selezione dei contenuti offerti
- in collaborazione con le imprese mediali e gli organi della società civile, la Confederazione deve intensificare la discussione sociale sulla gestione delle *fake news*, delle bolle di filtraggio e delle casse di risonanza. Le autorità preposte alla sicurezza (polizia, servizi segreti ed esercito) devono sviluppare – sotto il controllo del Parlamento – le competenze necessarie a identificare più rapidamente le campagne sistematiche di *fake news* che puntano alla manipolazione politica e a informare il pubblico in merito.

Infine, le raccomandazioni nel settore dell'**amministrazione** descrivono in concreto i requisiti peculiari dell'uso governativo dell'IA nell'esercizio della sovranità. Segue il testo delle raccomandazioni:

- la pubblica amministrazione deve definire criteri atti a determinare le modalità concrete di implementazione dell'uso responsabile dell'IA da parte dello stato

- la pubblica amministrazione deve garantire la qualità dei dati destinati all'uso dell'IA da parte dello stato.

Lo studio si conclude con un elenco di **lacune della ricerca**, la cui rimozione può favorire la corretta gestione dell'IA e delle relative innovazioni. Non interessano solo gli aspetti tecnici, come una migliore comprensione delle nuove forme di apprendimento automatico (IA spiegabile). La ricerca si rende necessaria anche nel campo delle scienze giuridiche (ad es. per gestire la regolamentazione di sistemi autonomi), in psicologia (ad es. condizioni per la fiducia nell'IA) e nelle scienze umane (ad es. cosa significa controllare un sistema di IA). Per sfruttare il potenziale dei sistemi di IA nella ricerca e nell'innovazione, si raccomanda agli istituti di ricerca di istituire centri dedicati all'IA per colmare meglio le lacune della ricerca qui menzionate.

Concludendo si sottolinea che lo studio mira a contribuire alla «demistificazione» del dibattito sull'IA, allontanandosi da paure e aspettative eccessive per avvicinarsi a un'analisi delle opportunità e dei rischi concreti di una tecnologia in sé promettente. Il team redazionale si augura che l'ampia presentazione dei numerosi sviluppi attuali nel campo dell'IA possa fornire alle lettrici e ai lettori prospettive e spunti di riflessione importanti.

1. Einleitung

In der Einleitung werden Auftrag, Zielsetzung und Methodik der Studie dargestellt. Es wird insbesondere ausgeführt, welche fünf Anwendungsbereiche von künstlicher Intelligenz (KI) genauer untersucht werden und wie sich die Studie von anderen Themen der digitalen Transformation abgrenzt. Vorgestellt wird auch das Team, welches die Studie umgesetzt hat.

1.1. Projektauftrag und Zielsetzung

1.1.1. Das Grundproblem

Die zunehmende Nutzung künstlicher Intelligenz (KI) ist ein Hauptmerkmal des digitalen Wandels, der die moderne Gesellschaft tiefgreifend verändert. Zwar ist die Anwendung von Informationstechnologie in der Gesellschaft nicht neu, und daraus resultierende Phänomene wie beispielsweise die Automatisierung von Produktionsprozessen werden seit Jahrzehnten untersucht (Henning et al. 1990; Verl et al. 1998). Auch künstliche Intelligenz ist bereits Ende der 1980er-Jahre in den Fokus der Technikfolgenabschätzung gelangt (Baggi 1989; Lebsanft & Gill 1987; Bürgi-Schmelz 1990). Die Besonderheit der gegenwärtigen Erfolge künstlicher Intelligenz resultiert allerdings aus der Kombination von Fortschritten im Bereich des maschinellen Lernens (ML), der Rechenleistung und der enorm gestiegenen Datenverfügbarkeit. Die dadurch ermöglichten KI-Systeme erzielen Lösungen für Probleme, bei denen herkömmliche Computerprogramme bislang gescheitert sind. Diese Fortschritte haben dazu geführt, dass KI-Systeme innert weniger Jahre beeindruckende Erfolge bei anspruchsvollen und mehrdeutigen Aufgaben wie Bilderkennung, Übersetzung natürlicher Sprache oder Spielen von Regelspielen erzielen konnten (siehe Abschnitt 2.2). Sie konkurrieren dabei nicht nur mit menschlichen Fähigkeiten, sondern übertreffen diese zuweilen auch. Diese Technologien verbessern sich rasant und führen zu Anwendungen, die zuvor nur Menschen vorbehalten waren wie z.B. das Führen von Fahrzeugen oder die Diagnose von Krankheiten. KI wird damit zu einer Basistechnologie (Jovanovic & Rousseau 2005) für eine grosse Bandbreite von Anwendungen.

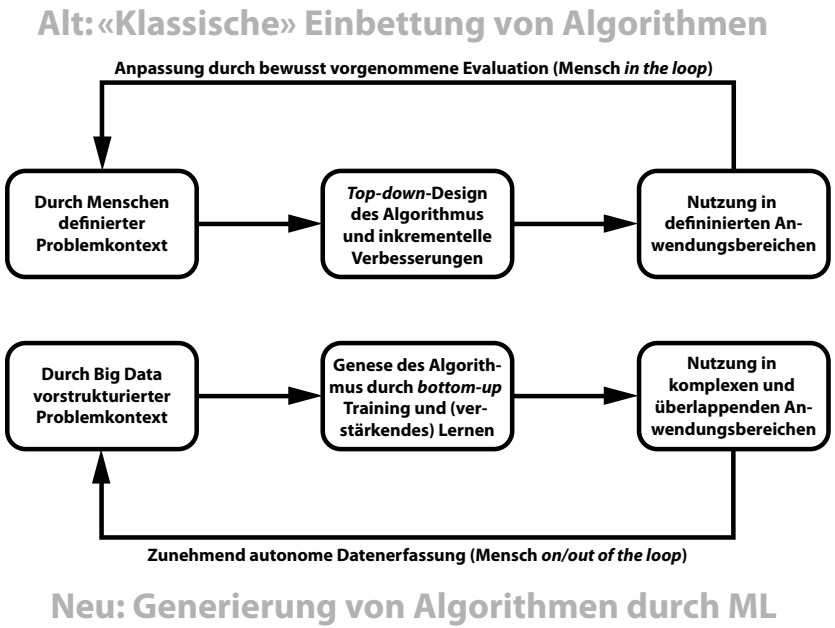

Abbildung 1: Idealtypisch beschriebene Veränderung der gesellschaftlichen Einbettung von Algorithmen mittels Nutzung neuer Formen von KI.

Der damit einhergehende Wandel ist tiefgreifend, weil sich die Art der Einbettung von Algorithmen in gesellschaftliche Systeme fundamental verändert, wie Abbildung 1 vereinfachend illustriert. Bislang definierten Menschen die algorithmisch zu bewältigenden Probleme explizit, kreierten die dafür nötigen Programme *top-down* und wendeten die Algorithmen in klar unterscheidbaren Bereichen an. Neu bilden grosse Mengen heterogener Daten die (oft nur unvollständig verstandene) Basis des Problemkontexts, maschinelles Lernen generiert daraus *bottom-up* den Algorithmus, der dann in zunehmend komplexen Anwendungsgebieten zum Einsatz kommt. Dabei agieren die Systeme vermehrt autonom und tauschen miteinander Informationen aus. Damit entzieht sich die «Rückkopplung» zwischen Problemlösung durch einen Algorithmus in einem definierten Anwendungsbereich und der daraus folgenden Anpassungen zunehmend der menschlichen Kontrolle. Daraus resultiert ein schleichender Wandel weg von der Entscheidungs*unterstützung* durch Algorithmen hin zur *Automatisierung* von Entscheidungen in Lebensbereichen wie z.B. Mobilität, Kreditvergabe, Auswahlverfahren bei Stellenbewerbungen

oder juristische Prüfungen. Nicht zuletzt deshalb wird die öffentliche Diskussion der Anwendung von KI oft von dystopischen Zukunftsszenarien dominiert.

Aus dieser Darstellung des Grundproblems werden sowohl Notwendigkeit wie auch Herausforderungen einer Beurteilung der künstlichen Intelligenz aus der Perspektive der Technikfolgenabschätzung ersichtlich. Dass ein derart tiefgreifender Wandel in der Art der Nutzung von Computertechnologie enorme gesellschaftliche Folgen hat, erscheint klar. Gleichzeitig wird deutlich, dass KI in eine ganze Reihe kontrovers diskutierter Themen rund um den digitalen Wandel eingebettet ist. So ist der Problemkomplex «Big Data» zu nennen, also die enorm gewachsene Verfügbarkeit von Daten, die den zentralen «Rohstoff» für das ML bilden (Flach 2012). Im Weiteren ist KI die zentrale Softwarekomponente für die Nutzung von Robotiksystemen in komplexen Anwendungen (Murphy 2000), wobei selbstfahrende Autos oder autonome Waffen prominent diskutierte Beispiele sind. Schliesslich basiert ein signifikanter Teil der zunehmenden Autonomie von KI-Systemen auf der Kommunikation zwischen Maschinen (Greengard 2015). Diese Problemkomplexe werfen teils eigene, von KI abgrenzbare Fragen auf; oft aber bestehen Überlappungen. Zudem ist – insbesondere im öffentlichen Diskurs – eine «Mystifizierung» des Themas ersichtlich (Kelly 2017), gepaart mit Wissenslücken hinsichtlich der tatsächlichen Möglichkeiten und Grenzen von KI-Systemen. Daher werden zahlreiche Aspekte des digitalen Wandels fälschlicherweise mit KI in Verbindung gebracht und die realen Fähigkeiten dieser Systeme oft überhöht.

Das in dieser Studie behandelte Grundproblem lässt sich demnach wie folgt auf den Punkt bringen:

> *Welche Folgen sind aus der Nutzung der neueren Formen von KI für die Unterstützung oder gar Automatisierung von Entscheidungsprozessen in relevanten gesellschaftlichen Bereichen zu erwarten und wie sind diese Veränderungen zu bewerten?*

Bei der Beantwortung dieser Frage ist darauf zu achten, dass erstens ein *realistischer Begriff* von künstlicher Intelligenz verwendet wird, zweitens definiert wird, was eine auf KI-Systemen basierte Entscheidung (im Folgenden kurz «KI-Entscheidung») ist und drittens *Abgrenzungen* zu anderen Aspekten des digitalen Wandels – soweit möglich – gezogen werden.

1.1.2. Ausdifferenzierung des Grundproblems

Die ethische, rechtliche und soziale Debatte rund um KI hat in den letzten Jahren einen enormen Aufschwung erlebt (siehe dazu auch 2.1). Grund dafür ist, dass KI eine Basistechnologie für die Automatisierung zahlreicher Prozesse ist. Der technologische Fortschritt, die Ausbildung qualifizierter Praktiker/-innen und der Wettbewerbsdruck beschleunigen die Verbreitung von KI. Entsprechend haben sich in den letzten Jahren verschiedene Diskurse entwickelt, die nachfolgend kurz dargestellt werden, weil sie Orientierungspunkte für diese Studie bilden:

- **Das Bias-Problem:** Daten bilden die zentrale Ressource für ML. Je nach Art des Entscheidungsproblems können sich aber in den Daten Einseitigkeiten oder Befangenheiten verbergen, die dann das Verhalten des Algorithmus prägen (Caliskan et al. 2017). Lernende KI-Systeme können deshalb durch entsprechende Lerndaten manipuliert und in die Irre geleitet werden. Das Bias-Problem ist relevant, weil Nutzerinnen und Nutzer oft nicht in der Lage sind, versteckte Einseitigkeiten in Trainings-Datensätzen zu identifizieren, die aus Millionen Einheiten bestehen.
- **Das Blackbox-Problem:** Im Unterschied zu «klassischen» Computeralgorithmen nutzen neuere ML-Technologien – insbesondere die sogenannten *deep neural networks* (siehe Abschnitt 2.2.4.2) – andere Programmiertechniken. Anstelle von expliziten Programmstrukturen, die zumindest prinzipiell nachvollziehbar sind, wird ein neuronales Netz zwar vorgegeben, dessen Konnektivität und Gewichtung der Verbindungen verändern sich aber über viele Trainingszyklen (ein Bilderkennungsalgorithmus wird z.B. mit Millionen von Bildern trainiert). Am Schluss ist selbst den Entwicklerinnen und Entwicklern nicht klar, wie der Algorithmus zur Lösung kommt, denn solche ML-Modelle sind Gleichungen, die keine offensichtliche physikalische oder logische Basis haben. Wie in Abschnitt 2.2.4.2 ausgeführt wird, erscheinen solche KI-Algorithmen als Blackbox, was für praktische Anwendungen von KI eine signifikante Einschränkung ist, wenn man verstehen will, wie ein System zu einer Entscheidung kommt (Pasquale 2015).
- **Das Fairness-Problem:** Nicht nur die Daten, sondern auch die Algorithmen selbst können implizite normative Annahmen enthalten. Wichtige Parameter werden vorab festgelegt und danach absichtlich oder auch unabsichtlich so konfiguriert, dass bestimmte Werte und Interessen gegenüber anderen privilegiert werden. Dies ist relevant, wenn KI-Systeme z.B. zur Beurteilung von

Personen im Strafvollzug eingesetzt werden (siehe dazu Abschnitt 3.5.2.2). Das Problem der Fairness von Algorithmen ist komplex, denn gegebene rechtliche Normen müssen in eine für Computerprogramme verständliche «Sprache» übersetzt werden. Mathematische Überlegungen zeigen zudem, dass sich gewisse Anforderungen an Algorithmen (z.B. bezüglich Genauigkeit und Fairness) nicht gleichzeitig erreichen lassen.

- **Das Vertrauens-Problem:** Die oben angesprochenen Probleme bezüglich Daten-Bias, Blackbox und algorithmischer Fairness kumulieren sich zur Frage, inwieweit Menschen, die KI-Systeme nutzen, diesen auch vertrauen (bzw. auf welcher Grundlage sie ihnen vertrauen oder misstrauen könnten). Die Erkenntnislage ist diesbezüglich uneinheitlich. Zum einen findet sich Evidenz, dass Menschen dazu neigen, den Ergebnissen automatisierter Entscheidungsfindung zu stark zu vertrauen, weil sie ein KI-Ergebnis als «objektiver» ansehen als jene eines Menschen (Jago & Laurin 2017). Andere Studien finden den gegenteiligen Effekt (Algorithmus-Aversion), wonach Menschen selbst dann einer menschlichen Entscheidung mehr vertrauen, wenn sie wissen, dass die «KI-Entscheidung» tatsächlich objektiver ist (Dietvorst et al. 2015). Diese Studien verweisen auf ein komplexes Vertrauensproblem hin, wenn Menschen sich zunehmend auf automatisierte Entscheidungen abstützen.

Diese vier Problemgruppen treten jeweils im Kontext einer konkreten KI-Anwendung auf. Es wird deshalb vom Einzelfall abhängen, wie sie beurteilt werden. Daneben finden aber auch Debatten auf gesellschaftlicher Ebene statt, die mit der Technologie als solcher zu tun haben und nur teilweise mit den oben genannten Problemgruppen zusammenhängen. Zu nennen sind hier folgende:

- **Ökonomische Auswirkungen:** Die ökonomischen Konsequenzen des digitalen Wandels nehmen im gesellschaftlichen Diskurs den grössten Platz ein – und der KI kommt hier angesichts des enorm breiten Anwendungspotenzials eine Schlüsselrolle zu. Denn im Unterschied zu bisherigen Automatisierungsschüben werden nun auch potenziell Tätigkeiten ersetzbar, die bislang eindeutig Menschen vorbehalten waren. So prognostizieren manche Studien, dass bis zu 50 % aller Berufe in den kommenden 20 Jahren automatisiert werden könnten und auch hoch qualifizierte Arbeit davon nicht verschont bliebe (siehe Abschnitt 3.1). Auch wenn solche Studien und damit das Ausmass des Verlusts an Arbeitsplätzen stark umstritten sowie das Potenzial zur Schaffung neuer Stellen unklar sind, dürfte kaum ein Berufsfeld von KI-Anwendungen

unberührt bleiben; Branchen wie Medien, Musik oder die Finanzwirtschaft haben bereits einen enormen Wandel erfahren. Es zeigen sich auch makroökonomische Effekte wie die Umkehrung von Globalisierungstrends (Rückkehr der automatisierten Industrieproduktion in die ursprünglichen Industrieländer; Mankiw 2016) oder verstärkte sozioökonomische Ungleichheit (Brynjolfsson & McAfee 2014). Diesbezüglich dürfte aber eine klare Zuordnung solcher Phänomene zu Entwicklungen im Bereich KI schwierig sein.

- **Das «Oligopol-Problem»:** Einen anderen ökonomischen Problemkomplex betrifft die relevanten Player in der Forschung und Entwicklung von KI-Systemen. Da die neuen Formen des ML stark datenbasiert sind, haben Unternehmen mit Zugriff auf enorm grosse Datensätze einen kompetitiven Vorteil (siehe Abschnitt 3.3.3.4). Es sind denn auch führende Technologieunternehmen aus China und den USA wie Alibaba, Amazon, Baidu, Facebook, Google und Microsoft, die ihre internen Geschäftsabläufe und Produkte rund um KI neu konzipieren. Das in der Internet-Ökonomie bekannte Prinzip «the winner takes it all» und der damit verbundenen Gefahr der Monopolbildung dürfte sich angesichts der grossen Ressourcen, welche die Entwicklung erfolgreicher KI-Systeme benötigt, noch verschärfen (Pemberton Levy 2016).
- **Geostrategische Fragen:** Ein weiterer Punkt betrifft geostrategische Fragen. China hat KI als zentrales Element für das strategische Ziel definiert, eine globale Führungsrolle in der Entwicklung neuer Technologien einzunehmen (Scott et al. 2017). Gleichzeitig ist KI ein mächtiges Instrument für die Unterstützung totalitärer Bestrebungen wie z.B. Massenüberwachung der Bevölkerung oder «Big Nudging». Dabei stellt sich die Frage, inwieweit die nationale Anwendung von KI-Technologien, die in Gesellschaften mit abweichenden sozialen Normen und demokratischen Traditionen entwickelt werden, ethische oder politische Probleme aufwirft. Auch militärische Nutzungen von KI fallen in diesen Themenkomplex, und Wissenschaftler warnen jetzt schon vor einem «KI-Wettrüsten» (Altmann & Sauer 2017).

Abbildung 2 zeigt die Ausdifferenzierung des Grundproblems in die gängigen Diskussionsfelder über die gesellschaftlichen Auswirkungen von KI, die für diese Arbeit leitend sind. Nachfolgend wird nun aber zuerst die Thematik der Studie eingegrenzt, und es wird in Anlehnung an den Auftrag von TA-SWISS verdeutlicht, welche Zielgruppen fokussiert werden.

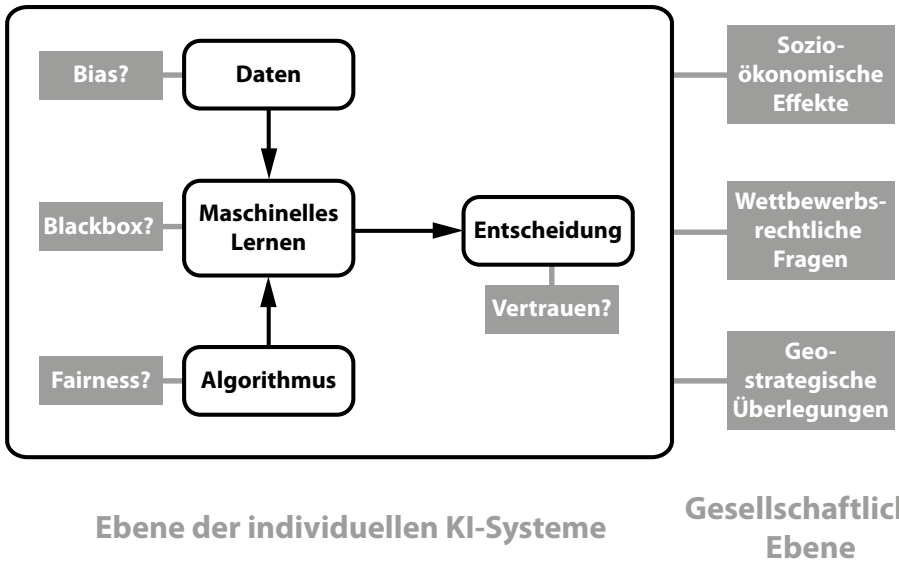

Abbildung 2: Gängige Diskussionsfelder bezüglich der gesellschaftlichen Auswirkungen von KI.

1.2. Eingrenzung und Zielgruppen der Studie

1.2.1. Eingrenzung der Fragestellung

In der Studie werden die Chancen und Risiken von KI für *exemplarische* Anwendungsfelder evaluiert. Der Fokus liegt dabei auf KI-Anwendungen, die Menschen bei der Entscheidungsfindung *unterstützen* oder gar *ersetzen* sollen. Folgende Anwendungsgebiete werden untersucht:

- **Arbeit:** In diesem Bereich werden volkswirtschaftliche Fragen (z.B. bezüglich der Veränderung des Arbeitsvolumens durch KI) als auch die möglichen Auswirkungen von KI auf den individuellen Arbeitsprozess (ohne dies berufsspezifisch aufzuschlüsseln) untersucht.
- **Bildung und Forschung:** In diesem Gebiet werden die Nutzung von KI für die Personalisierung des Lernens in der obligatorischen Schulzeit sowie für

die Förderung von Innovation in der Forschung (mit Fokus auf Naturwissenschaft und Technik) thematisiert.

- **Konsum:** In diesem Feld wird analysiert, inwieweit die Nutzung von KI durch Unternehmen das Verhältnis zum Endkunden (Kommunikation, Interaktion, Vertrauen) und damit auch das Verhalten der Konsumentinnen und Konsumenten beeinflusst. Es stellen sich zudem wettbewerbsrechtliche Fragen.
- **Medien:** In diesem Bereich wird analysiert, ob und inwieweit KI die bereits bestehenden Probleme der «Fake News» (Falschmeldungen) und der «Filterblasen» (Abschottung gegen konträre Meinungen) verschärft bzw. vermindert und damit den politischen Meinungsbildungsprozess beeinflusst.
- **Verwaltung und Gerichtsbarkeit:** In diesem Kontext wird vorab untersucht, wie die Rechtmässigkeit des staatlichen Handelns aufrechterhalten werden kann, wenn KI beispielsweise bei der Erhebung von Steuern oder im Strafrecht zunehmend zur Entscheidungsunterstützung herangezogen wird.

Die Untersuchungen in diesen Bereichen werden eingeführt durch technische Klärungen sowie die Analyse und Beurteilung ethischer und rechtlicher Fragestellungen (Kapitel 2). Für jedes dieser Themenfelder werden in Kapitel 3 spezifische Fragestellungen vorgestellt. Daraus resultieren generelle und bereichsspezifische Empfehlungen, wie aus Abbildung 3 ersichtlich wird.

Zeitlich liegt der Fokus auf den erwartbaren Entwicklungen in den nächsten fünf bis zehn Jahren. Utopische oder dystopische Szenarien rund um eine mögliche «Superintelligenz» sind kein Thema der Studie, weil die Diskussion möglicher Folgen stark von spekulativen Annahmen geprägt ist (etwa von der Annahme, eine solche «Superintelligenz» hätte eine Form von Bewusstsein).

Geografisch wird der Fokus auf die Schweiz gelegt. Durch die Einbindung internationaler Expertinnen und Experten sowie die Untersuchung von global relevanten unternehmerischen, politischen und wissenschaftlichen Einflüssen werden auch weltweite Entwicklungen und deren Auswirkung auf die Schweiz analysiert.

Die zentralen *Ergebnisse* der Studie sind (i) Empfehlungen an die Schweizer Politik bezüglich der Regulierung von KI-Anwendungen und der Unterstützung einer positiven Nutzung von KI sowie (ii) wissenschaftliche Erkenntnisse über Chancen und Risiken von KI-Anwendungen in den Bereichen Arbeit, Bildung, Konsum, Medien und Verwaltung. Einige Empfehlungen lassen sich auf die internationale Ebene skalieren und sind deshalb auch für solche Akteure relevant.

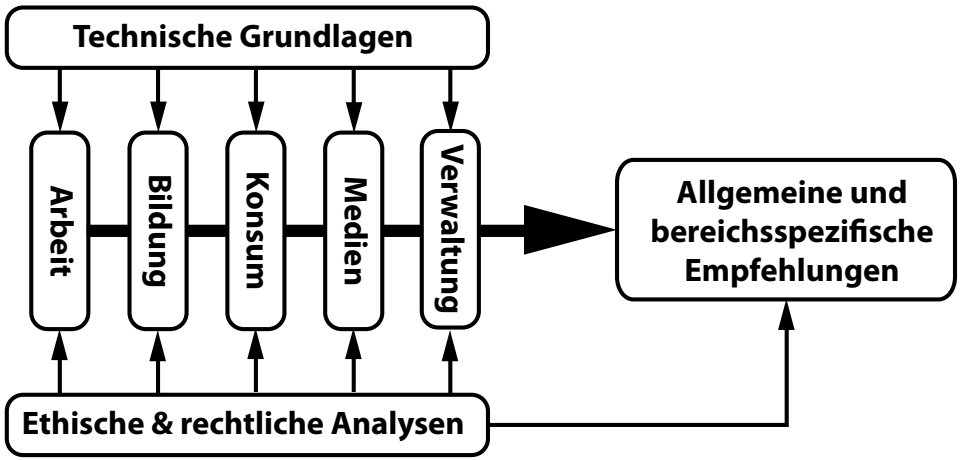

Abbildung 3: Übersicht über die in der Studie analysierten Themenfelder.

Wie in Abschnitt 1.1 vermerkt, überschneidet sich die Thematik KI mit anderen technikfolgenrelevanten Grundthemen des digitalen Wandels – namentlich Big Data, Robotik und Anwendungen in spezifischen, sensiblen Bereichen. Diese Themen werden wie folgt von der Zielsetzung der Studie abgegrenzt:

- Bezüglich *Technologie* liegt der Fokus auf jene Verfahren des ML, welche neue Herausforderungen bezüglich Verständlichkeit und Testung stellen, namentlich die Nutzung von Deep Learning. Doch auch andere Verfahren werden im technischen Teil vorgestellt und es wird an geeigneter Stelle darauf hingewiesen, wenn gängige KI-Anwendungen in den analysierten Themenbereichen auf anderen technischen Grundlagen beruhen.
- Bezüglich *Big Data* liegt der Fokus auf der konkreten Nutzung von Daten für KI-Systeme. Allgemeine Datenschutzfragen wie z.B. das «Recht auf Vergessen» oder die Kompatibilität von Datenschutzprinzipien wie Datensparsamkeit mit Big Data werden nicht thematisiert.
- Bezüglich *Robotik* liegt der Fokus auf den Risiken, welche sich generell aus der Nutzung von KI für die Steuerung von Robotiksystemen ergeben (z.B. bezüglich Haftpflicht). Konkrete Robotiksysteme wie z.B. Drohnen oder Fragen der Maschinenethik (z.B. bezüglich Mensch-Roboter-Interaktion) sind nicht Teil der Studie.

- *Anwendungsfelder* mit bereits etablierten Debatten (z.B. bezüglich autonomer Waffensysteme) oder solche, die in anderen TA-SWISS-Projekten behandelt werden (z.B. zum Thema autonomes Fahren), sind nicht Teil der Studie. Dies betrifft insbesondere auch Anwendungen der KI in der Medizin, die nicht Teil der von TA-SWISS vorgegebenen Aufgabenstellung waren.

Diese Einschränkungen sind notwendig, um eine vertiefende Analyse der spezifischen Folgen der Nutzung von KI-Technologie zu erreichen.

1.2.2. Zielgruppen

Zielgruppen der Studie sind jene Akteurinnen und Akteure, welche über Möglichkeiten verfügen, die weitere Entwicklung von KI in für die Gesellschaft positive Bahnen zu lenken, die KI nutzen oder von KI und deren Einflüssen tangiert werden. Konkret sind dies insbesondere:

- **Politik & Verwaltung:** Diese spielen eine entscheidende Rolle sowohl auf kommunaler, kantonaler, nationaler als auch internationaler Ebene, um die Potenziale von KI für die Gesellschaft zu nutzen und deren Wirkungen aus ethischer und rechtlicher Sicht zu reflektieren und entsprechend regulatorische Massnahmen zu treffen.
- **Wirtschaft:** Dies betrifft Unternehmen, welche KI nutzen, selbst in der Entwicklung tätig sind oder für deren Geschäftsmodell KI relevant sein könnte.
- **Bildung & Forschung:** Dies betrifft Bildungs- und Forschungsinstitutionen, die KI bereits anwenden oder zu deren Weiterentwicklung beitragen, insbesondere für eigene Bildungs- und Forschungsaktivitäten.
- **Zivilgesellschaft:** Dies betrifft Vereine und NGOs, die von KI tangiert werden oder KI direkt nutzen könnten.

Im Verlauf der Studie wurden die Arbeiten mit jenen von anderen Stakeholdern zu ähnlichen Themen abgestimmt. Insbesondere erfolgte ein enger Austausch mit der «interdepartementalen Arbeitsgruppe Künstliche Intelligenz» der Bundesverwaltung, welche Ende 2019 einen entsprechenden Bericht zuhanden des Bundesrates verfasst hat. Hierzu erfolgte ein Austausch von Zwischenresultaten und Textentwürfen. Zudem arbeitete das Team im Rahmen des Projekts «Künstliche Intelligenz in unserem Alltag: Wo wird sie eingesetzt und was erwarten wir von ihr?» mit den Akademien der Wissenschaften Schweiz, insbesondere mit der

Schweizerischen Akademie für Technische Wissenschaften (SATW) sowie der Stiftung Risiko-Dialog zusammen; insbesondere für die Durchführung einer Bevölkerungsumfrage. Schliesslich ist bezüglich der Umfrage auch die Zusammenarbeit mit SwissCognitive zu nennen (Kontakt via Begleitgruppe der Studie), einem in der Schweiz basierten Netzwerk von Unternehmen und weiteren Akteuren im Bereich KI.

1.3. Methodologie

Das Konsortium der Studie ist interdisziplinär zusammengesetzt. Dies erlaubt eine umfassende inhaltliche Abdeckung der in der Ausschreibung von TA-SWISS formulierten Fragestellungen. Methodisch wird eine transdisziplinäre Vorgehensweise gewählt (Jahn 2008), die sowohl Fragestellungen der beteiligten Forschenden aufgreift als auch jene der Expertinnen und Experten aus Bereichen der Verwaltung, Wirtschaft und Wissenschaft.

1.3.1. Instrumente und Arbeitsschritte

Zur Beantwortung der Fragestellungen wurde ein Methodenmix verwendet. Die Analyse in den Anwendungsbereichen stützte sich auf die jeweilige Fachexpertise in den Teams, die in Abschnitt 1.3.2 vorgestellt werden. Konkret sind folgende Instrumente und Arbeitsschritte in dieser Studie eingesetzt worden:

1. **Literaturanalyse:** Für jedes Anwendungsfeld erfolgte eine Auswertung der aktuellen Literatur, um eine Übersicht über derzeit diskutierte Fragen rund um die gesellschaftlichen Auswirkungen der Nutzung von KI zu erhalten. Die Literaturanalyse bildet die Grundlage für das zweite und dritte Kapitel des Berichts, in dem generelle technische, ethische und regulatorische Aspekte (Kapitel 2) sowie die interessierenden Fragestellungen pro Anwendungsgebiet (Kapitel 3) ausgearbeitet werden.
2. **Tiefeninterviews:** Mit diesem Instrument wurde ein tieferes Verständnis der Thematik pro Anwendungsbereich entwickelt. Tiefeninterviews wurden mit einer kleinen Anzahl an Teilnehmenden pro Anwendungsfeld (3–5) durchgeführt; die Auswahl der Fachpersonen erfolgte hierbei nach dem Urteil der Teams für die jeweiligen Bereiche.

3. **Zweiteilige Expertenumfrage:** Mithilfe der Erkenntnisse aus der Literaturanalyse und den Tiefeninterviews sind für alle Anwendungsbereiche detaillierte Fragebögen erarbeitet worden, um Ansichten und Einschätzungen zu bestimmten Anwendungsformen sowie deren Chancen und Risiken mithilfe einer breit angelegten Onlineumfrage erfassen zu können. Fachpersonen wurden international kontaktiert; die Umfrage erfolgte in Deutsch, Englisch und Französisch. Über 300 Personen haben valide Daten geliefert, welche eine Basis für mögliche Empfehlungen geliefert haben. Diese wurden dann in einer zweiten Umfrage erneut den Fachpersonen vorgelegt. Die Methodik der Umfrage wird im Anhang genauer erläutert.
4. **Zusatzumfrage:** In einer weiteren Umfrage sind auch die Auffassungen der allgemeinen Bevölkerung in die Studie eingeflossen. Dabei wurden Betroffene nach ihrer Nutzung von KI im Konsumbereich sowie bezüglich genereller ethischer Einschätzungen zur Nutzung von KI befragt. Ausführungen zur Methode finden sich ebenfalls im Anhang.
5. **Fokusgruppen-Workshops:** Die aus den Umfragen ermittelten Empfehlungen sind in zwei Fokusgruppen-Workshops in kleinerem Kreis intensiv diskutiert worden. Zum einen wurde im Juni 2019 ein Workshop mit einer Auswahl der an der Umfrage beteiligten Fachpersonen durchgeführt; zum anderen fand im August 2019 ein vergleichbarer Workshop mit Mitgliedern der Begleitgruppe statt. Die Ergebnisse der Arbeitsschritte 2 bis 5 werden in Kapitel 4 vorgestellt.

Ein zentrales Ergebnis dieser Arbeitsschritte bilden die Empfehlungen des Konsortiums, welche in Kapitel 5 vorgestellt werden.

1.3.2. Projekt-Konsortium

Die Breite der in der Ausschreibung angesprochenen Themen machte nicht nur ein interdisziplinäres Team unabdingbar. Dieses musste sich auch auf eine institutionelle Struktur abstützen können, in der Expertise auf einfache Weise hinzugezogen werden konnte. Deshalb figurieren die Digital Society Initiative der Universität Zürich zusammen mit der Abteilung Technologie und Gesellschaft der Eidgenössischen Materialprüfungs- und Forschungsanstalt und dem Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften als Konsortium. Diese Struktur erlaubte es, Fachkompetenz für die einzelnen Arbeitsschritte effizient beizuziehen. Zudem besitzt das Konsortium Erfahrung und

Kompetenz in den Bereichen Technikfolgenabschätzung und transdisziplinäre Methodik.

Im Projektbericht wird mittels Fussnoten in den Kapiteln 2 und 3 vermerkt, welche Autorinnen und Autoren die Inhalte der jeweiligen Abschnitte erarbeitet haben.

1.3.2.1. Digital Society Initiative – Universität Zürich

Die Digital Society Initiative (DSI) ist eine strategische Initiative der Universität Zürich zur Förderung der kritischen, interdisziplinären Reflexion und Innovation bezüglich aller Aspekte der Digitalisierung von Wissenschaft und Gesellschaft. Die DSI kann auf ein Netzwerk von über 300 Forschenden der Universität Zürich zurückgreifen, die zu Digitalisierungsthemen arbeiten. Sie verfügt damit über eine flexible und breite Struktur, um einen grossen Teil der für diese Studie relevanten Aspekte abzudecken. Forschende der DSI sind für folgende Bereiche der Studie zuständig:

- Ethik (Markus Christen & Markus Kneer)
- Konsum (Anne Scherer & Pascal Sutter)
- Medien (Tarik Abou-Chadi, Anita Gohdes & Hauke Licht)
- Öffentliche Verwaltung (Nadja Braun Binder)
- Recht (Florent Thouvenin, Luca Fábián, Damian George)
- Technologie (Abraham Bernstein & Daniele Dell'Aglio)

Zudem verantwortet die DSI die Schlussredaktion des Berichts (Markus Christen).

1.3.2.2. Abteilung Technologie und Gesellschaft – Empa, St. Gallen

Die Abteilung Technologie und Gesellschaft (Technology and Society Laboratory, TSL) ist am Departement Mobilität, Energie und Umwelt der Empa (Eidgenössische Materialprüfungs- und Forschungsanstalt) in St. Gallen angesiedelt. Ziel des TSL ist die Schaffung und der Transfer von Wissen zur Unterstützung der Gesellschaft in Richtung einer nachhaltigen Entwicklung. Die Forschungsgruppe Informatik und Nachhaltigkeit unter der Leitung von Lorenz Hilty entwickelt am TSL Computermodelle sowie softwarebasierte transdisziplinäre Methoden, um die Be-

wertung von technologischen Anwendungen zu unterstützen. Durch eine Partnerschaft mit dem Institut für Informatik an der Universität Zürich besteht im Hinblick auf die Digitalisierung in der Hochschulbildung und Forschung ein enger fachlicher Austausch. Forschende des TSL sind für folgende Bereiche zuständig:

- Bildung und Forschung (Clemens Mader & Claudia Som)
- Technologie (konzeptionelle Aspekte; Lorenz Hilty)

1.3.2.3. Institut für Technikfolgen-Abschätzung – Österreichische Akademie der Wissenschaften, Wien, Österreich

Das Institut für Technikfolgen-Abschätzung (ITA), angesiedelt an der Österreichischen Akademie der Wissenschaften (ÖAW), erforscht die Folgen neuer Technologien für Umwelt, Wirtschaft und Gesellschaft. Die Ergebnisse der wissenschaftlichen Arbeit unterstützen Politik, Verwaltung und Öffentlichkeit in technologiepolitischen Fragen. Das ITA betreibt interdisziplinäre Technikforschung, um die komplexen Wechselwirkungen von Technik und Gesellschaft aus verschiedenen Perspektiven zu verstehen, die Technologieentwicklung begleitend zu analysieren und durch Politik- und Gesellschaftsberatung zu einer sozialverträglichen Technologiepolitik beizutragen. Am ITA sind rund 20 Mitarbeitende tätig. Die Hälfte des wissenschaftlichen Teams kommt aus den Naturwissenschaften und aus technischen Fächern, die andere Hälfte aus den Geistes- und Sozialwissenschaften. Die Fächerpalette reicht von Philosophie, Soziologie, Politik- und Rechtswissenschaft über Ökonomie bis zu Informatik, Verfahrenstechnik, Biologie und Humanökologie. Forschende des ITA sind für folgende Bereiche zuständig:

- Arbeit (Johann Čas & Jaro Krieger-Lamina)
- Ethik (internationale Entwicklungen; Johann Čas & Jaro Krieger-Lamina)

2. Technische, ethische und rechtliche Grundlagen zu KI

Dieses Kapitel liefert das Fundament für die fünf Themenbereiche, die in dieser Studie behandelt werden. Abschnitt 2.1 beleuchtet die internationale KI-Debatte basierend auf einer quantitativen Medienanalyse. In Abschnitt 2.2 folgen Erläuterungen zu den technischen Grundlagen von KI mit dem Ziel, ein einheitliches Verständnis des diffusen Begriffs «künstliche Intelligenz» zu geben. Abschnitt 2.3 bietet eine Übersicht aktueller internationaler Initiativen zur Frage, wie man in ethischer, rechtlicher und sozialer Hinsicht mit der rasanten Entwicklung im Bereich KI umgehen soll. Abschnitt 2.4 erläutert anschliessend einige ethische Grundfragen, welche durch die Nutzung von KI gestellt werden. Generelle rechtliche Aspekte von KI unter primärer Bezugnahme auf das Schweizer Recht und unter Ausschluss verwaltungsrechtlicher Aspekte (diese werden im Folgekapitel behandelt) werden in Abschnitt 2.5 ausgeführt. Abschnitt 2.6 liefert schliesslich eine Einordnung des generellen Regulationsbedarfs in aktuell laufende Entwicklungen in der Schweiz basierend auf dem Bericht der «interdepartementalen Arbeitsgruppe Künstliche Intelligenz». Die in Kapitel 2 aufgeworfenen Punkte sind nur zum Teil in die Expertenumfrage und Empfehlungen eingeflossen; sie dienen vielmehr zur Einbettung der Studie in die nationale und internationale Diskussion zu KI.

2.1. Einführende Beobachtungen[1]

Es ist derzeit praktisch unmöglich, Medien zu konsumieren, ohne auf den Begriff der «künstlichen Intelligenz» zu stossen. Kaum ein Tag vergeht, in dem nicht ein «Sieg» einer KI über menschliche Expertise verkündet wird, sei dies nun Poker oder medizinische Diagnostik. Gleichzeitig verkünden Politikerinnen und Politiker regelmässig neue Investitionsoffensiven für die Förderung von KI in Forschung und Entwicklung. Wirtschaftsvertreter/-innen werden nicht müde, die Bedeutung dieser Technologie für den Wohlstand zu predigen, während Kritiker/-innen dystopische Szenarien von Überwachung und Kontrollverlust zeichnen.

[1] Dieser Abschnitt beruht auf Arbeiten von Markus Christen, DSI der Universität Zürich.

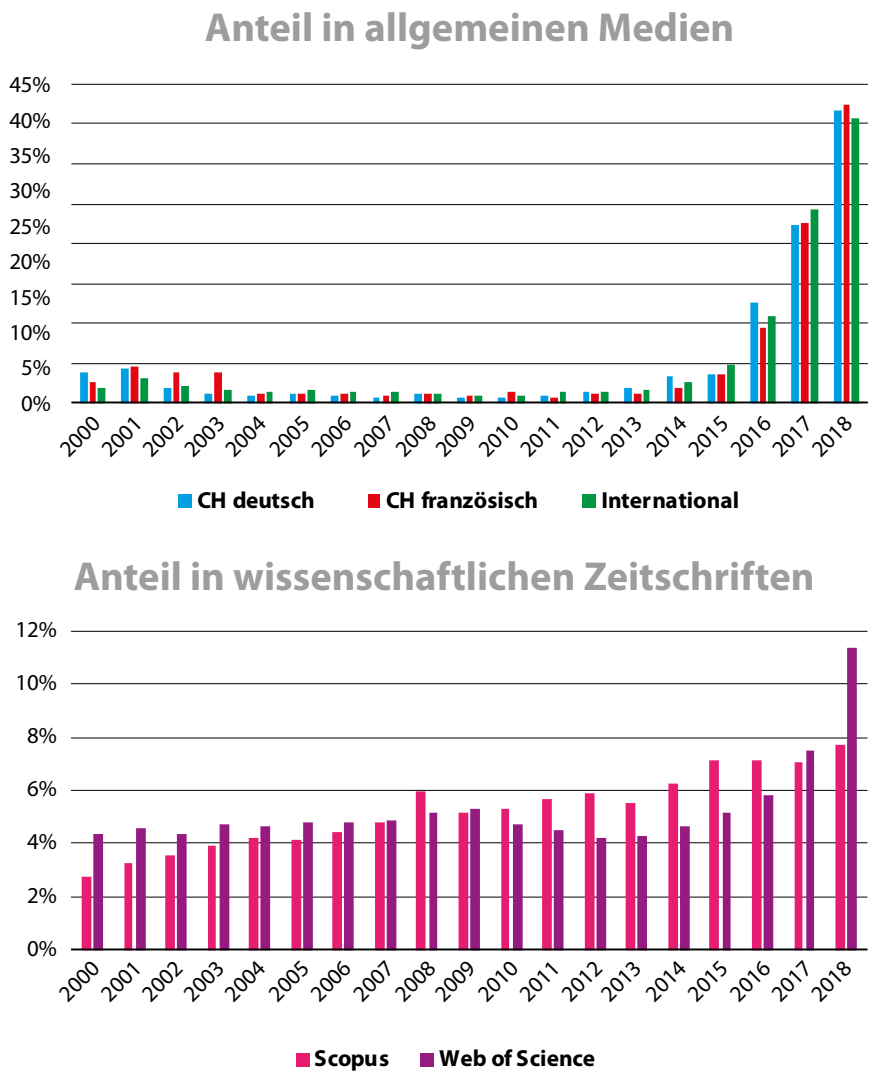

Abbildung 4: Ergebnisse bibliometrischer Untersuchungen zu KI in allgemeinen Medien und der Fachpresse. Gezeigt wird der prozentuale Anteil von KI-Beiträgen über den Suchzeitraum von 2000 bis 2018, reskaliert mit der Gesamtanzahl der Beiträge pro Jahr (Details siehe Fussnoten 2 und 6).

Eine bibliometrische Analyse bestätigt diese Alltagsbeobachtung (siehe Abbildung 4). Eine quantitative Erfassung von Beiträgen in Schweizer Medien, die das Stichwort «künstliche Intelligenz» bzw. «intelligence artificielle» enthalten, zeigt einen deutlichen Anstieg seit 2016, der sich sowohl in den Deutschschweizer und Westschweizer Medien zeigt als auch in der internationalen Presse.[2] Unter Verwendung der «Hype-Cycle»-Methodologie des Technologieberatungsunternehmens Gartner[3] lässt sich feststellen, dass zahlreiche KI-Technologien sich derzeit noch in der Phase der «übersteigerten Erwartungen» befinden; also gewissermassen «gehyped» werden. Es handelt sich aber nicht um den ersten solchen «Hype». Eine Google-Trend-Analyse[4] des Suchstichworts «artificial intelligence» verweist auf einen Aufmerksamkeits-Peak bereits vor 2004, was sich auch in der Medienanalyse widerspiegelt (lokales Maximum um 2001). Eine Google-Ngram-Analyse,[5] die sich auf die Häufigkeit des Stichworts «artificial intelligence» in der von Google erfassten Literatur (Bücher etc.) abstützt, verweist auf einen deutlich stärkeren Peak bereits Ende der 1980er-Jahre – also in der Zeit, als in der Schweiz Technikfolgenstudien zu «Expertensystemen» erschienen waren (damals bezeichnet als «forschungspolitische Früherkennung»).

---

2 Methodologie Medienanalyse: In der Datenbank «Faktiva» wurde in den Quellentypen «Nachrichten-/Presseagenturen», «Zeitschriften und Magazine» und «Zeitungen» in der Region «Schweiz» unter der Spracheinschränkung «deutsch» nach «künstliche Intelligenz» und der Spracheinschränkung «französisch» nach «intelligence artificielle» gesucht. Die Analyse wurde auf den Zeitraum von 2000 bis 2018 eingeschränkt. Als Referenzwert wurden jeweils alle in der Datenbank erfassten Artikel (quellen- und sprachspezifisch) ermittelt. Für die englischsprachigen Medien erfolgte keine regionale Einschränkung. Aufgrund der grossen Zahl von Medien wurden als Referenzwert alle Artikel erfasst, welche dem Suchausdruck «computer OR ‹information technology› OR ‹artificial intelligence› OR data» entsprechen. Die Suche wurde am 17. Juli 2019 durchgeführt.

3 Siehe https://www.gartner.com/en/research/methodologies/gartner-hype-cycle. Eine Übersicht zu verschiedenen KI-Technologien für das Jahr 2019 findet sich hier: https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/.

4 Google Trend erlaubt einen Überblick über die Häufigkeit von Suchstichworten bis 2004; die Methodologie der Erfassung hat sich allerdings im Zeitverlauf geändert, sodass die Daten hier nicht gezeigt und nur als Plausibilisierung erwähnt werden; siehe auch: https://trends.google.de/.

5 Google Ngram erlaubt einen Überblick über die Häufigkeit von Suchstichworten in der von Google Books erfassten Literatur. Präzise Informationen über die Grundlagen der Datenbank sind nicht ersichtlich; es ist anzunehmen, dass ein gewisser zeitlicher Delay in der Erfassung von Büchern besteht, sodass Aussagen über aktuellste Trends nicht möglich sind. Auch hier werden die Daten nicht gezeigt und nur als Plausibilisierung erwähnt; siehe: https://books.google.com/ngrams/info.

Ein bibliometrischer Blick auf die Fachliteratur[6] zeigt einen deutlichen Unterschied in der Publikationsdynamik im Vergleich zur Publikumspresse. Hier zeigt sich in den letzten zwei Jahrzehnten eine relativ konstante Präsenz der Thematik im Zeitverlauf, erst im «Web of Science» finden sich für 2018 erste Hinweise auf eine deutlich gestiegene Publikationstätigkeit.

Der «Hype»-Charakter rund um künstliche Intelligenz ist im historischen Zeitverlauf kein neues Phänomen. Die Ursprünge der KI gehen auf die 1950er-Jahre zurück (siehe Abschnitt 2.2.2) – und diese Anfangsphase war durch eine fast grenzenlose Erwartungshaltung im Hinblick auf die Fähigkeit von Computern geprägt. Diese Haltung wurde in der Vergangenheit regelmässig kritisiert und die hochgesteckten Erwartungen erfüllten sich jeweils nicht – man könnte gar die Entwicklung von KI als eine Kette bislang nicht eingelöster Versprechungen ansehen. Damit ist natürlich nicht gesagt, dass dies heute gleich sein wird. Die historische Erfahrung wie auch die hier skizzierte bibliometrische Evidenz zeigen aber, dass ein detaillierter und präziser Blick auf die Technologie nötig ist, um zwischen übertriebenen Erwartungen und Ängsten einerseits und wahrscheinlichen Entwicklungen andererseits unterscheiden zu können. Deshalb soll zuerst eine konzeptionelle Klärung und Präzisierung der technischen Grundlagen von KI erfolgen.

2.2. Begriffliche und technische Grundlagen[7]

Eine konzise Aufarbeitung der technischen Grundlagen der künstlichen Intelligenz bildet die Basis der Studie. Dies beinhaltet konzeptionelle Klärungen, eine kurze Darstellung historischer und aktueller KI-Anwendungen, eine Einführung in die wichtigsten technischen Grundlagen von KI sowie eine Übersicht zu generellen technischen Risiken von KI.

---

[6] Methodologie Fachliteratur-Analyse. In den Datenbanken «Scopus» und «Web of Science» (WoS; all databases) wurde nach dem Stichwort «artificial intelligence» im Zeitraum von 2000 bis 2018 gesucht (Scopus: Titel, Abstract, keywords / WoS: Topic). Als Referenzwert wurden alle Artikel erfasst, welche dem Suchausdruck «computer OR "information technology" OR "artificial intelligence" OR data» entsprechen. Die Suche wurde am 17. Juli 2019 durchgeführt. Hierzu muss ergänzt werden, dass die *Grey Literature* (z.B. ArXive), die ebenfalls wichtige Entwicklungen der KI-Forschung abdeckt, durch diese Suche nicht abgedeckt wird. Aus methodischen Gründen (mangelnde Vollständigkeit und damit Nachvollziehbarkeit) wurde darauf verzichtet.

[7] Dieser Abschnitt beruht auf Arbeiten von Daniele Dell'Aglio, Abraham Bernstein und Lorenz Hilti, Institut für Informatik der Universität Zürich.

2.2.1. Zum Begriff «künstliche Intelligenz» und «KI-Entscheidung»

Künstliche Intelligenz als Forschungsfeld wurzelt nicht in einer einheitlichen Disziplin, sondern kann von ihrem Ursprung her als Teilgebiet der sich ab den 1940er-Jahren entwickelnden Kybernetik verstanden werden (Rid 2016). Die 1956 am Dartmouth College in Hanover (New Hampshire) durchgeführte Konferenz «Dartmouth Summer Research Project on Artificial Intelligence» gilt als Gründungsanlass der KI, obgleich Vorstellungen eines maschinellen Denkens historisch gesehen noch weiter zurückreichen.

Der Begriff selbst besteht aus zwei Wörtern, «künstlich» und «Intelligenz». Während «künstlich» als Synonym für «von Menschenhand geschaffen» (Artefakt) verstanden werden kann, verweist «Intelligenz» auf einen komplexeren Sachverhalt. Das Merriam-Webster-Wörterbuch[8] schlägt zwei Definitionen von «Intelligenz» vor: «1) die Fähigkeit zu lernen oder zu verstehen oder mit neuen oder schwierigen Situationen umzugehen (Vernunft) und 2) die Fähigkeit, Wissen anzuwenden, um die eigene Umgebung zu manipulieren oder abstrakt zu denken, gemessen an objektiven Kriterien (wie Tests)»; das Oxford-Wörterbuch[9] umschreibt Intelligenz als «die Fähigkeit, Wissen und Fähigkeiten zu erwerben und anzuwenden». Aus diesen Definitionen kann künstliche Intelligenz grob als der Prozess des Verstehens und Lernens mittels eines Artefakts umrissen werden. Als Thema der psychologischen Forschung wird «Intelligenz» als noch komplexeres Phänomen verstanden, was an dieser Stelle aber nicht weiter ausgeführt werden kann.

In einem Standardwerk zu KI (Russel & Norvig 2010) wird «künstliche Intelligenz» anhand von vier Kategorien charakterisiert, die sich auf die Unterscheidungen «Denken» und «Handeln» sowie «Menschenähnlichkeit» und «Rationalität» abstützen; mit letzterem Paar wird zum Ausdruck gebracht, dass menschliches Denken und Handeln nicht notwendigerweise rational sein muss. Das bedeutet, dass unterschiedliche Zielsetzungen bezüglich der Schaffung von künstlicher Intelligenz bestehen, was sich auch auf die konkrete technische Umsetzung niederschlägt. Fokussiert man auf das Denken, will man mit KI die verschiedenen Aspekte der Denkfähigkeit modellieren. Ist Handeln im Zentrum, ist oft die Beantwortung praktischer Fragen das Ziel, und es spielt beispielsweise keine Rolle, ob das System in gleicher Weise wie ein Mensch zu seinen Schlüssen kommt. Der

[8] Siehe https://www.merriam-webster.com/dictionary/intelligence.

[9] Siehe https://en.oxforddictionaries.com/definition/intelligence.

Fokus auf «Rationalität» verweist auf einen idealen Standard des Denkens als Ziel – eine Fragestellung, welche die Philosophie seit Jahrhunderten beschäftigt. Dies impliziert, dass eine KI den Menschen diesbezüglich auch übertreffen kann oder soll. Ist schliesslich die «Menschenähnlichkeit» von Denken oder Handeln im Zentrum, so wird KI ein Instrument, um mehr über den Menschen selbst zu verstehen – was an tiefgreifende philosophische Probleme rührt, deren Beantwortung nicht Aufgabe dieser Studie ist.

Eine wichtige begriffliche Unterscheidung betrifft jene zwischen «genereller KI» (*artificial general intelligence* oder auch *strong AI*) und «angewandter KI» (oder auch *weak AI*). Ersteres betrifft die Idee der Schaffung einer KI, welche über ein breites Spektrum von Anwendungen intelligentes Denken und Handeln zeigt. Dies ist ein langfristiges Ziel eines Teils der KI-Forschung. Dabei bestehen sehr unterschiedliche Ansichten darüber, ob überhaupt und falls ja, wann dieses Ziel erreicht werden kann. Befürchtungen über eine «Superintelligenz», welche gar zu einer existenziellen Gefahr für die Menschen werden könne, sind an eine solche generelle KI geknüpft. Alle heutigen und in naher Zukunft zu erwartenden KI-Anwendungen dürften aber weiterhin in den Bereich der angewandten KI fallen. Diese Studie beschäftigt sich deshalb auch nicht mit dem Thema und den Risiken einer *strong AI*, sondern beschränkt sich auf angewandte KI.

Des Weiteren sollte begrifflich zwischen «künstlicher Intelligenz» als solcher und den einzelnen Technologien unterschieden werden, mit der man KI implementieren kann. Ein Beispiel für Letzteres ist maschinelles Lernen, was in Abschnitt 2.2.4.2 weiter ausgeführt wird. Solche *KI-Technologien* sind in der Regel nicht «intelligent» im oben beschriebenen Sinn; sie sind vielmehr sogenannte Basistechnologien (Jovanovic & Rousseau 2005), die nicht nur zahlreiche Lebensbereiche verändern, sondern die bereits eine Vielzahl an Branchen durchdrungen haben. Diese KI-Technologien realisieren – basierend auf einer Vielzahl möglicher Inputs (siehe Abbildung 5) – eine Reihe von grundlegenden Funktionen wie Mustererkennung in Daten, Klassifizierung von Input, Prognosen von Verhalten oder Synthesen wie z.B. die Schaffung von Sprache oder Musik.

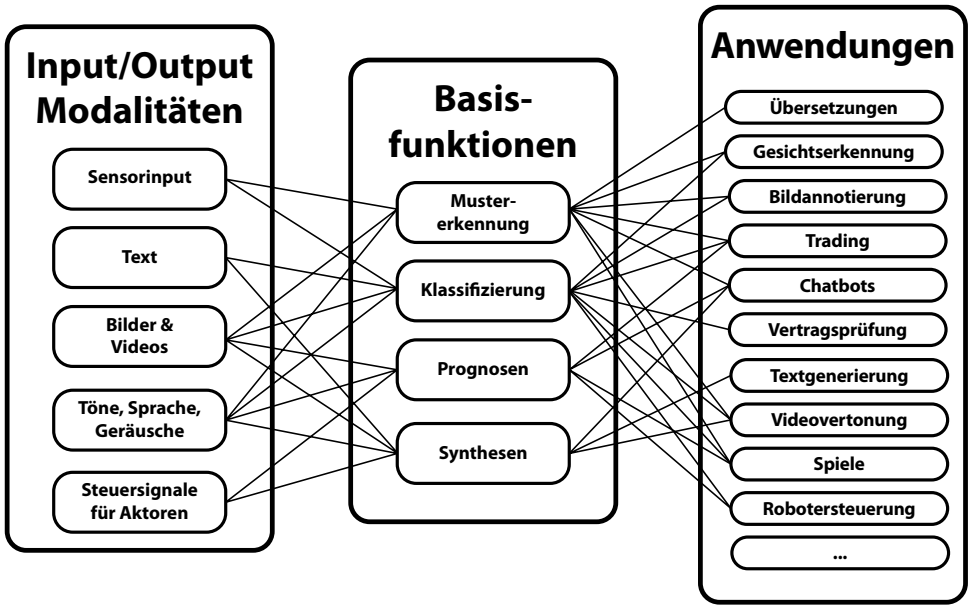

Abbildung 5: Input, Basisfunktionen und Anwendungen von heutigen KI-Systemen.

In der Regel ist es erst die Kombination von KI-Technologien zu einem *KI-System*, welche das gewünschte intelligente Verhalten zeigt (Interdepartementale Arbeitsgruppe 2019). Für heutige KI-Anwendungen müssen solche Systeme spezifisch für einen bestimmten Kontext definiert werden, was menschliche Expertise braucht; beispielsweise um die Struktur des Problems zu definieren, sodass dieses in bearbeitbare Aufgaben unterteilt und das Zusammenspiel der einzelnen KI-Komponenten untereinander geregelt werden kann.

Basierend auf dieser Zusammenstellung lässt sich besser verstehen, was als eine «*KI-Entscheidung*» aufgefasst werden kann. Zum einen könnte man darunter die Realisierung einer Basisfunktion wie z.B. eine Mustererkennung verstehen – beispielsweise wenn im Kontext des autonomen Fahrens eine Bilderkennungssoftware in der Umgebung ein Strassenschild erkennt. Ein derartiges Ergebnis ist aber nur in einem rudimentären Sinn eine Entscheidung; nachfolgend wird dafür der Begriff «KI-Ergebnis» verwendet. Die Kombination von KI-Technologien – die eine Folge menschlicher Designentscheidungen ist – kann dann aber als «Entschei-

dung» im engeren Sinn aufgefasst werden. Um im Beispiel des autonomen Fahrens zu bleiben, kann zwar eine KI-Systemkomponente ein Strassenschild mit einer Geschwindigkeitsbegrenzung erkennen; andere Systemkomponenten erkennen aber ein langsameres Fahrzeug vor einem oder berechnen mit Rückgriff auf Datenbanken die Wahrscheinlichkeit eines Rechtsabbiegens. Dies führt schliesslich zur Entscheidung, langsamer zu fahren, als an sich erlaubt wäre. Eine solche KI-Entscheidung ist aber in dreierlei Hinsicht abhängig von menschlichen Entscheidungen: Erstens sind menschliche Designentscheidungen bei der Kombination von KI-Technologien zu einem KI-System involviert. Zweitens fällt ein Mensch den strategischen Entscheid, das KI-System in einem bestimmten Kontext einzusetzen (z.B. anstelle eines menschlichen Entscheiders). Drittens schliesslich wird auf der «taktischen Ebene» entschieden, wann und auf welche Weise Menschen die Ergebnisse von KI-Entscheidungen überprüfen sollen. Dies kann zum einen die Einbindung des Menschen in den Entscheidungsprozess selbst sein; beispielsweise im Sinn, dass ein KI-System Vorschläge generiert, aus denen der Mensch auswählt (*human in the loop*; d.h. die Systeme haben eine Funktion als Entscheidungsunterstützung), der Mensch das System gewissermassen beobachtet und eine Art Vetorecht hat (*human on the loop*) oder dass die beiden Parteien im Sinn einer kooperativen Gruppe agieren (*human and machine in a group* oder *mixed-initiative system*). Zum anderen kann das eine regelmässige Überprüfung von KI-Entscheidungen sein, z.B. im Rahmen eines Audits. Es ist also klar, dass eine völlige Entkopplung zwischen menschlichen Entscheidungen und KI-Entscheidungen kein realistisches Szenario ist.

Zusammenfassend lassen sich nun folgende Definitionen festhalten:

> *«Künstliche Intelligenz» bezeichnet den Versuch, Verstehen und Lernen mittels eines Artefakts nachzubilden, wobei in erster Linie auf Denken bzw. Handeln fokussiert sowie ein rationales Ideal bzw. eine Nachbildung menschlicher Fähigkeiten angestrebt wird.*
>
> *«KI-Technologie» bezeichnet einzelne, in Computer implementierbare Funktionen für die Erreichung von künstlicher Intelligenz (z.B. maschinelles Lernen).*
>
> *«KI-System» bezeichnet eine strukturierte, kontextgebundene Kombination von KI-Technologien zwecks Erreichen von künstlicher Intelligenz.*

> *«KI-Entscheidungen» sind Schlussfolgerungen von KI-Systemen mit realweltlichen Auswirkungen, die auf der Ebene des Designs des Systems, der strategischen Ebene (Entscheid über Einsatz des Systems) und der taktischen Ebene (Ausgestaltung der Interaktion mit der Person, die das System nutzt) von menschlichen Entscheidungen abhängig sind.*

Zur Illustration sollen nachfolgend fünf Projekte beschrieben werden, die wichtige Meilensteine in der KI-Geschichte gesetzt haben. Damit soll aufgezeigt werden, dass sich die KI nicht auf eine bestimmte Herausforderung konzentriert, sondern sich mit einer Vielzahl von Problemen im Zusammenhang mit dem Verstehen und Lernen auseinandersetzt.

2.2.2. Meilensteine in der KI-Forschung

2.2.2.1. ELIZA

ELIZA gilt als einer der ersten Meilensteine in der KI-Forschung. Das von Joseph Weizenbaum (1966) realisierte System ist ein Chatbot, d.h. eine Software, die mit menschlichen Nutzerinnen und Nutzern kommuniziert, indem sie schriftliche Nachrichten mit diesen austauscht. Personen, die mit ELIZA interagierten, hatten dabei das Gefühl, dass sie mit einem anderen Menschen und nicht mit Software kommunizierten.

Technisch gesehen war der Prozess, der die Sätze von ELIZA formulierte, einfach: Das System extrahierte Schlüsselwörter aus Benutzersätzen und generierte aufgrund von fixen Regeln Antworten. Das Antwortspektrum von ELIZA ist dabei begrenzt; das System reformulierte in der Regel Aussagen der benutzenden Person. Wenn dieser zum Beispiel sagt: «Ich bin ohne Job», kann ELIZA antworten: «Wie lange bist du schon ohne Job?» oder «Erzähl mir mehr darüber». Diese Art von Gespräch gehört zu den einfachsten, die künstliche Intelligenz bewältigen kann. Es erfordert begrenztes Wissen über das Thema, und die Herausforderung besteht primär darin, Sätze zu formulieren, die Sinn ergeben und Thema des Gesprächs sind.

2.2.2.2. MYCIN

ELIZA hat keinen Zugang zu domänenspezifischem Wissen, sondern reagiert auf das Gegenüber. Künstliche Agenten mit domänenspezifischem Wissen auszustatten, war ein anspruchsvolles Problem und führte zur Entwicklung der sogenannten Expertensysteme, die ein umfassendes Wissen zu einem bestimmten Thema haben sollten. Während ELIZA also primär das Ziel «menschliches Handeln» (das Gegenüber soll das Gefühl haben, mit einem Menschen zu interagieren) fokussierte, will ein Expertensystem «rationales Denken» erreichen, um als Experte Entscheidungen wie medizinische oder technische Fachpersonen treffen zu können.

Die Machbarkeit von Expertensystemen wurde in den 1970ern gezeigt, als eine Gruppe von Forschern der Stanford University (Davis et al. 1977) das System MYCIN baute; ein Expertensystem, das Therapien (konkreter: die Dosierung von Antibiotika) anhand einer Reihe von Informationen über den Patienten festlegen sollte. Experimente zeigten dabei, dass die Leistung von MYCIN mit der von menschlichen Fachpersonen vergleichbar war.

Technisch gesehen bestand MYCIN aus einer Wissensdatenbank und einem (logikbasierten) Regelwerk. Die Regeln wurden durch die Benutzereingaben ausgelöst und führten zu Schlussfolgerungen über die Therapie. Ein wichtiges Merkmal von MYCIN war die Erklärbarkeit, d.h. man konnte genau nachvollziehen, wie das System zu seinen Vorschlägen gekommen ist.

2.2.2.3. Deep Blue

Spiele bieten ein ideales Umfeld, um das Verhalten einer künstlichen Intelligenz zu studieren. Spiele sind meist durch einen Satz klarer Regeln definiert, die eine hohe Anzahl möglicher Zustände erzeugen. Schach beispielsweise besteht aus einer Umgebung mit 64 Positionen (8x8-Raster) und 32 Figuren (16 weiss und 16 schwarz) mit festen Ausgangspositionen und definiertem Verhalten. Mit dieser Ausgangslage gibt es rund 10^{120} mögliche Schachspiele (zum Vergleich: das bekannte Universum zählt ca. 10^{22} Sterne). Alle möglichen Optionen durchzuprobieren, ist demnach nicht möglich. Das Schachspiel wurde deshalb ein prominentes Anwendungsgebiet für die Entwicklung von KI (Levinson et al. 1991).

Die Herausforderung, eine Maschine zu entwickeln, die Schach beherrscht, erreichte Mitte der 1990er-Jahre einen Wendepunkt. Damals gewann Deep Blue,

ein von IBM entwickeltes System, ein Spiel gegen Garry Kasparov, den damaligen Schachweltmeister (Campbell et al. 2002). Dies eröffnete einen ganzen Reigen an «Challenges» zwischen Mensch und Maschine, der bis heute anhält.

Technisch gesehen basiert Deep Blue auf einem Verfahren, das die Zahl der theoretisch möglichen Zustände auf eine kleinere Zahl der sinnvollen Züge beschränken konnte, die der Rechner sodann mit Bezug auf Gewinnwahrscheinlichkeiten durchrechnen konnte. Diese Strategie geht von der Voraussetzung aus, dass das Gegenüber rational spielt, um zu gewinnen, sodass es eine Verschwendung von Ressourcen wäre, sich auf Entwicklungen zu konzentrieren, bei denen die andere Person irrational spielen würde.

2.2.2.4. Watson

Das von IBM entwickelte Watson-System schaffte den Schritt von einem kontextgebundenen Expertensystem zu einem System zur Beantwortung allgemeiner Fragen zu diversen Themen (Ferrucchi et al. 2010). Wie im Fall von Deep Blue bildete ein Spiel die Testumgebung – in diesem Fall Jeopardy, eine der berühmtesten Quizshows in den USA. Watson forderte zwei der erfolgreichsten Spieler der Show heraus und wurde dabei wie die anderen menschlichen Spieler behandelt, d.h. Watson musste die Frage akustisch aufnehmen und sie sprachlich beantworten. Fragen konnten dabei zwischen jedem möglichen Thema variieren und auf knifflige Weise formuliert werden, einschliesslich Wortspiele und Idiome. Watson gewann das Spiel.

Die Entwicklung von Watson steht stellvertretend für mehrere Fortschritte in der KI-Forschung – namentlich die Integration verschiedener Fähigkeiten wie natürliche Sprachverarbeitung, Wissensrepräsentation und maschinelles Lernen in einem einzigen System.

2.2.2.5. AlphaGo

Eine weitere Testumgebung für KI bildete Go, ein chinesisches Spiel, das auf einem 19x19-Brett gespielt wird, auf dem zwei Spieler weisse bzw. schwarze Steine platzieren müssen. Wie beim Schach sind die Regeln einfach: Jeder Spieler kann einen seiner Steine auf (fast) jedem freien Platz des Brettes platzieren. Die Anzahl der möglichen Entwicklungswege des Spiels ist nochmals deutlich höher als beim Schach (10^{360}) – so wurde die Entwicklung eines KI-Systems, das erfolgreich Go

spielen kann, zu einer der grössten Herausforderungen, nachdem Deep Blue Kasparov besiegt hatte. Rund 20 Jahre später war es dann soweit, dass AlphaGo einen der weltbesten Go-Spieler, Lee Sedol, besiegte (Silver et al. 2016).

AlphaGo implementierte die heute gebräuchlichen Formen des maschinellen Lernens, d.h. AlphaGo wurde stärker, je mehr es spielte. Im Unterschied zu Deep Blue beschränkte sich AlphaGo nicht darauf, Strategien zu kopieren, die andere Spieler in ihrem Spiel verwendet hatten. Das KI-System war vielmehr in der Lage, neue erfolgreiche Strategien zu entwickeln.

2.2.3. Aktuelle und erwartbare KI-Anwendungen

ELIZA, Mycin, Deep Blue, Watson und AlphaGo setzten Meilensteine in der Entwicklung von KI-Systemen. Einige dieser Anwendungen bezeichnet man heute oft gar nicht mehr als KI – ein Hinweis darauf, dass der Begriff im öffentlichen Diskurs unscharf ist und man dazu neigt, verbreitete Fähigkeiten von Informationstechnologie nicht mehr als KI zu verstehen. Im Alltag sind aber KI-Technologien in zahlreichen Anwendungen präsent, die nicht in jedem Fall als «intelligent» im engeren Sinn gelten. Dies ist Ausdruck der Tatsache, dass KI-Technologien eine Basistechnologie (siehe Abschnitt 2.2.1) geworden sind. Nachfolgend soll kurz das aktuelle und absehbare Anwendungsspektrum von KI umrissen werden, wobei keines dieser Systeme als «generelle KI» im oben definierten Sinn zu werten ist.

2.2.3.1. Rechtschreibprüfungssysteme und Smart Keyboards

Moderne Textverarbeitungssysteme enthalten Software, die den Schreibprozess unterstützt. Diese überprüft die Rechtschreibung von Wörtern, bewertet die Grammatik, schlägt Synonyme vor und so weiter. Diese Systeme basieren in der Regel auf statistischen Ansätzen und implementieren damit KI-Technologien, die meist noch kein Lernen umfassen.

In Smartphones kommen lernfähige KI-Systeme zum Einsatz. Sie lernen beispielsweise, wie genau ein bestimmter Benutzer die Tasten drückt. Wenn etwa eine Person dazu neigt, auf dem Touch-Screen die N-Taste ganz links zu drücken, beginnt die Tastatur zu verstehen, dass die Person auch dann N meint, wenn sie ganz rechts auf (der neben N liegenden) Taste B drückt. Die Tastatur lernt zudem, welche Fehler die Person gewohnheitsmässig macht, korrigiert diese automatisch und vermag, das nächste Wort des Satzes vorherzusagen.

2.2.3.2. Maschinelle Übersetzung

Vor mehr als zehn Jahren wurde Google Translate eingeführt. Ursprünglich wurde ein rein statistisches System für die Übersetzung von Texten verwendet, welches gewissermassen die Wahrscheinlichkeit abschätzte, wie ein bestimmtes Wort im jeweiligen Kontext übersetzt werden sollte (Lopez 2008).

In jüngster Zeit wurde eine neue technische Implementierung für die maschinelle Übersetzung verwendet: die Nutzung neuronaler Netzwerke (siehe Abschnitt 2.2.4.2). Dies führte zu einem Quantensprung in der Qualität maschineller Übersetzung (Johnson et al. 2017). Heute verwenden alle gängigen Systeme wie DeepL oder Microsoft Translator diese Technologie.

2.2.3.3. Spiele

Spiele bilden weiterhin ein wichtiges Testfeld für KI-Systeme. Neu werden Kartenspiele für die KI-Forschung interessant, da sie eine Umgebung bieten, in der ein Teil der Informationen fehlt, da der Spieler in der Regel nicht weiss, welche Karten die anderen Spieler/-innen in der Hand haben. Dies hat die Entwicklung statistischer Methoden zur Bestimmung der Strategien und Taktiken vorangetrieben, die die Gewinnchancen maximieren. Menschliche Spieler/-innen zu besiegen, ist dabei jeweils die Benchmark, was jüngst für das Beispiel Poker geglückt ist.

KI wird aber zunehmend auch in kommerziellen Spielen eingesetzt, um den Spielern ein unterhaltsameres und realistischeres Erlebnis zu bieten. Videospiele können ihre Fähigkeiten automatisch an die Spieler/-innen anpassen; sie können die Muster lernen, die das Gegenüber verwendet, und ihnen entgegenwirken. Sie können auch mit den Spieler/-innen auf eine Weise interagieren, die rational erscheint oder dem menschlichen Verhalten ähnelt.

2.2.3.4. Empfehlungssysteme

Eine kommerziell weitverbreitete Nutzung von KI sind sogenannte Empfehlungssysteme, welche aus dem Verhalten von Nutzerinnen und Nutzern Rückschlüsse auf deren Präferenzen machen (Ricci et al. 2015). So analysieren beispielsweise E-Commerce-Websites (z.B. Amazon oder Zalando) die Browserverläufe von Kunden, um diesen neue Kaufempfehlungen vorzuschlagen. Mediendienstleister (z.B. Netflix oder YouTube) bieten ihren Kundinnen und Kunden neue Videos und

Filme an, die diese mutmasslich gerne ansehen würden. Soziale Netzwerke (z.B. Facebook oder Twitter) sortieren Nachrichten je nach Interesse, das die Nutzer/-innen an ihnen haben könnten. Auch die Personalisierung von Werbung beruht technisch gesehen auf ähnlichen Systemen. Damit geht eine Reihe kritischer Fragen einher, die in den Abschnitten Konsum (3.3) und Medien (3.4) erörtert werden.

2.2.3.5. Robotik

Robotik und künstliche Intelligenz waren schon immer zwei eng miteinander verbundene Bereiche. Die Robotik konzentriert sich auf das Design und die Entwicklung von physikalischen künstlichen Agenten (Robotern), die mit der physischen Welt interagieren können. Roboter verwenden Sensoren, um zu erkennen, was um sie herum passiert, und Effektoren, um Aktionen durchzuführen, die die Umwelt beeinflussen.

Um den Sensorinput und gespeichertes Wissen in Entscheidungen und Handlungen von Robotern zu überführen, wird KI zur immer wichtigeren Technologie. Zum Beispiel wird KI verwendet, um die Wahrnehmung der Umgebung zu unterstützen (wie die Identifizierung von Objekten und die Erkennung von Stimmen) oder um Aktionen zu planen (wie das Erreichen eines Ziels, das Hindernisse vermeidet, oder das Öffnen einer Tür).

Die Hauptanwendung von Robotik und KI liegt aktuell in der Industrie und künftig auch in der Automatisierung des Verkehrs (z.B. Züge und U-Bahnen). Selbstfahrende Autos bilden hier ein wichtiges künftiges Anwendungsfeld. Ein bekanntes Beispiel sind Roboterstaubsauger, die mittels KI das Haus kartografieren, in dem sie eingesetzt werden, um z.B. Hindernisse und Treppen zu umgehen.

2.2.3.6. Chatbots und virtuelle Assistenten

Chatbots sind Softwaresysteme, die mit Menschen kommunizieren, meist mittels Text. Sie finden heute unter anderem Anwendung in Messaging-Diensten wie z.B. Facebook Messenger, Skype und WeChat (einer chinesischen Messenger-Anwendung, ähnlich WhatsApp). Chatbots werden auch von Unternehmen verwendet, um einen Teil ihrer Kundenbeziehungen zu verwalten. Chatbots können beispielsweise mit Kundinnen und Kunden ins Gespräch kommen, um gängige

Probleme mit Produkten zu lösen, Reisen zu organisieren oder Essen zu bestellen.

Virtuelle Assistenten sind Softwaresysteme, die meist in natürlicher Sprache mit Menschen kommunizieren. Das Ziel von solchen Systemen wie z.B. Apples Siri, Google Assistant, Amazon Alexa und Microsoft Cortana ist es, zum persönlichen Assistenten von Menschen zu werden, der diese bei der Organisation ihres Alltags unterstützen soll. Sie bieten eine breite Palette von Funktionen, von der Beantwortung von Fragen bis hin zu Warnmeldungen, wenn jemand auf einen Stau zusteuert; von der Erinnerung an Termine bis hin zur Auswahl von Nachrichten, die eine Person interessieren könnten. Sie bilden im Bereich Konsum ein zunehmend wichtiges Instrument für Unternehmen, wie im Abschnitt 3.3.1.2 ausgeführt wird.

Persönliche Assistenten basieren auf einem breiten Satz von KI-Technologien, ähnlich wie Watson: Sie können die menschliche Stimme interpretieren, mit Menschen sprechen und nicht triviale Fragen beantworten. Zusätzlich zu Watson lernen sie menschliche Gewohnheiten kennen und passen ihre Funktionen an das Verhalten, das Sprechen und die Interaktion mit Menschen an. Man könnte sie als KI-Systeme verstehen, die erste Schritte in Richtung einer generellen KI machen.

2.2.3.7. Industrie 4.0

Die industrielle Automatisierung hat sich seit Kurzem auf die zunehmende Integration von KI-Technologien in die Herstellungsprozesse konzentriert, was zur Idee von Industrie 4.0 führt. Industrie 4.0 basiert auf vier Grundsätzen (Hermann et al. 2016): Vernetzung, Informationstransparenz, dezentrale Entscheidungen und technischer Support. KI-Technologien werden sich hier als zentral erweisen, um basierend auf den enormen Datenmengen, die in einer Industrie-4.0-Umgebung anfallen, aussagekräftige Modelle zu erstellen. Ebenso dürfte KI zunehmend in die Entscheidungsprozesse rund um die industrielle Fertigung (z.B. Erkennen von Verschleiss etc.) integriert werden und Menschen bei ihrer Arbeit unterstützen. In diesem Bereich werden in den kommenden Jahren voraussichtlich sehr viele KI-Anwendungen entwickelt, die kaum zu Kontroversen Anlass geben dürften (abgesehen vom Aspekt der Automatisierung menschlicher Tätigkeiten), weil dafür keine persönlichen Daten verarbeitet werden.

2.2.4. Funktionsprinzipien von KI

Der vorherige Abschnitt bot einen Überblick darüber, was KI ermöglichen kann. In diesem Abschnitt wird kurz vorgestellt, wie KI technisch funktioniert. Obgleich das Feld der künstlichen Intelligenz enorm ist und mit anderen Informatikbereichen wie die Theorie der Berechnung, dem Datenmanagement und der Softwareentwicklung überlappt, lassen sich dennoch Teilbereiche identifizieren und charakterisieren, die bestimmte Aufgaben oder spezifische Techniken im Rahmen der KI untersuchen. Dazu wird nachfolgend das «Sense-Think-Act»-Modell verwendet, ein klassisches Paradigma der Robotik und KI-Forschung (Albus 1991; siehe auch Abbildung 6).

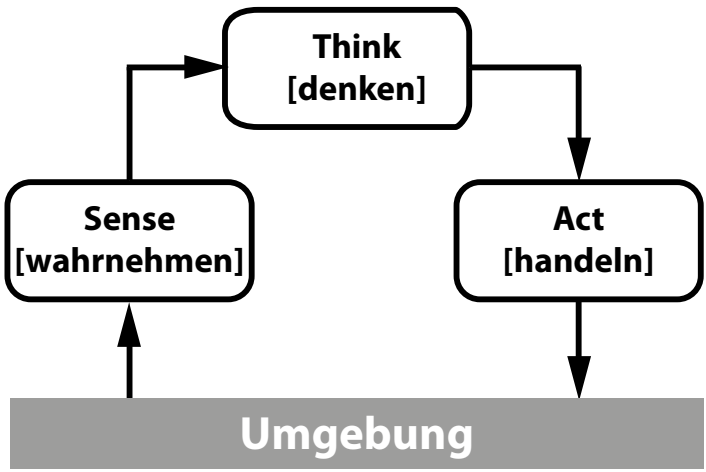

Abbildung 6: Das Sense-Think-Act-Modell.

In der Sense-Phase geht es um die Erfassung von Signalen aus der Umgebung. Dies können Daten aus einer physischen Umgebung sein, erfasst durch Sensoren, aber auch Daten aus einer virtuellen Umgebung wie z.B. Datenbanken oder anderen Arten von Datenquellen.

In der Think-Phase geht es darum, wie das System basierend auf den Daten aus der Umgebung eine Aufgabe lösen kann. Im klassischen Robotermodell besteht

die Aufgabe meist darin, eine Planungsaufgabe zu lösen, z.B. die beste Reihenfolge der Operationen zur Erreichung eines Ziels festzulegen. Im KI-Kontext umfasst sie auch Aufgaben wie Vorhersage oder Beantwortung von Fragen.

In der Act-Phase führt das System eine oder mehrere Aktionen aus, die sich auf die Umgebung auswirken. Abhängig von der Umgebung kann eine Aktion eine physische Interaktion oder die Produktion neuer Daten oder Informationen sein, die an einen Benutzer oder eine Datenbank gesendet werden. Da die Act-Phase die Umgebung verändert, beginnt der Zyklus meist von Neuem.

Die nachfolgende Übersicht ist nur skizzenhaft, für weitergehende Informationen kann z.B. auf Russell und Norvig (2010) zurückgegriffen werden. Zu beachten ist auch, dass bestimmte KI-Technologien zwar in der Regel ursprünglich für eine dieser drei Phasen entwickelt wurden, zunehmend aber auch in anderen Phasen Anwendung finden können. Zum Beispiel wird das maschinelle Lernen im Think-Bereich beschrieben, aber es wird derzeit erforscht, wie man maschinelles Lernen für die Sense-Phase nutzen kann. Zusätzlich werden zunehmend Methoden entwickelt, in welchen die drei Phasen verschmelzen. So verbinden reaktive Agenten, wie sie auch in der Robotik eingesetzt werden, zunehmend die Sense und Act-Phase direkt miteinander. Nachfolgend werden die einzelnen KI-Technologien gemäss ihrem Ursprung und ihrer grundlegenden Anwendung klassifiziert.

2.2.4.1. KI für die Sense-Phase: Sprach- und Objekterkennung

Sensing bezeichnet den Prozess, Informationen aus der Umgebung zu sammeln und für die weitere Verarbeitung zur Verfügung zu stellen – etwa wenn Menschen ihre Augen benutzen, um visuelle Informationen zu erfassen, ihre Ohren, um Geräusche einzufangen, und ihre Nase, um Gerüche wahrzunehmen. KI-Technologien kommen dabei dann zum Einsatz, wenn die Art der zu erfassenden Information eine gewisse Komplexität aufweist.

Ein erstes Beispiel dafür ist **natürliche Sprache**. In Textform gegebene Information kann für einen Computer unterschiedlich schwer zu verstehen sein. Ein einfaches Beispiel ist eine Tabelle. Durch Spalten und Zeilen ist die Information in einer klar definierten Struktur gegeben. Derart strukturierte Daten (wie z.B. auch Datenbanken, CSV- und JSON-Dateien) sind für Computer unproblematisch. Ganz anders verhält es sich bei einem natürlichsprachlichen Text, dessen Grammatik auf menschliches Sprachverständnis ausgerichtet ist. Für einen

Computer sind dies unstrukturierte Daten, die dieser in strukturierte Daten verwandeln muss, d.h. er muss die nützlichen Informationen aus dem Text extrahieren und in einem Format darstellen, das er verarbeiten kann. Mit diesem Problem beschäftigt sich die Disziplin des *natural language understanding* (NLU), einem Teilgebiet des *natural language processing* (Clark et al. 2010).

NLU-Lösungen lassen sich grob in zwei Arten einteilen: symbolisch und statistisch. Symbolbasierte Techniken stimmen den Text mit den Regeln der Sprache (z.B. den Grammatikregeln) ab. Dieser Ansatz ist in der Regel sehr aufwendig, da er die manuelle Modellierung aller Regeln erfordert, Grammatikregeln oft Ausnahmen haben und weil der gleiche Inhalt häufig in einer Vielzahl von möglichen Aussagen gepackt werden kann. Der statistische Ansatz nutzt Techniken, die gemeinsame Muster in den Daten identifizieren und Modelle erstellen, mit denen der Text «interpretiert» werden kann. Dieser Ansatz erfordert einen grossen Korpus an Textdaten, um die Modelle zu trainieren, die technisch oft in der Form von neuronalen Netzen realisiert werden (siehe Abschnitt 2.2.4.2). Die aktuelle Forschung konzentriert sich auf die Kombination der beiden Ansätze.

Eine weitere Datenkategorie, die KI für den Sense-Teil benötigt, sind **visuelle Informationen**. Es gibt hier verschiedene Aufgaben im Zusammenhang mit der Bildverarbeitung, wie z.B. die Objekterkennung (die Identifizierung der Objekte im Bild) und die Rekonstruktion einer 3-D-Welt (die Schaffung einer 3-D-Umgebung aus Bildern von verschiedenen Punkten des Raumes). Technisch gesehen wird eine Reihe von Ansätzen verwendet. Klassische Methoden verwenden eine hierarchisch organisierte Arbeitsweise, d.h. man beginnt mit der Erkennung von Kanten, der Texturanalyse und dem optischen Fluss, d.h. die Analyse von Bewegungen im Bild. Danach erfolgen Schritte wie die Bildsegmentierung, d.h. die Identifizierung von Pixelbereichen (die Punkte, aus denen das Bild besteht) mit gemeinsamen Merkmalen. Schliesslich wird das Ergebnis durch High-Level-Operationen verarbeitet, um die jeweilige Bildverarbeitungsaufgabe zu lösen. Auch hier eröffnet das Deep Learning (siehe Abschnitt 2.2.4.2) neue Möglichkeiten.

Eine weitere wertvolle Informationsquelle sind **Geräusche**. Wie in der Bildverarbeitung kann akustische Information analysiert werden, um eine Vielzahl von Aufgaben zu lösen, wie z.B. Spracherkennung (d.h. Umwandlung von Sprache in Text), Stimmerkennung (d.h. Identifizierung des Sprechers) und Musikerkennung (d.h. Identifizierung eines Songs). Grundlagen der Klangverarbeitung stammen aus der Signalverarbeitung, die eine Reihe von Instrumenten und Werkzeugen zur

Untersuchung und Analyse dieser Art von Daten bietet. Alternativ finden auch hier Deep-Learning-Techniken vermehrt Anwendung.

2.2.4.2. KI für die Think-Phase: maschinelles Lernen

Wie zu Beginn erläutert, ist die Fähigkeit, Daten auf rationale bzw. menschenähnliche Weise zu verarbeiten, ein Hauptziel von KI. Unterschiedliche Paradigmen führen zu verschiedenen technischen Umsetzungen: deduktives, induktives, abduktives und analoges Denken.

Deduktives Denken hat seine Wurzeln im antiken Griechenland. Eine der ersten Argumentationstechniken ist der von Aristoteles eingeführte Syllogismus, ein Katalog bestimmter Typen logischer Schlüsse. So kann man beispielsweise aus den Prämissen «Alle Katzen sind Säugetiere» und «Kitty ist eine Katze» den Schluss ziehen «Kitty ist ein Säugetier». Auch wenn das Beispiel banal klingt, verweist es doch auf einige interessante Eigenschaften von Wissen. Dieses kann sich auf Fakten über einzelne Individuen beziehen («Kitty ist eine Katze») oder aber gesetzesartige Festschreibungen bzw. «Axiome» festhalten («Alle Katzen sind Säugetiere»). Generiert man neue Fakten aus bestehenden Fakten und Axiomen, wird dies als deduktive Inferenz bezeichnet.

Solche deduktiven Argumentationen werden in der formalen Logik untersucht und in KI implementiert. Wissensrepräsentation und Argumentation ist ein Teilgebiet der KI, das sich auf Probleme im Zusammenhang mit dem deduktiven Denken konzentriert. Expertensysteme wie MYCIN beruhen auf solchen Ansätzen.

Die Hauptnachteile logikbasierter Inferenzsysteme liegen in der Nichtrobustheit der Logik (ein Fehler genügt, um Schlussfolgerungen zu verunmöglichen) und in der Tatsache, dass die Realität oft von Unsicherheit beeinflusst wird. Aus diesem Grund wurden probabilistische Argumentationen eingeführt. Diese modellieren die Unsicherheit basierend auf der Wahrscheinlichkeitstheorie.

Ein weiteres Paradigma ist **induktives Denken**, bei dem man aus einer Vielzahl von Fakten allgemeine Schlussfolgerungen erreichen will. Deduktives Denken führt zur sicheren Schlussfolgerung in dem Sinn, dass Folgerungen mit Sicherheit wahr sind, wenn die Prämissen dies auch sind. Induktives Denken kann zu falschen Schlussfolgerungen führen. Ein weiterer Unterschied betrifft die Rolle von Axiomen und Fakten. Im deduktiven Denken leiten Axiome den Inferenzprozess:

Sie führen zu neuen Fakten oder Schlussfolgerungen durch eine Abfolge von logischen Schritten. Im induktiven Denken leiten Fakten den Inferenzprozess und führen dazu, die Regeln zu erlernen, die die Realität regulieren.

Mehrere Informatikbereiche untersuchen induktives Denken. In der künstlichen Intelligenz ist das *maschinelle Lernen* (ML) das wichtigste Feld, das sich mit diesem Problem auseinandersetzt. Das maschinelle Lernen kann grob in drei verschiedene Teilbereiche unterteilt werden: überwachtes, unbeaufsichtigtes und Verstärkungslernen. Diese Trennung ist nützlich, um die grundlegenden Bausteine des maschinellen Lernens zu verstehen. Die aktuelle Forschung untersucht allerdings, wie man die verschiedenen Arten des Lernens miteinander kombinieren kann.

Beim sogenannten überwachten Lernen (*supervised machine learning*) wird der Algorithmus mit Daten trainiert, und ein menschlicher Beobachter bzw. Benchmark-Daten überwachen den Lernerfolg. Ein Beispiel dafür sind sogenannte Support-Vektor-Maschinen. Diese klassifizieren Datenpunkte, indem sie diese in einen hochdimensionalen Raum projizieren, wo eine Hyperebene konstruiert wird, welche verschiedene Datenklassen am besten trennt. Zum Beispiel wird der Algorithmus mit einer Folge von positiven und negativen Beispielen konfrontiert und lernt daraus, eine Ebene zu konstruieren, die die beiden Gruppen abgrenzt, sodass der Abstand zwischen der Ebene und dem nächstgelegenen Mitglied in jeder Gruppe maximiert wird.

Der Schlüssel zum Erfolg des überwachten Lernens liegt in der Identifizierung der Merkmale zur Beschreibung der Individuen. Diese Aufgabe, die als *feature engineering* bezeichnet wird, ist in der Regel Aufgabe der Experten des jeweiligen Anwendungskontextes. Die Situation änderte sich in den letzten Jahren, als zunehmend Deep Learning eingesetzt wurde (LeCun et al. 2015). Deep Learning kann als eine besondere Form des überwachten Lernens betrachtet werden, die auf einer technologischen Lösung namens «neuronaler Netze» basiert. Man kennt zwei Hauptklassen solcher Netze: *convolutional neural networks* (hier fliesst die Information nur in eine Richtung) und *recurrent neural networks* (diese Systeme haben eine Rückkopplungskomponente). Die resultierenden Modelle sind Gleichungen, die mathematische Funktionen mit einer oft enormen Zahl von Eingabeparametern berechnen. Im Gegensatz allerdings zu Gleichungen, die in den Naturwissenschaften verwendet werden und wo die Mathematik zur Beschreibung der physikalischen Welt verwendet wird, haben solche ML-Modelle keine offensichtliche physikalische oder logische Basis. Das bedeutet, dass Deep-Learning-Lösungen zwar Features extrahieren können, diese aber keine klare Semantik

haben. Durch Deep Learning geht deshalb die Erklärbarkeit verloren, d.h. das KI-System kann nicht erklären, warum es eine Person einer bestimmten Klasse zuordnet. Dies ist die technische Ursache des Blackbox-Problems.

Typische Probleme beim beaufsichtigten Lernen sind: (i) die Erstellung des Datensatzes (Sammeln von Input-Output-Paaren) kann kostspielig sein, (ii) die Ausgewogenheit der Trainingssätze (d.h. alle Ausgabekategorien sollten gleich dargestellt werden) kann nicht gewährleistet sein und (iii) overfitting, d.h. die Tatsache, dass die aus den Daten gewonnenen Modelle im Wesentlichen aus den Daten selbst bestehen und nicht verallgemeinert werden können.

Das unbeaufsichtigte maschinelle Lernen (*unsupervised machine learning*) zielt darauf ab, Zusammenhänge und Abhängigkeiten in Daten ausfindig zu machen und diese gegebenenfalls zu Merkmalen weiter zu verarbeiten, z.B. indem es aus Einkaufskörben erkennt, dass Spaghetti typischerweise mit Tomatensauce und Chianti gekauft werden. Beim unbeaufsichtigten Lernen besteht der Beispielsatz also aus Eingabedaten, ohne dass ein Hinweis auf die erwartete Leistung des Systems vorliegt. Ein typisches unbeaufsichtigtes Lernproblem ist Clustering, d.h. die Bildung von Gruppen von Elementen mit ähnlichen Eigenschaften. Clustering kann verwendet werden, um Kategorien zu identifizieren, die zu Beginn unbekannt sind. Es kann auch bei der Ausreissererkennung eingesetzt werden – der Identifizierung von Elementen, die ungewöhnlich sind. Beispielsweise kann die Analyse von Bankgeschäften mit Clustering-Techniken dazu führen, dass anomale Transaktionen identifiziert werden, die als Betrugsversuche infrage kommen.

Überwachtes und unbeaufsichtigtes Lernen setzen voraus, dass das System eine Reihe von Elementen beobachtet und aus ihnen lernt. Das *Verstärkungslernen* verfolgt einen anderen Ansatz: Der Agent soll durch Bewertung des Lernerfolgs lernen. Diese Art des Lernens orientiert sich an einer der häufigsten Lernweisen, die auch für den Menschen gilt: Man wird bestraft, wenn man etwas Falsches tut, und man wird gelobt, wenn man etwas Richtiges tut. Das Ziel des Verstärkungslernens ist es also, eine optimale Strategie zu erlernen, d.h. ein Entscheidungsverfahren, das eine möglichst optimale Aktion unter den möglichen auswählt. Dazu beginnt das System Aktionen auszuprobieren. Jede Aktion erhält eine Belohnung, und das Ziel des Agenten ist es, eine solche Belohnung zu maximieren.

Das Verstärkungslernen findet seine Hauptanwendung in Planungsproblemen wie der Navigation (der Agent muss ein Ziel erreichen, und die Belohnung kann mit Zeit oder Entfernung verknüpft werden) und Videospielen (der Agent muss spielen, und seine Belohnung ist mit dem Endergebnis des Spiels verknüpft). In letzter

Zeit wurde das Verstärkungslernen in Kombination mit Deep Learning eingesetzt: Derartige Paradigmen haben sich als sehr effektiv erwiesen, z.B. hat AlphaGo auf diese Weise gelernt, Go zu spielen.

Einige Nachteile des Verstärkungslernens sind (i) die Konstruktion der Belohnungsfunktion – bei komplexen Problemen kann es schwierig sein, zu entscheiden, wie der Agent belohnt werden soll –, und (ii) es kann aufwendig sein, die optimale Strategie zu erlernen. Braucht es Millionen Versuche, bis diese gefunden wurde, ist es oft nicht möglich, Verstärkungslernen einzusetzen.

Eine dritte Variante ist das **abduktive Denken**, das sich am besten aus dem Vergleich mit den beiden anderen Varianten erklären lässt. Beim deduktiven Denken kennt das System das Axiom und die kausale Tatsache; beim induktiven Denken generiert das System das Axiom aus einer Vielzahl von Fakten. Kennt nun aber ein System das Axiom (z.B. «Alle Katzen sind Säugetiere») und eine Schlussfolgerung («Kitty ist ein Säugetier»), könnte das System schliessen «Kitty ist eine Katze». Das ist die Idee hinter dem abduktiven Denken – der Argumentationsprozess bringt Informationen über Fakten (Neapolitan 1990). Logisch gesehen könnte der Schluss natürlich falsch sein (Kitty könnte ein Hund sein); d.h. das klassische Problem des abduktiven Denkens besteht darin, die wahrscheinlichste Ursache zu finden. Daraus folgt, dass abduktives Denken auf probabilistischen Frameworks aufbaut.

Abduktives Denken wird zur Lösung von Diagnoseproblemen eingesetzt, wie z.B. zur Erkennung von Defekten in komplexen Systemen oder bei Erkrankungen in der Medizin. Abduktives Denken ist auch bei der Bewertung des Inhalts von Wissensdatenbanken hilfreich, da es Ursachen für Fehler oder Inkonsistenzen identifizieren kann.

Analoges Denken (Bartha 2013) schliesslich ist ein weiteres typisches Argumentationsschema. Wie der Name schon sagt, geht es um Analogien, d.h. Ähnlichkeiten zwischen Konzepten, Problemen oder Systemen. Analoges Denken funktioniert unter der Annahme, dass zwei Elemente, die ähnlich sind, gemeinsame Merkmale aufweisen, sodass es möglich ist, unbekannte Merkmale des einen durch Analyse des anderen zu erschliessen. Eine typische Situation, in der analoge Argumentation angewendet wird, ist die Rechtsprechung (*case law*), in der Argumente auf früheren Fällen aufbauen.

In der künstlichen Intelligenz werden analoge Argumentationsfunde in eine fallbezogene Argumentation umgesetzt: Die Idee ist, dass ein System ein Problem

lösen kann, indem es eine Lösung anpasst, die es in der Vergangenheit zur Lösung eines ähnlichen Problems eingesetzt hat. Fallbasierte KI-Systeme speichern die Probleme und die Lösungen, mit denen sie in der Vergangenheit konfrontiert waren. Wenn sie ein neues Problem lösen müssen, rufen sie (i) das ähnlichste Problem ab, (ii) verwenden das Wissen aus der bisherigen Erfahrung wieder, (iii) überarbeiten die vorgeschlagene Lösung und (iv) speichern die neue Lösung im Speicher, um analoge Probleme in Zukunft zu lösen.

2.2.4.3. KI für die Act-Phase: Sprachsynthese

Um den Interaktionszyklus mit der Umgebung zu schliessen, sollte ein automatisierter Agent handeln. Während einige Aktionen einfach sind – das Schreiben von Daten in eine Datei, das Einschalten einer LED oder die Ausgabe eines Tons –, sind andere komplexer und können KI-Technologien beinhalten.

Ein erstes Beispiel ist die **Erzeugung natürlicher Sprache**, die *Natural Language Generation* (NLG). Dieses konzentriert sich auf die Transformation von Daten in einem Text. Eine der grössten Herausforderungen von NLG besteht darin, Text zu generieren, der nicht künstlich aussieht: Nicht nur sollte der generierte Text den Grammatikregeln der Zielsprache folgen, sondern er sollte auch die Vielfalt und die stilistischen Entscheidungen abbilden, die der Mensch beim Sprechen trifft. Zu diesen Wahlmöglichkeiten gehören die Verwendung von Konnektiven, die Zusammensetzung mit primären und sekundären Sätzen, die Wahl der richtigen Wörter (z.B. Synonyme) und die Orthografie. Typische NLG-Anwendungen sind die Erstellung von Textzusammenfassungen (z.B. Zusammenfassungen von Sportveranstaltungen), Chatbots und virtuellen Assistenten.

Die **Sprachsynthese** ist die Erzeugung von menschenähnlicher Sprache, die verständlich und natürlich klingt. Die einfachsten Sprachsyntheselösungen basieren auf einer Datenbank mit aufgezeichneten Wörtern, die zur Erstellung des Satzes zusammengesetzt werden. Dies führt zu Reden, die klar verständlich, aber künstlich klingen: Die Verkettung von zuvor aufgenommenen Klängen erlaubt es nicht, z.B. die Intonation anzupassen. Ausgefeiltere Lösungen erzeugen die Stimme ausgehend vom Text. In einem ersten Schritt wird der Text in prosodische Einheiten (d.h. Einheiten mit gleichen Lauteigenschaften) übersetzt, und dann werden diese Einheiten in Text übersetzt. Während zu Beginn Lösungen, die auf diesem Ansatz basierten, schwer verständliche Reden erzeugten, verbesserten die jüngs-

ten Fortschritte beim Deep Learning diese Techniken drastisch. Die Sprachsynthese wird im Verkehrswesen (z.B. zur Informationsvermittlung in Zügen und U-Bahnen), in der Usability (z.B. zur Unterstützung von Blinden bei der Nutzung von PCs oder Smartphones) und in virtuellen Assistenten eingesetzt.

2.2.5. Technische Risiken von KI

Wie bei jeder anderen Informatiklösung auch bestehen bei KI-Technologien technische Risiken, die in den Bereich der Cybersicherheit fallen (Sicherung der Integrität, der Vertraulichkeit und des Zugangs zu Daten bzw. zum System). Im Fall von Deep Learning werden aufgrund des Blackbox-Problems aber noch andere technische Risiken relevant, die hier kurz erläutert werden sollen.

Die oben skizzierten maschinellen Lernsysteme basierend auf *deep neural networks* sind eine Blackbox dahin gehend, dass sie Ergebnisse oder Empfehlungen liefern, ohne eine greifbare oder überprüfbare Erklärung darüber abzugeben, wie das Ergebnis erreicht wurde (Knight 2017). Diese Unerklärlichkeit ist von fundamentaler Bedeutung, da es derzeit keine zugrunde liegende Theorie gibt, die begründen kann, wie oder warum diese Modelle für eine bestimmte Art von Problem wirksam sind. Auch fehlt eine Basis, um ihre spätere Leistung vorherzusagen. Ein Modellentwickler beginnt mit einer riesigen Datenmenge und führt umfangreiche Berechnungen durch, um die Modellparameter für die Erstellung der besten Vorhersagen für diese Daten anzupassen, und wiederholt dann diesen Optimierungsprozess auf verschiedenen Datensammlungen, bis eine zufriedenstellende Vorhersagegenauigkeit erreicht ist. Daher sind *deep neural networks* derzeit nur schwer adäquat zu testen und relativ leicht zu täuschen, was aber bei anderen KI-Technologien nicht der Fall sein muss.

Das Blackbox-Problem hat wichtige praktische Konsequenzen. Herkömmliche Software-Testverfahren sind stark auf sogenannte Unit-Tests angewiesen, d.h. es werden die einzelnen Komponenten eines Softwaresystems validiert, bevor sie zu einem einheitlichen System zusammengefasst werden, das dann als komplette Einheit getestet werden kann. Die *deep neural network*-Modelle dagegen können nur als Ganze getestet werden. Da sie typischerweise über eine grosse Anzahl von Eingängen (bzw. Eingangsknoten) verfügen, ist es nicht möglich, selbst einfache Modelle umfassend zu testen, was die Frage offenlässt, wie sich ein solches ML-Modell in einer gegebenen Situation genau verhält. Zudem ermöglichen es sogenannte kontradiktorische maschinelle Lerntechniken – z.B. indem man einem

Signal Rauschen hinzufügt –, das System zu täuschen. So kann man Bilder von Verkehrssignalen beispielsweise durch das Hinzufügen von unscheinbarer Information, welche für den menschlichen Betrachter den Sinn des Bildes nicht verändern, derart manipulieren, dass der KI-Algorithmus diese völlig falsch interpretiert (Ackerman 2017). Das linke Bild unten wurde in einem Beispiel als 45-mph-Schild klassifiziert, das rechte Bild als Stopp-Signal (Abbildung 7). In ethischer Hinsicht erschwert dies die Sicherheitsprüfung (kann das System jemanden schädigen) sowie die Abschätzung der Fairness automatisierter Entscheidungen sowie die Zuschreibung von Verantwortung (siehe dazu Abschnitt 2.4).

Abbildung 7: Verfremdete Bilddaten, welche bei einem KI-System zu Fehlklassifikationen führten (Bildquelle: Eykholt 2018).

Derartige Erkenntnisse motivieren das neue Forschungsgebiet *explainable AI* (erklärbare KI), das insbesondere für unbeaufsichtigtes maschinelles Lernen relevant ist. Dazu existieren mehrere Ansätze. Eine Forschungsgruppe am deutschen Fraunhofer Heinrich-Hertz-Institut in Berlin beispielsweise entwickelt in Zusammenarbeit mit der Technischen Universität Berlin eine Art «Gehirnscan» für künstliche Intelligenz, die sogenannte *Layer-wise Relevance Propagation* (LRP; Binder et al. 2016). Diese Methode lässt den «Denkprozess» neuronaler Netze rückwärts ablaufen und macht so sichtbar, an welcher Stelle welche Gruppen von künstlichen Neuronen bestimmte Entscheidungen getroffen und wie stark diese zum Endergebnis beigetragen haben (Beuth 2017). Am Massachusetts Institute for

Technology untersuchen Forscher eine Blackbox, indem sie die Inputs – Satzteile oder Bildbereiche – immer wieder leicht verändern und beobachten, welchen Einfluss das auf den Output hat (Alvarez-Melis & Jaakkola 2017). Diese Methode ist z.B. geeignet, um einseitige Trainingsdaten oder im Algorithmus codierte Tendenzen zu identifizieren. Derartige Forschung wird wichtiger werden, um technische Risiken von KI-Systemen in konkreten Anwendungen zu identifizieren.

2.3. Die internationale Debatte zu KI[10]

Die beeindruckenden Fortschritte der letzten Jahre in der Entwicklung von KI haben jüngst zu einer geradezu fieberhaften Aktivität von staatlichen Organisationen, Berufsverbänden und Institutionen geführt, die sich mit den gesellschaftlichen Folgen von KI auseinandersetzen. Daraus resultierte eine nahezu unüberschaubare Anzahl an Initiativen und Bemühungen, welche die Entwicklung von KI kanalisieren und in geordnete Bahnen lenken wollen. In diesen spiegeln sich Hoffnungen und Befürchtungen in Bezug auf diese Entwicklung wider.

Die folgende Aufzählung ist nicht vollständig, sondern beispielhaft. Die Beispiele wurden basierend auf der politisch-gesellschaftlichen Bedeutung der Akteure oder eines herausragenden inhaltlichen Merkmals ausgewählt.

2.3.1. Internationale Initiativen

2.3.1.1. Die Vereinten Nationen

Die Vereinten Nationen haben die gesellschaftliche, politische und wirtschaftliche Relevanz von KI-Technologien erkannt und entsprechende Initiativen zur positiven Nutzung ihrer Potenziale entwickelt. Ein wesentliches Ziel dabei ist es, den möglichen Beitrag der KI zur Realisierung der Ziele für nachhaltige Entwicklung, besser bekannt unter der Abkürzung SDG (*Sustainable Development Goals*),[11] auszu-

[10] Dieser Abschnitt beruht auf Arbeiten von Johann Čas und Jaro Krieger-Lamina, Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften. Sie wurden bei den Recherchen von Susanne Hollin unterstützt.

[11] Siehe: https://sustainabledevelopment.un.org/.

loten und aktiv zu nutzen. Diese «AI for Good»-Initiative[12] wird von der International Telecommunications Union (ITU) koordiniert. Die ITU ist die auf Information- und Kommunikationstechnologien spezialisierte Teilorganisation der Vereinten Nationen. Etwa 30 weitere Teilorganisationen der Vereinten Nationen sind an der konkreten Umsetzung der Initiative beteiligt (ITU 2018).

2.3.1.2. Internationale Organisation für Arbeit

Die Internationale Organisation für Arbeit (*International Labour Organization*, ILO) hat sich in ihrem «Bericht zur Zukunft der Arbeit» mit dem Thema KI auseinandergesetzt. Dieser wurde von einer breiten Kommission 2017 bis 2019 erarbeitet (ILO 2019). Besonders hervorgehoben wird darin die Notwendigkeit, Autonomie und Kontrolle der Menschen über KI-Systeme sicherzustellen. Des Weiteren müsse die Datennutzung bei KI-Systemen im Arbeitskontext sowie die Verantwortlichkeit für algorithmische Prozesse geregelt werden.

2.3.1.3. Europarat

Der Europarat ist ein Forum für Debatten über allgemeine europäische Fragen. Bereits 1981 verfasste der Rat das «Übereinkommen zum Schutz des Menschen bei der automatischen Verarbeitung personenbezogener Daten» (Konvention 108), das von den 47 Mitgliedstaaten (darunter die Schweiz) und weiteren sechs Staaten ratifiziert wurde. In dem Übereinkommen wird der Schutz des Einzelnen vor Missbrauch bei elektronischer Verarbeitung personenbezogener Daten geregelt (Council of Europe 2018a). Von 2013 bis 2016 wurde ein Modernisierungsvorschlag der Konvention 108 erarbeitet (Council of Europe 2018b; man spricht von der Konvention 108+). Deren zwei Hauptziele sind die Bekämpfung datenschutzrechtlicher Probleme infolge der Nutzung neuer Informations- und Kommunikationstechnik und die Stärkung des Überwachungsmechanismus der Konvention. Es soll ein flexibler, transparenter und robuster multilateraler Rahmen geschaffen werden, um auch den grenzüberschreitenden Datenverkehr zu erleichtern und Missbrauch zu verhindern. Die wichtigsten Neuerungen wie Datenschutz durch Technikgestaltung (*privacy by design*) oder die Verpflichtung zur Meldung von Verstössen gegen den Datenschutz finden sich auch in der EU-Datenschutz-

[12] Siehe: https://aiforgood.itu.int/.

Grundverordnung. Im Zusammenhang mit KI und algorithmischer Entscheidungsfindung sind dabei von diesen Neuerungen insbesondere die zusätzlichen Garantien für die Betroffenen von Relevanz. Dazu gehören:

- das Recht, nicht einer Entscheidung unterworfen zu sein, die ausschliesslich auf automatischer Verarbeitung beruht und die Meinung des Betroffenen nicht berücksichtigt;
- das Recht zur Kenntnis der Logik des Entscheidungssystems, die als Grundlage für die Verarbeitung dient;
- das Recht, Einspruch gegen eine automatisierte Entscheidung zu erheben.

Um auf die besonderen Herausforderungen von KI Antworten geben zu können, hat das Konsultativkomitee der Konvention 108 am 25. Januar 2019 als Zusatz zur Konvention eigene Leitlinien zur künstlichen Intelligenz und zum Datenschutz veröffentlicht (Council of Europe 2019). Diese sollen sicherstellen, dass KI-Anwendungen das Recht auf Datenschutz nicht untergraben. Es wird betont, dass der Schutz der Menschenrechte, einschliesslich des Rechts auf Schutz personenbezogener Daten, eine wesentliche Voraussetzung für die Entwicklung oder Anwendung von KI-Systemen sein sollte; insbesondere wenn sie in Entscheidungsprozessen verwendet werden.

Zusätzlich zu den Bestimmungen von Konvention 108+ enthalten diese «Guidelines on Artificial Intelligence and Data Protection» eine Reihe von KI-spezifischen Empfehlungen. So sollen durch einen *Human-rights-by-design*-Ansatz mögliche Verzerrungen bei KI-Entscheidungen vermieden werden. Die Qualität, Art, Herkunft und Menge der verwendeten personenbezogenen Daten für das Training von KI-Systemen sowie mögliche negative Konsequenzen sollen prospektiv kritisch bewertet werden. Es wird angeregt, Ausschüsse von Fachleuten einzurichten und mit unabhängigen akademischen Institutionen zusammenzuarbeiten, die zur Gestaltung von menschenrechtsbasierten, ethisch und sozial orientierten KI-Anwendungen sowie zur Erkennung potenzieller Verzerrungen beitragen können. Alle Produkte und Services sollen so konzipiert sein, dass das Recht des Einzelnen gewährleistet wird, nicht ausschliesslich einer automatisierten Verarbeitung zu unterliegen. Ausserdem sollen die Betroffenen darüber informiert sein, wenn sie mit KI-Systemen interagieren.

Die Leitlinien für Gesetzgeber und Politik beinhalten die Einhaltung des Grundsatzes der Rechenschaftspflicht, die Einführung von Risikobewertungsverfahren und die Einführung von Verhaltenskodizes und Zertifizierungsmechanismen, um das

Vertrauen in KI-Produkte zu stärken. Die Rolle menschlicher Intervention in Entscheidungsprozesse und die Freiheit menschlicher Entscheidungsträger, sich nicht auf das Ergebnis der Empfehlungen der KI verlassen zu müssen, sollen erhalten bleiben. Aufsichtsbehörden sollen konsultiert werden, wenn KI-Anwendungen das Potenzial haben, die Menschenrechte und Grundfreiheiten der betroffenen Personen erheblich zu beeinträchtigen. Investiert werden sollte schliesslich in digitale Kompetenz und Bildung, professionelle Schulungen für KI-Entwickler und Forschung im Bereich der menschenrechtsorientierten KI.

2.3.1.4. Europäische Kommission

Die Europäische Kommission (EC) arbeitet auf vielen Ebenen an einer politischen Gestaltung des Themas «Künstliche Intelligenz». Ein Beispiel ist die «European AI Alliance»,[13] ein für alle offenes Forum, das eine Diskussion sämtlicher Aspekte der Entwicklungen und Auswirkungen von KI ermöglichen soll. Hervorzuheben wäre hier die von der EC eingerichtete «High Level Expert Group on Artificial Intelligence», die KI-Ethikrichtlinien erstellen soll und dabei auf den Input der KI-Allianz-Mitglieder zurückgreift (siehe Abschnitt 2.4.4). Darüber hinaus sollen die auf der Plattform geführten Diskussionen direkt zur europäischen Debatte über KI beitragen und in die Politikgestaltung der Europäischen Kommission einfliessen.

In ihrer Stellungnahme «Künstliche Intelligenz für Europa» (EC 2018) schlägt die EC einen europäischen Ansatz für den Umgang mit künstlicher Intelligenz vor, der auf drei Säulen basiert:

1. Technologieführerschaft übernehmen und die Akzeptanz im öffentlichen und privaten Sektor fördern.
2. Die Vorbereitung auf sozioökonomische Veränderungen durch KI vorantreiben.
3. Angemessene ethische und rechtliche Rahmenbedingungen sicherstellen.

Für letzteren Punkt hat die «High Level Expert Group» im Dezember 2018 einen Entwurf der KI-Ethikrichtlinien in eine Vernehmlassung gegeben; die finale Version wurde am 8. April 2019 publiziert (EC 2019). Im Bericht wird festgehalten, dass

[13] Siehe: https://ec.europa.eu/digital-single-market/en/european-ai-alliance.

vertrauenswürdige KI die Grundrechte, die geltenden Vorschriften und die europäischen Grundwerte respektieren und einem ethisch erstrebenswerten Ziel dienen sollte. Sie soll technisch robust und zuverlässig sein, da selbst bei guten Absichten ein Mangel an technischer Beherrschung unbeabsichtigten Schaden verursachen kann. Die Leitlinien richten sich an alle relevanten Interessengruppen, die KI entwickeln, einsetzen oder nutzen (Unternehmen, Organisationen, Forschende, öffentliche Dienste, Institutionen, Einzelpersonen oder andere Einrichtungen). Die Leitlinien sind primär nicht als Politikgestaltung oder Regulierung gedacht, sondern als Ausgangspunkt für eine Diskussion über eine vertrauenswürdige KI «made in Europe».

2.3.1.5. Der Europäische Datenschutzbeauftragte

Eine wichtige Institution auf europäischer Ebene ist die unabhängige Datenschutzbehörde der EU. Der (im August 2019 verstorbene) Datenschutzbeauftragte (European Data Protection Supervisor, EDPS) Giovanni Buttarelli hat zum Umgang mit künstlicher Intelligenz jüngst eine Reihe von Vorschlägen gemacht (Buttarelli 2018). Erstens sei eine Erarbeitung der ethischen Rahmenbedingungen für die Nutzung von KI dringend notwendig, wobei die Frage nach der Rechenschaftspflicht eine grosse Rolle spiele. Ausserdem müsse das öffentliche Interesse an der Nutzung grosser Datenmengen und KI geklärt werden. Zweitens müssten Regulierungen wie die Datenschutz-Grundverordnung (DSGVO) vervollständigt werden. Die DSGVO beinhalte nur ein Minimum an Standards im Umgang mit personenbezogenen Daten. Der EDPS spricht sich für eine ausgewogene Balance zwischen einer innovativen Wiederverwendung personenbezogener Daten, die der Kontrolle der Benutzenden unterliegt, und einem hohen Schutzniveau für Grundrechte aus. Der dritte Punkt ist, dass die EU alle Instrumente, insbesondere das Kartellrecht, nutzen müsse, um das Internet zu dezentralisieren, damit die Menschen mehr Freiheit und Wahlmöglichkeiten im Internet erhalten. Der EDPS betont die Gefahr, dass einige Firmen so gross werden, dass sie die Demokratie bedrohen könnten.

2.3.1.6. Die OECD

Die «Organization for Economic Cooperation and Development» (OECD) ist eine internationale Organisation und umfasst 36 Mitgliedstaaten, die meist zu den Ländern mit hohem Pro-Kopf-Einkommen gehören. In Zusammenhang mit künstlicher

Intelligenz (KI) besonders erwähnenswert ist das Projekt «Going Digital».[14] Dieses soll politischen Entscheidungsträgern in allen relevanten Politikbereichen helfen, die digitale Revolution besser zu verstehen. Ein Teil dieses Projekts setzt sich mit KI auseinander. Am 22. Mai 2019 hat die OECD die «Recommendation of the Council on Artificial Intelligence» beschlossen. Die G-20-Staaten haben am 9. Juni darauf basierende Prinzipien für den Umgang mit KI beschlossen.

Die Empfehlungen der OECD sind:

- KI sollte den Menschen und der Erde durch Wachstum und nachhaltige Entwicklung nützen.
- KI-Systeme müssen so gestaltet werden, dass sie Gesetze, Menschenrechte, demokratische Werte und Diversität wahren. Darüber hinaus sollten sie angemessene Sicherheitsvorkehrungen vorsehen, z.B. die Intervention durch Menschen, um eine gerechte Gesellschaft sicherzustellen.
- Es sollte auf verantwortungsvolle Art Transparenz hergestellt werden, um Menschen zu ermöglichen, die Systeme zu verstehen und deren Entscheidungen zu hinterfragen.
- KI-Systeme müssen auf eine sichere und robuste Art und Weise funktionieren. Potenzielle Risiken sollten kontinuierlich bewertet werden.
- Organisationen und Individuen, die KI entwickeln, vertreiben, implementieren oder betreiben, sollten dafür verantwortlich sein, dass die Systeme nach den zuvor beschriebenen Prinzipien arbeiten und funktionieren.

2.3.1.7. IEEE

Das Institute of Electrical and Electronics Engineers (IEEE) ist der weltweit grösste technische Berufsverband; er widmet sich dem technischen Fortschritt zum Wohle der Menschheit und zählt über 420 000 Mitglieder in mehr als 160 Ländern. Die «IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems»[15] wurde 2016 gegründet. Diese Initiative möchte Akteure aus unterschiedlichen technologischen und wissenschaftlichen Communities zusammenbringen, um aktuelle Themen zu erkennen und einen Konsens dazu zu fördern (IEEE

[14] Siehe: https://www.oecd.org/going-digital/.

[15] Siehe: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

2017). Die IEEE Global Initiative soll sicherstellen, dass alle Interessengruppen, die an der Entwicklung und dem Design autonomer und intelligenter Systeme beteiligt sind, ausgebildet, geschult und befähigt werden, ethische Überlegungen zu priorisieren, damit diese Technologien zum Wohle der Menschheit weiterentwickelt werden.

Beruhend auf den Prinzipien der Menschenrechte, des Wohlergehens, der Rechenschaftspflicht, der Transparenz und der Sensibilisierung für Missbrauch haben mehrere Hundert internationale Mitglieder aus der IEEE Global Initiative einen Kodex zum Umgang mit «autonomous and intelligent systems» (gemeint sind technische Systeme, die weitgehend ohne die Notwendigkeit menschlicher Intervention ihre Aufgaben erfüllen können) entworfen. Dieses Dokument ist nicht als verbindliche Verhaltensnorm oder Berufsethik gedacht, auch nicht als politische Positionierung oder formaler Report der IEEE. Es soll vielmehr als Arbeitswerkzeug und Nachschlagewerk dienen. Diese Arbeit zielt darauf ab, eine öffentliche Diskussion voranzutreiben sowie die Schaffung von Standards und den damit verbundenen Zertifizierungsprogrammen zu unterstützen.

Neben diesem Kodex unter dem Titel «Ethically Aligned Design» identifiziert und empfiehlt die IEEE Global Initiative Projekte zu Standards im ethischen Umgang mit «autonomous and intelligent systems». Das Ziel der IEEE Global Initiative ist es, dass «Ethically Aligned Design» und die davon inspirierten IEEE-Standards Einblicke und Empfehlungen liefern, die in den kommenden Jahren zu einer wichtigen Referenz für die Arbeit von KI-Entwicklern werden.

2.3.2. Nationalstaatliche Strategien

Auch auf nationalstaatlicher Ebene findet sich eine Vielzahl an Initiativen, die hier nur summarisch dargestellt werden (siehe Tabelle 1 auf der Folgeseite).

Speziell bei den Länderinitiativen zeigen sich – neben dem Konsens über die erwünschte Nützlichkeit von KI für die Menschen, sowohl individuell als auch auf gesellschaftlicher Ebene – auch andere Schwerpunkte. So legen manche Staaten den Fokus auf die nationale Wirtschaftsentwicklung und/oder Exzellenz in der Forschung, andere wollen über die Implementierung in der staatlichen Verwaltung einen Nutzen für die Bürgerinnen und Bürger schaffen, und wieder andere wollen auf dem Gebiet erklärterweise eine internationale Vorreiterrolle übernehmen, somit auch nationale Anstrengungen wie Wirtschaftsförderung und Forschungsfinanzierung in diese Richtung lenken.

Hervorzuheben ist hier die KI-Initiative Chinas, die von der *Beijing Academy of Artificial Intelligence* (BAAI) verabschiedet wurde, weil sie dem sehr nahekommt, was es an europäischen Initiativen zu dem Thema gibt. So wird beispielsweise der Schutz der Privatsphäre, der Würde des Menschen und der Schutz von Minderheiten vor Diskriminierung hervorgehoben, was von westlichen Beobachtern so nicht erwartet wurde und als «Gesprächsangebot» an andere Staaten gedeutet wird, um vielleicht eine weltweit einheitliche Regelung zum Thema KI voranbringen zu können. Es bestehen hier aber auch Widersprüche zur faktischen Nutzung von KI-Technologien in China (etwa für das «Sozialkreditsystem») und deren Einsatz für Massenüberwachung und Unterdrückung von Minderheiten in der Provinz Xinjiang im Westen Chinas.

Tabelle 1: Summarischer Überblick inhaltlicher Schwerpunkte nationaler KI-Strategien.

Beispiel-länder	Führung in Forschung	Förderung Wirtschaft & Wachstum	Klärung des rechtlichen Rahmens	Ethische Rahmen-bedingungen	Orientierung an Gemein-wohl	Transparenz & Nachvoll-ziehbarkeit	Sicherheit	Öffentliche Diskussion
China				X	X	X	X	
Dänemark	X	X		X	X			
Deutschland	X		X	X	X	X		
Frankreich	X			X	X	X		X
Grossbritannien				X	X	X		
Luxemburg	X	X			X			
Russland	X	X	X					
Singapur				X		X		
USA		X	X	X	X	X	X	X
Verein. Arab. Emirate	X	X			X			

2.3.3. Akademische Initiativen: Montrealer Erklärung

Viele Forschungseinrichtungen beschäftigen sich weltweit mit der Entwicklung von künstlicher Intelligenz – entweder aus technischer Sicht oder aus einer Metaperspektive, als Forschung über die Entwicklung der Technik, der Rahmenbedin-

gungen, die Technikfolgen oder Ähnliches. Aufgrund der stark partizipativ angelegten Entwicklungsmethode sticht die Montrealer Erklärung aus der Menge der akademischen Initiativen heraus.

Diese «Montrealer Erklärung für eine verantwortungsvolle Entwicklung von künstlicher Intelligenz» ist eine Initiative der Universität Montreal.[16] Die Arbeiten an der Erklärung wurden 2017 gestartet; sie zielt darauf ab, die öffentliche Debatte anzuregen und eine progressive und integrative Ausrichtung der Entwicklung von KI zu fördern. Ziele sind es, einen ethischen Rahmen zu gestalten, um die technologische Entwicklung dahin gehend auszurichten, dass alle davon profitieren, sowie ein nationales und internationales Forum zu eröffnen, in dem Diskussionen über eine gerechte, integrative und ökologisch nachhaltige KI-Entwicklung stattfinden können.

Die Montrealer Erklärung umfasst zehn Prinzipien. Diese Prinzipien wurden von einer Gruppe von Fachpersonen für Ethik, Recht, öffentliche Ordnung und künstliche Intelligenz in einem Zeitraum von drei Monaten und im Austausch mit über 500 Bürgern, Experten und Interessengruppen erarbeitet. Obwohl diese Prinzipien global anwendbar sind, ist die Erklärung regional ausgerichtet.

Im Weiteren erwähnenswert ist die Initiative *AI4People* (Floridi et al. 2018), eine vom Atomium – European Institute for Science, Media and Democracy – ausgehende Initiative zur Etablierung eines Forums über die sozialen Auswirkungen der künstlichen Intelligenz.

2.3.4. Private Initiativen

Viele NGOs und NPOs beschäftigen sich mit dem Einsatz von KI. Im Folgenden sind einige davon beispielhaft dargestellt.

2.3.4.1. The Toronto Declaration

Im Fokus dieser Erklärung steht der Schutz des Rechts auf Gleichheit und Nichtdiskriminierung von Einzelpersonen und Gruppen in Systemen des maschinellen Lernens. Gestartet wurde sie im Mai 2018 von den Organisationen Amnesty Inter-

[16] Siehe: https://www.montrealdeclaration-responsibleai.com/.

national und Access Now (2018). Beide setzen sich für die Einhaltung der Menschenrechte ein, Access Now vor allem im digitalen Bereich. Die in der Erklärung dargelegten Menschenrechtsstandards sollen eine Grundlage für die Entwicklung ethischer Rahmenbedingungen im Bereich des maschinellen Lernens bilden.

2.3.4.2. The Future Society

Die Future Society koordinierte eine siebenmonatige öffentliche Bürgerbefragung, um den Aufstieg, die Dynamik und die Folgen von KI besser zu verstehen. Über die Onlineplattform Assembl und durch 20 globale Veranstaltungen wurde von September 2017 bis März 2018 eine vielfältige Community von über 2000 Teilnehmenden aus der ganzen Welt mit über 3300 Beiträgen in fünf Sprachen (Englisch, Chinesisch, Französisch, Japanisch und Russisch) zusammengestellt, um neue Perspektiven für die gesellschaftliche Steuerung von KI zu entwickeln. Ziel ist u.a. die Umsetzung einer globalen Politikempfehlung auf Grundlage der Beiträge.

2.3.4.3. OpenAI

OpenAI ist eine Non-Profit-Forschungseinrichtung, die 2015 in San Francisco gegründet wurde (Brockman et al. 2015). Die Mission von OpenAI ist es, eine sichere KI zu entwickeln und zu gewährleisten, dass die Vorteile von KI so breit und gleichmässig wie möglich verteilt werden und der Menschheit als Ganzes Nutzen bringen. OpenAI wird von Einzelpersonen und Unternehmen gesponsert. Zu den Einzelpersonen zählen Elon Musk, Sam Altman, Greg Brockman, Reid Hoffman, Jessica Livingston und Peter Thiel. Die beteiligten Unternehmen sind YC Research, Infosys, Microsoft, Amazon Web Service und das Open Philantropy Project.

2.3.4.4. The Public Voice – AI Universal Guidelines

Die Public Voice Coalition wurde 1996 vom Electronic Privacy Information Center gegründet, um die Beteiligung der Öffentlichkeit an Entscheidungen über die Zukunft des Internets zu fördern. Im Oktober 2018 veröffentlichte die Organisation Richtlinien für den Umgang mit KI.[17] Ziel ist es, dass diese Leitlinien in ethische Normen aufgenommen, in nationales Recht und internationale Vereinbarungen

[17] Siehe: https://thepublicvoice.org/ai-universal-guidelines/.

einfliessen und in die Gestaltung von Systemen integriert werden. Vor allem Transparenz und Rechenschaftspflicht für diese Systeme sollen gefördert werden. Ein weiterer Schwerpunkt liegt darauf, sicherzustellen, dass die Menschen die Kontrolle über die von ihnen geschaffenen Systeme behalten. Die Richtlinien bauen auf früheren Arbeiten aus Wissenschaft, Thinktanks, NGOs und internationalen Organisationen auf.

2.3.4.5. SwissCognitive

SwissCognitive ist eine Schweizer Initiative, die als «Global AI Hub» eine Plattform für den Wissens- und Erfahrungsaustausch und die Diskussion über die Entwicklung, Ergebnisse und Auswirkungen von KI-Technologien bietet. Als Reaktion auf die Komplexität von KI sowie ihrer Potenziale und Risiken bringt SwissCognitive Branchen, Unternehmen, Führungskräfte und Technologiefachleute zusammen und schafft eine gemeinsame Basis für den Austausch.

2.3.4.6. Kommerzielle Unternehmen

Zahlreiche Firmen beschäftigen sich ebenfalls mit der Entwicklung und dem Einsatz von KI, was auch zu einigen Initiativen bezüglich der Gestaltung der Auswirkungen von KI auf die Gesellschaft geführt hat. Weltweit dominierende Akteure aus der IT-Industrie wie Amazon, Facebook, Google, IBM und Microsoft haben sich 2018 zu einer «Partnership on AI» zusammengeschlossen, um Best-Practice-Beispiele für Anwendungen künstlicher Intelligenz zu studieren und zu entwickeln, das öffentliche Verständnis über KI zu fördern und als Plattform für die Diskussion von gesellschaftlichen und ethischen Problemen zu dienen.[18]

Singapurs Zentralbank hat als Hilfestellung eine Reihe von Grundsätzen für den Einsatz von KI und Datenanalysen bei der Entscheidungsfindung in Bezug auf die Bereitstellung von Finanzprodukten und -dienstleistungen veröffentlicht. Die Grundsätze behandeln die Themen Fairness, Verantwortung, Ethik und Transparenz (Monetary Authority of Singapore 2018).

[18] Siehe: https://www.partnershiponai.org/.

2.3.5. Zusammenfassung

Diese Zusammenstellung macht deutlich, dass trotz der Vielzahl an Initiativen überlappende Ziele erkennbar sind. Einen hervorragenden Überblick über die verschiedenen Inhalte und deren Gewichtung liefern dazu Fjeld et al. (2019). Sie haben 32 Richtlinien für KI untersucht und darin 47 Prinzipien gefunden, die in acht Kategorien gruppiert wurden. Dadurch war es möglich, (auch grafisch) zu veranschaulichen, in welchen Punkten sich die verschiedenen Papiere unterscheiden und worin weitgehend Einigkeit zu bestehen scheint. Die daraus abzuleitenden Erkenntnisse decken sich weitgehend mit der Analyse in dieser Studie.

Mehr als bei anderen technologischen Entwicklungen der jüngeren Vergangenheit stellen nicht kommerzielle Akteure in ihren Richtlinien für die Entwicklung von und den Umgang mit KI die Würde des Menschen in den Mittelpunkt. Die Technik soll zum Wohle der Menschen genutzt werden und einen Beitrag zur positiven Entwicklung der Gesellschaft(en) leisten. Die schon eingangs angesprochene Rückbesinnung auf die Grundlagen des Menschseins lassen sich vermutlich damit erklären, dass der Technologie zugetraut wird, das Selbstverständnis von uns Menschen tiefgreifend zu verändern.

Die der KI zugeschriebene Macht löst Befürchtungen und den Ruf nach Kontrolle und Transparenz aus; nicht nur nach Transparenz in den Entscheidungsgrundlagen, sondern auch nach Transparenz über den Einsatz von KI. Noch öfter werden aber Verantwortung und Zurechenbarkeit gefordert. Das umschliesst vor allem die Fragen nach Haftung, Erklärbarkeit, Vorhersagbarkeit und Monitoring. Aber auch der Schutz der Privatsphäre und die informationelle Selbstbestimmung haben in den verschiedenen Dokumenten hohes Gewicht.

Generell lässt sich feststellen, dass alle Dokumente den Menschen in den Mittelpunkt stellen und darauf abzielen, die Kontrolle des Menschen über die Maschinen sicherzustellen und seine Würde zu schützen. Dies scheint Konsens zu sein.

2.4. Generelle ethische Aspekte von KI[19]

Wie die Darstellung der internationalen Initiativen (siehe Abschnitt 2.3) deutlich gemacht hat, spielen ethische Aspekte eine prominente Rolle in der Debatte um die gesellschaftlichen Auswirkungen von KI. Das thematische Spektrum ist dabei sehr breit, und die Übergänge zu völkerrechtlichen Aspekten wie z.B. den Menschenrechten sowie zu rechtlichen Prinzipien wie z.B. Transparenz und Privatsphäre sind fliessend. Eine kürzlich erschienene Übersichtsarbeit (Jobin et al. 2019) zu 84 veröffentlichten Richtlinien zu AI zeigt zwar eine globale Konvergenz zu fünf ethischen Prinzipien (Transparenz, Gerechtigkeit und Fairness, Nichtschaden, Verantwortung und Privatsphäre). Es bestehen aber erhebliche Unterschiede in Bezug darauf, wie diese Prinzipien interpretiert werden, warum sie als wichtig erachtet werden, zu welchem Thema, Bereich oder zu welchen Akteuren sie gehören und wie sie umgesetzt werden sollten.

Auch in der Technikethik hat sich das Interesse an ethischen Fragen, die von KI aufgeworfen werden, in jüngster Zeit verstärkt. Darüber hinaus hat die Zusammenarbeit zwischen Fachleuten aus Computerwissenschaften und Ethik zugenommen, z.B. zur Frage, was es bedeutet, dass Algorithmen «fair» sein sollen.

Dieser Abschnitt liefert eine Einführung in einige der aktuellen Debatten mit Fokus auf jene Aspekte, die für die neueren KI-Technologien relevant sind. Nach einer Übersicht über ethische Kernthemen gemäss dem Vorschlag der *European Group of Ethics in Science and New Technologies* folgen Erläuterungen zur «Fairness» von KI-Entscheidungen und zur Frage, wie sich KI künftig auf Verantwortung und Rechenschaftspflichten in ethischer Hinsicht auswirken könnten. Der Abschnitt schliesst mit einer detaillierteren Darstellung der aktuell von der *High Level Expert Group on Artificial Intelligence* der EU erstellten KI-Ethik-Richtlinien (siehe dazu auch Abschnitt 2.3.1.4).

Das mögliche Feld von ethischen Fragen von KI ist damit nicht vollständig abgedeckt. In ethisch-anthropologischer Hinsicht kann man sich beispielsweise fragen, in welchem Sinn die Nutzung von KI das Handlungsspektrum des Individuums erweitert und damit als eine Form des «human enhancement» angesehen werden kann. Eng damit verbunden ist auch die Frage, ob aus der heutigen KI heraus sich

[19] Dieser Abschnitt beruht auf Arbeiten von Markus Christen und Markus Kneer von der Digital Society Initiative der Universität Zürich sowie von Johann Čas und Jaro Krieger-Lamina vom Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften.

dereinst eine «Superintelligenz» entwickeln könnte, die dann entweder in Konkurrenz zu den Menschen treten oder aber die Machtausübung einer Elite perfektionieren würde. Solche Fragestellungen werden oft im Kontext der «starken KI» diskutiert und die entsprechenden Antworten sind von spekulativen Annahmen geprägt. Sie fallen wie bereits erwähnt nicht in den Fokus dieser Studie.

2.4.1. Ethische Kernthemen gemäss der EGE

Die Europäische Gruppe für Ethik der Naturwissenschaften und der Neuen Technologien (EGE) ist ein unabhängiges, multidisziplinäres Organ, das die Europäische Kommission seit 1991 zu allen Aspekten der Politik und Gesetzgebung berät, in denen ethische, gesellschaftliche und Grundrechtsaspekte mit der Entwicklung von Wissenschaft und neuen Technologien zusammenhängen. Jeder EU-Mitgliedstaat hat einen nationalen Ethikrat oder eine gleichwertige Einrichtung, ebenso wie viele Drittländer. Die EGE fungiert dabei als wichtiger Bezugspunkt für die 28 nationalen Ethikräte in der EU.

In ihrer Erklärung zu künstlicher Intelligenz, Robotik und autonomen Systemen (EGE 2018) setzt sich die EGE dafür ein, dass ein Prozess zur Schaffung eines gemeinsamen internationalen ethischen und rechtlichen Rahmens für die Produktion, Konstruktion, Verwendung und Steuerung von KI, Robotik und autonomen Systemen eingeleitet werden soll. Basierend auf einer Auswertung bestehender Initiativen schlägt die EGE folgende Klassifizierung ethischer Orientierungspunkte für KI vor, welche die bisherige Debatte gut abbildet:

- **Die Würde des Menschen:** Menschen sollen wissen, ob und wann sie mit einer Maschine oder einem anderen Menschen interagieren und das Recht haben, bestimmte Aufgaben einer Maschine oder aber einem Menschen zu übertragen.
- **Autonomie:** Transparenz und Vorhersagbarkeit von Handlungen/Handlungsdispositionen autonomer Systeme sollen es Menschen ermöglichen, in Systeme einzugreifen, wenn sie dies aus moralischen Gründen für notwendig erachten.
- **Verantwortung:** Systeme sollten dem Wohl der Gesellschaft und der Umwelt dienen. Zum Schutz grundlegender Werte sollten Forschung, Konstruktion und Entwicklung von KI einem *Ethics-by-Design*-Ansatz folgen, d.h. gewisse Grundwerte sollten sich im Design der Systeme widerspiegeln.

- **Gerechtigkeit, Gleichbehandlung und Solidarität:** KI sollte zu globaler Gerechtigkeit und gleichberechtigtem Zugang zu Nutzen und Vorteilen beitragen. Diskriminierende Verzerrungen sind zu vermeiden, und die entsprechende Ausbildung zu diesen Fragen soll speziell in den MINT-Fächern gefördert werden.
- **Demokratie:** Schlüsselentscheidungen sollten das Ergebnis demokratischer Debatten und öffentlichen Engagements sein. Dazu ist ein öffentlicher Dialog notwendig, in dem jeder über die Risiken und Chancen aufgeklärt ist.
- **Rechtsstaatlichkeit und Rechenschaftspflicht:** Investitionen in Lösungen für eine faire und eindeutige Zuordnung von Verantwortlichkeit und wirksame rechtsverbindliche Mechanismen, inklusive Systeme zur Schadensbegrenzung, sind notwendig.
- **Sicherheit, Schutz, körperliche und geistige Unversehrtheit:** Die Sicherheitsdimensionen «äussere Sicherheit für ihre Umgebung und die Benutzer», «Zuverlässigkeit und interne Belastbarkeit» und «psychische Sicherheit im Zusammenhang mit der Interaktion zwischen Mensch und Maschine» müssen von KI-Entwicklern berücksichtigt werden.
- **Datenschutz und Privatsphäre:** In Bezug auf die Sammlung von Daten müssen die Datenschutzvorschriften eingehalten werden und die Rechte auf Schutz personenbezogener Daten und Schutz der Privatsphäre im Vordergrund stehen. Zwei neue Rechte sollen mit einbezogen werden – «das Recht auf sinnvollen zwischenmenschlichen Kontakt» und das «Recht, nicht profiliert, gemessen, analysiert, angeleitet oder angestossen zu werden».
- **Nachhaltigkeit:** Der Mensch soll grundlegende Voraussetzungen für das Leben auf unserem Planeten sicherstellen, das Wohlergehen der Menschheit schützen und die Umwelt für künftige Generationen bewahren. Entsprechend sollen Strategien zur Verhinderung der negativen Auswirkungen künftiger Technologien auf das menschliche Leben und die Natur den Vorrang von Umweltschutz und Nachhaltigkeit sicherstellen (EGE 2018, S. 22). Die Vereinten Nationen haben dieses Anliegen in Form der Ziele für nachhaltige Entwicklung konkretisiert.[20]

[20] Siehe dazu auch die Ausführungen der Nachhaltigkeitskommission der Universität Zürich; zugänglich unter: https://www.uzh.ch/de/about/basics/sustainability.html.

Die EGE fordert systematische Überlegungen und Forschungsinitiativen zu diesen ethischen, rechtlichen und governance-relevanten Aspekten von KI-Systemen, die in der Lage sind, ohne menschliche Kontrolle zu agieren. Ausserdem wird die Wichtigkeit eines öffentlichen Engagements und einer öffentlichen Auseinandersetzung mit dem Thema thematisiert. Die EGE fordert die Kommission auf, eine Untersuchung über bestehende Rechtsinstrumente einzuleiten, um den dargelegten Problemen wirksam zu begegnen und gegebenenfalls neue Steuerungs- und Regelungsinstrumente einzuführen. Eine Gefahr sieht die EGE schliesslich in der Verlagerung der Entwicklung und Nutzung von KI in jene Regionen, die niedrigen ethischen Standards folgen, und erkennt deshalb Handlungsbedarf.

2.4.2. KI-Entscheidungen und Fairness

Eine zunehmende Nutzung von KI-Technologien zur Unterstützung oder gar Automatisierung von Entscheidungen in sensitiven Bereichen wie Strafverfolgung, Mitarbeiterbeurteilung oder Diagnosen in der Medizin wirft unweigerlich die Frage nach der «Fairness» bzw. der Diskriminierung allfälliger Personengruppen auf (Danaher 2016, 2017; Binns 2018; Just & Latzer 2017). Die Befürchtung ist hier, dass sich bestehende Formen sozialer Ungleichheit und Diskriminierung durch solche Praktiken verstärken oder gar neue Formen von nicht gerechtfertigter Ungleichbehandlung entstehen könnten (Barocas & Selbst 2016; Ferguson, 2017).

In rechtlicher Hinsicht gibt es durchaus eine Handhabe gegen «direkte» Diskriminierung, z.B. wenn das Geschlecht als Kriterium für eine Stellenvergabe verwendet würde. Technisch gesehen kann man direkte Diskriminierung verhindern, indem man den Algorithmus derart gestaltet, dass er bestimmte Datencharakteristika (z.B. Informationen zu Geschlecht oder sozialem Status) systematisch ignoriert, was zuweilen die Genauigkeit des Algorithmus beeinflusst (Hardt et al. 2016). Es ist dann aber möglich, dass diese Merkmale aus anderen Merkmalen, mit denen sie korreliert sind, abgeleitet werden können (Pedreschi et al. 2008; Barocas & Selbst 2016).

Das Problem liegt aber noch tiefer, wie jüngere Forschungen gezeigt haben (Dwork et al. 2011). KI-Systeme, die für Entscheidungen herangezogen werden, können beispielsweise so programmiert werden, dass Mitglieder verschiedener Gruppen (z.B. Männern und Frauen) die gleiche Wahrscheinlichkeit auf eine positive Vorhersage (z.B. ein guter Kandidat für eine Stelle) haben. Das Problem ist

nun aber, dass die Idee, «Mitgliedern verschiedener Gruppen die gleiche Behandlung anzubieten», mehrere mögliche Interpretationen hat. Anstatt eine «gleiche Wahrscheinlichkeit auf eine positive Vorhersage» zu fordern, kann man z.B. verlangen, dass sich der Anteil der Klassifizierungsfehler erster und zweiter Ordnung (also: x wird fälschlicherweise einer Gruppe G zugeordnet, ein Fehler erster Ordnung, bzw. fälschlicherweise nicht zugeordnet, ein Fehler zweiter Ordnung) über diskriminierungsrelevante Gruppen nicht unterscheiden dürfen (Berk et al. 2017). So könnte beispielsweise verlangt werden, dass ein KI-System zur Beurteilung von Anstellungen gut geeignete Frauen gleich wahrscheinlich fälschlicherweise aussondert wie gut geeignete Männer – solche Klassifizierungsfehler wird es schliesslich immer geben. Mathematische Überlegungen zeigen allerdings, dass sich gewisse Anforderungen an Algorithmen (z.B. bezüglich Genauigkeit und Fairness) prinzipiell nicht gleichzeitig erreichen lassen (Berk et al. 2017; Kleinberg et al. 2016). Selbst gutmeinende Entscheidungsträger stehen bei der Definition eines diskriminierungsfreien Prädiktors deshalb vor schwierigen Entscheidungen.

Der öffentlich diskutierte Fall des COMPAS-Algorithmus (siehe auch Abschnitt 3.5.2.2) veranschaulicht dieses Thema in der Praxis.[21] ProPublica, eine Organisation für investigativen Journalismus in den USA, hat im Jahr 2016 eine Analyse zu einem Instrument namens COMPAS der US-Firma Northpointe publiziert. Dieses liefert Richterinnen und Richtern, die über eine frühzeitige Haftentlassung einer Person im Strafvollzug zu befinden haben, eine Einschätzung über dessen Rückfallprognose (Chouldechova 2016). ProPublica hat kritisiert, dass COMPAS rassistische Prognosen machen würde, obgleich die Information, welcher Rasse diese Person angehört, nicht in die Berechnung einfliesst: Das System attestiert Afroamerikanern, die keinen Rückfall begehen, eine fast doppelt so hohe Rückfallquote wie Weissen. Der Algorithmus verletzt damit die folgende Intuition von Fairness: Menschen, die nicht rückfällig werden, sollten unabhängig von der Rasse die gleiche Wahrscheinlichkeit haben, dass ihnen die Bewährung (zu Unrecht) verweigert wird.

Diese Analyse machte Schlagzeilen. Wie kann es sein, dass Menschen, die bezüglich der für das Problem relevanten Eigenschaft gleich sind – nämlich nicht rückfällig zu werden –, derart unterschiedlich beurteilt werden? Also weniger häufig als risikoarm eingestuft werden, wenn sie schwarz sind?

[21] Die folgenden Ausführungen basieren auf Loi (2018).

Ethisch ist die Verwendung statistischer Prognosen für solche Fälle gerechtfertigt, vorausgesetzt, dass die Prognose unvoreingenommen ist. Aber was genau bedeutet «unvoreingenommen»? ProPublica verband mit der Analyse die Vermutung, dass das Resultat auf eine diskriminierende Praxis hinweist, beispielsweise weil mehrheitlich weisse Softwareentwickler/-innen zu wenig Sorgfalt bei der Programmierung des Algorithmus walten liessen und vielleicht gar implizit rassistisch gewesen seien. Wäre dies der Grund gewesen, so wäre dies ein Fehlverhalten mit gravierenden Konsequenzen für das Leben der Betroffenen. Man könnte das Problem dann etwa durch bessere Schulung oder mehr Diversität bei den Fachpersonen angehen.

Die wahre Ursache des Problems ist aber komplizierter. Die Mathematiker, die das Problem nach den ProPublica-Enthüllungen analysiert hatten, konnten zeigen, dass eine Form diskriminierender Verzerrung unvermeidlich ist – selbst wenn man mit bestem Gewissen programmiert und die Daten frei von Verzerrung sind (Chouldechova 2016). COMPAS wurde auf Diskriminierung getestet – doch man erfüllte ein anderes Kriterium für Fairness: Personen, denen die Bewährung verweigert (oder gewährt) wird, sollten die gleiche Wahrscheinlichkeit auf Rückfall haben. Dies erreicht man mittels sogenannter Kalibrierung.

Gemeint ist damit Folgendes: Wenn der Algorithmus z.B. eine Rückfallwahrscheinlichkeit (Score) von 0,3 für einen Straftäter angibt, dann ist damit die Erwartung verbunden, dass im Schnitt 30 % aller Personen, bei denen der Algorithmus diesen Wert ermittelt, erneut eine Straftat begeht. Die Erwartung ist, dass dies auch für alle Untergruppen der betroffenen Personengruppen gilt: Wenn also von allen weissen Straftätern mit einem Score von 0,3 deren 30 % tatsächlich rückfällig werden, bei den schwarzen Straftätern mit demselben Score von 0,3 aber deren 50 %, dann würde dies als diskriminierend gelten – der gleiche Risiko-Score hätte je nach Rasse eine unterschiedliche Bedeutung. Um dem entgegenzuwirken, müsste man je nach Rasse unterschiedliche Scores als Schwellenwert für eine Entlassung definieren – was offensichtlich rassistisch wäre. Um das zu vermeiden, führt man eine solche Kalibrierung durch, die sicherstellt, dass die Risiko-Scores für jede sensitive Gruppe (in diesem Fall die Rasse) die gleiche statistische Bedeutung haben. Es hat sich also gewissermassen gezeigt, dass das verwendete KI-System «gleichzeitig fair und unfair» ist (Angwin & Larson 2016).

Dieses Problem verweist auf eine grundsätzliche Schwierigkeit, wenn KI-Systeme mit der guten Absicht entwickelt werden, gesetzliche Anforderungen wie Nichtdiskriminierung und übergeordnete ethische Werte wie Fairness zu implementieren.

Die Gesetze lassen meist Interpretationsspielraum und die ethischen Werte sind komplex und können auf unterschiedliche Weise verstanden werden. Der Diskurs um ethische Fragen liefert grundsätzlich keine endgültigen Antworten. Im Fall eines KI-Systems müssen diesbezüglich aber Festlegungen getroffen werden – ethische Intuitionen müssen gewissermassen expliziert und «festgenagelt» werden. Die Folgefrage ist dann, wer solche Entscheidungen treffen soll.

Noch viel grundlegender sollte in diesem Beispiel darüber nachgedacht und diskutiert werden, ob Entscheidungen der Justiz sich auf prognostizierte – und damit nicht ausgeführte, ja nicht einmal geplante – Handlungen von Personen stützen sollen. Dieses Grundsatzproblem stellt sich unabhängig von der Frage der Prognosegüte und einer allfälligen Diskriminierung durch den Prognosealgorithmus.

2.4.3. KI-Entscheidungen, Rechenschaft und Verantwortung

Ein zweites, grundlegendes ethisches Problem bezüglich der Nutzung von KI in (sensiblen) Entscheidungsbereichen betrifft die Frage nach der moralischen Verantwortung und der damit verbundenen Rechenschaftspflicht für die involvierten Akteure. Diese geht über das rein rechtliche Problem der Haftung (siehe dazu Abschnitt 2.5.1) hinaus. Relevant ist zum einen die Frage, wie sich die Verteilung von Verantwortung bei der Mensch-Maschine-Interaktion verändert, wenn Entscheidungskompetenz zunehmend an KI-Systeme delegiert wird. Soll man von «hybrider Verantwortung» sprechen, oder werden *responsibility gaps* auftreten, d.h. Situationen, in denen niemandem vernünftigerweise Verantwortung übertragen werden kann? Eng damit verknüpft ist die Frage, welche moralischen Verpflichtungen die Designer, Trainerinnen, Besitzer und Nutzerinnen von KI-Systemen haben.

Ein relevantes Konzept ist hier die sogenannte *meaningful human control* (bedeutungsvolle menschliche Kontrolle, MHC) von KI-Systemen, sofern diese für sensitive Entscheidungen herangezogen werden. Dieses Konzept ist vorab bezüglich des Einsatzes von KI im Sicherheitsbereich entwickelt worden, d.h. in Kontexten, in denen KI-Systeme bewaffnet sind (Crootof 2016). Aus ethischer Hinsicht interessant ist hierbei vorab der Zusammenhang zwischen Kontrolle und Verantwortung. Nach Ansicht der meisten Philosophinnen und Philosophen ist Kontrolle über das eigene Handeln eine notwendige Voraussetzung für Verantwortung (Morse 1994, Fischer 2000). Für von KI-gesteuerte autonome Systeme hat sich hierbei eine Debatte um die angebliche Öffnung von Verantwortungslücken nach dem

Einsatz solcher Systeme herauskristallisiert (Sparrow 2007, Roff 2013), obwohl das Phänomen des Auftretens solcher Lücken als solches umstritten ist (Leveringhaus 2016).

Eng mit den Fragen bezüglich Verantwortung und Kontrolle verbunden ist jene des Vertrauens in KI-Systeme. Das Konzept des Vertrauens wird in verschiedenen Disziplinen (Philosophie, Psychologie, *Behavioral Economics*) untersucht. Was KI-Systeme betrifft, kommen die unterschiedlichen Blickwinkel im recht neuen Bereich *human-robot interaction* zusammen. Die Übersichtsarbeit von Hoff & Bashir (2015) analysiert 127 Artikel zu diesem Thema und gibt einen Überblick über jene Aspekte, die das Vertrauen von Menschen in Roboter aller Art beeinflussen können. Metaanalysen der Literatur (Hancock et al. 2011; Schaefer et al. 2016) heben hervor, dass das Vertrauen in Roboter eher von Eigenschaften Letzterer als von menschlichen und situativen Faktoren abhängt.

Die aktuelle Debatte rund um MHC entwickelt sich aktuell vorab im Kontext autonomer Waffensysteme. Der sich abzeichnende Standard von MHC zielt darauf ab, die Grundprinzipien des humanitären Völkerrechts zu wahren, d.h. Unterscheidung von Kombattanten und Nichtkombattanten, militärische Notwendigkeit und Verhältnismässigkeit des Einsatzes. Der Begriff MHC wird dabei auch kritisiert, weil er eher einen formalen als einen substanziellen Konsens verkörpert (Crootof 2016), eher politischen als rechtlichen Zwecken dient (Marauhn 2018) oder gar undefiniert ist (Canellas 2015). Rechtsexperten haben versucht, MHC mit Bedeutung zu füllen (Horowitz 2015: MHC erfordert sachkundige und bewusste Entscheidungsfindung, spezifisches Waffendesign und -testen sowie Bedienungsschulung), während Informatiker gezeigt haben, dass alle bisher vorgeschlagenen Definitionen immer noch zu einem Missverhältnis zwischen Kontrolle und Verantwortung führen könnten (Cannellas 2015), was der oben beschriebenen Verantwortungslücke von Sparrow weitere Beachtung verschafft (Sassoli 2014).

Denkt man diese Fragen in Richtung einer sich entwickelnden «starken KI» weiter, könnten sich dereinst auch Fragen nach dem moralischen Status von KI-Systemen selbst stellen. Unter welchen Umständen sollen KI-Systeme als moralische Akteure (*moral agents*) angesehen werden können, weil sie vielleicht für andere Akteure (u.a. Menschen) sorgen könnten? In die gleiche Richtung zielen Fragen, unter welchen Umständen KI-Systeme als moralisch berücksichtigenswert (*moral patients*) angesehen werden können (Bostrom & Yudkowsky 2014). Diese Themen liegen aber ausserhalb des Fokus dieser Studie.

2.4.4. Ethikrichtlinien der *High Level Expert Group*

Der bislang ambitionierteste Versuch einer Schaffung von Ethikrichtlinien für KI stammt von der *High Level Expert Group on Artificial Intelligence* der EU. Diese veröffentlichte im Dezember 2018 einen Entwurf der KI-Ethikrichtlinien, die finale Version erschien im April 2019. Im Working Paper wird erwähnt, dass vertrauenswürdige KI die Grundrechte, die geltenden Vorschriften, die Grundprinzipien und Grundwerte respektieren und einem ethisch erstrebenswerten Ziel dienen solle. Sie soll technisch robust und zuverlässig sein, da selbst bei guten Absichten ein Mangel an technischer Beherrschung unbeabsichtigten Schaden verursachen kann. Die Leitlinien richten sich an alle relevanten Interessengruppen, die KI entwickeln, einsetzen oder nutzen, wie Unternehmen, Organisationen, Forschende, öffentliche Dienste, Institutionen, Einzelpersonen oder andere Einrichtungen. Die Leitlinien sind nicht als Beitrag zur Politikgestaltung oder Regulierung gedacht, sondern als Ausgangspunkt für eine Diskussion über eine vertrauenswürdige KI «made in Europe». Das Dokument ist dabei in drei Kapitel aufgeteilt:

Grundrechte, Grundsätze und Werte: Die KI sollte mit einem ethischen Zweck entwickelt, eingesetzt und verwendet werden, der auf den Grundrechten, gesellschaftlichen Werten und den ethischen Prinzipien von Wohltätigkeit, Nichtschaden, Autonomie der Menschen, Gerechtigkeit und Erklärbarkeit beruht und diese reflektiert. Dies ist entscheidend, um auf eine vertrauenswürdige KI hinzuarbeiten. Besondere Aufmerksamkeit gilt Situationen, in denen schutzbedürftigere Gruppen wie Kinder, Menschen mit Behinderungen oder Minderheiten beteiligt sind, oder Situationen mit Macht- oder Informationsasymmetrien, wie beispielsweise zwischen Arbeitgebern und Arbeitnehmern oder Unternehmen und Konsumenten. Die negativen Folgen, die KI nach sich ziehen kann, müssen ebenso beachtet werden.

Umsetzung einer vertrauenswürdigen KI: Leitlinien für die Realisierung einer vertrauenswürdigen KI müssen bereits in der frühesten Designphase integriert werden. Es müssen Informationen über Grenzen und Fähigkeiten von KI-Systemen für Interessengruppen bereitgestellt werden, dabei spielt die Rückverfolgbarkeit eine grosse Rolle. Ausserdem sollten Interessengruppen bei der Gestaltung und Entwicklung eines KI-Systems miteinbezogen werden. Nachvollziehbarkeit und Rechenschaftspflicht von KI-Systemen sind von grosser Bedeutung und müssen beachtet und die Nachvollziehbarkeit muss erleichtert werden. Grundlegende Spannungen zwischen verschiedenen Zielen sollen dokumentiert werden (Transparenz kann die Tür zum Missbrauch öffnen; das Erkennen und Korrigieren von

Verzerrungen könnte im Widerspruch zum Datenschutz stehen). Im Weiteren werden die Notwendigkeit von Aus- und Fortbildung sowie die Förderung von Forschung und Innovation thematisiert.

Bewertung einer vertrauenswürdigen KI: Eine Bewertungsliste für vertrauenswürdige KI bei der Entwicklung, Bereitstellung oder Nutzung soll angewendet und an den spezifischen Fall angepasst werden, in dem das System eingesetzt wird. Bei der Sicherstellung einer vertrauenswürdigen KI geht es um einen kontinuierlichen Prozess der Identifizierung von Anforderungen, der Bewertung von Lösungen und der Gewährleistung verbesserter Ergebnisse über den gesamten Lebenszyklus des KI-Systems. Die Liste beinhaltet wichtige Fragestellungen, welche geklärt werden sollen, zu den Themen: Verantwortlichkeit, Data Governance, Design für alle, Steuerung der KI-Autonomie, Nichtdiskriminierung, Respekt vor der menschlichen Autonomie, Respekt vor dem Datenschutz, Robustheit, Sicherheit, Transparenz. Um die praktische Umsetzung der Bewertungsliste zu erleichtern, werden in der endgültigen Fassung vier spezielle Anwendungsfälle von KI diskutiert, die auf der Grundlage der Beiträge der Expertengruppe und der Mitglieder der Europäischen KI-Allianz ausgewählt wurden: (1) Diagnose und Behandlung im Gesundheitswesen, (2) Autonomes Fahren/Bewegen, (3) Versicherungsprämien sowie (4) Profilerstellung und Strafverfolgung.

Die Leitlinien nennen sieben Anforderungen an KI-Systeme, damit diese als vertrauenswürdig eingestuft werden können:

1. **Menschliche Kontrolle:** KI-Systeme sollte die Menschen befähigen, fundierte Entscheidungen treffen zu können und ihre Grundrechte zu fördern. Gleichzeitig müssen angemessene Aufsichtsmechanismen gewährleistet sein, die durch die Ansätze *human-in-the-loop*, *human-on-the-loop* und *human-in-command* erreicht werden sollen.
2. **Technische Robustheit und Sicherheit:** KI-Systeme müssen robust, sicher und zuverlässig sein und reproduzierbare Ergebnisse liefern. Es sollten Notfallpläne vorliegen für den Fall, dass die Systeme versagen. Damit soll sichergestellt werden, dass unbeabsichtigte Schäden minimiert bzw. verhindert werden können.
3. **Datenschutz:** Neben der Gewährleistung der uneingeschränkten Achtung der Privatsphäre und des Datenschutzes müssen Data-Governance-Mechanismen garantiert sein, die der Qualität und Integrität der Daten Rechnung tragen und einen legitimierten Zugang zu Daten gewährleisten.

4. **Transparenz:** Die Geschäftsmodelle für Daten, Systeme und KI sollten transparent sein, was unter anderem Mechanismen für Rückverfolgbarkeit beinhalten kann. KI-Systeme und ihre Entscheidungen sollen in einer Weise erläutert werden, die an die jeweiligen Interessengruppen angepasst ist. Der Mensch muss sich bewusst sein, dass er mit einem KI-System interagiert, und er muss über Möglichkeiten und Grenzen des Systems informiert sein.
5. **Vielfalt, Nichtdiskriminierung und Fairness:** Unfaire Verzerrungen müssen vermieden werden, da sie mehrere negative Auswirkungen haben könnten: von der Marginalisierung von Minderheiten bis hin zur Verschärfung von Vorurteilen und Diskriminierung. Zur Förderung der Vielfalt sollten KI-Systeme für alle zugänglich sein und relevante Interessengruppen während ihres gesamten Lebenszyklus einbeziehen.
6. **Gesellschaftliches und ökologisches Wohl:** KI-Systeme sollten allen Menschen zugutekommen, auch zukünftigen Generationen. Daher muss sichergestellt sein, dass sie nachhaltig und umweltfreundlich sind. Darüber hinaus sollten sie die Umwelt berücksichtigen, und ihre gesellschaftlichen Auswirkungen sollten sorgfältig geprüft werden.
7. **Verantwortlichkeit:** Es sollten Mechanismen geschaffen werden, um die Verantwortung und Rechenschaftspflicht für KI-Systeme und deren Ergebnisse zu gewährleisten. Dabei spielt die Auditierbarkeit, die eine Bewertung von Algorithmen, Daten und Designprozessen ermöglicht, eine Schlüsselrolle, insbesondere in kritischen Anwendungen.

2.5. Rechtsfragen bei der KI-Nutzung durch Private[22]

Die Nutzung von KI stellt eine ganze Reihe rechtlicher Fragen. Diese sollen hier mit Fokus auf Private (Unternehmen etc.) ausgeführt werden; KI-Anwendungen in Verwaltung und Gerichtsbarkeit bilden Gegenstand einer eigenständigen Betrachtung. Unternehmen nutzen KI teils intern, insbesondere in der Produktion, aber auch extern im Rahmen der von ihnen angebotenen Dienste und Produkte. In rechtlicher Hinsicht werfen diese Anwendungen ganz unterschiedliche Fragen auf.

[22] Dieser Abschnitt beruht auf Arbeiten von Damian George, Luca Fábián und Florent Thouvenin vom Lehrstuhl für Informations- und Kommunikationsrecht der Universität Zürich.

In der rechtswissenschaftlichen Diskussion wurden dabei bisher vor allem die Auswirkungen von KI im Bereich des Haftungs-, Immaterialgüter- und Datenschutzrechts näher untersucht. Besondere Aufmerksamkeit wird in aktuellen Debatten auch dem Diskriminierungspotenzial von KI geschenkt. Auf diese Bereiche wird nachfolgend näher eingegangen. Nicht thematisiert werden allfällige strafrechtliche Fragestellungen, die sich aus der Nutzung von KI durch Private ergeben können.

Haftungsrecht ist betroffen, wenn KI-Systeme in Entscheidungsprozesse involviert sind, bei denen Dritte zu Schaden kommen. Dies schliesst die Frage ein, ob und inwieweit KI-Systeme der Produktehaftung unterstehen (Cerka et al. 2015). Teilweise werden aber auch weitergehende Fragen aufgeworfen, z.B. ob und inwiefern autonomen Systemen zumindest eine beschränkte Rechtsfähigkeit zugebilligt werden soll, um sie als Akteure rechtlich erfassen zu können (Cerka et al. 2017). Eng mit dem Haftungsrecht verbunden ist auch die Frage, inwieweit präventive Massnahmen rechtlich gefordert sind, was z.B. eine Bewilligung bestimmter Einsatzformen von KI (etwa beim autonomen Fahren) beinhalten kann.

Mit Blick auf das *Immaterialgüterrecht* geht es zum einen um adäquaten Schutz für KI-Systeme, die teilweise mit Daten trainiert werden, die ihrerseits rechtlichen Bestimmungen unterliegen. Andererseits sind gewisse KI-Systeme auch in der Lage, «Neues» zu schaffen (siehe Abschnitt 2.2.1). Dies wirft die Frage auf, ob Werke der Literatur und Kunst sowie Erfindungen, die durch den Einsatz von KI entstanden sind, durch Urheberrechte bzw. Patente geschützt werden können und wem die Rechte gegebenenfalls zustehen (Schafer et al. 2015).

Datenschutzrecht ist insbesondere deshalb tangiert, weil neuere KI-Technologien entscheidend auf den Zugriff auf grosse Datenmengen angewiesen sind. Dieser Punkt wird bereits an anderer Stelle in diesem Bericht aufgegriffen (Konsum im Abschnitt 3.3.3.5 und Verwaltung im Abschnitt 3.5.3.1). An dieser Stelle folgen deshalb nur allgemeine Bemerkungen zu datenschutzrechtlichen Fragen bei der Nutzung von KI durch Private.

Die Nutzung von KI-Systemen kann schliesslich zu einer Ungleichbehandlung von Personen führen, beispielsweise wenn diese Systeme gewissen Kunden günstigere Preise offerieren als anderen. Nicht jede Ungleichbehandlung ist aber rechtlich als Diskriminierung zu werten. Eine solche liegt grundsätzlich nur dann vor, wenn die betroffenen Personen wegen bestimmter Merkmale ungleich behandelt werden, etwa wegen ihres Alters, ihres Geschlechts oder ihrer sexuellen Orientierung, und wenn diese Ungleichbehandlung sachlich nicht gerechtfertigt ist. Hinzu

kommt, dass das verfassungsrechtliche *Diskriminierungsverbot* in den Rechtsbeziehungen unter Privaten nicht unmittelbar wirkt; Private sind also nicht zur Gleichbehandlung verpflichtet. Die Behörden haben aber dafür zu sorgen, dass die Grundrechte, soweit sie sich dazu eignen, auch unter Privaten wirksam werden. Das gilt auch für das Diskriminierungsverbot. Entsprechend gibt es einige wenige Gesetze, welche Diskriminierungen durch Private verbieten. Deshalb werden nachfolgend die Grundzüge des verfassungsrechtlichen Diskriminierungsverbotes skizziert. Zum Diskriminierungspotenzial im Kontext staatlicher Entscheidungen vgl. Abschnitt 3.5.3.4.

2.5.1. Haftungsrecht

Eine Konstante in der Entwicklung von KI ist die zunehmende Autonomie der KI-Systeme. Verursacht die autonome Entscheidung einer KI einen Schaden, stellt sich die Frage, wer dafür zu haften hat. Die Rechtsordnung geht bisher axiomatisch davon aus, dass am Anfang einer Handlung die Entscheidung eines Subjekts steht. Tatbestandsmerkmale und Rechtsfolgen knüpfen alle an den Entscheidungen dieses Subjekts an. Da KI-Systeme aber zunehmend eigenständige und unvorhersehbare Entscheidungen treffen, wird dieses Konzept der Zurechnung infrage gestellt. Dies ist das Grundproblem, das KI-Systeme im Haftungsrecht aufwerfen.

2.5.1.1. Bestehende Konzepte

Im geltenden Haftungsrecht wird zwischen vertraglicher und ausservertraglicher (deliktischer) Haftung unterschieden. Die vertragliche Haftung beruht auf der Verletzung einer vertraglichen Pflicht, während die ausservertragliche Haftung auf der Verletzung einer allgemeinen, von der Rechtsordnung auferlegten Pflicht beruht (Huguenin 2014). Die ausservertraglichen Haftungsansprüche werden weiter in Verschuldenshaftungen, gewöhnliche Kausalhaftungen und scharfe Kausalhaftungen unterteilt.

Im **Vertragsrecht** haftet der Schuldner für vorsätzliche oder fahrlässige Vertragsverletzungen. Verursacht ein KI-System eine Vertragsverletzung, stellt sich demnach die Frage, ob der Einsatz des KI-Systems fahrlässig war. Im Zivilrecht wird grundsätzlich jedes Abweichen vom Verhalten eines durchschnittlich vernünftigen

Menschen als fahrlässig angesehen, d.h. es gilt ein objektivierter Fahrlässigkeitsmassstab (Huguenin 2014). Da von einem sorgfältig tätigen Dienstleister erwartet wird, dass er die Wirkungen und Risiken seiner technischen Werkzeuge versteht, wird die zunehmende Verbreitung von KI-Systemen hohe Anforderungen an das Verständnis und damit an die Weiterbildung der Mitarbeiter der betroffenen Dienstleister stellen (Meuldijk & Wattenhofer 2017; Vokinger et al. 2017). Setzt sich der Einsatz von KI-Systemen bei einer Dienstleistungserbringung durch, könnte der Rückgriff auf KI zum Standard des sorgfältigen Dienstleisters werden. Im Zusammenhang mit der Arzthaftung wird etwa diskutiert, ob der verbreitete Einsatz von KI in der Radiologie dazu führt, dass die Nutzung von KI von einem durchschnittlich sorgfältigen Arzt oder einer durchschnittlich sorgfältigen Ärztin erwartet wird (Vokinger et al. 2017). Zudem wird davon ausgegangen, dass ein durchschnittlich sorgfältiger Arzt seine Patienten vor der Behandlung über den Einsatz von KI aufzuklären hat (Vokinger et al. 2017). Inwiefern die Verbreitung von KI sich auf die Sorgfaltspflichten von Dienstleistern auswirkt, wird in der Literatur auch im Zusammenhang mit Anwälten und Wirtschaftsprüfern erörtert, wobei die Diskussion hier erst am Anfang steht (Meuldijk & Wattenhofer 2017; Burrus 2016). Setzt sich eine solche Erwartung durch, kann das vertragliche Haftungsrecht dazu führen, dass der Einsatz von KI für die sorgfältige Erbringung einer Leistung vorausgesetzt und damit für die betroffenen Dienstleister faktisch unumgänglich wird.

Haftung bedingt nicht notwendigerweise das Bestehen eines Vertragsverhältnisses zwischen den Akteuren. Solche Fälle werden durch die **ausservertragliche Verschuldenshaftung** oder durch **ausservertragliche Kausalhaftungen** geregelt. Die Grundnorm der Verschuldenshaftung im Schweizer Recht ist Art. 41 OR[23]. Gemäss Art. 41 OR haftet ein Verursacher für Schäden, die er kausal, widerrechtlich und schuldhaft verursacht. Führt der Einsatz von KI zu einem schädigenden Ereignis und musste der Verursacher nach der allgemeinen Lebenserfahrung mit dem Schadenseintritt rechnen, ist Kausalität zu bejahen. In den meisten Fällen ist dies unproblematisch, denn wer KI eine Tätigkeit ausführen lässt, die ein Schadensrisiko mit sich bringt, weil er sie zum Beispiel ein Fahrzeug steuern oder einen Menschen operieren lässt, wird regelmässig damit rechnen müssen, dass dies einen Schaden begünstigen kann (Rosenthal 2008). Da KI-Systeme nicht deliktsfähig sind, können sie jedoch nicht selber haften, auch wenn ein Schaden auf ihre autonome Entscheidung zurückzuführen ist (Müller 2014).

[23] Bundesgesetz betreffend die Ergänzung des Schweizerischen Zivilgesetzbuches (Fünfter Teil: Obligationenrecht) vom 30. März 1911 (SR 220).

Mangels Deliktsfähigkeit kann dem Betreiber oder Entwickler der KI deren Verhalten auch nicht über die sogenannte Geschäftsherren- oder Hilfspersonenhaftung zugerechnet werden (Freytag 2016).

Es stellt sich daher die Frage, wann die Betreiber/-innen oder Entwickler/-innen von KI-Systemen haften. Eine Verschuldenshaftung ist gegeben, wenn ihnen der Vorwurf gemacht werden kann, sie hätten bei Entwicklung oder Einsatz von KI fahrlässig oder gar vorsätzlich gehandelt. Im Vordergrund wird dabei die Fahrlässigkeit stehen. Diese ist hier gegeben, wenn die Entwickler oder Betreiber nicht die nach den Umständen gebotene Sorgfalt haben walten lassen, die ein durchschnittlich sorgfältiger Mensch in dieser Situation an den Tag legen würde (Huguenin 2014).

Mit Blick auf den *Entwickler* einer KI liesse sich argumentieren, dass eine KI nur dann sorgfältig programmiert sei, wenn sie in für sie nicht einschätzbaren Situationen passiv bleibt (Rosenthal 2008). Neuere KI-Systeme unterscheiden sich von deterministischer Software aber gerade dadurch, dass sie nicht einfach vordefinierten Anweisungen folgen, sondern aus bestehenden Daten Regeln für ihnen unbekannte Situationen entwickeln können. Daher ist eine derartige Programmierung nicht unsorgfältig (Müller 2014). Man könnte dem Hersteller aber auch den Vorwurf machen, dass er das KI-System nicht genügend getestet habe, wenn dieses mit einer Situation überfordert war (Rosenthal 2008).

Dem *Betreiber* der KI wiederum könnte fahrlässiges Verhalten vorgeworfen werden, wenn er gegen den Gefahrensatz verstösst. Gemäss diesem ungeschriebenen Rechtsprinzip hat derjenige, der einen gefährlichen Zustand schafft oder aufrechterhält, die zur Vermeidung eines Schadens erforderlichen Massnahmen zu treffen (Freytag 2016). Erforderlich sind aber nur diejenigen Massnahmen, die einem durchschnittlich sorgfältigen Menschen in der gleichen Situation angemessen erschienen wären. Ein gewisses Restrisiko ist also hinzunehmen und eine permanente Überwachung von KI-Systemen wird nicht gefordert, da dies faktisch einem Verbot gleichkäme. Das Schädigungsrisiko kann aber in gewissen Einsatzgebieten so hoch sein, dass von einem vernünftigen Betreiber erwartet wird, auf den Einsatz von KI ganz zu verzichten (Müller 2014; Rosenthal 2008).

Selbst wenn dem Verursacher des Schadens kein fahrlässiges Verhalten vorgeworfen werden kann, ist es immer noch möglich, dass er aufgrund einer **ausservertraglichen Kausalhaftung** für den Schaden einstehen muss. Kausalhaftungen greifen, wie der Name sagt, verschuldensunabhängig bereits aufgrund der

kausalen Verursachung eines Schadens. Bei sogenannten milden Kausalhaftungen kann sich der Schädiger immerhin von der Haftung befreien, wenn er den Nachweis erbringt, die übliche Sorgfalt eingehalten zu haben bzw. dass der Schaden auch beim Einhalten dieser Sorgfalt eingetreten wäre. Bei scharfen Kausalhaftungen ist ein solcher Entlastungsbeweis hingegen nicht möglich.

Im Zusammenhang mit KI und Robotern wird in der Literatur eine Haftung nach dem Bundesgesetz über die Produktehaftpflicht (PrHG)[24] diskutiert. KI-Systeme sind eine Software, und ob diese für sich genommen unter den Produktebegriff des Art. 3 PrHG fällt, ist umstritten (Hess 2016). Wenn aber eine KI-Software in eine bewegliche Sache wie z.B. ein Fahrzeug oder einen Roboter integriert wird, ist das PrHG anwendbar. Anders als nach den allgemeinen Regeln des Haftpflichtrechts haftet der Hersteller nach dem PrHG für alle Schäden, die sein fehlerhaftes Produkt verursacht. Ein Produkt gilt dabei bereits dann als fehlerhaft, wenn es die legitimen Sicherheitserwartungen der Allgemeinheit nicht erfüllt.[25] In der Literatur wird hervorgehoben, dass sich Fehler auch bei «klassischer» deterministischer Software statistisch gesehen zwar nicht vermeiden liessen, dies aber keinen Einfluss auf die Sicherheitserwartungen der Allgemeinheit habe (Lohmann & Müller-Chen 2017). Weist nun ein (nicht deterministisches) KI-System Programmierfehler auf, wird es daher nicht den Sicherheitsanforderungen der Allgemeinheit genügen und als fehlerhaft zu qualifizieren sein. Als milde Kausalhaftung gestattet es das PrHG dem Hersteller allerdings, sich von der Haftung zu befreien. Er kann unter anderem nachweisen, dass nach den Umständen davon auszugehen sei, das Produkt sei ursprünglich fehlerfrei in den Verkehr gebracht worden (Art. 5 Abs. 1 lit. b PrHG). Sofern der Hersteller Sicherheitsmassnahmen getroffen hat, um die ihm bekannten unerwünschten Reaktionen des KI-Systems zu verhindern, soll er daher gemäss einem Teil der Lehre nicht haften, wenn die KI autonom Fehlentscheidungen trifft (Hänsenberger 2018). Der Umfang der zu ergreifenden Massnahmen hängt vom Schädigungspotenzial des Produkts und von dessen vernünftigerweise zu erwartendem Gebrauch ab.

Andere Autoren halten die Anforderungen an diesen Sorgfaltsbeweis für zu hoch, als dass er dem Hersteller gelingen würde (Rosenthal 2008). Der Einwand des Entwicklungsrisikos (Art. 5 Abs. 1 lit. e PrHG), welches den Hersteller vor einer Haftung für Schäden schützt, die nach Stand der Wissenschaft und Technik nicht

[24] Bundesgesetz über die Produktehaftpflicht vom 18. Juni 1993 (SR 221.112.944).

[25] Bundesgericht (BGE) 133 III 81, E. 3.1.

erkannt werden konnten, greift bei KI-Systemen jedenfalls dann nicht, wenn diese laufend lernen und aufgrund der bei der Verwendung gewonnenen Erfahrung weiterentwickelt werden. Hinzu kommt, dass nach heutigem Stand der Wissenschaft bekannt ist, dass KI-Systeme unerwartete Outputs generieren können; entsprechend ist zu verlangen, dass die Hersteller dieses Gefahrenpotenzial beherrschen können (Lohmann & Müller-Chen 2017). Die Hersteller werden deshalb grundsätzlich kausal für von KI-Systemen verursachte Schäden haften. Zu berücksichtigen ist aber, dass nur für Personenschäden und CHF 900.– übersteigende Sachschäden für Sachen im Privatgebrauch gehaftet wird (Art. 1 Abs. 1 und Art. 6 Abs. 1 PrHG). Für Vermögensschäden, Schäden am Produkt selbst und Schäden an Sachen, die gewerblich oder beruflich genutzt werden (z.B. wenn ein Industrieroboter andere Maschinen beschädigt), besteht keine Haftung nach PrHG (Freytag 2016).

Sofern das PrHG für den verursachten Schaden keine Anspruchsgrundlage darstellt, kommt eine Kausalhaftung des Geschäftsherrn nach Art. 55 OR infrage. Der Geschäftsherr haftet hiernach für diejenigen Schäden, die seine subordinierten Hilfspersonen bei der Besorgung der ihnen übertragenen Geschäfte verursachen. Grundsätzlich würde der Hersteller eines KI-Systems haften, wenn dieses aufgrund seiner Weisungen fehlerhaft hergestellt wurde. Dem Hersteller stünde aber der Beweis offen, dass er seine Sorgfaltspflichten erfüllt hat bzw. der Schaden auch bei Erfüllung der Sorgfaltspflichten eingetreten wäre. Im Zusammenhang mit KI-Systemen wird in der Literatur die Pflicht zur zweckmässigen Organisation des Betriebs hervorgehoben, die Kontrollsysteme, klare Kompetenzzuordnungen und entsprechende Pflichtenhefte verlange (Hänsenberger 2018). Da es nicht möglich sei, alle durch das KI-System erlernbaren Handlungsoptionen ex ante durchzuprüfen, sei mit der ausreichenden Kontrolle der Softwarearchitektur der Sorgfaltspflicht Genüge getan. Andernorts wird aber darauf hingewiesen, dass zumindest für Produkte, die Personenschäden verursachen können, hohe Anforderungen an die zweckmässige Organisation gestellt werden und diese Entlastung kaum gelingen dürfte (Rosenthal 2008).

Aufgrund der erzielten Fortschritte im Bereich des autonomen Fahrens werden in der Literatur insbesondere strassenverkehrsrechtliche Haftungsfragen aufgeworfen.[26] Das Strassenverkehrsgesetz (SVG)[27] unterwirft den Halter eines Motorfahrzeugs einer scharfen Kausalhaftung für Schäden, die aus dem Fahrzeugbetrieb resultieren (Art. 58 Abs. 1 SVG). Eine durch den Betrieb eines Motorfahrzeugs geschädigte Person muss lediglich ihren Schaden sowie dessen kausale Verursachung durch den Betrieb des Motorfahrzeugs beweisen (Dähler 2018). Dieses strikte Haftungsregime wird mit einem Versicherungsobligatorium und einem Direktforderungsrecht ergänzt (Art. 63 Abs. 1 und Art. 65 Abs. 1 SVG), die geschädigte Person kann hier also direkt gegen die Versicherung des Halters klagen. Art. 59 SVG sieht zwar Entlastungsbeweise vor, diese können aber nur dann geführt werden, wenn nicht die fehlerhafte Beschaffenheit des Fahrzeugs den Schaden verursacht hat. Verursacht ein Motorfahrzeug durch einen KI-Systemfehler einen Schaden, muss somit der Halter hierfür haften (Lohmann & Müller-Chen 2017). Dies entspricht weitgehend der Situation im Luftfahrtgesetz, denn auch hier haftet der Halter einer autonom fliegenden Drohne kausal für Schäden (Art. 64 LFG); und auch hier muss sich der Halter für diese Schäden versichern oder gleichwertige Sicherheiten beim Bundesamt für Zivilluftfahrt hinterlegen (Art. 70 Abs. 1 LFG) (Hänsenberger 2017). Allerdings lassen sich die Halter von Drohnen oft nicht identifizieren, weshalb in der Literatur eine Registrierungs- und Kennzeichnungspflicht für Kleindrohnen gefordert wird (Christen et al. 2018).

2.5.1.2. Analogien

Sollte keine bestehende Haftungsgrundlage direkt auf KI-Systeme anwendbar sein, könnte man eine Norm mit anderen Tatbestandsvoraussetzungen per Analogieschluss anwenden. In der Literatur wird vor allem die Tierhalterhaftung nach Art. 56 OR als prüfenswert bezeichnet. Hiernach haftet derjenige, der Interesse und Nutzen an einem Tier hat, für dessen unberechenbares Verhalten. Der Schaden muss sich aus der Verwirklichung einer typischen Tiergefahr ergeben. Der Halter kann sich aber von dieser Haftung befreien, wenn er beweist, dass er alle nach den Umständen gebotene Sorgfalt in Verwahrung und Beaufsichtigung des Tieres eingehalten hat.

---

[26] TA-SWISS hat zum autonomen Fahren eine eigene Studie publiziert; siehe dazu https://www.ta-swiss.ch/themen-projekte-publikationen/mobilitaet-energie-klima/selbstfahrende-autos/.

[27] Strassenverkehrsgesetz vom 19. Dezember 1958 (SR 741.01).

Die Literatur gibt jedoch zu bedenken, dass KI-Systeme zwar unberechenbar seien, die typische Tiergefahr aber aufgrund der Lebendigkeit von Tieren nicht mit einer typischen «Robotergefahr» vergleichbar sei. Zudem trage der Analogieschluss der Tatsache zu wenig Rechnung, dass bei autonomen Systemen eine Aufgabendelegation erfolge. Die Analogie wird daher skeptisch betrachtet, die Tierhalterhaftung könnte aber nach vereinzelt vertretener Auffassung Pate für eine spezifische Regelung der Haftung autonomer Systeme stehen (Lohmann 2017).

2.5.1.3. Neue Konzepte

Sollten die bestehenden haftungsrechtlichen Konzepte sowie Analogien nicht ausreichen, um die Haftung von KI-Systemen zu klären, müssten neue Konzepte entwickelt werden. Eine Möglichkeit besteht in der – allerdings recht weitgehenden – Schaffung einer **allgemeinen Gefährdungshaftung**. Wer eine besonders gefährliche, aber von der Rechtsordnung geduldete Tätigkeit betreibt, der würde gemäss der allgemeinen Gefährdungshaftung für die Verwirklichung des charakteristischen Risikos dieser Tätigkeit haften (Widmer & Wessner 1999). Die allgemeine Gefährdungshaftung würde der Tatsache Rechnung tragen, dass der Gesetzgeber die technische Entwicklung nicht vorhersehen kann, und sie wäre insbesondere für KI-Systeme relevant (Fellmann & Werro 2013). Art. 50 OR des Vorentwurfs zum Bundesgesetz über die Revision und Vereinheitlichung des Haftpflichtrechts sowie Art. 60 des OR 2020-Forschungsprojekts (siehe zu OR 2020 Huguenin & Hilty 2013 sowie www.or2020.ch) sahen eine solche Haftung vor, sind aber nie Gesetz geworden.

Noch weiter geht die Idee, das Zurechnungsproblem mit einem **eigenen rechtlichen Status für KI-Systeme** zu lösen. Gemäss dieser Ansicht führt die zunehmende Delegation von Aufgaben an KI-Systeme zu einer Verantwortungslücke, die es mit der Anerkennung eines rechtlichen Status zu füllen gelte (Beck 2017; Ebneter 2010). Das EU-Parlament hat die EU-Kommission 2017 dazu aufgefordert, langfristig den Status einer elektronischen Person für «ausgeklügelte» Robotersysteme zu schaffen, die für ihre verursachten Schäden haften sollen.[28] Andere Autoren halten ein Tätigwerden des Gesetzgebers (noch) nicht für nötig. Sollte der Bedarf nach einer neuen Rechtsform bestehen, sei aber eine Anpassung des

[28] Europäisches Parlament, Bericht mit Empfehlungen an die Kommission zu zivilrechtlichen Regelungen im Bereich Robotik, 27. Januar 2017, A8-0005/2017, Empfehlung 59 lit. f.

Rechts der Kapitalgesellschaften an die Bedürfnisse von KI-Systemen einem neugeschaffenen rechtlichen Status vorzuziehen (Beck 2017; Ebneter 2010).

2.5.1.4. Präventive Massnahmen

Haftungsnormen wirken ex post, nachdem ein Schaden eingetreten ist. Der Gesetzgeber reguliert aber viele Lebenssachverhalte nicht nur ex post, sondern ex ante mittels präventiver Massnahmen. Daher sind viele gefährliche Tätigkeiten bewilligungspflichtig oder unterstehen behördlicher Aufsicht. Kommen in diesen Bereichen KI-Systeme zum Einsatz, stellt sich demnach die Frage, ob das bisherige Regime angepasst werden muss. Da KI-Systeme sehr viele Anwendungsfelder betreffen, kann nachfolgend nur eine kleine Auswahl der betroffenen Bewilligungsregime erfolgen.

Fortgeschritten und detailliert ist die Finanzmarktregulierung von KI-Systemen. So wird den Handelsteilnehmern vorgeschrieben, dass sie ihre Algorithmen angemessen zu testen haben.[29] Wann autonomes Fahren marktreif ist, kann derzeit noch nicht abgesehen werden. Momentan sind autonome Fahrzeuge nicht allgemein zugelassen und so dürfen Testfahrten mit autonomen Fahrzeugen nur dank einer Ausnahmebewilligung für neue technische Erscheinungen des Eidgenössischen Departements für Umwelt, Verkehr, Energie und Kommunikation gemäss Art. 106 Abs. 5 SVG erfolgen.[30] Auf internationaler Ebene wurde die Wiener Strassenverkehrskonvention an die Herausforderungen des automatisierten Fahrens angepasst, sodass autonome Fahrzeuge als beherrschbar und daher konventionskonform gelten, wenn sie entweder speziellen Vorschriften entsprechen oder der Fahrzeugführer sie übersteuern kann.[31] Eine Vernehmlassungsvorlage zur Revision des SVG wird wohl in Kürze publiziert (Häberli & Müller 2018).

---

[29] Art. 31 Abs. 2 lit. e Finanzmarktinfrastrukturverordnung vom 25. November 2015 (SR 958.11). Siehe dazu auch Monsch (2018) und Contratto (2014).

[30] Siehe dazu auch Zurkinden (2017) und Lohmann (2016). Solche Testfahrten finden seit Sommer 2016 mit autonomen Bussen in Sion statt; siehe dazu https://www.postauto.ch/de/projekt-smartshuttle.

[31] Art. 8 Abs. 5bis Wiener Übereinkommen über den Strassenverkehr (SR 0.741.10). Siehe auch Zurkinden (2017).

Im Recht der Gesundheitsversorgung und Krankenversicherung wirft schliesslich die Qualifikation von KI-Systemen als Medizinprodukte Fragen auf. Während Arzneimittel ein behördliches Zulassungsverfahren durchlaufen, werden Medizinprodukte lediglich in einem Konformitätsbewertungsverfahren geprüft, das je nach Risikobewertung des Produkts unterschiedlich ausgestaltet ist.[32] Die dieser Zweiteilung zugrunde liegende Annahme, dass Medizinprodukte risikoärmer seien als Arzneimittel, wird durch künstlich intelligente Medizinprodukte infrage gestellt, da diese zunehmend ärztliche Aufgaben übernehmen. Daher wird über erhöhte Anforderungen an die klinische Bewertung solcher Medizinprodukte nachgedacht (Vokinger et al. 2017).

2.5.2. Immaterialgüterrecht

Die immaterialgüterrechtlichen Fragestellungen, die sich aus der Nutzung von KI ergeben, lassen sich grob in drei Gruppen einteilen: Zunächst ist zu klären, ob KI als solche Gegenstand immaterialgüterrechtlichen Schutzes sein kann. Sodann stellt sich die Frage, ob die von KI generierten Ergebnisse schutzfähig sind. Schliesslich interessiert die rechtliche Einordnung von Schutzrechtsverletzungen im Zusammenhang mit der Nutzung von KI.

2.5.2.1. KI als Gegenstand immaterialgüterrechtlichen Schutzes

KI-Systeme können mit dem Patent- und/oder dem Urheberrecht geschützt werden. Entsprechend drängt sich eine nach diesen Rechtsgebieten gesonderte Betrachtungsweise auf.

Einem **patentrechtlichen Schutz** von KI steht in erster Linie entgegen, dass gemäss Art. 52 Abs. 2 lit. a und c des Europäischen Patentübereinkommens (EPÜ)[33] mathematische Methoden und Computerprogramme von der Patentierbarkeit ausgeschlossen sind (Hetmank & Lauber-Rönsberg 2018). Diese Ausnahmen sind auch unter dem schweizerischen Patentgesetz (PatG)[34] anerkannt (Heinrich 2018; Briner 2006; Bertschinger 2002). Mathematische Methoden und Computer-

[32] Siehe Art. 8 ff. (Arzneimittel) und Art. 45 ff. (Medizinprodukte) des Bundesgesetzes über Arzneimittel und Medizinprodukte, Heilmittelgesetz (SR. 812.21).

[33] Europäisches Patentübereinkommen (SR 0.232.142.2).

[34] Bundesgesetz über die Erfindungspatente vom 25. Juni 1954 (SR 232.14).

programme sind jedoch nach Art. 52 Abs. 3 EPÜ bloss «als solche» vom Patentschutz ausgeschlossen, womit sogenannte «computerimplementierte Erfindungen» der Patentierung grundsätzlich zugänglich bleiben. Gemeint sind damit Erfindungen, die Computer, Computernetze oder andere programmierbare Vorrichtungen umfassen, wobei mindestens ein Merkmal der Erfindung mit einem Computerprogramm realisiert wird.

Die Abgrenzung schutzbegründender technischer Anteile computerimplementierter Erfindungen von nicht zu berücksichtigenden nicht technischen Beiträgen ist indes mit erheblichen Schwierigkeiten verbunden und entsprechend umstritten (Hetmank & Lauber-Rönsberg 2018). So hat beispielsweise das deutsche Bundespatentgericht (BPatG) in einigen Urteilen die Patentierbarkeit von Simulationsverfahren mit Verweis auf fehlende technische Überlegungen verneint.[35] Offener zeigt sich die Beschwerdekammer des Europäischen Patentamts (EPA), die solche Simulationsverfahren als grundsätzlich patentierbar beurteilt, wenn sie die Ausführung von Aufgaben ermöglichen, «die für eine moderne Ingenieurstätigkeit typisch sind».[36] Diese mit Blick auf computerimplementierte Erfindungen teilweise als zu liberal kritisierte Patentierungspraxis manifestiert sich beispielsweise auch in der jüngsten Ausgabe der «Richtlinien für die Prüfung im Europäischen Patentamt» (EPA 2018). Dort heisst es zwar zunächst, Rechenmodelle und Algorithmen, wie sie im Rahmen von künstlicher Intelligenz und maschinellem Lernen eingesetzt werden, seien «per se von abstrakter mathematischer Natur, unabhängig davon, ob sie anhand von Trainingsdaten ‹trainiert› werden können» (EPA 2018, G.II.3.3.1), was ihrer Patentierbarkeit entgegenstehen würde. Gleich darauf werden jedoch Beispiele angeführt, bei denen die Verwendung von maschinellem Lernen einen technischen Beitrag darstelle. Dies sei etwa der Fall bei der Verwendung von neuronalen Netzen in einem Herzüberwachungsgerät zur Identifizierung von unregelmässigem Herzschlag und bei der Klassifizierung von digitalen Bildern, Videos etc. auf der Grundlage von Low-level-Merkmalen (wie Kanten oder Pixelattributen für Bilder), nicht jedoch bei der Klassifizierung von Textdokumenten

[35] BPatG, Beschluss vom 13. September 2016 – 17 W (pat) 20/14 – Kollisionsbestimmung; Beschluss vom 26. Mai 2014 – 23 W (pat) 8/10; Beschluss vom 29. November 2017 – 18 W (pat) 11/15 – Simulationsvorrichtung zur Roboteranwendung; für weitere Nachweise siehe (Hetmank & Lauber-Rönsberg 2018).

[36] EPA, Entscheid vom 13. Februar 2006 – T-1227/05 – Schaltkreissimulation I/Infineon Technologie.

aufgrund ihres Inhalts, weil es sich dabei nicht um einen technischen, sondern um einen linguistischen Zweck handle.

Die Patentierbarkeit von Algorithmen und Modellen künstlicher Intelligenz kann damit nicht abschliessend beurteilt werden und bedarf der Prüfung im Einzelfall. Für computerimplementierte, KI-basierte Erfindungen dürften jedoch vermehrt Patente erteilt werden, insbesondere durch das EPA. Hervorzuheben ist indes das hohe Freihaltebedürfnis, das an den Lösungsideen und -prinzipien künstlicher Intelligenz besteht. Dieses Freihaltebedürfnis ist umso grösser, je mehr sich KI in die Richtung einer «allgemeinen» bzw. «starken» KI entwickelt. Lässt man die Patentierung solcher Algorithmen zu, könnte dies die Monopolisierung von Wissen und Innovationsfähigkeit begünstigen, was unbedingt zu verhindern ist.

Im Vergleich zur Frage nach dem Patentschutz von KI scheint sich diejenige nach dem **urheberrechtlichen Schutz** *prima facie* einfacher beantworten zu lassen. Denn nach Art. 2 Abs. 3 URG[37] gelten auch Computerprogramme als Werke, womit sie – bei Erfüllen der Schutzvoraussetzungen – urheberrechtlichen Schutz geniessen. Subsumiert man demnach in einer bestimmten Programmiersprache zum Ausdruck gebrachte KI-Algorithmen unter den Begriff des Computerprogramms, lässt sich die urheberrechtliche Schutzfähigkeit von KI recht unproblematisch bejahen (Hetmank & Lauber-Rönsberg 2018).

Allerdings stellt sich bei näherem Hinsehen die Frage, ob KI tatsächlich vom Begriff des «Computerprogramms» im Sinne von Art. 2 Abs. 3 URG erfasst wird. Zwar hat der Gesetzgeber mit Blick auf die rasche technische Entwicklung auf eine gesetzliche Definition des Begriffs des «Computerprogramms» verzichtet[38] und auch im Schrifttum hat sich bis anhin keine Definition durchzusetzen vermocht (Calame 2006; Marly 2018). Den verschiedenen Definitionen ist jedoch allen gemeinsam, dass sie Computerprogramme – oft mit Verweis auf die Mustervorschriften zum Schutz von Computerprogrammen der WIPO aus dem Jahr 1978[39] – eher eng umschreiben, beispielsweise als «eine Folge von Befehlen, die von einem

[37] Bundesgesetz über das Urheberrecht und verwandte Schutzrechte vom 9. Oktober 1992 (SR 231.1).

[38] Siehe BBl 1989 III 477 ff., 522.

[39] Siehe § 1 (i) der Mustervorschriften der WIPO (WIPO Publikation Nr. 814), wo es heisst: «computer program means a set of instructions capable, when incorporated in a machine-readable medium, of causing a machine having information-processing capabilities to indicate, perform or achieve a particular function, task or result».

Datenverarbeitungssystem verarbeitet werden kann und deren Zweck es ist, das Datenverarbeitungssystem zu betreiben, bestimmte Funktionen auszuführen oder bestimmte Aufgaben zu lösen» (Berger 2004, S. 27; ähnliche Definitionen – ebenfalls mit Verweis auf die Mustervorschriften der WIPO – vertreten z.B. Barrelet & Egloff [2008] sowie Neff und Arn [1998]).

Geht man von einer solchen Definition des Computerprogramms aus, sind der Erfassung von KI durch das Urheberrecht enge Grenzen gesetzt. So kann wohl nicht mehr von einer «Folge von Befehlen» gesprochen werden (Hartmann & Prinz 2018), weil die Funktionalität von KI-Technologien – konkret im Fall der praktisch bedeutsamen künstlichen neuronalen Netze – nebst einem statischen Befehlssatz auch aus der Topologie des Netzes und schliesslich aus der Parametrisierung des trainierten Netzes besteht (siehe dazu Abschnitt 2.2.4.2). Auch die Festlegung auf «bestimmte Funktionen» oder «bestimmte Aufgaben» erweist sich als problematisch, weil ein Charakteristikum von KI gerade in ihrer Ergebnisoffenheit und mangelnden Determiniertheit besteht (Hartmann & Prinz 2018). Dies dürfte namentlich relevant werden, wenn dereinst KI-Systeme eingesetzt werden sollten, die der «starken KI» zugeordnet werden können, weil diese Systeme zweifellos nicht mehr (ausschliesslich) der Erfüllung «bestimmter Aufgaben» dienen.

Die vorstehenden Ausführungen machen deutlich, dass eine Erweiterung des bis anhin vertretenen Begriffs des Computerprogramms erforderlich ist, wenn auch KI-Systeme vom urheberrechtlichen Schutz erfasst werden sollen. Eine solche Begriffserweiterung ist möglich (Hartmann & Prinz 2018), aber nicht zwingend. Sie dürfte allerdings dem Willen des historischen Gesetzgebers entsprechen, welcher den Begriff des Computerprogramms für künftige Entwicklungen offenhalten wollte, um damit auch neuen Formen von Computerprogrammen Schutz durch das Urheberrecht zu gewähren.[40]

Ein weiteres Problem ergibt sich daraus, dass KI-Systeme veränderlich sind und sich – insbesondere im Fall von sogenannten *Online-learning*-Systemen (vgl. dazu Hartmann & Prinz 2018) – kontinuierlich weiterentwickeln. Der Gegenstand des urheberrechtlichen Schutzes ist damit zu einem bestimmten Zeitpunkt möglicherweise nicht derselbe wie zu einem späteren Zeitpunkt, und es stellt sich die für das Urheberrecht zentrale Frage, ob die weiterentwickelte Version noch dem ursprünglichen Urheber zurechenbar ist. Ist die Zuordnung zu einer bestimmten Per-

[40] BBl 1989 III 477 ff., 522.

son als Urheber nicht möglich, weil sich das System sozusagen selbst weiterentwickelt hat, kann die Weiterentwicklung nach heutiger Rechtslage auch nicht urheberrechtlich geschützt sein.

Unabhängig von den vorstehenden Ausführungen stellt sich die Frage nach der praktischen Relevanz eines urheberrechtlichen Schutzes von KI. Denn regelmässig dürften die einer KI zugrunde liegenden Ideen und mathematischen Konzepte einen wesentlichen Teil des Werts solcher Systeme ausmachen. Das Urheberrecht schützt jedoch lediglich die konkrete Ausprägung einer KI, nicht die ihr zugrunde liegende Lösungsidee.

2.5.2.2. KI-generierte Ergebnisse

Bei der Beurteilung der Schutzfähigkeit KI-generierter Ergebnisse steht die Frage im Vordergrund, wer aus rechtlicher Sicht als Schöpfer dieser Ergebnisse betrachtet wird und ob diese Eigenschaft allenfalls auch einem KI-System zuteilwerden kann. Auch diese Frage ist für das Patent- und Urheberrecht unterschiedlich zu beantworten.

Im **Patentrecht** können nach dem schweizerischen PatG und nach dem EPÜ gemäss ganz herrschender Meinung nur natürliche Personen Erfinder sein (Heinrich 2018; Münch & Herzog 2002; Krasser & Ann 2016; Osterrieth 2015; Blok 2017). Dieser Auffassung ist beizupflichten. Liessen sich mittels einer extensiven Gesetzesauslegung allenfalls noch juristische Personen als Erfinder erfassen, ist die Anerkennung von KI-Systemen als Erfinder schon wegen der fehlenden Rechtsfähigkeit ausgeschlossen.

Allerdings ist es nach heutiger Auffassung unerheblich, wie Erfindungen zustande kommen (Bertschinger 2002; Krasser & Ann 2016; Hetmank & Lauber-Rönsberg 2018; Osterrieth 2015; für das US-amerikanische Recht siehe Abbott 2016). Geschützt sind namentlich auch Zufallserfindungen oder solche, die auf Geistesblitzen basieren (Pedrazzini & Hilti 2008; Osterrieth 2015). Eine subjektive Leistung des Erfinders ist demnach nicht erforderlich, obschon beispielsweise die Schutzvoraussetzung der «erfinderischen Tätigkeit» nach EPÜ ein solches Erfordernis vermuten lassen würde. Entscheidend ist lediglich die Erfindungshöhe im Sinne des Nichtnaheliegens nach Art. 1 Abs. 2 PatG bzw. Art. 56 EPÜ. Vor diesem Hintergrund vertritt die Mehrheit der Autoren die Ansicht, dass KI-generierte Erfindungen einem menschlichen Erfinder im Rechtssinn zugeordnet werden können und damit grundsätzlich patentierbar sind. Erfinder einer KI-generierten Erfindung

(in der älteren Literatur teilweise als «Computererfindung» bezeichnet) ist demnach diejenige natürliche Person, die eine solche Erfindung zuerst zur Kenntnis nimmt und als Lösung eines technischen Problems begreift (Münch & Herzog 2002; Krasser & Ann 2016; Hetmank & Lauber-Rönsberg 2018; Melullis 2015).

Diese Lösung ist dogmatisch zwar vertretbar, sie birgt aber die Gefahr willkürlicher Zuordnungsergebnisse (Abbott 2016; Fraser 2016). Bei Erfindungen, die tatsächlich zu einem hohen Grad unabhängig von menschlicher Einflussnahme und damit autonom durch KI generiert werden, erscheint es daher insgesamt als nicht zeitgemäss, Menschen als Erfinder zu erfassen (Hetmank & Lauber-Rönsberg 2018).

Unabhängig von der Frage nach dem Erfinder im Rechtssinn ergibt sich eine weitere KI-bezogene Implikation: Ob die zentrale patentrechtliche Schutzvoraussetzung des Nichtnaheliegens (Art. 1 Abs. 2 PatG bzw. Art. 56 EPÜ) erfüllt ist, bemisst sich nach der durchschnittlichen Fachperson. Diese ist eine hypothetische Denkfigur.[41] Sie hat die Fähigkeiten einer durchschnittlich gut ausgebildeten, logisch denkenden, jedoch nicht kreativen Fachperson aus dem betreffenden Gebiet der Technik (Heinrich 2018) und sie verfügt über die für dieses Gebiet üblichen Mittel für routinemässige Arbeiten und Versuche.[42] Mittelfristig wird die Fachperson bei der Entwicklung von Lösungen für technische Probleme auf (durchschnittliche) KI zurückgreifen können (Abbott 2016; Blok 2017; Fraser 2016; Hetmank & Lauber-Rönsberg 2018). Da KI auf Dauer mutmasslich die erfinderischen Fähigkeiten von Menschen in den Schatten stellen dürfte, wird eine wachsende Anzahl von Erfindungen die Schwelle des Nichtnaheliegens nicht mehr erreichen und damit von der Patentierbarkeit ausgeschlossen sein (Hetmank & Lauber-Rönsberg 2018; Abbott 2016; Samore 2013; Krasser & Ann 2016).

Das **Urheberrecht** schützt nach Art. 1 Abs. 1 URG Werke, d.h. geistige Schöpfungen der Literatur und Kunst. Erforderlich ist dabei der «Ausdruck einer Gedankenäusserung»,[43] womit klargestellt ist, dass als Urheber i.S.v. Art. 6 URG – wie im Patentrecht als Erfinder – ausschliesslich natürliche Personen in Betracht kommen (Von Büren & Meer 2014). Es stellt sich folglich die Frage, ob hinter Werken, die originär und autonom von KI-Systemen geschöpft wurden, ähnlich wie im Patentrecht, Menschen als Urheber erblickt werden können.

---

41 Diese zwar bloss im EPÜ ausdrücklich erwähnte Denkfigur ist auch unter dem schweizerischen PatG anerkannt; siehe dazu Bertschinger (2002) und Heinrich (2018).

42 EPA (2018) G.7.3.; siehe dazu auch Blok (2017).

43 BGE 130 III 168, E. 4.5.

Die Literatur unterscheidet dabei den unproblematischen Fall des Einsatzes von Computern als Werkzeuge des Urhebers von der Schaffung eines Werkes durch einen Computer «in eigener Regie ohne menschliche Steuerung» (Von Büren & Meer 2014). Nur im zweiten Fall läge kein Werk eines Menschen vor, weshalb die Schutzfähigkeit verneint werden müsste. Zugleich wird jedoch teilweise apodiktisch festgehalten, die Entstehung «reiner» Maschinenwerke sei unwahrscheinlich, da «*immer* [Hervorhebung hinzugefügt] ein Mensch der Maschine Befehle erteilt, Programme eingibt und Parameter definiert» (Von Büren & Meer 2014, S. 61; siehe auch Barrelet & Egloff 2008). Diese Prämisse wird allerdings zunehmend an Gültigkeit verlieren, wie das Beispiel von «Google Deep Dream» (Castellano 2018) und der sich ausbreitende «Robo Journalism» (Graff 2018) zeigen.

Wo genau die Grenze zwischen KI als Hilfsmittel und KI als «in eigener Regie» agierender Urheber gezogen werden soll, ist umstritten (Hetmank & Lauber-Rönsberg 2018; Schönberger 2018). Gewisse Autoren scheinen eine beliebige menschliche Mitwirkung, die das Ergebnis beeinflusst, ausreichen lassen zu wollen (z.B. von Büren & Meer 2014). Nach dieser Meinung ist ein Mensch bereits dann als Urheber eines Werks zu betrachten, wenn dieses «unter bewusster Verwendung von Zufallsprinzipien» entstanden ist (Barrelet & Egloff 2008) oder wenn der Mensch «bei mehreren im Wege der Aleatorik entstandenen Erzeugnissen zumindest eine Auswahl treffen und damit eine der verschiedenen möglichen Versionen konkret als Werk bestimmen» muss (Schulze 2018).

Eine blosse Selektionsentscheidung kann allerdings kaum für die Gewährung urheberrechtlichen Schutzes ausreichen (Hetmank & Lauber-Rönsberg 2018). Wo zwischen dem Programmierer oder Bediener einer KI und dem KI-generierten Ergebnis keine schöpferische Beziehung mehr besteht, ist demnach urheberrechtlicher Schutz abzulehnen.

2.5.2.3. Schutzrechtsverletzungen

Mit Blick auf erfinderische bzw. schöpferische KI stellen sich schliesslich auch Fragen im Zusammenhang mit Schutzrechtsverletzungen. Relevant ist zum einen die Zuweisung der Haftung für von KI autonom begangene Schutzrechtsverletzungen (WEF 2018); hierzu kann auf die allgemeinen Ausführungen zu Haftungsfragen in Abschnitt 2.9.1 verwiesen werden.

Zum anderen ist zu klären, wie das Urheberrecht damit umgehen soll, dass viele Formen von KI enorme Mengen von Daten für den Trainingsprozess benötigen.

Zumindest ein Teil dieser Daten ist regelmässig urheberrechtlich geschützt, so etwa bei Fotografien, die für das Trainieren einer Bilderkennungssoftware verwendet werden. Diese Daten müssen zur Verwendung durch KI in der Regel vervielfältigt werden, was grundsätzlich eine Urheberrechtsverletzung darstellt (Art. 10 Abs. 2 lit. a URG).[44] Für die Weiterentwicklung von KI könnte dies eine erhebliche Hürde darstellen, weshalb Schönberger das Konzept eines «non-expressive use» (S. 163 f.) und die Anwendbarkeit der Schranke von Art. 5 Abs. 1 Richtlinie 2001/29/EG (vorübergehende Vervielfältigung) bzw. Art. 24*a* URG im Hinblick auf das Training von KI diskutiert. Als Lösung drängt sich künftig wohl die Anwendung der neuen urheberrechtlichen Schranke für die Verwendung von Werken zum Zweck der wissenschaftlichen Forschung (Art. 24d URG) auf. Voraussetzung ist allerdings, dass der Begriff der wissenschaftlichen Forschung weit ausgelegt wird, damit auch die Entwicklung von KI zu kommerziellen Zwecken erfasst werden kann.

2.5.3. Datenschutzrecht

Da die Entwicklung vorab neuerer Formen von KI-Technologien auf den Zugriff auf grosse Datenmengen angewiesen ist, stellen sich auch zahlreiche Herausforderungen für das Datenschutzrecht. Ziel des Datenschutzrechts ist es, die Persönlichkeit und die Grundrechte von Personen, über die Daten bearbeitet werden, zu schützen. Ausschlaggebend für die Anwendbarkeit des Datenschutzrechtes ist dabei die Tatsache, dass Personendaten (in Abgrenzung zu Sachdaten) bearbeitet werden. Als Personendaten gelten alle Angaben, die sich auf eine bestimmte oder bestimmbare Person beziehen.

Für die Qualifikation als Personendatum spielt es keine Rolle, wie die Daten beschafft worden sind. Es kann also um Daten gehen, welche die betroffene Person selbst zur Verfügung gestellt hat (beispielsweise durch Eingabe in einem Webformular), um Daten, die über eine betroffene Person gesammelt wurden (beispielsweise durch Tracking ihrer Internetaktivitäten), oder um Daten, die einer Person aufgrund ihrer Onlineaktivitäten und Verhaltensweisen zugewiesen werden. In diesem dritten Fall werden die Daten bisweilen als abgeleitete Daten oder *inferred data* bezeichnet. Diese Begriffe, die nicht einheitlich verwendet werden, sollen

[44] Siehe dazu Schönberger (2018), 162 f., der das Beispiel einer KI erwähnt, die mit rund 11 000 Romanen «gefüttert» wurde, um ihr die Bildung von Sätzen in natürlicher Sprache beizubringen.

zum Ausdruck bringen, dass es sich um Daten handelt, die aus der Analyse anderer Daten oder aus der Kombination verschiedener Daten gewonnen worden sind.

2.5.3.1. Datenminimierung, Erkennbarkeit und Zweckbindung

Die kommerziell erfolgreichsten KI-Systeme basieren auf der Technik des maschinellen Lernens (ML). ML wiederum beruht zu einem grossen Teil auf der Inferenzstatistik bzw. schliessenden Statistik, d.h. dem statistischen Verfahren, bei dem aufgrund von Untersuchungen von Stichproben Schlüsse in Bezug auf die Charakteristika der zugehörigen Grundgesamtheit getroffen werden (siehe dazu Abschnitt 2.2.4.2). Eine KI, die mit ML-Software entwickelt wurde, entscheidet nicht nach einer vorgegebenen Programmierung, sondern nach den erlernten Mustern. Diese Verfahren sind zwar nicht neu, mit der heutigen Verfügbarkeit an digitalen Daten und der stetig steigenden Rechenleistung können aber Datensets in bisher ungeahnter Grösse verarbeitet werden. Diese Entwicklung ist unter dem Stichwort Big Data bekannt, ein Phänomen, das in der datenschutzrechtlichen Literatur rege diskutiert wurde (für einen Überblick siehe Weber & Thouvenin 2014; Kuner et al. 2012).

Big Data Analytics und KI sind nur möglich, wenn grosse Mengen an Daten gesammelt und ausgewertet werden können. Diese Vorgehensweise ist aber – bei der Verarbeitung von Personendaten – mit dem datenschutzrechtlichen Grundsatz der Datenminimierung nicht in Vereinbarung zu bringen. Auch die Grundsätze der Zweckbindung und Erkennbarkeit, nach welchen Personendaten nur zu einem im Voraus angegebenen oder aus den Umständen ersichtlichen Zweck gesammelt werden dürfen, stehen Big Data Analytics diametral entgegen (Thouvenin 2014; Kuner et al. 2017). Das schweizerische Datenschutzgesetz (DSG)[45] lässt zwar inhaltlich ziemlich weit gefasste Zweckbindungen zu, dies kann den grundlegenden Konflikt aber nicht beheben, sondern nur relativieren. Im Entwurf des revidierten Datenschutzgesetzes (E-DSG)[46] wird die Zweckbindung denn auch gelockert,

[45] Bundesgesetz über den Datenschutz vom 19. Juni 1992 (SR 235.1).

[46] Entwurf zur Totalrevision des Bundesgesetzes über den Datenschutz und die Änderung weiterer Erlasse zum Datenschutz vom 15. September 2017, BBl 2017, 7193 ff.

indem die Bearbeitung zu kompatiblen Bearbeitungszwecken erlaubt wird (Rosenthal 2017). Dies entspricht weitgehend der Lösung in der EU.[47] Ob Big Data Analytics mit dem ursprünglichen Zweck kompatibel sind, wird sich daher anhand der inhaltlichen Verbindung der Zwecke, dem Erhebungsanlass der Daten, der Art der Daten, den möglichen Folgen für die Betroffenen und dem Vorhandensein geeigneter Garantien bestimmen.

2.5.3.2. Transparenz

Mit der zunehmenden Automatisierung von Entscheidungen stellt sich die Frage, ob und allenfalls wie die Transparenz der Entscheidungsprozesse von KI-Systemen hergestellt werden kann. Diese Diskussion wird auch – aber nicht nur – im Datenschutzrecht geführt. Der Grundsatz der Transparenz ist im europäischen Datenschutzrecht (DSGVO) ausdrücklich geregelt.[48] Er findet sich als Grundsatz der **Erkennbarkeit** (Art. 4 Abs. 4 DSG) auch im schweizerischen Recht.

Der Grundsatz der Transparenz enthält eine *ex ante* und eine *ex post* Perspektive: Einerseits müssen betroffene Person vorab über die Datenverarbeitung informiert werden (Informationspflichten); andererseits haben sie die Möglichkeit, die Datenverarbeitung im Nachhinein zu überprüfen (Auskunftsrecht). Die Information muss dabei in klarer, leicht zugänglicher und verständlicher Sprache erfolgen.[49] Im europäischen Recht werden die umfassenden Informationen, welche den betroffenen Personen zur Verfügung gestellt werden müssen, in den Art. 13 und 14

---

[47] Art. 5 Abs. 1 lit. b i.V.m. Art. 6 Abs. 4 DSGVO.

[48] So hält Art. 5 Abs. 1 lit. a der DSGVO fest, dass personenbezogene Daten «auf rechtmässige Weise, nach Treu und Glauben und in einer für die betroffene Person nachvollziehbaren Weise verarbeitet werden», was den engen Zusammenhang zwischen Transparenz und Fairness (im Deutschen: Treu und Glauben) verdeutlicht. Der englische Text spricht von «processed lawfully, fairly and in a transparent manner (‹lawfulness, fairness and transparency›)».

[49] Gegebenenfalls mit Visualisierung; siehe Erwägungsgründe 39 und 58. Siehe Art. 12 DSGVO – mit dem Titel «Transparente Information, Kommunikation und Modalitäten für die Ausübung der Rechte der betroffenen Person» –, welcher die Modalitäten für eine transparente Information der betroffenen Personen festlegt. Die Informationen müssen schriftlich oder gegebenenfalls auf elektronischem Weg zur Verfügung gestellt werden und in verständlicher und leicht zugänglicher Form vorliegen; insbesondere dann, wenn die für die Datenverarbeitung Verantwortlichen sich an Kinder wenden.

DSGVO definiert. Zusammenfassend verlangen diese, dass die für die Verarbeitung Verantwortlichen Informationen über sich selbst (wer), die Quantität und Qualität der verarbeiteten Daten (was), den Zeitpunkt (-rahmen) der Verarbeitungstätigkeit (wann), den Grund (warum) und den Zweck der Verarbeitung (wofür) offenlegen (Frenzel 2018).

In der Schweiz gehen die Informationspflichten weniger weit, insbesondere für private Datenbearbeiter.[50] Allerdings ist davon auszugehen, dass im revidierten Schweizer DSG Informationspflichten für Private und Bundesorgane Eingang finden werden. Das E-DSG vom September 2017 erwähnt zumindest in Art. 17 (Informationspflicht bei der Beschaffung von Personendaten) und in Art. 23 (Auskunftsrecht), dass eine «transparente Datenbearbeitung» zu gewährleisten ist. Der Umfang der Informationspflicht scheint allerdings nicht wesentlich erweitert worden zu sein.[51]

Transparenz kann nicht nur durch proaktive Informationspflichten erreicht werden, sondern auch über Individualrechte, insbesondere über das **Auskunftsrecht**. Für den Einsatz von KI steht im Vordergrund, dass Art. 15 Abs. 1 lit. h DSGVO den betroffenen Personen das Recht vermittelt, beim Vorliegen von automatisierten Entscheidungen «aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen» zu erhalten. Welcher Detaillierungsgrad hier erforderlich ist, ist in der Literatur allerdings umstritten.[52] Neu soll

[50] Informationspflichten, welche auf europäischer Ebene einen hohen Stellenwert haben (siehe Ausführungen oben), sind im Schweizer Recht nur eingeschränkt vorhanden. So bestehen Informationspflichten für Private nur beim Beschaffen von besonders schützenswerten Personendaten oder Persönlichkeitsprofilen (Art. 14 DSG). Die minimal zu kommunizierenden Informationen halten sich dabei in Grenzen: So müssen der Inhaber der Datensammlung, der Zweck des Bearbeitens und die Kategorien der Datenempfänger (wenn eine Bekanntgabe vorgesehen ist) angegeben werden (Art. 14 Abs. 2 DSG). Bundesorgane sind stets zur Information über die Bearbeitung von Personendaten verpflichtet; neben den oben erwähnten Umständen umfasst die Informationspflicht von Bundesorganen auch einen Hinweis auf das Auskunftsrecht und die Folgen einer Weigerung der betroffenen Person, die verlangten Personendaten herauszugeben (Art. 18*a* Abs. 2 DSG).

[51] Siehe Art. 17 E-DSG, welcher zusätzlich in Abs. 3 festhält, dass die Kategorien der bearbeiteten Personendaten dem Betroffenen mitgeteilt werden müssen.

[52] Gemäss Thouvenin et al. (2017) ist es ausreichend, wenn Beispiele der möglichen Auswirkungen auf die betroffene Person aufgezählt werden. Mendoza und Bygrave (2017) verlangen, dass eine Bank, die automatisiert Einzelentscheide aufgrund eines Scoring-Wertes der betroffenen Person trifft, diese darüber aufklären muss, dass sie für gewisse Kredite nicht infrage kommt. Teilweise wird gar eine Offenlegung des Algorithmus gefordert.

in der Schweiz das im E-DSG erweiterte Auskunftsrecht den Betroffenen die Möglichkeit geben, Informationen zum Vorliegen von automatisierten Einzelentscheidungen sowie zur Logik, auf der die Entscheidung beruht, zu erhalten (Art. 23 Abs. 2 lit. f E-DSG). Ein Recht auf Auskunft über die Tragweite und die angestrebten Auswirkungen der automatisierten Entscheidung ist aber bis anhin nicht vorgesehen (Thouvenin et al. 2018).

Umstritten ist insbesondere, ob die DSGVO ein sogenanntes «Recht auf Erklärung» enthält (Wachter et al. 2017; Selbst & Powles 2017; Edwards & Veale 2018; Casey et al. 2019; Kaminski 2019; Goodman & Flaxman 2017). Ein solches Recht kann wohl dem Wortlaut von Art. 13 Abs. 2 lit. f und Art. 15 Abs. 1 lit. h DSGVO entnommen werden, wonach den betroffenen Personen zumindest dann aussagekräftige Informationen über die Logik sowie die Bedeutung und die beabsichtigten Folgen einer Datenverarbeitung zur Verfügung zu stellen sind, wenn solche Entscheidungen Rechtswirkungen auf sie haben oder sie erheblich beeinträchtigen. Die Formulierungen «aussagekräftige Informationen über die involvierte Logik» und «Tragweite und angestrebte Auswirkungen» sind dem Konzept einer «Erklärung» ähnlich, zu welcher die betroffene Person über das Auskunftsrecht in Art. 15 DSGVO Zugang hat (Selbst & Powles 2017).

Noch ungeklärt ist, ob der Verantwortliche vor der Entscheidung die Funktionalitäten des Systems erklären muss oder ob sich die Informationen auf ein spezifisches Szenario beziehen müssen (Edwards & Veale 2018). Unklar ist auch, was als «aussagekräftige Information» gelten kann. Die Beurteilung wird hier wohl aus der Perspektive der betroffenen Person vorzunehmen sein (Kuner et al. 2012). Wenig hilfreich wird es sein, wenn KI-Provider ihre Algorithmen oder detaillierte technische Beschreibungen der maschinellen Lernprozesse offenlegen müssen. Daher ist der Begriff «Erklärung» in der Literatur auch auf Kritik gestossen. Einerseits wird kritisiert, dass es technisch unmöglich oder zumindest sehr schwierig ist, die Entscheidungsfindung von komplexen ML-Prozessen nachzuvollziehen, wenn diese auf vielschichtigen Deep-Learning-Algorithmen bzw. neuronalen Netzen beruhen. Andererseits werden damit unrealistisch hohe Anforderungen an automatisierte Entscheidungen gestellt, die deutlich weiter gehen als diejenigen bei menschlichen Entscheidungen, bei welchen die zugrunde liegende Logik (wenn es sie überhaupt gibt) in aller Regel auch nicht transparent ist und eine Begründung oftmals nur nachgeschoben wird (Thouvenin et al. 2018; Zerilli et al. 2018; Kuner et al. 2012).

Gemäss einem Bericht der Berkman Klein Center Working Group, «Accountability of AI Under the Law», soll der Beobachter dank einer Erklärung feststellen können, ob ein bestimmter Input entscheidend war oder zumindest Einfluss auf den Output hatte (Doshi-Velez et al. 2017). Nach dieser Definition sollten die Informationen es der betroffenen Person ermöglichen, die wichtigsten Faktoren einer Entscheidung zu bestimmen oder zu verstehen, wie bestimmte Faktoren eine Entscheidung verändern. Andere Autoren scheinen diesem Ansatz zu folgen, wenn sie argumentieren, dass die Betroffenen wissen wollen, wie unterschiedliche Faktoren gewichtet wurden, um die endgültige Entscheidung zu treffen (Zerilli et al. 2018; Edwards & Veale 2018).

Welche Informationen für den Betroffenen aufbereitet werden müssen, wird derzeit intensiv diskutiert.[53] Die Debatte hat sich dabei in Richtung eines «Rechts auf eine vernünftige Schlussfolgerung» entwickelt. Nach diesem Ansatz sind Schlussfolgerungen, Vorhersagen und Annahmen, die sich auf eine Person beziehen oder sich auf sie auswirken, personenbezogene Daten im Sinn des Datenschutzrechts. Wachter und Mittelstad (2019) argumentieren, dass der derzeitige Rechtsrahmen die betroffenen Personen nicht genügend vor risikoreichen Inferenzanalysen schützt. Während ein «Recht auf Erklärung» bedingt, dass eine Entscheidung stattfand, würde ein «Recht auf eine angemessene/vernünftige Schlussfolgerung» bereits im Vorfeld bestehen, sodass der für die Datenverarbeitung Verantwortliche prüfen müsste, ob eine Schlussfolgerung angemessen ist. Es müsste offengelegt werden, warum bestimmte Daten benötigt werden, um eine Schlussfolgerung zu ziehen, warum diese Schlussfolgerungen notwendig sind, um einen bestimmten Verarbeitungszweck oder eine bestimmte Entscheidung zu erreichen, und ob die Daten und Methoden, mit denen die Schlussfolgerungen gezogen werden, statistisch zuverlässig sind.

2.5.3.3. Automatisierte Entscheidungen

Ein zentraler Anwendungsbereich von KI sind automatisierte Entscheidungen. Diese sind nach der DSGVO grundsätzlich verboten, wenn sie auf der Bearbeitung

[53] Siehe dazu eine Studie von Binns et al. (2019), welche aber keine klare *best practice* aufzeigt.

von Personendaten beruhen (Art. 22 DSGVO).[54] Namentlich darf eine betroffene Person keiner Entscheidung unterworfen werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in einer ähnlichen Weise erheblich beeinträchtigt, wenn sie ausschliesslich auf einer automatisierten Verarbeitung von Personendaten beruht. Ob dieses Verbot nur bei nachteiligen Entscheidungen greift, ist umstritten (Thouvenin et al. 2018).

Vom Verbot bestehen zudem verschiedene Ausnahmen. Namentlich kann durch die ausdrückliche Einwilligung der betroffenen Person eine automatisierte Entscheidung zulässig sein (Art. 22 Abs. 2 lit. c DSGVO); dieser Ausnahmetatbestand wurde in der Literatur allerdings kritisiert, weil eine informierte Einwilligung in die Entscheidungsprozesse komplexer KI-Systemen utopisch erscheint (Kuner et al. 2012). Die Ausnahmetatbestände sind zudem mit weiteren Auflagen verbunden: So muss der Verantwortliche angemessene Massnahmen treffen, um die berechtigten Interessen der betroffenen Person zu wahren (Art. 22 Abs. 3 DSGVO). Dies umfasst mindestens ein Recht auf Erwirkung des Eingreifens einer Person seitens des Verantwortlichen sowie ein Recht auf Darlegung des eigenen Standpunkts und auf Anfechtung der Entscheidung.

Das DSG enthält derzeit noch keine Regelung zu automatisierten Entscheidungen. Dies soll sich aber mit der laufenden Revision des DSG ändern, welche die revidierte Konvention 108 des Europarats vom 18. Mai 2018 ratifiziert (siehe dazu Abschnitt 2.3.1.3). Anders als die Konvention 108 oder die DSGVO beschränkt sich der Entwurf des DSG aber – zu Recht – auf die Einführung einer Informationspflicht über automatisierte Entscheidungen (Art. 19 des E-DSG mit dem Titel «Informationspflicht bei einer automatisierten Einzelentscheidung»). Sind diese weiterführenden Informationspflichten erfüllt, hat die betroffene Person noch die Möglichkeit, ihren Standpunkt darzulegen oder zu verlangen, dass die Entscheidung von einer natürlichen Person überprüft wird (Art. 19 Abs. 2 E-DSG). Eine Anfechtungsmöglichkeit der Entscheidung selbst ist bei Entscheidungen von privaten Datenbearbeitern jedoch grundsätzlich ausgeschlossen (Thouvenin et al. 2018).

[54] Thouvenin et al. (2018) zeigen auf, dass die dogmatische Konstruktion der Bestimmung umstritten ist (siehe hierzu auch Pagallo 2018). So kann Art. 22 DSGVO als *Verbotsnorm* angesehen werden (Dovas 2017; Noto La Diega 2018; Mendoza & Bygrave 2017). Andere Autoren (Kamlah 2018; Vedder & Naudts 2017) sehen die Bestimmung als ein *Individualrecht*.

2.5.4. Diskriminierungsverbot

Gemäss Art. 8 Abs. 2 der Bundesverfassung darf «niemand […] diskriminiert werden, namentlich nicht wegen der Herkunft, der Rasse, des Geschlechts, des Alters, der Sprache, der sozialen Stellung, der Lebensform, der religiösen, weltanschaulichen oder politischen Überzeugung oder wegen einer körperlichen, geistigen oder psychischen Behinderung». Die in der Verfassung aufgezählten Merkmale sind nicht abschliessend. In der Rechtsprechung können weitere Kriterien anerkannt werden, um neu gefährdete Gruppen zu schützen (Kiener et al. 2018, § 36 Rn. 15). Von besonderer Relevanz ist das verfassungsrechtliche Diskriminierungsverbot im Kontext des KI-Einsatzes durch staatliche Behörden (vgl. Abschnitt 3.5.3.4).

In Rechtsbeziehungen zwischen Privaten entfaltet das verfassungsrechtliche Diskriminierungsverbot allerdings keine unmittelbare Wirkung. Eine Ausnahme stellt in diesem Kontext der Anspruch auf gleichen Lohn gemäss Art. 8 Abs. 3 Satz 3 BV dar. Dieser Anspruch besteht sowohl gegenüber einem öffentlichen als auch einem privaten Arbeitgeber (Kiener et al. 2018, § 36 Rn. 104). Im Übrigen gilt für das verfassungsrechtliche Diskriminierungsverbot, dass dieses lediglich nach Massgabe von Art. 35 Abs. 1 BV («Die Grundrechte müssen in der ganzen Rechtsordnung zur Geltung kommen.») und Art. 35 Abs. 3 BV («Die Behörden sorgen dafür, dass die Grundrechte, soweit sie sich dazu eignen, auch unter Privaten wirksam werden.») Wirkung entfaltet. Im Kontext der Nutzung von KI-Systemen unter Privaten wird dies regelmässig die Frage aufwerfen, ob der Gesetzgeber tätig werden muss, um Diskriminierung zu verhindern.

Bisweilen wird vorgeschlagen, die Frage der Diskriminierung im Kontext des Datenschutzrechts anzugehen, zumal die DSGVO die Begriffe Transparenz und Fairness miteinander verknüpft. Der Begriff Fairness wird in Bezug auf KI insbesondere dann verwendet, wenn die Auswertungsprozesse zur Diskriminierung führen (Barocas & Selbst 2016). Verzerrungen der Auswertungsprozesse können bereits beim Entwurf von KI-Systemen und bei der Auswahl von Trainingsdaten einfliessen, welche dann bestehende Vorurteile in automatisierte Entscheidungsprozesse einbetten (Kuner et al. 2012). So kann beispielsweise die Unterrepräsentation einer Minderheit in Trainingsdaten zur künftigen Diskriminierung dieser Gruppe in KI-basierten Einstellungsverfahren oder beim Credit-Scoring führen. Die Identifizierung und Kontrolle solcher Verzerrungen ist eine Herausforderung bei der Gestaltung und Bewertung der Fairness von maschinellen Lernprozessen. Ob dieser Herausforderung allerdings mit den Mitteln des Datenschutzes zu begegnen ist,

wird jedoch in der Literatur und auch in dieser Studie bezweifelt (George et al. 2018; siehe auch die Ausführungen in Abschnitt 3.5.3.1 bzw. 3.5.3.4 sowie Empfehlung 2).

2.6. Der bundesrätliche Expertenbericht

Im Dezember 2019 wurde der Bericht der «interdepartementalen Arbeitsgruppe Künstliche Intelligenz» an den Bundesrat übergeben. Unter dem Titel «Herausforderungen der künstlichen Intelligenz» wurden – nebst allgemeinen Einführungen und Betrachtungen – insbesondere die Auswirkungen von KI auf verschiedene Politikbereiche des Bundes untersucht. Da offenkundig Bezüge zur vorliegenden Studie bestehen, wird der Bericht nachfolgend summarisch vorgestellt.

Vertreterinnen und Vertreter aller sieben Departemente waren Teil der Arbeitsgruppe, die folgende Themen bearbeitete (der Bericht nennt 17 Themenbereiche, die hier teilweise zusammengefasst wurden):

- Aussenpolitik (Vertretung in internationalen Gremien inkl. *Digital Europe Programme*)
- Volkswirtschaft (Arbeitswelt, Industrie und Dienstleistungen)
- Bildung, Wissenschaft und Forschung
- Medien und Öffentlichkeit
- Verkehrspolitik (insbesondere automatisierte Mobilität)
- Gesundheitswesen
- Finanzwirtschaft
- Landwirtschaft
- Energie, Klima und Umwelt
- Verwaltung und Justiz sowie allgemeiner Rechtsrahmen inkl. Daten- und Immaterialgüterrecht
- Cybersicherheit und Sicherheitspolitik.

Aufgrund der Breite des Themenfeldes und der offensichtlichen Überschneidung mit den in dieser Studie untersuchten Bereichen wurden Informationen zwischen

den Projektgruppen ausgetauscht. Die Zielsetzungen waren verschieden, denn die bundesrätliche Arbeitsgruppe fokussierte auf die Auswirkungen von KI und möglichen Handlungsbedarf auf Themensetzung und Tätigkeit der Verwaltung des Bundes. Nachfolgend sollen die wichtigsten Gemeinsamkeiten und Unterschiede (sofern diese die gleichen thematischen Gebiete betreffen) zwischen den beiden Studien kurz hervorgehoben werden.

Eine wichtige Erkenntnis des bundesrätlichen Berichts ist, dass sich das zentrale Prinzip eines technologieneutralen Rechtsetzungs- und Regulierungsansatzes auch im rasch verändernden technologischen Umfeld von KI bewährt hat. So wird empfohlen, dass der Bund auch weiterhin eine grundsätzlich technologieneutrale Politik verfolgen sollte, welche technologiespezifische Regulierungen möglichst zu vermeiden versucht. Dies entspricht unserer ersten Empfehlung, wonach ein generelles «KI-Gesetz» angesichts der enormen Breite der Anwendungen dieser Basistechnologie sinnlos ist. Eine Regulierung ohne Erfahrung mit konkreten Problemen, d.h. eine Regulierung von «diffusen Risiken» im Kontext rascher technischer Entwicklungen erhöht die Gefahr, dass die Regulierung an den effektiven Problemen vorbei zielt und damit ins Leere geht.

Eine generelle Empfehlung des bundesrätlichen Berichts bezieht sich auf die Sicherstellung des Informations- und Wissensaustausches zwischen den Akteuren. Dies betrifft erstens die *Beobachtung der technologischen Entwicklung*, wofür unter anderem die bestehenden Monitoringaktivitäten des Bundesamtes für Statistik genutzt und allenfalls weiterentwickelt werden sollen. Zweitens sollen der *Informations- und Wissensaustausch sowie die Koordination im internationalen Umfeld* verstärkt werden. Dazu soll unter anderem die vom Bundesamt für Kommunikation für die Vorbereitung des UNO-Weltgipfels zur Informationsgesellschaft geschaffene «Plateforme Tripartite» genutzt werden. Diese «Plateforme Tripartite» soll als interdisziplinäres nationales Kompetenznetzwerk zu internationalen KI-Themen und -Prozessen dienen, welches in der Lage ist, Wissen und Erfahrungen auch horizontal zu vernetzen und so kohärente Positionen der Schweiz auf internationaler Ebene sicherzustellen. Drittens sollen auf Basis des Berichts strategische Leitlinien für den Umgang mit KI auf Ebene des Bundes erarbeitet werden. Dabei ist künstliche Intelligenz nicht als isolierte Technologie zu betrachten, sondern als ein zentrales Element der fortschreitenden Digitalisierung von Wirtschaft und Gesellschaft. Daher soll die KI-Politik als wesentlicher Bestandteil der Strategie «Digitale Schweiz» weitergeführt werden. Damit soll die Zusammenarbeit aller Ebenen der Verwaltung mit der Wirtschaft, der Zivilgesellschaft, der Politik und der Wissenschaft gefördert werden.

In der vorliegenden Studie wird vorgeschlagen, die Förderung des Informations- und Wissensaustausches zwischen den Akteuren durch entsprechende Forschungsförderung zu unterstützen. Damit können neue KI-Anwendungen wie auch deren kritische Reflexion und somit die Inhalte für diesen Erfahrungsaustausch wie auch die nötige Expertise geschaffen werden. Dies wird in Abschnitt 5.3 weiter ausgeführt.

Schliesslich kommt der bundesrätliche Bericht zum Schluss, dass nur ein geringer *zusätzlicher* Handlungsbedarf seitens des Bundes bestünde. Zwar identifiziert der Bericht grossen Klärungs- und teilweise auch Handlungsbedarf (insgesamt wurden 35 prioritiäre Aktionsfelder dargestellt). Aber weil in der Bundesverwaltung bereits Gremien bzw. Prozesse existieren würden, um den absehbaren KI-Herausforderungen begegnen zu können, wurden lediglich vier zusätzliche Aktionsfelder identifiziert, für welche bestehende Kompetenzen oder Gefässe nicht als ausreichend angesehen wurden:

1. **Weiterentwicklung des Völkerrechts angesichts der künstlichen Intelligenz:** Diese Forderung betrifft die zahlreichen internationalen Aktivitäten zur Schaffung von Industriestandards sowie ethischen Prinzipien bzw. KI-spezifischen Regeln. Hier solle die Schweiz prüfen, inwiefern dadurch Völkerrecht geschaffen wird und welche Auswirkungen dieses auf die Schweiz haben könnte. Das Eidgenössische Departement für auswärtige Angelegenheiten soll diesbezüglich per Ende 2020 einen Bericht vorlegen.
2. **Beeinflussung von Medien und Öffentlichkeit durch KI-Systeme:** Plattformen und andere Intermediäre können KI-Anwendungen für kommerzielle oder politische Zwecke instrumentalisieren oder selbst für diese Zwecke instrumentalisiert werden. Dadurch kann die öffentliche Meinungs- und Willensbildung insbesondere im politischen Bereich beeinflusst werden. Das Eidgenössische Departement für Umwelt, Verkehr, Energie und Kommunikation soll bis im Frühling 2021 einen Bericht vorlegen, der spezifische Massnahmen zu dieser Thematik prüft.
3. **Künstliche Intelligenz in der Verwaltung:** Dieser Aspekt betrifft schliesslich die Nutzung von KI in der Bundesverwaltung, wobei insbesondere verhindert werden soll, dass fragmentierte Lösungen in den einzelnen Verwaltungsstellen geschaffen werden. Um dies zu verhindern, könnte eine gemeinsame Anlaufstelle bzw. ein Kompetenznetzwerk mit speziellem Fokus auf technische Aspekte für die konkrete Anwendung von KI in der Bundesverwaltung geschaf-

fen werden. Das Eidgenössische Finanzdepartement soll deshalb in Zusammenarbeit mit den anderen Departementen den Mehrwert und die Machbarkeit einer gemeinsamen Anlaufstelle prüfen. Diese Stelle soll eine beratende Funktion haben und gegebenenfalls auch Kompetenzen für andere Basistechnologien (z.B. Blockchain oder IoT) aufbauen.

4. **Strategische Leitlinien für eine KI-relevante Politik:** Angesichts der raschen Entwicklungen und der breiten Diskussionen zum Einsatz von künstlicher Intelligenz bedarf es auf Ebene des Bundes strategischer Leitlinien. Diese sollen aus dem bundesrätlichen Bericht abgeleitet werden. Dabei ist künstliche Intelligenz nicht als isolierte Technologie zu betrachten, sondern als ein wesentlicher Bestandteil der fortschreitenden Digitalisierung von Wirtschaft und Gesellschaft. Die Strategie «Digitale Schweiz» soll daher die KI-Politik als wesentlichen Bestandteil berücksichtigen.

Diese Empfehlungen decken sich teilweise mit den Empfehlungen dieser Studie. Die Beobachtung der Weiterentwicklung des Völkerrechts ist Teil der Forschungsempfehlungen. Der Thematik «Beeinflussung von Medien und Öffentlichkeit durch KI-Systeme» weist die vorliegende Studie ebenfalls eine prominente Rolle zu, wobei einige weiter gehende Empfehlungen vorgebracht werden. Bezüglich des Themenfeldes «Verwaltung» gehen die hier formulierten Empfehlungen weiter als lediglich die Forderung nach einer Abstimmung zwischen den einzelnen Departementen in Form einer Anlaufstelle. Die vierte Empfehlung des bundesrätlichen Berichts deckt sich sinngemäss mit jenen (insbesondere der ersten Empfehlung) des vorliegenden Berichts. Demnach sind die Herausforderungen je nach Politikbereich sehr unterschiedlich und eine Gesamtstrategie ist daher nicht angezeigt. Der bundesrätliche Bericht identifiziert allerdings Erfolgsfaktoren im Umgang mit neuen Technologien. Um diese auch im Kontext KI zu nutzen, setzt der Bundesrat auf strategische Leitlinien. In diesem Sinn macht auch die vorliegende Studie weiter gehende Empfehlungen, die sich aber nicht primär an die Bundesverwaltung richten, sondern weitere gesellschaftliche Akteure einbeziehen.

3. Stand des Wissens in den fünf Themenfeldern

Das dritte Kapitel gibt eine thematische Einführung in die für diese Studie relevanten Themenfelder Arbeitswelt, Bildung und Forschung, Konsum, Medien sowie öffentliche Verwaltung. Ziel dieser Abschnitte ist es, eine Übersicht zur Diskussionslage und zum Stand des Wissens in den einzelnen Bereichen zu geben. Angesichts der Breite der einzelnen Themenfelder kann dies nicht umfassend geschehen; vielmehr sollen gezielt jene Fragestellungen herausgearbeitet werden, die in der Expertenumfrage genauer betrachtet wurden.

3.1. KI und die Arbeitswelt[55]

Unter den vielen Lebensbereichen, die von KI beeinflusst und verändert werden, zählen die Auswirkungen auf die Arbeitswelt zu den am meisten diskutierten und am heftigsten umstrittenen Themen. Wenngleich die möglichen Veränderungen der Qualität, der Quantität und der Organisation von Arbeit nicht unmittelbar eine der Kernfragen der künstlichen Intelligenz berührt, nämlich die Verschiebung von Entscheidungen vom Mensch auf die Maschine, so werden davon doch zentrale Fragen der Stellung des Menschen in Gesellschaft und Wirtschaft aufgeworfen.

Arbeit und die damit geschaffenen Produkte, Dienstleistungen und Werte sind die Basis menschlicher Existenz. Für die überwiegende Mehrheit der Menschen stellen Einkommen aus Arbeit die einzige oder überwiegende Quelle dar, um ihr Leben und das ihrer Familie zu bestreiten. Für viele ist die Arbeit darüber hinaus ein zentraler Lebensinhalt; eine Anerkennung als nützliches Mitglied der Gesellschaft ist oft an eine aktive Teilnahme am Erwerbsleben geknüpft. Arbeit trägt aber auch indirekt zur Existenzsicherung bei, denn Steuern und Beiträge aus Arbeitseinkommen finanzieren zu einem Grossteil Staatshaushalte, das Sozial- und Rentensys-

[55] Dieser Abschnitt beruht auf Arbeiten von Johann Čas und Jaro Krieger-Lamina, Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften. Sie wurden bei den Recherchen von Susanne Hollin unterstützt.

tem. Inwieweit diese zentrale Rolle von Arbeit als Existenzgrundlage und Sinnstiftung angesichts von KI und Digitalisierung längerfristig Bestand haben kann, ist eine Frage, die in diesem Abschnitt immer wieder aufgeworfen wird. Für den Zeithorizont dieser Studie von gut fünf Jahren kann aber davon ausgegangen werden, dass sich diese Beziehungen nicht grundlegend verändern werden.

Nachfolgend wird die Thematik «Arbeitswelt» unter Bezugnahme auf aktuelle Literatur aus zwei Perspektiven beleuchtet. Zum einen geht es um die makroökonomische Betrachtungsweise – also um vermutete Auswirkungen von KI auf Beschäftigung, Lohnstruktur und dergleichen. Zum anderen geht es um die Auswirkungen von KI auf die Gestaltung der Arbeit selbst, wobei dies aber nur aus einer generellen Perspektive geschehen kann, weil die konkreten Auswirkungen je nach Berufsfeld sehr unterschiedlich sein dürften. Die Resultate dieser Analysen dienten als Grundlage für die Expertenumfrage. Kein Thema dieser Studie sind Struktur und Perspektiven der KI-Industrie in der Schweiz. Eine solche Marktstudie übersteigt den Rahmen dieser Arbeit.

Schliesslich ist es notwendig, zwei Einschränkungen zu erwähnen. Erstens sind die Gegebenheiten und Entwicklungen am Arbeitsmarkt von einer Vielzahl von Faktoren beeinflusst. Dazu zählen die Globalisierung, Konjunktur und Wirtschaftswachstum, Marktmacht von Unternehmen und die Position von Gewerkschaften, staatliche Regulierung und Sozialpolitik oder demografische Entwicklungen, um einige wesentliche Faktoren neben dem technischen Fortschritt zu nennen. Zweitens ist – gerade im Bereich Arbeitswelt – der Begriff «Künstliche Intelligenz» trotz des inflationären Gebrauchs ungenau und mehrdeutig und dessen Verwendung umstritten. In der Mehrzahl der Studien werden generellere Begriffe wie Automatisierung oder Digitalisierung verwendet. Selbst Berichte, die sich im Titel explizit auf KI beziehen (siehe z.B. Frontier Economics 2018), schliessen in ihrer Analyse Automatisierung, Roboter, Informations- und Kommunikationstechnologien und Digitalisierung als relevante Begriffe ein. Entsprechend ist es schwierig, die Effekte von KI von solchen anderer Einflussfaktoren klar abzugrenzen.

3.1.1. Makroökonomische Effekte von KI auf die Arbeitswelt

3.1.1.1. Quantitative Wirkungen auf den Arbeitsmarkt

Eine Initialzündung der Debatte, welche Effekte KI auf die Arbeitswelt haben kann, bildete eine von Frey und Osborn (2013) verfasste Studie. Darin kommen die Autoren zum Ergebnis, dass 47 % der Berufe in den USA ein hohes Risiko besitzen, durch die Computerisierung ersetzt zu werden. Für das grosse Interesse, auf das diese Studie gestossen ist, dürften einerseits die hohen Werte verantwortlich sein, wonach fast die Hälfte der Beschäftigung in den USA in den nächsten 20 Jahren potenziell automatisierbar wäre. Andererseits dürfte auch der Umstand dazu beigetragen haben, dass in Europa die Auswirkungen der Finanzkrise 2008 und der folgenden Wirtschaftskrise, und hier insbesondere auf den Arbeitsmärkten, noch deutlich spürbar waren und sich keine fundamentale Trendwende abzeichnete.

Es sind aber nicht nur reale Entwicklungen auf den Arbeitsmärkten – anhaltende hohe Arbeitslosenraten, die in einigen europäischen Ländern bis in die Gegenwart vorherrschen –, welche diese Debatten befeuern. Daneben sind sie auch Ausdruck von Überlegungen, inwieweit die Digitalisierung – und in diesem Zusammenhang insbesondere die KI – zu fundamentalen, historisch nicht vergleichbaren Transformationen (Brynjolfsson & McAfee 2011) von Wirtschaft und Arbeit führt.

Beide Entwicklungen haben dazu beigetragen, dass in einer Vielzahl von Studien versucht wurde, die Analysen von Frey und Osborn (2013) auf andere Länder und Regionen zu übertragen, die Ergebnisse grundsätzlich zu bestätigen, zu widerlegen oder zu relativieren, zum Beispiel durch kritische Auseinandersetzungen mit dem gewählten methodischen Ansatz (z.B. Arntz et al. 2016). Lovergine und Pellero (2018) geben einen aktuellen, nicht vollständigen Überblick über solche Studien (Arntz et al. 2016; Bonin et al. 2015; Bowles 2014; Chang & Huynh 2016; Frey et al. 2016; Frey & Osborne 2013; Manyika et al. 2017; World Bank Group 2016; WEF 2016; Wolter et al. 2015), welche Prognosen über die Potenziale, Arbeitsplätze bzw. Tätigkeiten zu automatisieren und einzusparen, beinhalten (siehe Tabelle 2).

Tabelle 2: Überblick von Studien mit Abschätzungen über das Ausmass des durch Automatisierung verursachten Arbeitsplatzverlustes (adaptiert von Lovergine & Pellero 2018, S. 75).

Quelle	Region, Sektor, Zeitraum	Kernaussage
Frey & Osborn 2013	USA, alle Sektoren, 10–20 Jahre	Etwa 47 % der gesamten Beschäftigung in den USA laufen Gefahr, in den nächsten 20 Jahren durch Automatisierung ersetzt zu werden.
Bruegel & Bowles 2014	EU-28, 10–20 Jahre	In den nächsten Jahrzehnten werden die EU-Länder aufgrund des technologischen Fortschritts zwischen 47 % (Schweden) und über 60 % (Rumänien) der Arbeitskräfte verlieren (Diese Studie repliziert die Arbeit von Frey und Osborne).
Bonin et al. 2015	Deutschland, alle Sektoren	Deutschland wird rund 12 % der Arbeitsplätze verlieren.
Wolter et al. 2015	Deutschland, verarbeitende Industrie (bis 2030)	Etwa 60 000 Arbeitnehmer/-innen werden bis 2030 ihre Arbeitsplätze in der verarbeitenden Industrie verlieren (Differenz zwischen dem Verlust von 420 000 und der Schaffung von 360 000 neuen Arbeitsplätzen dank Robotik, KI und neuen Technologien).
OECD (Arntz et al. 2016)	21 OECD-Länder	Im Durchschnitt sind 9 % der Arbeit in den 21 untersuchten OECD-Ländern automatisierbar. Der Anteil der Arbeitnehmer/-innen mit der höchsten Wahrscheinlichkeit, durch Automatisierung ersetzt zu werden, ist in Deutschland und Österreich höher (12 % der Erwerbsbevölkerung) und niedriger in Estland, Finnland, Belgien und Korea (6–7 %).
Citibank (Frey et al. 2016)	Über 50 Länder und Regionen	Der Durchschnitt der automatisierbaren Arbeitsplätze steigt auf 57 %, mit Spitzenwerten von 69 % in Indien und 77 % in China.
WEF 2016	15 grosse Industriestaaten und Schwellenländer	Zwei Millionen neue Arbeitsplätze werden geschaffen und sieben Millionen verschwinden, was zu einem negativen Saldo von netto über fünf Millionen Arbeitsplätzen führen wird. Auf der Ebene der Berufsgruppen werden sich die Verluste vorab auf den Bereich Verwaltung (4,8 Mio.) und Produktion (1,6 Mio.) konzentrieren. Die Bereiche Finanzen, Management, IT und Ingenieurwesen werden diese Verluste teilweise kompensieren.
ILO (Chang & Huynh 2016)	ASEAN-5, 20 Jahre	56 % der Beschäftigten sind einem hohen Automatisierungsrisiko ausgesetzt (32 % einem mittleren Risiko und 12 % einem geringem Risiko).
World Bank 2016	Die nächsten Jahrzehnte	2/3 aller Arbeitsplätze in Entwicklungsländern könnten für die Automatisierung anfällig sein.
PWC 2017	Grosse Industriestaaten, 2030	Die Studie schätzt, dass das Vereinigte Königreich (30 %) einen geringeren Anteil an Arbeitsplätzen mit potenziell hohem Automatisierungsrisiko hat als die USA (38 %) und Deutschland (35 %), aber mehr als Japan (21 %).
McKinsey Global Institute 2017	46 Länder (repräsentieren etwa 80 % der globalen Arbeitskräfte)	Weniger als 5 % der Berufe können vollständig automatisiert werden; bei etwa 60 % von ihnen können 30 % der Tätigkeiten automatisiert werden. Diese Prozentsätze betreffen etwa 1,2 Milliarden Arbeitnehmer/-innen und 14,6 Billionen Dollar Löhne. China, Indien, Japan und die Vereinigten Staaten machen mehr als die Hälfte der Beschäftigten und Löhne aus. Pro Jahr könnten ca. 0,8–1,4 % der globalen Produktivität automatisiert werden.

Auf eine detaillierte Analyse und Auseinandersetzung mit den Ergebnissen der zahlreichen Einzelstudien wird hier bewusst verzichtet. Eine solche Analyse würde einerseits den Rahmen dieser Arbeit sprengen und andererseits kaum relevante Ergebnisse für eines der wesentlichen Ziele der vorliegenden Untersuchung hervorbringen, nämlich Handlungsoptionen und Empfehlungen für die Schweiz zu entwickeln und zu diskutieren. Dennoch ist es wichtig, grundsätzliche und vergleichende Aspekte dieser Studien zur Einschätzung der arbeitsmarktwirksamen Einsparungspotenziale anzusprechen und zu erläutern. Ebenso soll angesprochen werden, warum diese Fokussierung auf Prognosen von Automatisierungspotenzialen von menschlicher Arbeit nicht nur wenig relevant, sondern auch kontraproduktiv bei der Entwicklung von Optionen sein können, welche den Menschen in den Mittelpunkt stellen.

Natürlich hat der hohe Anteil von 47 % der Beschäftigten in den USA, welche laut der Studie von Frey und Osborn potenziell von Automatisierung betroffen sein sollen, den Bedarf geweckt zu untersuchen, ob diese Abschätzung gerechtfertigt ist. Kritische Auseinandersetzungen mit den Ergebnissen der Studie von Frey und Osborn konzentrieren sich auf das Ausmass der Einsparpotenziale. So wird in der Vergleichsstudie für die OECD-Länder (Arntz et al. 2016) darauf hingewiesen, dass ein berufsbasierter Ansatz die Situation zwangsläufig falsch wiedergebe, da nicht alle Aufgaben und Tätigkeiten innerhalb von Berufen gleichermassen automatisierbar seien. Dementsprechend wurde die Studie von Frey und Osborne adaptiert und mit einem *Task-based*-Ansatz anstatt eines Berufsansatzes auf 21 OECD-Länder übertragen. Dabei wurden die unterschiedlichen Tätigkeitsbereiche innerhalb einzelner Berufe berücksichtigt. Angesichts des Umstands, dass es in fast allen Berufen Tätigkeiten gibt, die in naher Zukunft nicht von Maschinen übernommen werden können, kamen Arntz et al. zum Ergebnis, dass im Durchschnitt 9 % der Beschäftigten in den einbezogenen OECD-Ländern einem hohen Risiko ausgesetzt sind, ihre Arbeit durch Automatisierung zu verlieren. Als Berufe mit einem hohen Risiko werden solche verstanden, bei denen der Anteil der automatisier-baren Tätigkeiten mindestens 70 % ausmacht. Innerhalb der untersuchten Länder beträgt das Risiko zwischen 6 % (Korea oder Estland) und 12 % (Deutschland und Österreich). Für Arbeitnehmende mit geringem Einkommen erhöht sich das für Österreich geschätzte Arbeitsplatzrisiko auf 38 % (Arntz et al. 2016, S. 33–34). Die Schweiz wurde leider in diese Untersuchung nicht einbezogen.

Arntz et al. nennen weitere Gründe, warum die Schätzung von Frey und Osborn zu hoch ausfalle. Erstens sei die Diffusion neuer Technologien ein langwieriger Prozess, zweitens können sich Arbeitnehmende durch den Wechsel von Aufgaben auf veränderte technologische Ausstattung einstellen, drittens schaffe der technologische Wandel durch die Nachfrage nach neuen Technologien und durch eine höhere Wettbewerbsfähigkeit auch zusätzliche Arbeitsplätze.

In einer Studie des österreichischen Wirtschaftsforschungsinstituts werden u.a. die Ergebnisse von sieben Studien bezüglich des Automatisierungspotenzials von Berufen oder Tätigkeiten aus den Jahren 2013–2016 zusammengefasst (Peneder et al. 2016, S. 111). Die in diesen Studien gemachten Einschätzungen zur potenziellen Betroffenheit bewegen sich, ähnlich wie in Tabelle 2, zwischen 59 % und 12 %; sofern ausgewiesen, bewegen sich die entsprechenden Werte für Österreich zwischen 54 % und den genannten 12 %. Hinsichtlich der Bezugsgrösse der Automatisierungspotenziale beträgt die Spanne zwischen 59 % und 36 % bezogen auf Berufe und zwischen 15 % und 12 % bezogen auf Tätigkeiten.

Die beträchtlichen Schwankungsbreiten der geschätzten Einsparungspotenziale und entsprechend grossen Differenzen zwischen den Studien, selbst für die gleichen Wirtschaftsräume, sind nicht der Hauptgrund dafür, exakten Zahlen keine übertriebene Bedeutung zuzumessen. Dennoch soll zumindest kurz auf mögliche Erklärungen eingegangen werden, um die teilweise sehr gewichtigen Unterschiede bei den resultierenden Einsparungspotenzialen verstehen zu können. Damit soll auch möglichen Fehlinterpretationen vorgebeugt werden, welche aufgrund der grossen Unterschiede die Existenz solcher Potenziale insgesamt infrage stellen könnten. So sind Unterschiede allein aufgrund uneinheitlicher, nicht exakt definierter oder definierbarer Technologien, von divergierenden Voraussetzungen und Ausgangsbedingungen in den betrachteten Wirtschaftsräumen, verschiedener Datenquellen oder statistischer Ungenauigkeiten unvermeidbar. Selbstverständlich führen unterschiedliche Forschungsansätze und Methoden normalerweise auch zu divergierenden Ergebnissen, wobei vor allem die Unterschiede zwischen berufs- und tätigkeitsbezogenen Analyseansätzen hervorstechen.

Diese gravierenden Unterschiede bei der Einschätzung der Automatisierungspotenziale – hier werden für berufsbezogene Ansätze bis zu fünfmal so hohe Prozentsätze ausgewiesen wie für Berechnungen bzw. Schätzungen auf Basis von einzelnen Tätigkeiten – weisen auf eine systematische Über- bzw. Unterschätzung je nach gewählter Methode hin. Genauso wenig wie eine hohe Wahrscheinlichkeit bei einzelnen Berufen den Schluss zulässt, dass alle Tätigkeiten, die mit diesen

Berufen zusammenhängen, automatisierbar sind, genauso wenig bedeutet eine geringere Wahrscheinlichkeit bei einzelnen Tätigkeiten, dass die damit verbundenen Arbeitsplätze zur Gänze bestehen bleiben. Plausibler wäre es anzunehmen, dass die einzelnen untersuchten Tätigkeiten entsprechend ihres Potenzials und ihrer Wahrscheinlichkeit zur Rationalisierung auch tatsächlich automatisiert werden. Offen bleibt dabei, inwieweit die freigewordene Zeit für andere Tätigkeiten verwendet wird oder die Anzahl der Arbeitsplätze entsprechend abnimmt.

Ein weiterer Grund, der zu grossen Unterschieden in den Studien zu den Automatisierungspotenzialen beiträgt, dürfte in der Festlegung von Grenzwerten zu finden sein, mit denen ein «hohes» Automatisierungsrisiko definiert wird. Willkürliche Grenzwerte führen zu willkürlichen Resultaten, wie die Autoren einer für Österreich durchgeführten Studie selbst anmerken. Die in dieser Studie ausgewiesenen 9 % der Beschäftigten basieren auf einem Grenzwert von > 70 %. Die Autoren schreiben: «Würde man die gewählte Grenze geringfügig darunter bei 60 % ansetzen, würde der Anteil der Beschäftigten, die von einer hohen Automatisierungswahrscheinlichkeit betroffen sind, auf 39,5 % ansteigen» (Nagl et al. 2017, S. 16). Analysen, welche die Werte für einzelne Berufe oder Tätigkeiten für ganze Berufsgruppen aggregieren würden, wären aussagekräftiger und nicht von einer mehr oder weniger zufälligen oder willkürlichen Festlegung von Grenzwerten abhängig.

Letztlich können Debatten über die korrekten Einschätzungen zu den Einsparungspotenzialen auch die Identifizierung konkreter Massnahmen und die Beurteilung von deren Dringlichkeit erschweren, denn «selbst die Schätzungen am unteren Ende der Skala können gravierende Auswirkungen auf einzelne Berufsgruppen oder Qualifikationsstufen haben» (Čas et al. 2017, S. 41). Die Konzentration auf die Bestimmung korrekter Werte birgt aber nicht nur die Gefahr, die Betroffenheit auf individueller Ebene oder von bestimmten Gruppen nicht angemessen zu betrachten. Auch auf der Ebene einzelner Volkswirtschaften bringt sie eine Tendenz mit sich, die Automatisierungspotenziale weniger als gesellschaftliche und wirtschaftliche Chance, sondern überwiegend als Bedrohung im Sinne von Verlusten an Arbeitsplätzen und Einkommensmöglichkeiten zu interpretieren. Dies zeigt sich einerseits an den bereits erwähnten Studien, welche zum Teil wesentlich geringere Einsparungspotenziale schätzen als die Analysen von Frey und Osborn, andererseits an der Kritik, dass diese Studien sich einseitig mit den Möglichkeiten der Einsparung von menschlicher Arbeitskraft befassen und die Entstehung neuer und zusätzlicher Arbeit durch den technischen Fortschritt ausblenden.

3.1.1.2. Kompensationseffekte

Die im vorherigen Abschnitt teilweise sehr hoch geschätzten Rationalisierungspotenziale haben nicht nur eine wissenschaftliche Debatte über die möglichen Arbeitsplatzverluste befeuert. Sie haben auch dazu beigetragen, Problemstellungen der durch KI induzierten Arbeitslosigkeit in Öffentlichkeit und Politik mehr Augenmerk zu schenken. Dabei stellt sich nicht nur die Frage, welche Tätigkeiten automatisiert werden können, welche Berufe durch KI wegfallen bzw. von ihr durchgeführt werden können oder wie sich diese Entwicklungen auf die Beschäftigung insgesamt auswirken werden. Thema ist auch, wie viel an Beschäftigung durch die Entwicklung von KI selbst angeregt wird, welche neuen Berufe entstehen, welche Qualifikationen dafür erforderlich sind oder wie durch neue Produkte oder Dienstleistungen zusätzliche Arbeitsplätze generiert werden könnten.

Es wird allerdings nur in wenigen Studien versucht, die positiven Beschäftigungseffekte zu quantifizieren. Dazu zählen etwa die Berichte des WEFs (2016; WEF 2018b). Während der Bericht aus dem Jahr 2016 noch von einem überwiegenden Nettoverlust ausgeht (siehe Tabelle 2), werden im Bericht aus dem Jahr 2018, basierend auf der Extrapolation von Umfragedaten von Grossunternehmen, für die Periode bis zum Jahr 2022 beträchtliche Nettozuwächse auf globaler Ebene für möglich gehalten (WEF 2018b, S. viii). Wie von den Autoren selbst betont, sind diese Ergebnisse mit Vorsicht zu interpretieren. Die gewählte Methode, der kurze Zeitrahmen sowie die im Vergleich zur nur zwei Jahre älteren Studie konträren Ergebnisse sind einige Gründe, dieser quantitativen Einschätzung wenig Vertrauen zu schenken. Dies gilt auch für die wenigen Studien oder Szenarien generell, welche einen positiven Gesamteffekt als wahrscheinlich ansehen. So wird auch in einer Studie aus dem Jahr 2017, die unterschiedliche Beschäftigungsszenarien für Österreich entwickelt, ein «Industrie 4.0 Frontrunner-Szenario» skizziert, in welchem neue Produkte und Dienstleistungen sowie durch eine gestiegene Wettbewerbsfähigkeit entsprechend steigende Exporte dazu beitragen, die durch die Automatisierung entstehenden Beschäftigungsverluste mehr als wettzumachen. Das Eintreten dieser Entwicklung wurde aber im Rahmen der Studie selbst als sehr wünschenswert, aber wenig wahrscheinlich eingeschätzt (Dinges et al. 2017, S. 45). Dennoch sind Überlegungen zu potenziellen positiven Beschäftigungseffekten wichtig, um das Spektrum an Handlungsoptionen umfassend analysieren zu können.

Allerdings ist dieser verständliche Wunsch, quantitative Einschätzungen auch zu den Potenzialen zusätzlicher und neuer Beschäftigung zu erhalten, aus mehreren

Gründen nicht sinnvoll zu erfüllen. Erstens sind die im vorherigen Abschnitt skizzierten Zahlen zu den Automatisierungspotenzialen nur als theoretische Potenziale zu verstehen. Sie geben an, welche Tätigkeiten aufgrund der zu erwartenden technischen Fortschritte von KI oder von Maschinen anstelle von Menschen ausgeführt werden könnten. Sie haben aber keine oder nur wenig Aussagekraft im Sinne von Prognosen bzw. Aussagen darüber, in welchem Ausmass tatsächlich in welchem Zeitraum Maschinen Menschen ersetzen werden. Sie sind eher als Maximalwerte zu verstehen, die erreicht werden könnten, wenn sich die Rationalisierungsmassnahmen auch wirtschaftlich lohnen, wenn sie durch die Politik unterstützt werden und nicht durch negative öffentliche Reaktionen be- oder verhindert werden. Sie geben in diesem Sinn eher einen Rahmen vor, der unter gewissen Bedingungen erreicht werden kann. Für die Art und das mögliche Ausmass an neu entstehenden Tätigkeiten, Berufen und Arbeitsplätzen ist selbst das Abstecken von solchen Rahmen nicht möglich.

Ein offensichtlicher Grund dafür ist, dass sich neu entstehende Bedürfnisse und neue Produkte oder Serviceleistungen bzw. die Richtung, die Innovationen einschlagen, nicht oder nur sehr eingeschränkt vorhersagen lassen. Hier fehlen, im Gegensatz zu Automatisierungspotenzialen, Daten und Vorstellungen, um seriöse und wissenschaftlich fundierte Analysen überhaupt ins Auge fassen zu können. Ein weiterer Faktor, der aussagekräftiger und verlässliche Einschätzungen verhindert, ist die Abhängigkeit der Beschäftigung von einer Vielzahl von politischen und wirtschaftlichen Entwicklungen, von denen sich einige der Beeinflussung und Gestaltung durch (nationale) Politik weitgehend entziehen. Als Ausweg wird vielfach auf historische Entwicklungen und Vergleiche verwiesen, die zeigen sollen, dass technischer Fortschritt langfristig immer mit steigender Beschäftigung verbunden war (siehe z.B. Bundesrat 2017; Nagl et al. 2017).

Die direkten Effekte des Ersatzes menschlicher Arbeitskraft durch KI bezogen auf eine konkrete Anwendung werden aus ökonomischer Perspektive in der Regel negativ sein. Der Einsatz von KI anstelle von Menschen muss sich, sofern es sich nicht um Fehlinvestitionen handelt, allein aus betriebswirtschaftlicher Rationalität in geringeren Kosten niederschlagen und daher zu einer Abnahme des Beschäftigungsvolumens führen. Als direkter Effekt gilt der Saldo, der sich aus den Einsparungen durch den Einsatz der KI als Prozessinnovation und das für die Entwicklung und Einführung dieser KI-Anwendung notwendige Arbeitsvolumen ergibt. Unter Wettbewerbsbedingungen werden sich Investitionen in KI zum Zweck der Automatisierung nur durchsetzen, wenn damit längerfristig geringere Kosten bei der Erbringung von Dienstleistungen oder der Produktion von Gütern verbunden

sind. Wenn man dabei noch berücksichtigt, dass in der Regel für die Entwicklung von KI-Anwendungen hohe Qualifikationen notwendig sein werden, während eher Routinetätigkeiten durch diese Technologien übernommen werden, dürfte der Nettoeffekt beim Beschäftigungsvolumen negativ sein.

Mit der Entwicklung und dem Einsatz von KI ist aber natürlich auch eine Reihe von Sekundäreffekten verbunden, von denen neue oder zusätzliche Beschäftigungsmöglichkeiten ausgehen können. Im Sinne von KI als Prozessinnovation wird sie zu Produktivitätserhöhungen führen. Diese könnten sich wiederum in höheren Löhnen oder geringeren Kosten für Produkte und Dienstleistungen niederschlagen und damit über eine steigende Nachfrage Beschäftigungsverluste kompensieren. Ähnliches gilt für verbesserte Positionen im Wettbewerb zwischen Unternehmen oder Nationen. Hier ist allerdings einschränkend anzumerken, dass eine gestiegene Wettbewerbsfähigkeit global keinen Zuwachs an Beschäftigung bringt, da eine bessere Position im Wettbewerb auch eine Verschlechterung für Konkurrenten bedeutet. Wie erwähnt, erlaubt KI natürlich nicht nur effizientere Prozesse oder verbesserte Produkte, sondern auch die Entwicklung von gänzlich neuen Dienstleistungen oder Geschäftsfeldern. Aus bereits genannten Gründen sind aber bezüglich der daraus resultierenden Beschäftigungswirkungen keine seriösen und verlässlichen Schätzungen möglich.

3.1.1.3. Übertragbarkeit historischer Erfahrungen

Angesichts der grossen Automatisierungspotenziale und der grundsätzlichen Unmöglichkeit, zukünftige Beschäftigungszuwächse verlässlich zu identifizieren und zu quantifizieren, liegt es nahe, sich auf historische Entwicklungen und fehlende Evidenz dafür, dass uns die Arbeit ausgehe, zu berufen (Tichy 2016). Wenngleich unbestreitbar ist, dass der technische Fortschritt eine zentrale Grundlage unseres langfristig stetig gewachsenen Wohlstands ist und dass trotz des stetigen Produktivitätszuwachses über lange Perioden hin die Arbeit nicht weniger, sondern mehr geworden ist, ist doch der Rückgriff und die Berufung auf historische Entwicklungen aus zumindest drei Gründen zu hinterfragen. Erstens spricht ein Vergleich der Rahmenbedingungen und der betrachteten Technologien dafür, dass es diesmal anders sein könnte. Der Rückgriff auf historische Entwicklungen bezieht sich oft auf die ersten Wellen der Automatisierung und den Einsatz von Industrierobotern im letzten Jahrhundert. Diese Perioden waren durch hohe Wachstumsraten, Vollbeschäftigung und damit verbunden eine Zunahme der Lohnquote und tendenziell abnehmenden Einkommensunterschieden sowie Verkürzungen der Arbeitszeit

charakterisiert – Faktoren, die sich in dieser Form in vielen Volkswirtschaften so nicht wiederfinden. Zweitens steht die Behauptung, dass es in der Vergangenheit immer so gewesen sei, auf empirisch wackeligen Beinen. Sie trifft nur zu, wenn man lange Zeiträume aggregiert betrachtet; die grundsätzlich positive Entwicklung der Beschäftigung wurde und wird aber immer wieder durch längere Perioden abnehmender Beschäftigung und hoher Arbeitslosigkeit unterbrochen. Drittens könnten gerade die Schlussfolgerungen, welche aufgrund von (a)historischen Vergleichen gezogen werden, dazu führen, dass die mit der KI verbundenen positiven Möglichkeiten sich nicht oder nur teilweise erfüllen können.

Vergangene technische Revolutionen waren dadurch gekennzeichnet, in erster Linie manuelle Arbeiten durch Maschinen zu unterstützen oder mittels Roboter zu ersetzen. Mit technischen Innovationen bei Informations- und Kommunikationstechnologien und der Computerisierung hat sich diese Unterstützung auf nicht manuelle Tätigkeiten ausgedehnt, blieb aber im Wesentlichen auf die Erledigung von oder die Assistenz bei Routineaufgaben beschränkt. Neue Formen der Digitalisierung dringen in praktisch alle Lebensbereiche ein, kaum eine private oder berufliche Aktivität bleibt davon unbeeinflusst; mit KI werden auch Aufgaben automatisierbar, welche lange Zeit als unbestrittene Domäne menschlicher Intelligenz und Entscheidungshoheit galten. Damit öffnen sich im Dienstleistungssektor, der lange Zeit als Auffangbecken für abnehmende Beschäftigung im Produktionssektor fungiert hat, gleichermassen Automatisierungspotenziale und es sind auch Tätigkeiten, die eine sehr hohe Qualifikation erfordern, zum Beispiel im Bereich Finanzdienstleistungen, medizinische Diagnosen oder juristische Entscheidungen, von einer Unterstützung und damit möglicherweise auch verbundenen teilweise Übernahme durch KI-Technologien betroffen. Aber auch die ökonomischen Rahmenbedingungen haben sich verändert oder sind im Begriff, sich durch KI zu verändern (Korinek & Stiglitz 2017). So gibt es für die letzten Jahrzehnte keine Evidenz dafür, dass Produktivitätsgewinne sich in der Entwicklung der Löhne niedergeschlagen haben, wofür unter anderem eine längerfristige Akzeptanz von hohen Arbeitslosenraten – teilweise auch technologisch bedingt – und der damit verbundene Druck auf die Lohnentwicklung verantwortlich gemacht werden können. In den folgenden Absätzen wird auf die Frage, inwieweit historische Vergleiche in diesem Bereich relevante Erkenntnisse für die Zukunft versprechen, noch näher eingegangen.

Bei den historischen Vergleichen lassen sich zwei Typen voneinander unterscheiden: einerseits Studien, welche die langfristigen Entwicklungen seit Beginn der

industriellen Revolution betrachten, und andererseits Analysen, die sich auf spezifische Technologien beschränken, zum Beispiel die Einführung von früheren Generationen von Robotern in der zweiten Hälfte oder gegen Ende des 20. Jahrhunderts (Frontier Economics 2018). Zum ersten Typ ist kritisch anzumerken, dass diese, wenngleich Aussagen bezüglich der langfristigen Tendenz zutreffend sind, Zwischenphasen mit hoher Arbeitslosigkeit und deren fatale Konsequenzen für die Menschheit ausblenden – zu erinnern ist etwa an die Auswirkungen der Weltwirtschaftskrise zu Beginn der 1930er-Jahre. Beim zweiten Typ der Vergleiche fehlt eine Überprüfung, ob die *ceteris paribus* Voraussetzung bezüglich der oben genannten ökonomischen Rahmenbedingungen (Wachstumsraten, Einkommensverteilung etc.) – eine Standardvoraussetzung für die Durchführung von ökonomischen Vergleichen – zutrifft. Da sich diese Vergleiche, zumindest implizit, auf Phasen mit stabilen Wachstumsraten der Wirtschaft, mit geringer Arbeitslosigkeit bzw. Vollbeschäftigung, mit stabilen oder steigenden Anteilen der Löhne in der funktionalen Einkommensverteilung und mit wiederholten Arbeitszeitverkürzungen beziehen, fehlt einer Übertragung auf heutige Verhältnisse jegliche Grundlage. Die Wachstumsraten sind, zumindest in Europa, wesentlich geringer als im Vergleichszeitraum, die meisten europäischen Staaten sind von einem Wiederreichen einer Vollbeschäftigung weit entfernt und der Anteil der Löhne am Gesamteinkommen ist seit Längerem im Sinken begriffen.

Wird das Argument, dass der technische Fortschritt in der Vergangenheit zu keinen längerfristigen Problemen für die Beschäftigung geführt hat, für die Zukunft fortgeschrieben, so kann dies zum Schluss führen, dass auch die durch die Digitalisierung und KI entstehenden Automatisierungspotenziale problemlos zu bewältigen sein werden. Vordergründig mag diese Aussicht beruhigend wirken, sie bringt aber beunruhigende Konsequenzen mit sich. Die Hoffnung oder Erwartung, dass Arbeitsplatzverluste in einzelnen Bereichen automatisch durch neue Beschäftigung kompensiert werden wird, verleitet dazu, untätig zu bleiben und keine Massnahmen vorzubereiten oder vorzuschlagen. Diese Beruhigung auf politischer Ebene wird bei den Betroffenen, deren Arbeitsplätze gefährdet sind oder zumindest bei Teilen der Bevölkerung im Allgemeinen, die von den grossen Einsparungspotenzialen wissen, eher Verunsicherung hervorrufen. Sie können nicht darauf vertrauen, dass ihre Sorgen ernst genommen werden und dass Wirtschaft und Gesellschaft auf diese Herausforderung vorbereitet sind. Eine passive Herangehensweise bedeutet aber zumindest implizit, dass auf eine langsame Entwicklung ohne radikale Veränderungen gesetzt wird. Man hofft darauf, dass die Automatisierungspotenziale nur zu einem Teil realisiert werden und dass dieser

Prozess langsam abläuft, anstatt darauf zu setzen, die Rationalisierungspotenziale und die damit verbundenen Produktivitätszuwächse und Wettbewerbsvorteile so gut als möglich zu nutzen. Ob und inwieweit mit solch einer Strategie auch längerfristig Beschäftigung gesichert werden kann, ist aus mehreren Gründen fraglich. Zumindest für Hochlohnländer scheinen Versuche, Wettbewerbsfähigkeit auf andere Art als durch Nutzung technischer Potenziale, etwa durch niedrige Löhne, zu erreichen, wenig erstrebenswert. Ebenso wenig ist auf andere, in der Vergangenheit wirksame Kompensationsmechanismen Verlass. Produktivitätsgewinne schlagen sich seit längerer Zeit nicht mehr in steigenden Löhnen nieder und daher auch nicht auf eine steigende gesamtwirtschaftliche Nachfrage. Ein zentraler Faktor für dieses Auseinanderklaffen ist sicher in einer auf globaler Ebene inakzeptablen Arbeitslosigkeit und damit verbundenen Lohndruck zu suchen. Ohne Gegenmassnahmen wird sich dieser Effekt über die Automatisierungspotenziale von Digitalisierung und KI weiter verstärken.

3.1.1.4. Zusammenfassung der makroökonomischen Perspektive

Wie in den anderen Bereichen, die in dieser Studie aufgegriffen und diskutiert werden, sind auch in der Arbeitswelt grosse Umwälzungen durch KI zu erwarten. Je nach Standpunkt können diese Veränderungen als Bedrohung empfunden oder als Chance wahrgenommen werden. Eine wesentliche Voraussetzung, um die positiven Effekte realisieren zu können, wird wohl darin liegen, den Menschen und seine Bedürfnisse in den Mittelpunkt zu stellen. Anlässlich ihres 100-jährigen Bestehens seit ihrer Gründung 1919 hat die Internationale Arbeitsorganisation einen Bericht (ILO 2019) verfasst, der genau diese Aufgabe erfüllen will, nämlich durch eine menschenzentrierte Agenda zu einer positiven Zukunft beizutragen. Zu den Kernzielen zählen eine Zukunft der Arbeit mit Würde, sozialer Sicherheit und Gerechtigkeit. Es wird an die im Jahr 1919 formulierten fundamentalen Rechte erinnert, welche ausreichende Löhne für eine adäquate Lebensführung, eine Begrenzung der Arbeitszeit und den Schutz von Sicherheit und Gesundheit am Arbeitsplatz umfassen. Technologien sollen genutzt werden, um zu mehr Wahlmöglichkeiten und einer besseren Vereinbarkeit von Arbeit und persönlichem Leben zu führen. In Bezug auf KI bedeutet das, dass Entscheidungen, die das Arbeitsleben betreffen, in letzter Konsequenz von Menschen getroffen werden.

Im Bericht «Zukunft der Arbeit, Zukunft der Gesellschaft» der *European Group on Ethics in Science and New Technologies* (EGE 2018) wird darauf hingewiesen, dass die technologische Entwicklung – nebst anderen Faktoren – es notwendig

machen wird, traditionelle Konzepte von Arbeit, Beschäftigung, Kapital, Identität, Gerechtigkeit, Solidarität und soziale Sicherheit und ihre gegenseitige Bezogenheit neu zu denken. Unter starkem Bezug auf ethische und grundrechtliche Anforderungen wird verlangt, zu einem breiten Verständnis von Arbeit zu gelangen, das sowohl bezahlte als auch unbezahlte Tätigkeiten umfasst. Das Innovationspotenzial technischer Entwicklungen und sozialer Arrangements solle genutzt werden, um die europäischen Wirtschaften zum Vorteil aller zu stärken. Neben individueller Weiterbildung soll auch die gesellschaftliche Fortentwicklung als kollektive Verantwortlichkeit wahrgenommen werden. In einem breiten gesellschaftlichen Dialog sollen Möglichkeiten der Entkoppelung von sozialer Sicherheit und Beschäftigung entwickelt werden, wobei insbesondere Ungleichheiten innerhalb und zwischen Gesellschaften angesprochen und vermindert werden sollen.

Der Blick auf die Literatur zeigt, dass die vielfältigen Herausforderungen und Chancen der Digitalisierung im Allgemeinen und von KI im Speziellen ebenso viele Antworten erzeugt haben. Diese reichen von einer weitgehenden Negierung, dass Probleme bestehen, bis hin zur Vorstellung, dass die Menschheit in absehbarer Zukunft ihr Leben weitgehend befreit von Mühen der Arbeit gestalten können wird. Die vielfältigen Antwortmöglichkeiten und Handlungsoptionen betreffen auch sehr unterschiedliche Ebenen. Diese reichen von Reaktionsmöglichkeiten einzelner Individuen, zum Beispiel welche Aus- oder Weiterbildungswege beschritten werden sollen, um auch in Zukunft am Arbeitsmarkt reüssieren zu können, bis hin zu den oben skizzierten nationalen oder globalen Konzepten, wie die Menschen die durch die Technik gebotenen Möglichkeiten im Sinne einer positiven Zukunft (ILO 2019) und gerechteren Gesellschaft (EGE 2018) bestmöglich nutzen können.

Die in der Literatur genannten Massnahmen zur positiven Lenkung der makroökonomischen Auswirkungen von KI, insbesondere der Automatisierung, lassen sich in drei grundsätzliche Strategien zusammenfassen:

1. **Passives Abwarten:** Diese Option ist eine Konsequenz aus der Erwartung, dass zumindest kurzfristig keine gravierenden Auswirkungen auf das Gesamtvolumen der Beschäftigung zu befürchten seien. Dementsprechend wird von der Nutzung von KI kein besonderer Handlungsbedarf verursacht. Natürlich kann es auch in diesem Fall zu grossen Verwerfungen bei einzelnen Tätigkeiten oder Berufen kommen. Diese können aber vom bestehenden Sozialsystem bewältigt oder beispielsweise durch spezifische Weiterbildungsmassnahmen abgefedert werden.

2. **Aktive Vorkehrungen:** Bei dieser Option wird davon ausgegangen, dass durch die Nutzung von KI zumindest mittelfristig so grosse Einsparungspotenziale realisiert werden, dass Anpassungen bei der Arbeitszeit erforderlich werden. In Analogie zu den Maastricht-Vereinbarungen der Europäischen Union bezüglich Staatsverschuldung könnte dann auf EU-Ebene eine Verpflichtung zu staatlichen Eingriffen eingeführt werden, falls über längere Perioden hinweg eine zu vereinbarende Rate der Arbeitslosigkeit überschritten wird. Dabei gibt es keinen Automatismus, der die Art der Massnahme bestimmt, aber eine politische Verantwortlichkeit, konkret und nachweislich wirksam tätig zu werden. Eine mögliche Option unter vielen anderen wäre eine Verkürzung der Arbeitszeit.
3. **Grundlegender Umbau:** Grundannahme dieser Option ist die Erwartung, dass die Digitalisierung generell und die Nutzung von KI im Besonderen derart grosse Produktivitätszuwächse und Einsparungspotenziale mit sich bringt und zu so gravierenden Veränderungen führen werde, dass eine Entkopplung von Arbeit und Einkommen notwendig werde; beispielsweise in Form einer Grundabsicherung ohne Verpflichtung, Erwerbsarbeit zu leisten. Dies käme einer Erweiterung der Forderung nach Mindestlöhnen in Form eines bedingungslosen Grundeinkommens gleich, welche allen Bürgern und Bürgerinnen eine würdige Existenz ermöglichen soll.

Auf den ersten Blick bestehen zwischen den ersten beiden Optionen und für den Zeitpunkt (Mitte 2019), zu dem diese Studie für die Schweiz erstellt wird, kaum Unterschiede in Bezug auf konkret durchzuführende Aktivitäten. Dennoch lassen sich Differenzen hinsichtlich der gebotenen Anreize und längerfristigen Wirkungen feststellen. Auf der Seite der Arbeitnehmer/-innen wäre mit der zweiten Variante eine Garantie verbunden, bei Bedarf auch in Zukunft eine Arbeit finden zu können, wenngleich möglicherweise im geringeren Ausmass. Mit der Vermeidung von hoher Arbeitslosigkeit würde der durch ein permanentes Überangebot an Arbeitskräften entstehende Druck auf die Löhne verringert werden. Mit dem Wegfall von Befürchtungen vor KI-bedingter Arbeitslosigkeit wäre vermutlich eine höhere Akzeptanz von Automatisierungen verbunden, was nicht zuletzt auch die Investitionssicherheit für Unternehmen verbessern könnte. Für eine Investitionsentscheidung sind die technischen Möglichkeiten und Kosten von Automatisierungsmassnahmen ebenso entscheidend wie die Kosten der menschlichen Arbeit, welche sie unterstützen oder ersetzen sollen (Decker et al. 2017). Stabile Erwartungen über Lohnzuwächse entsprechend der Produktivitätsentwicklung bieten demgemäss

Anreize für die Wirtschaft, den Einsatz von KI zu erhöhen, und tragen daher auch zur langfristigen Sicherung der Wettbewerbsfähigkeit bei.

Der Unterschied zwischen der zweiten und dritten Option besteht in der Art, wie Einkommen verteilt bzw. umverteilt werden. Bei den aktiven Vorkehrungen liegt das Schwergewicht bei einer gerechteren Verteilung der Chancen, Einkommen zu erzielen, indem das Ausmass an unfreiwilliger Arbeitslosigkeit begrenzt wird. Damit wird auch ein Beitrag geleistet, die Finanzierbarkeit bestehender Sozialsysteme, die zu einem grossen Teil von der Besteuerung von Einkommen aus Arbeit abhängen, zu garantieren, ohne dass dafür ein radikaler Reformbedarf für die nähere Zukunft erforderlich wäre. Die Einführung eines bedingungslosen Grundeinkommens basiert im Gegensatz dazu nicht auf einer gerechteren Verteilung von Einkommenschancen, sondern auf einer gleichmässigeren Verteilung der insgesamt erzielten Wirtschaftsleistung und somit auf Umverteilung als Leitprinzip. Auf internationaler Ebene finden sich einige wenige Experimente, um die Wirkung eines bedingungslosen Grundeinkommens zu prüfen, doch von einer realen Umsetzung ist diese Massnahme noch weit entfernt. In diesem Sinne bilden die drei skizzierten Handlungsoptionen auch unterschiedliche Zeithorizonte ab. Bei der ersten bleibt der Blick auf die Gegenwart gerichtet, bei der zweiten orientiert er sich an der näheren Zukunft und leistet damit möglicherweise auch Schritte ein, um für eine längerfristig mögliche radikale Entkopplung von Arbeit und Einkommen vorbereitet zu sein, was Option 3 entsprechen würde.

Die drei genannten Optionen bilden auch die Debatte in der Schweiz ab, wo beispielsweise 2016 über ein bedingungsloses Grundeinkommen abgestimmt wurde (76,9 % der Stimmenden und alle Kantone lehnten damals die Initiative ab). Die Diskussion um Rahmenbedingungen, die zur Erreichung solcher Ziele beitragen, wird auch hierzulande auf unterschiedlichsten Ebenen geführt werden und es wird Initiativen auf individueller, unternehmerischer, branchenspezifischer, Kantons-, und Bundesebene benötigen. Welche konkreten Herausforderungen auf den einzelnen Subebenen bestehen und wie diese bewältigt werden sollen, sollte wahrscheinlich am besten von den Betroffenen selbst bzw. zwischen den beteiligten Akteuren auf individueller, politischer oder sozialpartnerschaftlicher Ebene verhandelt und entschieden werden. Zu den Handlungsmöglichkeiten der Schweiz als Nation wäre anzumerken, dass natürlich auch die Schweiz als Teil der globalen Wirtschaft von deren Entwicklungen betroffen oder durch internationale Vereinbarung in ihrer Autonomie gebunden ist. Es sprechen aber einige Fakten dafür, dass die Schweiz im Vergleich zu vielen anderen Staaten mehr Freiheitsgrade besitzt.

Dazu zählen etwa eine eigene Währung, eine relativ lose Verbindung zur Europäischen Union, aber auch ökonomische Fakten wie ein hoher Wohlstand und eine geringe Arbeitslosigkeit. In diesem Sinn besitzt die Schweiz sowohl mehr Spielräume als auch weniger Dringlichkeit bei ihren Möglichkeiten, auf Herausforderungen zu reagieren bzw. sich auf diese vorzubereiten.

3.1.2. Effekte von KI auf die Gestaltung der Arbeit selbst

Wie oben ausgeführt, betreffen die Auswirkungen einer durch KI induzierten Automatisierung ein breites Feld an Berufen, wobei aber die Effekte auf die einzelnen Berufe sehr unterschiedlich sein dürften. Eine Analyse solcher Effekte auf der Ebene der Branchen oder gar einzelner Berufe wird hier nicht behandelt. Dennoch gibt es einige in der Literatur diskutierte generelle Effekte auf die Gestaltung der Arbeit selbst, die auch in der Expertenumfrage aufgenommen und diskutiert wurden. Namentlich handelt es sich hier um die Frage der Selektion von Arbeitskräften im Anstellungsprozess sowie die Überwachung von Arbeitskräften zum Zweck der Kontrolle oder der individuellen Karriereplanung (z.B. Beförderung). Indirekte Effekte von KI sind auch der Grad der Autonomie in der Arbeitsgestaltung sowie Zu- oder Abnahme bestimmter Formen von Arbeitsverhältnissen (z.B. mehr projektbasiertes Arbeiten und damit verbunden eventuell prekäre Formen der sozialen Absicherung). Diese indirekten Effekte dürften dabei aber aus der Digitalisierung generell resultieren und eine Abschätzung der Rolle von KI selbst ist hier schwierig. Nachfolgend wird eine kurze Übersicht zu diesen Themen gegeben.

3.1.2.1. Selektion von Arbeitskräften

Mittlerweile gibt es verschiedene Systeme am Markt, die den Personalverantwortlichen Unterstützung bei der Auswahl und dem Management der Angestellten versprechen. Die Rekrutierung kann ein zeitaufwendiger Prozess sein, der mit vielen anscheinend leicht zu automatisierenden Tätigkeiten eingeleitet wird. Der Vergleich aller Bewerberinnen und Bewerber im Hinblick auf die Erfüllung ausgeschriebener Kriterien und Kompetenzen wird oft schon automatisiert durchgeführt. Dabei sind Systeme im Vorteil, die die Struktur zur Erfassung aller Angaben bei der Bewerbung vorgeben. Dadurch wird der Bewerbungsprozess häufig über das Ausfüllen umfangreicher Fragebögen auf der Webseite der Firma (oder des beauftragten Recruiting-Dienstleisters) gestartet.

Bisweilen gehen Art und Umfang der Befragung schon in Richtung Assessment-Center. Es werden nicht nur biografische Fakten abgefragt, sondern Tests durchgeführt, die weitere Kompetenzen abklären sollen. Die gesammelten Ergebnisse führen zu einer KI-basierten Shortlist, mit der die Verantwortlichen weiterarbeiten.

Interessant ist in solchen Prozessen vor allem die Frage, ob durch die maschinelle Vorauswahl eine Objektivierung im Bewerbungsprozess stattfindet oder bestehende Diskriminierungen fortgeführt werden. Die Auseinandersetzung mit den Kriterien, die zur Anstellung einer Person führen, macht oft bestehende Vorurteile erst sichtbar. Darin liegt eine der Chancen im Einsatz von KI-Systemen: Bestehende diskriminierende Praktiken können offengelegt werden und als Ausgangspunkt für eine Reflexion über zugrunde liegende Entscheidungskriterien dienen. Zugleich besteht aber auch die Gefahr, dass durch die Verwendung von KI-Systemen die Erfahrung menschlicher HR-Verantwortlicher ausgeblendet wird – etwa das intuitive Gefühl dafür, ob jemand gut ins Team passen wird – und damit nicht mehr zum Tragen kommt.

Manche Systeme geben nach Beantwortung der Fragen auch Tipps an die Bewerbenden, z.B. darüber, für welche Positionen ihr Kompetenzprofil gut passen würde, wer der richtige Ansprechpartner im Unternehmen sei, welche Punkte noch offen sind bzw. bei einem persönlichen Gespräch Thema sein könnten.

3.1.2.2. Überwachung und Karriereplanung von Arbeitskräften

Überwachung am Arbeitsplatz ist grundsätzlich in vielen Ländern streng reglementiert. Die Grenze zwischen dem Erlaubten und Nichterlaubten verläuft oft zwischen einer Leistungsbeurteilung, die allgemein als zulässig erachtet wird, und einer Überwachung bzw. Beurteilung des Verhaltens, die untersagt ist. Die Grenze ist allerdings verschwommen, was auch die Durchsetzung der Regulierung durch Arbeitsinspektorate erschwert. Die neuen technischen Möglichkeiten haben in den vergangenen Jahren immer wieder dazu geführt, dass Überwachungssituationen entstanden sind. Beispielsweise können elektronische Zutrittskontrollen – vor allem, wenn sie Arbeitsräume von Toiletten und Pausenräumen trennen – vermeintlich viel darüber aussagen, wie viel jemand arbeitet. Ähnlich verhält es sich mit Videoüberwachung, die Gangbereiche kontrolliert, die aber benutzt werden müssen, um beispielsweise Toiletten zu erreichen. Auch teilautomatisierte Fertigungsanlagen können genauen Aufschluss darüber geben, wie viel Zeit jemand für einzelne Arbeitsschritte benötigt. Bis hin zur individuellen Qualitätskontrolle lassen

sich so viele Daten über Mitarbeitende erheben, die manchem geeignet erscheinen mögen, um die Nützlichkeit eines Angestellten für ein Unternehmen zu bewerten. Natürlich greift die Erfassung einer Person, ihrer Arbeitsleistung und ihres «Wertes» für ein Unternehmen aufgrund rein quantitativer Daten viel zu kurz. Das hindert manche Arbeitgeber aber nicht daran, die Vergabe von Leistungsboni u.Ä. an diese Werte zu koppeln und die Beschäftigten so unter Druck und in eine gegenseitige Konkurrenzsituation zu bringen.

Innerbetriebliche Aus- und Weiterbildung ist ein Thema, das vor allem in grossen Konzernen systematisch bearbeitet wird. KI-Systeme können hier das Kompetenzprofil einzelner Arbeitnehmerinnen und Arbeitnehmer als Ausgangslage für einen Weg zu einem Jobprofil aufnehmen und den Personalverantwortlichen die geeigneten Weiterbildungsmassnahmen, individuell zugeschnitten für alle Angestellten, vorschlagen. Damit wird jedoch in die Autonomie der Beschäftigten eingegriffen. Nicht mehr sie selbst oder sie in einem Aushandlungsprozess mit ihrem unmittelbar Vorgesetzten entscheiden, in welche Richtung sich die Karriere entwickeln soll. Vielmehr gibt ein System, das die vorhandenen menschlichen und finanziellen Ressourcen zu optimieren versucht, vor, wohin der berufliche Werdegang führt. Auch wenn solche Systeme nicht tatsächlich die Entscheidung darüber treffen, entwickeln die von ihnen erstellten Vorschläge im betrieblichen Umfeld eine Eigendynamik. Es wird schwierig oder zumindest begründungspflichtig, sich gegen so einen Vorschlag zu entscheiden. Auch durch die Zuschreibungen, die Menschen maschinellen Systemen gegenüber machen, wie vermeintliche Objektivität und dergleichen, entwickeln die Vorschläge ein besonderes Gewicht.

Während noch umstritten ist, ob der Einsatz von künstlicher Intelligenz zu weniger oder mehr Jobs für Menschen führen wird, weist die EGE (2018) darauf hin, dass der Vorteil nicht schon dadurch entstehe, dass es mehr Jobs gibt, die dann auch von jenen wahrgenommen werden könnten, die ihre bisherige Aufgabe gerade an eine Maschine verloren haben. Die verschiedenen Ansätze zum Erlernen neuer Fähigkeiten (*skilling*, *upskilling* oder *reskilling*) dürften nicht die Verantwortung für das Fortkommen auf einem stark im Umbruch begriffenen Arbeitsmarkt dem Individuum aufbürden. Es seien hier massive Investitionen in Bildung gefragt, von denen auch jene profitieren müssen, die nicht zu den sogenannten Digital Natives gerechnet werden. Die EGE hält fest, dass die Veränderungen einen Einfluss auf den Arbeitsmarkt haben werden. Sie macht jedoch nicht nur die technologischen Veränderungen für die Probleme, die entstehen werden, verantwortlich. Hinter allem stünden politische Entscheidungen. Ein Umdenken im Bereich der traditio-

nellen Konzepte von Arbeit, Beschäftigung, Kapital, Identität, Gerechtigkeit, Solidarität und sozialer Sicherheit sei notwendig. Anstatt der individuellen Weiterqualifizierung strebt die EGE eine gesellschaftliche Weiterqualifizierung an.

3.1.2.3. Indirekte Effekte von KI auf die Gestaltung der Arbeit

Die Digitalisierung ermöglicht die zunehmende Flexibilisierung der Arbeit (Bundesrat 2017). Gleicher Meinung sind auch Apt et al. (2016) in ihrer Foresight-Studie. Click- und Crowdworking sind neue Formen digital vermittelter Arbeitsteilung, die über Onlineplattformen umgesetzt werden. Die Bedeutung dieses Bereichs wird vor allem in den Dienstleistungen, der Kreativwirtschaft und im Bereich wissensintensiver Arbeit zunehmen. Dieser Trend führt zu einer Flexibilisierung der Arbeitsverhältnisse. Durch künstliche Intelligenz wird es möglich, auch wissensintensive Arbeiten zu automatisieren, indem sie in Einzelaufgaben zerlegt und auf technische Hilfsmittel übertragen werden können.

Ein oft übersehener Aspekt dieses Trends ist, dass die Entwicklung künstlicher Intelligenz und die Verbesserung und Überprüfung von generierten Algorithmen auch umgekehrt viel menschlicher Arbeit bedarf, welche über digitale Plattformen vermittelt und organisiert wird. Die bekannteste dieser Plattformen ist die von Amazon unter dem Namen Mechanical Turk oder MTurk betriebene Internetplattform, welche Mikrojobs an Mikrojobber vergibt. MTurk ist eine von vielen Plattformen, über die menschliche Hilfsarbeiten gehandelt werden, die zur Entwicklung von KI-Algorithmen benötigt werden, oft aber auch hinter vermeintlich von KI-Systemen erbrachten Leistungen stehen. Spracherkennungssoftware wird um menschliche Kontrolle und Korrektur verbessert, die Entwicklung autonomer Fahrzeuge bedarf menschlicher Aufsicht bei der Entwicklung von Bilderkennungssystemen. Da diese Arbeiten über globale Netzwerke vermittelt wird, gelten für sie weder Mindestlöhne noch Formen der sozialen Absicherung. Betschon (2019) spricht in diesem Zusammenhang von einem KI-Prekariat, welches mit miserabler Bezahlung und unter menschenunwürdigen Bedingungen an der Entwicklung von Technologien beteiligt ist oder Entscheidungen zu treffen hat, die unsere Sicherheit oder unsere politische Entwicklung in relevante Weise beeinflussen können. Dazu gehören zum Beispiel autonome Fahrzeuge oder das Aussortieren von Fake News.

Bezüglich des indirekten Effekts von KI auf die konkreten Tätigkeiten hält die ILO (2017) fest, dass es oft nicht ganze Jobs sein werden, die durch den Einsatz von KI verloren gehen, sondern eher einzelne Tätigkeiten, wodurch sich der Arbeits-

alltag der Betroffenen ändern wird und sie mit intelligenten Maschinen zusammenarbeiten werden (vgl. Mensch-Maschine-Team; Apt et al. 2016). Vor allem bei nicht erklärbaren Entscheidungen/Empfehlungen ist es wichtig, dass diejenigen, die mit dem System arbeiten, in der Lage sind, die Plausibilität der Ergebnisse abzuschätzen, um grobe Fehleinschätzungen sofort zu erkennen.

Frey und Osborne (2013) wiederum halten in ihrer Studie fest, dass Zeit für andere Tätigkeiten frei wird. Aus dem Bericht geht im Weiteren hervor, dass bereits jetzt eine Arbeitsplatzpolarisierung zu beobachten sei, da die Zahl der Arbeitsplätze sowohl in gering qualifizierten als auch in hoch qualifizierten Berufen zugenommen hätte, während dies bei mittelqualifizierten Routineberufen, zumindest in den Industrieländern, nicht der Fall sei. Der technologische Wandel betrifft nicht alle Menschen gleichermassen, das Risiko eines Arbeitsplatzverlustes bei Routine- und manuellen Arbeiten ist Studien zufolge hoch, auch in einigen Dienstleistungsbranchen. Viele der betroffenen Menschen würden niedriger bezahlte Arbeit oder Jobs, für die sie überqualifiziert sind, annehmen müssen. Dies kann dazu führen, dass neben der Langzeitarbeitslosigkeit auch prekäre Beschäftigungsverhältnisse zunehmen werden. Es stelle sich insgesamt die Frage, wie die Produktivitätsgewinne aus neuen Technologieformen wie KI geteilt werden.

Die erwarteten Veränderungen der nachgefragten Kompetenzen und Fähigkeiten erfordern gut abgestimmte Bildungs- und Ausbildungseinrichtungen. Die eigentliche Herausforderung des technologischen Wandels besteht laut ILO darin, herauszufinden, wie man Unternehmen bei der Transformation unterstützen und den Übergang von alten zu neuen Arbeitsplätzen (räumlich und in Bezug auf Kompetenzen) erleichtern kann und wie man Produktivitätsgewinne gerecht verteilt.

Die Verkürzung der Arbeitszeit in den letzten Jahrzehnten war nur durch eine Intensivierung der Arbeit erreichbar, die gewissermassen verdichtet wurde (Jürgens et al. 2017). Das gelang hauptsächlich durch höheres Arbeitstempo, die Nutzung von Automatisierung und den Einsatz von Informations- und Kommunikationstechnologie. Bei einer weiter zunehmenden Automatisierung und einer Zusammenarbeit mit intelligenten Maschinen sei anzunehmen, dass die Produktivität durch weiterhin steigende Verdichtung und Geschwindigkeit zunehmen werde.

Wie von Frontier Economics (2018) zusammengefasst, lässt sich festhalten, dass der Einfluss von KI nicht nur von der Technologie selbst, sondern auch von kulturellen, wirtschaftlichen und sozialen Faktoren abhängt. Ausserdem geht damit eine Polarisierung der Arbeit in Bezug auf Jobs für Niedrigqualifizierte und Hochqualifizierte einher. Historisch gesehen ist die Beschäftigung insgesamt nicht zurück-

gegangen, aber es gab Verschiebungen von der Fertigung zum Dienstleistungsbereich und Einkommensverluste für verdrängte gering qualifizierte Arbeitskräfte. Mehrere Studien deuten darauf hin, dass KI eine signifikante Minderheit der bestehenden Arbeitsplätze betreffen könnte und dass Berufe, die von gering qualifizierten Arbeitnehmer/-innen ausgeübt werden, eher betroffen sind als die von hoch qualifizierten Fachpersonen. Potenzielle Arbeitsplatzverluste werden kurzfristig durch Ausgleichsmechanismen teilweise kompensiert werden. Dies kann jedoch die Ungleichheit erheblich verstärken, insbesondere wenn die Arbeitgeber/-innen über eine beträchtliche Marktmacht verfügen.

3.1.3. Fazit: Themenauswahl für die Expertenumfrage

Aus der Literaturrecherche und der ersten Interviewrunde ergaben sich Leitthemen, die in der Expertenumfrage abzuklären waren. Es handelt sich dabei um Themen, die an vielen Stellen und in mehreren Quellen als Schlüsselfaktoren genannt wurden. Über diese lagen zum Teil einheitliche und zum Teil unterschiedliche Einschätzungen vor. Die Umfrage sollte hier weitere Beurteilungen bringen, mit einem Fokus auf die Entwicklung in der Schweiz. Es ging daher besonders um:

- Fragen der Verteilung von verfügbarer Arbeit und (in der zweiten Runde) um Massnahmen zur Vermeidung einer potenziell langanhaltenden hohen Arbeitslosigkeit, die Entwicklung des Beschäftigungsvolumens und die Bedeutung menschlicher Arbeitsleistung.
- Verteilung der Produktivitätsgewinne und Zunahme der Produktivität.
- Polarisierung der Löhne und Kompetenzen.
- Unterschiede für verschiedene Wirtschaftssektoren, besonders den Dienstleistungssektor.
- Überwachung und Kontrolle der Arbeitnehmer/-innen.
- Zunahme instabiler/prekärer Arbeitsverhältnisse.
- Entwicklung von individueller Arbeitszeit und Arbeitsbelastung.

3.2. KI in Bildung und Forschung[56]

Bildung und Forschung sind ein wesentlicher Motor für die Innovations- und Wirtschaftskraft einer Nation. Um dabei international eine Vorreiterrolle zu spielen, muss ein Land die strukturellen und inhaltlichen Rahmenbedingungen der Bildung und Forschung entsprechend der gesellschaftlichen Herausforderungen weiterentwickeln. Mit KI kommen auf die Gesellschaft neue Herausforderungen zu. Es ergeben sich aber auch neue Möglichkeiten, KI für Innovationsprozesse in Bildung und Forschung zu nutzen.

Nachfolgend werden zuerst Anwendungen von KI-Systemen in der Bildung aufgezeigt. Es folgt eine Gesamtsicht zur Debatte, welche Kompetenzen für den Umgang mit KI vermittelt werden sollen. Ein Überblick bildungspolitischer Aspekte auf internationaler und nationaler Ebene und eine Einschätzung der Nutzung von KI im Bereich Forschung und Innovation runden die Darstellung ab.

3.2.1. KI-Anwendungen in der Bildung

Die ersten Anwendungen von KI reichen zurück bis in die 1970er-Jahre, als durch die Nutzung von Computern das Interesse anstieg, diese auch für Lernzwecke und besonders für die Lernbetreuung (Tutoring) zu nutzen, um zeitaufwendige und oftmals nicht zugängliche Interaktionen zwischen Menschen durch Maschinen möglichst zu ersetzen (Bloom 1984). Man begann zunehmend Systeme zu entwickeln, die ebenso effektiv arbeiten sollen wie die menschliche Betreuung in der Ausbildung (Roll & Wylie 2016). Ein Fokus der Nutzung von KI ist dabei die Personalisierung des Lernens, d.h. Inhalte und Kompetenzen sollen auf individuelle Ansprüche und Fähigkeiten zugeschnitten vermittelt werden. Die grossen IT-Konzerne wie Google und Apple stehen dabei aktiv hinter der Förderung von KI im Bildungssystem. Eine Vielzahl von Unternehmen bietet zudem neue Tools für KI-unterstütztes virtuelles Lernen an. In den USA herrscht dabei ein offener Zugang zu neuer Technologie im Bildungsalltag (Mead 2016). Pilotschulen wie im Silicon Valley (Altschool 2017) bieten etwa in allen Schulklassen KI-unterstützte Lernformate an, im Zuge derer der Lernfortschritt der Schülerinnen und Schüler durch das KI-System ausgewertet und gesteuert wird. Vergleichbare Initiativen und Ansätze stehen in der Schweiz noch aus.

[56] Dieser Abschnitt beruht auf Arbeiten von Clemens Mader und Claudia Somm, Abteilung Technologie und Gesellschaft, Empa.

Nachfolgend werden KI-Anwendungen in der Bildung gemäss Holmes et al. (2019) in den Bereichen «Administration», «KI für Lernende» und «KI für Lehrende» klassifiziert und besprochen.

3.2.1.1. KI-Anwendungen zur Unterstützung der Administration

Administrationsorientierte KI-Systeme sind den Anwendungen im Management zuzuordnen, und sie sollen dazu beitragen, administrative Abläufe in Schulen effizient zu automatisieren (Villanueva 2003), also z.B. Stundenpläne zu organisieren, die Übermittlung von Hausarbeiten zu melden oder Abwesenheiten zu kontrollieren. Einige Systeme nutzen Data-Mining, um Informationen für die Administration, aber auch für die Lehrenden und mitunter auch Studierenden zu generieren. Beispielsweise können schriftliche Prüfungen über KI-Anwendungen beurteilt und zugleich Stärken und Schwächen der Lernenden dem Lehrenden mitgeteilt werden (James et al. 2008; Marsh et al. 2006). In der Schuladministration können diese Daten wiederum in der Ressourceneinteilung für die weitere Programmgestaltung genutzt werden. Abbildung 8 zeigt auf, wie auf Ebene der schulischen Verwaltung Daten gesammelt, analysiert und daraus Voraussagen getroffen und Entscheide unterstützt werden können.

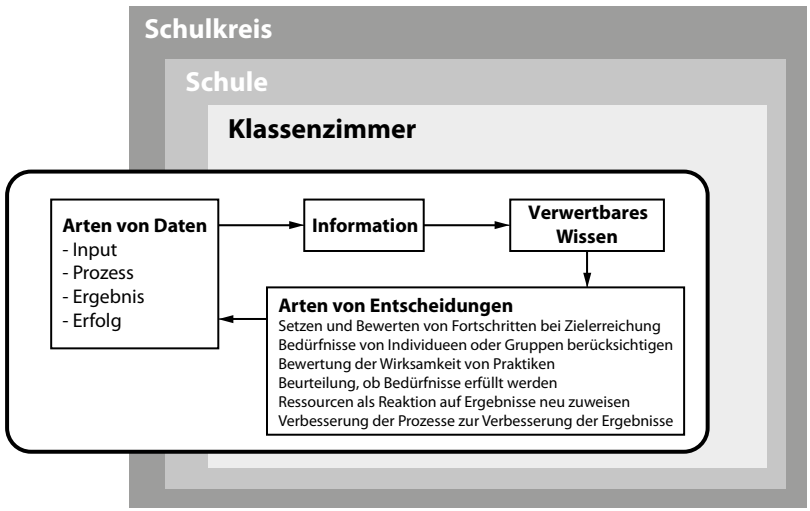

Abbildung 8: Datenunterstützte Entscheidungsfindung in der Bildung. Quellen und Typen der Daten sowie Typen der Entscheidungen (Marsh et al. 2006).

3.2.1.2. KI-Anwendungen zur Unterstützung der Lernenden

KI-Anwendungen zur Unterstützung von Lernenden haben bisher die grösste Aufmerksamkeit erhalten. Diese Anwendungen versprechen individualisiertes, personenbezogenes Lernen. Sie stützen sich auf Daten der Lernenden und erzeugen Analysen zu Potenzialen und Schwächen der Lernenden. Darauf beruhend kann KI die Lernprogramme anpassen. Nachfolgend sollen einige solche Beispiele vorgestellt werden.

Intelligente Tutorsysteme (ITS): Ein Beispiel für ein intelligentes Tutorsystem ist Squirrel AI[57]. Derartige Tutorsysteme sind wohl die meist genutzte KI-Anwendung im Bildungsbereich (Holmes et al. 2019). Für die Studierenden werden Schritt für Schritt individualisierte Tutorials für Fächer wie Mathematik oder Physik aufbereitet. Im Laufe der Arbeit der Studierenden mit dem System passt sich dieses an die Studierenden an und kann die Aufgaben entsprechend stellen.

Dialogbasierte Tutorsysteme (DBTS): DBTS nutzen *Natural Language Processing*, um Gespräche zwischen den Lernenden und dem KI-System zu simulieren. Die DBTS fokussieren auf das Fragestellen und nicht wie IT-Systeme auf Instruktionen. Konversationen zwischen den Studierenden und dem DBTS-System sollen die Selbsterkenntnis der Studierenden fördern. Ein Beispiel für ein DBTS kommt von IBM Watson: Watson beginnt mit tiefen Argumentationsfragen und bietet dann Hinweise, die den Studenten helfen, eine Antwort zu geben, die mit einer Reihe von Aussagen oder Wissenskomponenten übereinstimmt. Das System besteht aus sechs Komponenten, um diese Funktionalität zu erreichen: Domänenmodell, Dialoginhalt, Klassifizierung der natürlichen Sprachreaktion, Fragenbeantwortung, Lernmodellierung und Dialogmanagement (Chang et al. 2018). Tabelle 3 zeigt einen Ausschnitt einer solchen Konversation.

Explorative Lernsysteme (ELS): Eine Alternative zu ITS und DBTS wird durch explorative Lernsysteme geboten. Beim explorativen Lernen werden Lernende angehalten, selbst Rückschlüsse und Erfahrungen durch Experimentieren und Austausch zu sammeln. Die Rolle von KI-Systemen in diesem Lernzugang ist, diese Lernerfahrungen durch Analyse der Lernwege zu steuern. Dies geschieht durch Feedback zu Erlerntem, aber auch dem aktiven Vorbeugen einer mit explorativem Lernen assoziierten kognitiven Überforderung, die wiederum zu vermindertem langfristigem Lernen führt.

[57] Siehe: www.squirrelaius.com.

Autonome Schreibbewertungssysteme: Diese Programme nutzen *Natural Language Processing*, um während des Schreibvorgangs ein Feedback zu geben. Die Systeme werden unterschieden in formative Systeme, die die Studierenden direkt im Schreibprozess unterstützen sollen, und in summative Systeme, die eine automatisierte Bewertung anbieten. Die Entwicklung der Systeme fusst auf der Motivation, Kosten der Bewertung zu senken, hat sich aber in den vergangenen Jahren durch die direkte Rückmeldung an die Studierenden während des Schreibprozesses auch zu einem Lerntool entwickelt. Als Beispiel kann hier Open Essayist[58] genannt werden, ein von der Open University UK entwickeltes Tool. In dieser Software wird eine computerbasierte Analyse der wesentlichsten Teile und Schlüsselwörter des Texts verfasst.

Tabelle 3: Ausschnitt einer Konversation zwischen dem DBTS-System Watson und einem Lernenden (Chang et al. 2018).

Watson: What is the nature of gender differences?
Learner: it is common for gender to be defined by sex
Watson: Looks like I have something different ... How do sociologists define sex?
Learner: biological differences
Watson: Well said. Sex refers to the biological differences between men and women. There are 2 key facts left in this review. What is gender?
Learner: idk
Watson: No problem. I'll help you get this! Gender refers to socially created differences between men and women. How are gender differences primarily constructed?
Learner: socially
Watson: Very good! Gender differences are primarily socially constructed, not biologically, constructed.

[58] Siehe: https://cloudworks.ac.uk/cloud/view/9622.

Sprachlern- und Übersetzungssysteme: Sprachlernsysteme, wie beispielsweise die App Duolingo, nutzen intelligente Tutorsysteme, die kombiniert mit einem Spracherkennungssystem auf die individuellen Bedürfnisse der Lernenden beim Erlernen der Sprache eingehen. Aussprache, Vokabular und Grammatik werden je nach Bedürfnis geübt. Zwischen der App und den Lernenden können Dialoge in der zu lernenden Sprache geführt werden. Eine weitere Anwendung der Sprachlernsysteme sind Übersetzungsprogramme wie DeepL oder Google Translate. Diese können Wörter, Sätze und ganze Dokumente übersetzen und sogar auf den grammatikalischen oder sprachlichen Ausdruck achten. Den Lernenden können Vorschläge für die Übersetzung gemacht werden. Ein weiterer Nutzen ist dabei auch die Senkung von Sprachbarrieren für Lernmaterialien aus unterschiedlichen Sprachregionen der Erde.

Chatbots: Chatbots sind darauf ausgelegt, automatisch auf Fragen und Mitteilungen schriftlich oder auch verbal zu reagieren. Sie nutzen dabei Regeln und Schlüsselwörter, um vorprogrammierte oder über das KI-System individuell generierte Antworten zu liefern. In der Bildung werden Chatbots genutzt, um Informationen zu Kursen zu vermitteln, in Aufnahmeprozessen die FAQs zu beantworten oder auch direkt den Lernprozess zu unterstützen, etwa das Sprachenlernen (Holmes et al. 2019). Weitere Anwendungsfelder können Nachhilfeprogramme sein. Bei GoStudent.org[59] werden Lernende mit den idealen Nachhilfelehrern zusammengebracht und die FAQs werden aus den bereits gesammelten und bestehenden Daten über ein KI-System beantwortet.

Virtual und Augmented Reality: Virtuelle Realität und Augmented (erweiterte) Reality sind zwei Anwendungsfelder der Digitalisierung in der Bildung, die oftmals aber nicht notwendigerweise KI nutzen, um individualisierte Angebote für die Nutzer/-innen zu ermöglichen. In entlegenen Regionen Chinas werden etwa in Grundschulen, in denen nur einzelne Lehrpersonen für mehrere Fächer zur Verfügung stehen, über virtuelle Klassenzimmer Lehravatare für einzelne Fächer eingesetzt. Die vor Ort zur Verfügung stehenden Lehrkräfte begleiten die über KI gesteuerten virtuellen Lehravatare in der Vermittlung von Inhalten. Über Video- und Audiosysteme können Schüler/-innen den Lehravataren Fragen stellen oder von diesen auch gezielt angesprochen werden. Vertiefende Inhalte können somit weit skalierbar in Schulklassen ohne Fachlehrkräfte angeboten werden. Augmented-Reality-Systeme wiederum können den Schülerinnen und Schülern über die Mobiltelefone

[59] Siehe: http://www.gostudent.org.

Inhalte zu den live abgebildeten Bildern liefern. Dies können historische Gebäude in Städten sein oder auch Pflanzen und Tiere in der Natur. KI-Systeme verbessern die Erkennung der Bilder. Sie können aber auch in Verbindung mit Geo- und Nutzerdaten individuell auf die Lernbedürfnisse zugeschnittene Informationen filtern und anzeigen.

Lerngruppen-Management-Systeme: Diese Systeme werden genutzt, um das Lernen in Lerngruppen zu organisieren und über KI-Anwendungen individuelle Bedürfnisse in Gruppen zu berücksichtigen. So werden Lernende mit zueinander passenden Profilen miteinander gruppiert. Lernende, die etwa in Mathematik Schwächen haben, können mit Mathematik-Tutoren aus anderen Klassen, Schulen oder auch Ländern in Kontakt gebracht werden.[60]

Roboter: Roboter werden in erster Linie in der Schule eingesetzt, wenn im Falle einer körperlichen Beeinträchtigung der oder die Lernende nicht persönlich am Schulunterricht teilnehmen kann. Der Roboter nimmt anstelle des Lernenden im Unterricht teil und dient als Medium für Bild und Sprache in beide Richtungen. KI-Systeme werden zudem bei Menschen mit Autismus eingesetzt. Der Roboter ist dabei für den Lernenden im Lerndialog einfacher als Gegenüber zu handhaben als Personen, deren Reaktionen für die von Autismus betroffenen Personen nicht einschätzbar sind. Roboter unterstützen dabei die Vermittlung von Lerninhalten als auch von sozialen Kompetenzen.[61]

3.2.1.3. KI-Anwendungen zur Unterstützung der Lehrenden

KI-Anwendungen, welche die Lehrenden in ihrer Arbeit gezielt unterstützen, sind bisher eher rar. Dieses Manko wurde jedoch in den vergangenen Jahren erkannt und so kommen nach und nach auch Lernanwendungen mit erweiterten Funktionen für die Lehrenden auf den Markt. Führend ist dabei Century[62], ein Start-up aus England, das KI-unterstützte Lernsoftware für Schulen anbietet und zugleich auch den Mehrwert für die Lehrenden unterstreicht. Das Unternehmen berichtet von beträchtlichen Zeitersparnissen der administrativen Arbeit bei den Lehrenden von wöchentlich bis zu sechs Stunden. Das System unterstützt dabei die Lehrenden

[60] Siehe: https://thridspacelearning.com.

[61] Siehe *Kaspar the social robot*: https://www.herts.ac.uk/kaspar/the-social-robot.

[62] Siehe: https://www.century.tech/about-us/.

in der Beurteilung, in der Lerngruppenzusammensetzung und der individuell gezielten Förderung der Lernenden.

3.2.2. Kompetenzen für den Umgang mit KI

3.2.2.1. KI-Kompetenzen für Lernende

Wie bereits im Abschnitt Arbeitswelt ausgeführt, ist es für die wirtschaftliche Entwicklung wesentlich, Kompetenzen für den Umgang mit KI zu fördern. Diese sind auch nötig, um den künftigen beruflichen Anforderungen zu genügen. Dies stellt die Frage, welche Kompetenzen auf welchen Stufen erworben werden sollen.

Diese Frage ist eingebettet in den breiteren Kontext der Festlegung von *digital skills* bzw. *digital literacy*, die als nötig erachtet wird, um den digitalen Wandel zu bewältigen. Darunter wird die Fähigkeit verstanden, sich in vielerlei Hinsicht in einem technologischen Umfeld zurechtzufinden, also z.B. die Fähigkeit, mittels digitaler Medien Informationen zu suchen und zu beurteilen (Mohammadyari & Singh 2015). Die UNESCO definiert digitale Kompetenz allgemein als «die Fähigkeit, auf Informationen sicher und angemessen über digitale Geräte und vernetzte Technologien zur Teilnahme am wirtschaftlichen und sozialen Leben zuzugreifen, sie zu verwalten, zu verstehen, zu integrieren, zu kommunizieren, zu bewerten und zu erstellen» (Antoninis & Montoya 2018). Der Begriff «digitale Kompetenz» bezieht sich also auf das gesamte soziotechnische System, welches mit dem Einsatz einer Technologie einhergeht (Martin & Medigan 2006).

Es existieren nun verschiedene Vorschläge, welche Kompetenzen dies genau sein sollen. Das von der EU entwickelten Framework DigComp[63] nennt eine Reihe von Kompetenzen, die heute benötigt würden, um digitale Technologien in einem kritischen, zuverlässigen, kollaborativen und kreativen Umfeld zu nutzen, damit der Einzelne seine Ziele im Zusammenhang mit Arbeit, Lernen, Freizeit, Integration und Beteiligung in der digitalen Gesellschaft erreichen könne (Vuorikari et al. 2016). Dieser Rahmen ist in fünf Kompetenzbereiche gegliedert:

- Informations- und Datenkompetenz
- Kommunikation und Zusammenarbeit

[63] Siehe: https://schools-go-digital.jrc.ec.europa.eu.

- Erstellung digitaler Inhalte
- Sicherheit
- Problemlösung.

Dabei ist sich die Fachwelt einig, dass die digitale Kompetenz kontextabhängig ist. In Anerkennung dieser Tatsache entwickelte die *Global Alliance to Monitor Learning* (GAML) eine Methodik zur Kartierung von Lernpfaden, um Länder, Sektoren, Gruppen und Einzelpersonen bei der Ausarbeitung von Strategien und Plänen zur Erreichung ihrer eigenen Ziele bezüglich der digitalen Kompetenz zu unterstützen, die auf den Bedürfnissen und Prioritäten ihrer spezifischen Länder- und Wirtschaftskontexte basieren (GAML 2018).

Entsprechend ist es nicht einfach, festzulegen, welche spezifischen KI-Kompetenzen (Bucher 2019) denn nun Lernenden vermittelt werden sollen. Die beiden europäischen Komitees für informatikbezogene Themen – *Informatics Europe* und EUACM – heben diesbezüglich die Notwendigkeit hervor, dass die technologischen Hintergründe, auf denen KI-Systeme basieren, bewusst gemacht werden sollten (Larus et al. 2018). Doch die für KI wohl relevanteste *digital skill* dürfte das sogenannte *Computational Thinking* (CT) sein. CT soll es den Lernenden ermöglichen, in einer von KI geprägten Gesellschaft deren Potenziale zu nutzen und mögliche Risiken zu erkennen. Die US-amerikanische *Computer Science Teachers Association* und die *International Society for Technology in Education* definieren CT für Bildungsinstitutionen als einen Problemlösungsprozess mit folgenden Merkmalen (ISTA & CSTA 2011):

- Probleme so zu formulieren, dass ein Computer und andere Hilfsmittel eingesetzt werden können, um sie zu lösen.
- Daten logisch organisieren und analysieren.
- Darstellung von Daten durch Abstraktionen wie Modelle und Simulationen.
- Automatisierung von Lösungen durch algorithmisches Denken (eine Reihe von geordneten Schritten).
- Identifizierung, Analyse und Implementierung möglicher Lösungen mit dem Ziel, die effizienteste und effektivste Kombination von Schritten und Ressourcen zu erreichen.
- Verallgemeinerung und Übertragung dieses Problemlösungsprozesses auf eine Vielzahl von Problemen.

Diese Fähigkeiten werden durch eine Reihe von Dispositionen oder Einstellungen unterstützt und verbessert, die wesentliche Dimensionen des CT sind. Diese beinhalten: Sicherheit im Umgang mit Komplexität; Hartnäckigkeit bei der Arbeit mit schwierigen Problemen; Toleranz für Mehrdeutigkeit; die Fähigkeit, mit offenen Problemen umzugehen; und die Fähigkeit, mit anderen zu kommunizieren und zu arbeiten, um ein gemeinsames Ziel oder eine gemeinsame Lösung zu erreichen.

In diesen Fähigkeiten sind Schnittstellen mit den Kompetenzen für nachhaltige Entwicklung sichtbar. Die Kompetenzentwicklung kann somit nur in einem gesamtschulischen Kontext betrachtet werden; eine gesonderte Kompetenzentwicklung für den Umgang mit künstlicher Intelligenz wäre wenig sinnvoll.

3.2.2.2. KI-Kompetenzen für Lehrende

Das Vermitteln von Kompetenzen braucht seinerseits Kompetenzen, die den Lehrenden abverlangt werden. Bezüglich *digital skills* hat die UNESCO das «ICT Competency Framework for Teachers» publiziert (UNESCO 2018). Dieses unterscheidet zwischen Wissensaneignung, Wissensvertiefung und Wissensentwicklung. Diese Wissensprozesse sollen seitens der Lehrenden auf die folgenden Kompetenzbereiche angewandt werden können:

- Verstehen von Informations- und Kommunikationstechnologien in Bildung
- Curriculum und Assessment
- Pädagogik
- Anwendung der digitalen Fähigkeiten
- Organisation und Administration
- Professionelle Weiterbildung der Lehrenden.

Für die Lehrenden geht es somit stark um aktive Wissensprozesse, um IKT nicht nur in der Lehre zu vermitteln, sondern auch selbst anwenden zu können. Dies ist bezüglich der Nutzung von KI-Systemen als Lehrtools von hoher Wichtigkeit.

3.2.2.3. KI-Kompetenzen im Schweizer Lehrplan 21

Wie in der generellen Debatte ist auch in der Schweiz die Frage nach KI-Kompetenzen eingebettet in die Vermittlung von digitalen Kompetenzen – im Fall der

Deutschschweiz betrifft das vorab den zwischen 2010 bis 2014 von der Deutschschweizer Erziehungsdirektoren-Konferenz ausgearbeiteten «Lehrplan 21»; in der Westschweiz und im Tessin wurden zeitlich verschoben ähnliche Lehrplanprojekte durchgeführt. Ziel des Lehrplans ist es, den gesellschaftlichen Umständen gerecht zu werden und die Basis zu einer zeitgemässen Herangehensweise an die Bildung zu legen (EDK 2014). Den Schwerpunkt bilden nicht Lehrziele je Schulstufe, sondern es wurden Kompetenzen definiert, die in den Fachbereichen erreicht werden sollten. Die Kompetenzen sollen einen Bezug zu realen Lebenssituationen aufweisen und sind auch überfachlich, wie etwa sozial und methodisch, angelegt.

Medien und Informatik sollen schon in der Primarschule behandelt werden, sind jedoch nicht als eigenes Fach zu finden, sondern als begleitendes Modul integriert. Die Kompetenzen dazu sollen in Verbindung mit den Fachbereichen aufgebaut werden. Die Einführung der Informatik stellt damit ab der ersten Schulklasse eine wesentliche Änderung gegenüber dem auslaufenden Lehrplan dar. Aufgrund dieser Tatsache gibt es gemäss Sieber (2017) einen Mangel an Kompetenzen unter dem Lehrpersonal und eine entsprechend starke Nachfrage für Weiterbildungen im Bereich Medien und Informatik (Jaquemart & Kobler 2017).

3.2.3. KI und Bildungspolitik

National wie auch international wurde auf politischer Ebene bezüglich der Integration von KI-Systemen in der Bildung vielfältig reagiert. Einige relevante Policies, Deklarationen und politische Prozesse werden hier aufgeführt.

3.2.3.1. Internationale Entwicklungen: UNESCO

Die UNESCO hat eine Reihe von Erklärungen, Kommuniqués und Strategiepapieren veröffentlicht, die die neuesten Anwendungen von KI-Systemen im Bereich Bildung und Forschung hervorheben.

Die **Erklärung von Qingdao** über Informations- und Kommunikationstechnologien (IKT) im Bildungswesen wurde zum Abschluss der Konferenz über IKT für die Bildungsagenda 2030 in Qingdao, China, verabschiedet (UNESCO 2015). Der Text ist die erste globale Erklärung zu IKT im Bildungswesen, die zur Koordination der internationalen Entwicklung in den nächsten 15 Jahren dienen soll. Die Erklärung von Qingdao legt dar, wie Technologie eingesetzt werden kann, um Bildungs-

ziele für Gerechtigkeit, Zugang, Qualität und lebenslanges Lernen sowie nachhaltige Entwicklung zu erreichen. Der Text hebt die verschiedenen Möglichkeiten hervor, wie die Technologie die auf dem Weltbildungsforum vorgeschlagene globale Bildungsagenda für die nächsten 15 Jahre unterstützen kann. Um das Ziel einer integrativen und gerechten Qualität der Bildung und des lebenslangen Lernens bis 2030 zu erreichen, müssen die IKT genutzt werden, um die Bildungssysteme, die Verbreitung von Wissen, den Zugang zu Informationen, die Qualität und das effektive Lernen sowie eine effizientere Erbringung von Dienstleistungen zu stärken (UNESCO 2019b). KI wird in dieser Erklärung noch nicht explizit erwähnt.

Eine Publikation der UNESCO (2019c) widmet sich spezifisch dem **Gender Gap** bezüglich *digital skills*. Sie besteht aus drei Teilen, einem Policy Paper und zwei Think Pieces. Das Policy Paper hebt die Beständigkeit des Gender Gaps im Bereich der digitalen Kompetenzen und Fähigkeiten hervor und zeigt Lösungs- und Handlungsmöglichkeiten durch Bildungsaktivitäten auf. Das erste Think Piece verweist auf den Befund, dass gerade jene Länder Europas mit der grössten Gleichstellungsrate zwischen Männern und Frauen die geringsten Anteile an Frauen mit Abschlüssen in Informatik oder Ähnlichem aufweisen. Länder mit den geringsten Gleichstellungsraten wie z.B. arabische Länder verzeichnen die höchsten Abschlussraten von Frauen in Informatik und Technologie. Das zweite Think Piece untersucht, wie KI-Assistenten, die als junge Frauen gestaltet werden (z.B. durch die Wahl der Stimme), schädliche Geschlechterverzerrungen aufrechterhalten. Es enthält Empfehlungen, um sicherzustellen, dass die anhaltende Verbreitung von digitalen Assistenten die geschlechtsspezifischen Unterschiede nicht vergrössert.

Als Ergebnis der «International Conference on Artificial Intelligence and Education» vom 18. bis 19. Mai 2019 in Peking, China, veröffentlichte die UNESCO den **Beijing Consensus on AI and Education** (UNESCO 2019). Dieser macht zuhanden der Politik folgende Empfehlungen:

- Planen Sie die künstliche Intelligenz in der Bildungspolitik als Reaktion auf die Chancen und Herausforderungen, die KI-Technologien bieten, von einem gesamtstaatlichen, multi-Stakeholder- und sektorübergreifenden Ansatz, der es auch ermöglicht, lokale strategische Prioritäten zur Erreichung der SDG-4-Ziele festzulegen.
- Unterstützen Sie die Entwicklung neuer Modelle, die durch KI-Technologien für die Durchführung von Bildung und Ausbildung ermöglicht werden, bei denen der Nutzen die Risiken deutlich überwiegt, und den Einsatz von KI-Tools,

um Systeme für lebenslanges Lernen anzubieten, die ein personalisiertes Lernen jederzeit, überall und für jedermann ermöglichen.

- Erwägen Sie, gegebenenfalls relevante Daten zu verwenden, um die Entwicklung einer evidenzbasierten Politikplanung voranzutreiben.
- Stellen Sie sicher, dass KI-Technologien eingesetzt werden, um Lehrer zu befähigen, anstatt sie zu ersetzen, und entwickeln Sie geeignete Programme zum Kapazitätsaufbau, damit Lehrer mit KI-Systemen zusammenarbeiten können.
- Bereiten Sie die nächste Generation von Arbeitskräften mit den Werten und Fähigkeiten für Leben und Arbeit vor, die in der KI-Ära am wichtigsten sind.
- Fördern Sie eine gerechte und integrative Nutzung von KI unabhängig von Behinderung, sozialem oder wirtschaftlichem Status, ethnischem oder kulturellem Hintergrund oder geografischem Standort, wobei der Schwerpunkt auf der Gleichstellung der Geschlechter liegt sowie auf der Gewährleistung einer ethischen, transparenten und prüffähigen Verwendung von Bildungsdaten.

3.2.3.2. Nationale Akteure

Das **Staatsekretariat für Bildung, Forschung und Innovation** hat 2016 erstmals den Bericht «Forschung und Innovation in der Schweiz» publiziert. Er enthält einen Schwerpunkt zu Informations- und Kommunikationstechnologien (Kapitel 12). Dieser fokussiert auf die Schweizer Forschungslandschaft; in Abschnitt 12.4 wird jedoch auch ein Blick auf die «Herausforderung für die Bildung» gelegt. Darin wird erläutert, dass die berufliche Grundbildung eine Hauptrolle bei der Zunahme der ausgebildeten Personen im IKT-Bereich seit 1990 hat. Hier lag die Wachstumsrate seit 2009 bei 8,3 % pro Jahr. 2016 waren dies 2448 eidgenössische Fähigkeitszeugnisse auf dem IKT-Bereich gegenüber 1495 im Jahr 2009 (+ 63,7 %). Darüber hinaus lag hier das Wachstum an den Fachhochschulen bei 60,7 % und an den Universitäten und ETH bei 21,6 %.

Im «Aktionsplan Digitalisierung im BFI-Bereich in den Jahren 2019 und 2020» (SBFI 2017) werden für acht Aktionsfelder im Bereich Bildung und Forschung konkrete Massnahmen geplant. Hervorgehoben werden dabei die Förderung der digitalen Kompetenzen, die Nutzung von IKT im Bildungssektor, die Anpassung des Bildungssystems an die Anforderungen des Marktes sowie eine verstärkte Zusammenarbeit zwischen den Bildungsinstitutionen. KI als Thema wird hierbei zwar

nicht explizit hervorgehoben; doch die genannten Massnahmen unterstützen natürlich auch das Ziel, Risiken der Nutzung von KI im Bildungsalltag entgegenzuwirken und Lernende auf einen von KI geprägten Markt vorzubereiten.

Die Verantwortung für Bildung tragen in der Schweiz die Kantone. Diese koordinieren ihre Arbeit auf nationaler Ebene und bilden mit der **Schweizerischen Konferenz der kantonalen Erziehungsdirektoren** (EDK) eine politische Behörde. Im Juni 2018 beschloss die EDK die Entwicklung einer Strategie für den Umgang mit Wandel durch Digitalisierung im Bildungswesen (EDK 2018). Dem folgend wurden im Juni 2019 Massnahmen zur Digitalisierungsstrategie beschlossen (EDK 2019). Wenngleich auch hier nicht spezifisch auf KI eingegangen wird, so werden doch wesentliche Herausforderungen angesprochen, die die Digitalisierung in der Bildung mit der Nutzung von Anwendungen künstlicher Intelligenz gemeinsam haben. So werden Massnahmen zur kohärenten Datennutzung, zur digitalen Transformation der Schulen, zur Förderung digitaler Kompetenzen der Lernenden, zur Stärkung der Rolle der Lehrpersonen in der Digitalisierung, zur Nutzung der Potenziale der Digitalisierung in der Bildung, zum Umgang mit neuen Akteuren sowie zur digitalen Transformation des Bildungsbehördennetzwerkes beschrieben.

Ende Oktober 2019 lancierte die EDK «edulog», die Föderation der Identitätsdienste im Bildungsraum Schweiz. Damit werden digitale Identitätslösungen der Kantone im Bildungsbereich auf einer nationalen Ebene zusammengeführt. Ziel ist es, Lernenden, Lehrenden und Mitarbeitenden der Schulverwaltungen der obligatorischen Schule und der Sekundarstufe II einen vereinfachten Zugang zu den Onlinediensten für Schule und Unterricht zu ermöglichen. Zugleich schützt die Plattform die persönlichen Daten und sichert die digitalen Zugänge zu Plattformen von Drittanbietern (edulog 2019).

Im Weiteren hat der Bundesrat 2018 eine **interdepartementale Arbeitsgruppe «Künstliche Intelligenz»** eingesetzt. Diese hat Ende 2019 dem Bundesrat eine Übersicht über bestehende Massnahmen, eine Einschätzung zu neuen Handlungsfeldern sowie Überlegungen zu einem transparenten und verantwortungsvollen Einsatz von künstlicher Intelligenz unterbreitet. Ein Kapitel betrifft das Bildungswesen. Der Bericht kommt zum Schluss, dass die Herausforderungen der Digitalisierung für die Bildung von den entsprechenden Gremien aufgenommen worden sind, die bereits eine Vielzahl von Massnahmen lanciert haben.[64] Diese

[64] Eine Übersicht findet sich im Aktionsplan im Bereich Bildung, Forschung und Innovation in den Jahren 2019–2020: https://www.sbfi.admin.ch/sbfi/de/home/das-sbfi/digitalisierung.html.

Arbeiten umfassen auch die Thematik KI. Nach Ansicht der Arbeitsgruppe stellt sich demnach derzeit kein Bedarf für weitere, über diese Massnahmen hinausgehenden Abklärungen oder Gremien.

3.2.4. KI-Anwendungen in der Forschung

3.2.4.1. Wichtigste Trends in der KI-Forschung

Wie in Abschnitt 2.2 ausgeführt, kann die Forschung zu KI auf eine jahrzehntelange Geschichte zurückblicken. Welche Themen und Trends diesen Forschungsbereich in naher Zukunft prägen werden, wurde mehrfach untersucht. So hat das Fraunhofer Institut in seiner Studie für Deutschland drei wesentliche Aktionsfelder ausgemacht (Döbel et al. 2018): Bild- und Videoanalyse, Text- und Sprachverarbeitung sowie Audiodatenverarbeitung aus heterogenen Quellen. Die Studie hat auch Forschungsziele und deren Relevanz bewertet (Tabelle 4).

Tabelle 4: Zukünftige Forschungsthemen und die Bewertung der Experten (Fraunhofer Institut 2018).

Forschungsziele	Forschungsansätze	Relevanz (1: höchste Relevanz
Verbesserung der Akzeptanz		
Nachvollziehbarkeit	Erklärbare KI Erkennung von Diskriminierung Adversial Training Robustes Lernen	1
Ausbau der Fähigkeiten		
Lernen mit zusätzlichem Wissen	Grey-Box-Modelle Lernen mit symbolischem Wissen	2
Kollaboration	Interaktives Lernen vom Menschen Meta-Lernen	4
Anpassungsfähigkeit und Flexibilität	Multitask-Lernen Transfer-Lernen Lebenslanges Lernen Multimodales Lernen	5

Datennutzung		
Lernen mit wenig Daten	Lernen aus Simulationen One-Shot- und Zero-Shot-Lernen Unüberwachtes Lernen von Labels	3
Lernen mit sehr grossen Datenmengen	Verteilte Algorithmen Lernen aus Datenströmen Lernen in Quantencomputern	6

Eine Studie von Elsevier (2018) basierend auf Publikationstrends kommt zum Schluss, dass sich die KI-Forschung derzeit in sieben Cluster gruppiert:

- Suche und Optimierung
- Unscharfe Systeme (*fuzzy systems*)
- Maschinelle Verarbeitung natürlicher Sprache und Wissensrepräsentation
- Computer Vision
- Maschinelles Lernen und probabilistisches Schliessen
- Planung und Entscheidung
- Neurale Netzwerke.

Anwendungen, wie selbstfahrende Fahrzeuge oder Roboter, sind dabei im Cluster «Planung und Entscheidung» eingegliedert. Die Studie weist darauf hin, dass der starke Fokus der Publikationen über lernbasierte Systeme den KI-Hype vor 15 Jahren über wissensbasierte Systeme ablöst. Dies zeigt auch die Wachstumsrate an KI-Publikationen von 12,9 % jährlich während der vergangenen fünf Jahre.

Die Analyse belegt die führende Rolle der USA und Chinas in der KI-Forschung. Zwar dominieren chinesische Publikationen in der Anzahl, werden aber weniger oft zitiert als die aus den USA, die wiederum auf einer starken internationalen Zusammenarbeit beruhen. Die Analysten von Elsevier vermuten daher, dass chinesische Forscher/-innen international weniger kooperieren. Für Europa hält die Studie fest: «Während Europa mehr Forscherbewegungen von der internationalen Wissenschaft nach Europa sieht als umgekehrt, ist es im regionalen Vergleich mit einem starken Nettoabfluss von akademischem Talent in die internationale Industrie konfrontiert.» Die Schwerpunkte der KI-Forschung in Europa liegen in der Suche und Optimierung, *fuzzy systems*, sowie maschinelle Verarbeitung natürlicher Sprache und Wissensrepräsentation.

Die Schweiz zählt gemäss dieser Studie umgerechnet auf die Bevölkerungsdichte mit 5516 Veröffentlichungen im Untersuchungszeitraum deutlich mehr KI-bezogene Publikationen pro Kopf als Deutschland (25 310) oder Frankreich (21 188). Gewichtet nach der Auswirkung der Zitierungen liegt die Schweiz sogar an der Weltspitze (Elsevier 2018).

Zur Förderung von KI-Forschung stehen in der Schweiz mehrere Instrumente zur Verfügung: Bereits 1998 hatte der Schweizerische Nationalfonds das Nationale Forschungsprogramm «Künstliche Intelligenz und Robotik» (NFP 23) mit einem Gesamtfördervolumen von CHF 12 Mio. ausgeschrieben. Mit dem neuen NFP 77 «Digitale Transformation» wurde 2019 erneut ein fünfjähriges Programm mit diesmal CHF 30 Mio. Gesamtfördersumme ausgeschrieben, das auch KI-Themen beinhaltet. Im Rahmen des Aktionsplans «Digitalisierung im BFI-Bereich 2019–2020» setzt Innosuisse CHF 24 Mio. für das Impulsprogramm «Fertigungstechnologien» ein. In diesem Rahmen wird auch die Anwendung von KI gefördert. Zudem unterstützt Innosuisse zehn nationale thematische Netzwerke, welche die Forschungs- und Unternehmenswelt zusammenführen sollen, so unter anderem die «Swiss Alliance for Data-intensive Services».

3.2.4.2. Nutzung von KI in der Forschung

Neben der Forschung zu KI ist für die Innovationsentwicklung die Nutzung von KI in der Forschung von grosser Relevanz. Durch starke Rechnerleistung und grosse Datensammlungen können Forschende heute KI zur Entwicklung neuer Erkenntnisse nutzen. Im Folgenden werden beispielhaft drei Anwendungen von KI in der Forschung unterschiedlicher Disziplinen aufgeführt.

Astrophysik: Im April 2019 konnten Forscher eines internationalen Konsortiums im Rahmen des *Event-Horizon-Telescope*-Projektes die Daten von Teleskopen aus der ganzen Welt so kombinieren, dass diese ein gemeinsames Bild und somit die erste Aufnahme eines schwarzen Loches ergaben. Ohne KI wäre ein Teleskop mit einem Durchmesser von 10 000 km nötig gewesen, um diese Entdeckung zu ermöglichen (National Science Foundation, 2019).

Medizin: In der Entwicklung von Medikamenten stossen Chemiker immer wieder an ihre Grenzen, wenn sie neue Wirkstoffmoleküle zielgerichtet entwerfen und auswählen müssen. Um die Moleküle mit der besten Wirkung und den wenigsten unerwünschten Nebenwirkungen auswählen zu können, braucht es hochspezialisiertes Expertenwissen, das über viele Jahre angeeignet werden muss. KI kann

sehr viel effizienter grosse Datenmengen analysieren und reproduzierbare Ergebnisse liefern. An der ETH Zürich beispielsweise werden derartige Versuche bereits durchgeführt (Schneider 2019).

Materialwissenschaften: An der Empa wird am *Advanced-Materials-Processing*-Labor KI genutzt, um in Kooperation mit dem Unternehmen Selfrag das Verfahren zum Brechen und Zerkleinern von Beton durch Blitzentladungen in seine Ursprungsbestandteile (Kiesel, Sand und Zement) zu optimieren. Der Algorithmus wertet dabei Ton und Bilddaten live aus und kann im Bereich von Millisekunden das Verfahren optimieren und anpassen. Dieser Algorithmus lässt sich auch auf weitere Verfahren wie etwa das Löten durch Laser übertragen (Klose 2018).

Abgesehen von diesen für einen spezifischen Forschungsbedarf entwickelten KI-Algorithmen, nutzt die Wissenschaft eine Reihe weiterer KI-Anwendungen disziplinübergreifend, um bestimmte Methoden und Verfahren in der Forschung effizienter zu gestalten:

Text- und Data-Mining: Durch Text- und Data-Mining können grosse Mengen an (Text-)Daten auf Zusammenhänge und auf die Wesentlichkeit einer bestimmten Fragestellung hin durchsucht werden. In der Forschung werden heute immer mehr Publikationen und Daten in kürzeren Abständen veröffentlicht. Für die Forschenden sind diese Mengen zunehmend schwierig zu überblicken. Text- und Data-Mining stellen somit heute bereits eine wesentliche Unterstützung für Forschende dar.

Plagiatserkennung: In der Plagiatserkennung spielten KI-Anwendungen bereits eine wichtige Rolle. Sie unterstützen die wissenschaftlich Verantwortlichen in der Textanalyse, beim Aufspüren ähnlich klingender Phrasen und Sätze, nicht referenzierter Information, falscher statistischer Ergebnisse sowie auch modifizierter Daten (Enago Academy 2018).

Review von Facharbeiten: Um Begutachtungsverfahren zu unterstützen, helfen KI-Anwendungen in der Analyse der Referenzen, in der Verifizierung von statistischen Daten sowie auch bei der Suche nach für die Anforderungen des Papers benötigten Gutachter (Price et al. 2013).

Hypothesengenerierung: Ein Team von IBM arbeitet seit Längerem an einem Algorithmus, der aus durch Text-Mining, Visualisierung und Analyse von extrahierten Daten neue Hypothesen generieren soll (Spangler et al. 2014).

3.2.5. Fazit: Themenauswahl für die Expertenumfrage

Zusammenfassend können für den Bereich Bildung und Forschung folgende Herausforderungen skizziert werden, die Thema der Expertenumfrage waren:

Bildung:

- **Datenschutz:** Nachdem in der Vergangenheit die Daten der Schüler/-innen meist physisch an der Schule gelagert wurden, werden über KI-Anwendungen persönliche Daten der Schüler/-innen analysiert, in anonymisierter Form an Firmen weitergegeben und für eine kaum definierbare Zeit gespeichert. Wie sollen Schüler/-innen, Schulen und auch die Gesellschaft damit umgehen? Wie kann vor einem Missbrauch geschützt und auch das «Vergessen» der Daten ermöglicht werden? Wer darf welche persönlichen Daten für wie lange speichern, besitzen oder abfragen? Ein aktuelles Fallbeispiel dazu wird in der Ausgabe von 2. Juli 2019 im Onlinemagazin «Republik» geschildert (Fichter 2019). Der Fall zeigt auf, wie Google mit der Schülersoftware G Suite bereits heute Daten von Schweizer Schülern sammelt und das Datenschutzrecht wie auch den Gerichtsstand der Schweiz umgeht.
- **KI-Kompetenzen:** Die Aus- und Weiterbildung von heutigen und zukünftigen Arbeitskräften unter besonderer Berücksichtigung von Kompetenzen, die zur Entwicklung und den Umgang von und mit KI-Anwendungen beiträgt, ist eine zentrale Herausforderung. Neben den inhaltlichen Kriterien, die weithin als KI-Kompetenzen diskutiert werden, stellt sich die Frage, wann welche der Kompetenzen bereits unter Lernenden aufgebaut werden sollen und wie die Kompetenzen zur Ausbildung unter den Lehrenden verbreitet werden.
- **Privatunternehmen:** Der Einfluss privater Unternehmen wird auf das Bildungswesen durch die Nutzung von KI-Anwendungen weiter zunehmen, und dies in einem besonders sensiblen Bereich, der zum einen stark im öffentlichen Interesse und bisher auch in der öffentlichen Hand lag und zum anderen die Nutzung persönlicher Daten umfasst. Durch die Aufbereitung und Sammlung von persönlichen Daten durch Unternehmen begeben sich Bildungseinrichtungen zudem in Abhängigkeiten, die zukünftig die Auswahl der Anbieter von KI-Anwendungen in der Bildung einschränken könnten.

Forschung:

- **Sicherheit und Eigentum:** Durch die Nutzung von KI-Anwendungen in der Forschung stellen sich auch hier wesentliche Fragen der Datensicherheit und

des Eigentums. Es ist beispielsweise unklar, wie Unternehmen mit noch nicht veröffentlichten Texten, die etwa über öffentlich kostenlos zugängliche Übersetzungsprogramme übersetzt werden, umgehen. Des Weiteren stellt sich die Frage nach der Reproduzierbarkeit von neuen Entwicklungen und Transparenz des Algorithmus. Siehe auch Abschnitt 2.5.2.2 zu rechtlichen Rahmenbedingungen KI-generierter Ergebnisse.

- **Interdisziplinarität:** In der KI-Forschung bedarf es besonders für die Entwicklung von Anwendungen der interdisziplinären Zusammenarbeit. Entsprechend müssen Kompetenzen der Zusammenarbeit zwischen Disziplinen aufgebaut sowie auch Rahmenbedingungen und Anreizsysteme dazu geschaffen werden.

Die Studie des Fraunhofer Instituts für Deutschland (Döbel 2018) wie auch die Studie des House of Lords (2018) für Grossbritannien weisen auf einen grossen heutigen und zukünftig steigenden **Fachkräftemangel** im Bereich der KI-Forschung hin. Dies gilt auch für die Schweiz, wie dem Bericht des SBFI (2016) und der dort enthaltenen Befragung zu entnehmen ist. Wenngleich die Schweiz mit 5 % der IKT-Fachkräfte am gesamten Arbeitsmarkt eine der höchsten Dichten an Informatikfachkräften ausweist, so übersteigt die Nachfrage aus Wirtschaft und Forschung das Angebot an in der Schweiz ausgebildeten Fachkräften deutlich. Die Zuwanderungsquote an IKT-Experten aus dem Ausland beträgt gemäss SECO (2016) 14,2 % und übersteigt damit die gesamtwirtschaftliche Quote von 10,5 %.

3.3. KI und Konsum[65]

Eine der ökonomisch relevantesten Anwendungen von KI-Systemen betrifft deren Nutzung an der Schnittstelle zwischen Anbietern von Produkten und Dienstleistungen und deren Kundinnen und Kunden. Im Wesentlichen dienen die Algorithmen dabei dem Ziel, Entscheidungen und Prozesse rund um den Konsum zu unterstützen, zu optimieren oder zu automatisieren (Mari 2019). Prominente KI-Anwendungen in der Kunde-Anbieter-Interaktion stellen dabei digitale Assistenten

[65] Dieser Abschnitt beruht auf Arbeiten von Anne Scherer, Assistenzprofessorin am Lehrstuhl für Marketing und Marktforschung an der Universität Zürich, mit Unterstützung von Pascal Sutter und Alexandra Hofer.

und Chatbots sowie Empfehlungssysteme und Personalisierungsalgorithmen dar. Bereits heute unterstützen etliche digitale Empfehlungssysteme und virtuelle Assistenten wie Siri, Alexa und Google Assistant Nutzer/-innen entlang des Konsum- und Kaufentscheidungsprozesses und nehmen damit massgeblich Einfluss auf die Informationen, Produkte und Dienstleistungen, die tagtäglich konsumiert werden. Solche Systeme dürften demnach die alltäglichste Form der Interaktion von Menschen mit KI sein, wenn auch vielen dies nicht bewusst ist. Das liegt nicht zuletzt daran, dass Algorithmen zur Kundensegmentierung und -profilierung oder auch zur automatisierten Preisgestaltung immer häufiger im Hintergrund Anwendung finden, ohne dass explizit darauf verwiesen wird (Gentsch 2018).

Durch die wachsende Anzahl digitaler Fussabdrücke können mithilfe von KI-Systemen selbst ohne aktive Befragung der Konsumentinnen und Konsumenten detaillierte und stellenweise auch intime Persönlichkeitsprofile erstellt werden (Kosinski et al. 2013). Gemäss aktueller Studien können beispielsweise anhand einiger Hundert Likes auf Facebook (Youyou et al. 2015) oder eines Profilbildes Persönlichkeitseigenschaften mit erstaunlich hoher Genauigkeit eingeschätzt werden (Segalin et al. 2017). Neben diesen relativ statischen Persönlichkeitsprofilen ermöglicht KI zudem den Unternehmen, zusehends das aktuelle Empfinden der Personen einzufangen, um damit das Angebot und das Kundenerlebnis dynamisch und häufig voll automatisiert darauf anzupassen und damit zu personalisieren.

Dies wirft die Frage auf, inwieweit Konsumentinnen und Konsumenten eine derart weitgreifende Personalisierung durch KI-Systeme überhaupt bemerken, eine solche wünschen und wie sie diese konkret wahrnehmen. Da für die Betroffenen meist nicht ersichtlich ist, wie KI-Systeme funktionieren, stellt das Vertrauen in diese Systeme und deren Empfehlungen eine zentrale Herausforderung für Unternehmen dar. In diesem Abschnitt soll dieses breite Feld der Nutzung von KI im Konsumbereich genauer vorgestellt werden. Zu diesem Zweck werden zuerst einige der prominentesten Anwendungsfelder etwas detaillierter vorgestellt. Danach folgt eine Übersicht über die Vorteile von KI-Systemen aus Unternehmens- und Konsumentensicht. Schliesslich werden die wichtigsten Herausforderungen von KI im Konsumbereich aufgelistet, welche aktuell in der Literatur diskutiert werden.

3.3.1. Anwendungsfelder von KI im Konsumbereich

3.3.1.1. Personalisierung und Empfehlungssysteme

Generell stellen Empfehlungssysteme Softwareprogramme dar, die den Konsumentinnen und Konsumenten helfen sollen, verfügbare Produkte und Informationen im Internet zu sortieren (Cooke et al. 2002). Schafer, Konstan und Riedl (2002) definieren Empfehlungssysteme als «any system that provides a recommendation, prediction, opinion, or user-configured list of items that assists the user in evaluating items» (Schafer et al. 2002, S. 43).

Gerade aufgrund der wachsenden Auswahl an Produkten und der Flut an Informationen hat die Nachfrage nach Personalisierung und Empfehlungssystemen stetig zugenommen. Empfehlungssysteme helfen Konsumentinnen und Konsumenten, die Informationsflut zu überblicken, Unsicherheiten zu vermeiden und effizient zu einer Entscheidung zu gelangen; Unternehmen werden befähigt, hochpersonalisierte Dienstleistungen zu geringen Kosten anzubieten und damit Kundinnen und Kunden langfristig an das Unternehmen zu binden. Unternehmen wie Google, Apple, Facebook und Amazon (GAFA) treiben daher die Entwicklung von Personalisierungsalgorithmen und Empfehlungssystemen massiv voran. Der Streaminganbieter Netflix allein investiert jährlich Millionen Dollar in die Weiterentwicklung seines Empfehlungssystems und schätzt, dass damit Einsparungen in Höhe von jährlich 1 Milliarde US-Dollar möglich sind (Gomez-Uribe & Hunt 2016).

Empfehlungssysteme sind nicht neu. Frühe Empfehlungssysteme haben Nutzer/-innen explizit zu den persönlichen Präferenzen befragt, um anschliessend auf dem Markt verfügbare Produkte zu filtern und diese nach dem spezifischen Profil des Einzelnen zu sortieren. Fokus aktueller Diskussionen ist jedoch eine ganz neue Generation von Empfehlungssystemen: Anstelle von einmaligen Eingaben stützen sich diese modernen KI-Systeme ausschliesslich auf beobachtete Daten, die über längere Zeiträume gesammelt wurden. Diese modernen Empfehlungssysteme, die maschinelles Lernen anstelle von regelbasierten Algorithmen verwenden, decken Muster in riesigen Datenmengen auf, um Prognosen für Konsumentenpräferenzen zu erstellen, die «Customer Experience» anzupassen und personalisierte Empfehlungen anzubieten (Murray & Häubl 2009).

Nutzer/-innen begegnen diesen Systemen heute tagtäglich, sei es, wenn sie eine empfohlene Wiedergabeliste auf Spotify hören, Filme auf Netflix ansehen, Informationen auf Google suchen oder Produkte bei Amazon einkaufen. Im Gegensatz

zu Input-basierten Empfehlungssystemen geben neue KI-Systeme Empfehlungen immer häufiger auch unaufgefordert ab. Zum Beispiel werden Konsumentinnen und Konsumenten auf Amazon auf Produkte hingewiesen, die für sie interessant sein könnten («Nutzer, welche diesen Artikel bestellt haben, bestellten auch ...»), und die Seite von Facebook oder YouTube wird kontinuierlich von Empfehlungssystemen strukturiert. Durch eine Einschränkung der gegebenen Informationen und eine Anpassung der Inhalte auf den situativen und personalen Kontext nehmen moderne Empfehlungssysteme starken Einfluss auf die Entscheidungsfindung der Konsumentinnen und Konsumenten.

Während traditionelle merkmalsbasierte Empfehlungssysteme zur Abgabe von einmaligen Empfehlungen konzipiert wurden (Murray & Häubl 2009) und keine Daten aus vergangenen Interaktionen einflossen, lernen moderne KI-Systeme aus Daten, die im Laufe der Zeit von einer breiten Nutzerbasis gesammelt wurden. Folglich sind diese modernen Empfehlungssysteme bewusst so konzipiert, dass sie wiederholte Interaktionen mit den Konsumenten fördern. Dies wird auch erreicht, indem die Interaktion mit den Systemen natürlicher gestaltet wird. Viele moderne Empfehlungssysteme haben sich daher zu anthropomorphen Agenten entwickelt, welche menschenähnliche Merkmale wie Stimme, Bewegung oder morphologische Ähnlichkeit nutzen (eine Übersicht der Merkmale bieten Epley et al. 2007). Insgesamt können diese modernen KI-Systeme zwar immer bessere Vorhersagen und Empfehlungen abgeben, werden aber auch – gerade durch den fehlenden aktiven Input – für die Konsumentinnen und Konsumenten zu einer intransparenten und unerklärlichen Blackbox.

3.3.1.2. Chatbots und digitale persönliche Assistenten

Chatbots und digitale persönliche Assistenten sind Softwareprogramme, die mit Menschen in natürlicher Sprache einen Dialog führen können (Shawar & Atwell 2005; Dale 2016). Wie in Abschnitt 2.2.2 dargelegt, finden sich Versuche zur Entwicklung solcher Systeme bereits in der frühen Phase der KI-Forschung, wie z.B. das System ELIZA. Diese frühen Systeme zeichneten sich jedoch durch einen stark eingeschränkten Dialogfluss aus, da nach Schlüsselwörtern gesucht und regelbasierte Antworten gegeben wurden. Diese Programme sind in der Lage, einfache Fragen zu beantworten, die einfachen Regeln folgen (z.B. «Wie wird das Wetter heute in Zürich?»). Da diese Systeme für einfache Anwendungen oft ausreichend sind, sind sie auch heute noch häufig in Gebrauch.

Die neue Generation von Chatbots und digitalen Assistenten ermöglicht durch KI eine intuitivere Interaktion mit den Konsumenten (Følstad & Brandtzæg 2017). Es werden Spracherkennungs- und Parsing-Algorithmen eingesetzt, um die Eingaben der Nutzer/-innen zu verstehen. Mithilfe semantischer Technologien werden nicht nur die Wörter, sondern auch der Kontext interpretiert (Mitchell et al. 1994). Neben textbasierten Chatbots geht hierbei die Entwicklung weiter hin zu Sprachassistenten, welche auch eine natürliche Spracheingabe des Konsumenten verstehen und mit Sprache antworten. Die Erwartungen an derartige KI-basierte Sprachassistenten sind hoch, da sie die bisher natürlichste und einfachste Interaktion mit Softwareprogrammen erlauben (Grigore et al. 2016). Zuweilen werden diese Interaktionen so natürlich, dass es für Konsumenten kaum noch erkennbar ist, ob sie mit einer Maschine oder mit einem Menschen interagieren (Dale 2016).

Sprachassistenten sind heute in Millionen von Geräten integriert: Über 700 Millionen iPhone-Nutzerinnen und Nutzern steht «Siri» zur Verfügung, 400 Millionen Nutzer/-innen können mit dem Google-Assistenten interagieren und weitere 400 Millionen können mit Microsoft Cortana sprechen (Boeing 2018). Während digitale Assistenten zunächst vornehmlich auf Smartphones integriert waren, sind sie heute auch auf PCs, Tablets, Smartwatches oder smarten Lautsprechern wie dem Amazon Echo verfügbar. Diese sogenannten *conversational interfaces* werden als eine der wichtigsten Entwicklungen an der Kundenschnittstelle gesehen. Laut dem Marktforschungsunternehmen Ovum wird die Zahl der Geräte mit digitalen Assistenten die Zahl der Menschen im Jahre 2021 übertreffen (Ovum 2017), wobei geschätzte 1,8 Milliarden Menschen weltweit solche Dienste nutzen werden (Ali 2018).

Während Chatbots und digitale Assistenten anfangs nur für eingeschränkte Lebensbereiche entworfen wurden (vgl. Ada 2019 in der Gesundheitsversorgung oder Amy 2016 für Terminvereinbarungen), entwickeln sich die Assistenten immer mehr zu einem persönlichen Berater für alle Lebenslagen (Gensch 2019). Dawar und Bendle (2018) gehen sogar davon aus, dass Konsumentinnen und Konsumenten in Zukunft *einen* vertrauten KI-Berater für sämtliche Lebensbereiche haben werden. Den Forschern zufolge müssen Unternehmen damit in Zukunft ihr Marketing stärker auf diese sogenannten *metabots* ausrichten, da diese einen Engpass hin zum Konsumenten darstellen. Aktuell kann der Wettbewerb zwischen den verschiedenen digitalen Assistenten als oligopolistisch bezeichnet werden, da den vielen Nachfragern nur wenige Anbieter gegenüberstehen (Siri von Apple, Google Assistant von Google, Alexa von Amazon, Cortana von Microsoft und Bixby von Samsung; Bundesverband Digitale Wirtschaft 2017).

3.3.2. Vorteile von KI im Konsumbereich

Der Einsatz von künstlicher Intelligenz bringt im Konsumbereich verschiedene Vorteile mit sich. Die wesentlichen Chancen sowohl für Unternehmen als auch für Konsumenten sollen im Folgenden kurz skizziert werden.

3.3.2.1. Vorteile für Unternehmen

KI-Systeme eröffnen Unternehmen etliche Wege, um **Kosten einzusparen sowie Gewinne zu steigern**. Durch einen angemessenen Einsatz künstlicher Intelligenz können Hersteller laut einer Studie von BCG (2018) beispielsweise ihre Verarbeitungskosten um bis zu 20 % reduzieren; die Effizienz, Flexibilität und der Time-to-Market lassen sich optimieren; und es können innovative, auf Kundinnen und Kunden zugeschnittene Produkte mit tieferer Vorlaufzeit lanciert und damit zusätzliche Umsätze erzielt werden. Gerade im Marketing trägt KI dazu bei, in folgenden Bereichen Kosten einzusparen und zusätzliche Gewinne zu erzielen (Gentsch 2019; Pradeep et al. 2018):

1. Detaillierte Kundensegmentierung und -profilierung: Durch KI können effizient (neue, attraktive) Kundensegmente identifiziert werden.
2. Planung, Optimierung und Personalisierung von (Loyalitäts-)Angeboten und Werbung: Die Marketingkommunikation kann durch KI gewinnbringend und zielgerichtet angepasst werden.
3. Dynamische Preisgestaltung weiter personalisieren: Preise können kontextabhängig und vollautomatisiert auf die individuellen Preisbereitschaften angepasst werden.
4. Automatisierung des Kundenservice: Kontakt mit Kunden kann durch Chatbots und digitale Assistenten effizienter, günstiger und gewinnbringender gestaltet werden.

Wie eine Studie von Adobe (2018) zeigt, zeichnen sich gerade erfolgreiche Unternehmen durch den Einsatz von KI im Marketing aus. Demnach würden die erfolgreichsten Unternehmen mit einer Wahrscheinlichkeit von 28 % KI für Marketing einsetzen, die anderen nur mit 12 % Wahrscheinlichkeit.

Im Weiteren können KI-Systeme in allen Phasen des **Marketingprozesses** Entscheidungen unterstützen: in der Situationsanalyse, in der Marketingstrategie, bei

Marketingmixentscheidungen sowie in der Implementation und der Kontrolle. Ein Beispiel für den Einsatz künstlicher Intelligenz im Bereich Marketingmixentscheidungen bietet der Automobilhersteller Volkswagen. Die Mediaplanung wird vollumfänglich durch KI-Systeme unterstützt, die voraussagen, welche Investitionen in welchen Medien sinnvoll sind und welches Vorgehen die besten Renditen bringt (Torcasso 2018). In der Implementierungsphase unterstützen Algorithmen das Schalten von personalisierter Werbung, das Lancieren einer Webseite oder auch das Erstellen, Personalisieren und Senden von Marketingkampagnen per E-Mail (Gentsch 2019).

Während in manchen Bereichen bereits seit Jahren Algorithmen erfolgreich im Einsatz sind (z.B. für die Preisgestaltung oder Kundensegmentierung), unterstützen neuartige KI-Systeme auch immer mehr Kreativbereiche. In der Werbebranche werden so vermehrt «Rapid Advertising Development»-Methoden angewandt und in der Produktentwicklung mithilfe von Algorithmen zeitnah die besten Prototypen identifiziert (Pradeep et al. 2018). Auch im Kontakt mit Konsumentinnen und Konsumenten kommen Algorithmen häufiger zum Einsatz, beispielsweise wenn Kundenrückmeldungen per Sentiment-Analyse automatisch an den geeigneten Kundendienst weitergeleitet werden (Future Customer 2018).

Auch auf die **Marktforschung** wird KI einen erheblichen Einfluss haben. So kann traditionelle Marktforschung, die nur einen kleinen Teil der Bevölkerung abdeckt und den Konsum nur indirekt messen kann, durch Daten zum realen Verbrauch ergänzt und sogar ersetzt werden. Kunden müssen nicht länger direkt zu ihren Einstellungen befragt werden; vielmehr können Einstellungen und Absichten, wie z.B. die Kundenzufriedenheit oder ein Wechselwunsch, durch KI-Systeme gemessen und sogar vorhergesagt werden. Dabei können Verhaltensmuster und -vorhersagen auch mithilfe von Datenquellen erstellt werden, die über die digitalen Fussabdrücke in sozialen Netzwerken hinausgehen (Azucar et al. 2018). Die Analyse des Whatsapp-Verhaltens, ohne Einblick in den Inhalt der Nachrichten, reicht beispielsweise für ein überzeugendes demografisches Profil (Rosenfeld et al. 2018) und die Einschätzung der Emotionslage (Ghosh et al. 2017) der Nutzer. Ebenso erlauben Ortungsdaten von Smartphonenutzern in Kombination mit einer Kategorisierung der besuchten Orte (Mohamed & Abdelmoty 2017) oder der Mobilfunk-Metadaten (de Montjoye et al. 2013) ein detailliertes Nutzerprofil. Das Start-up beyond verbal zeigt, wie Stimmaufzeichnungen zur Analyse des Gesundheitszustandes genutzt werden können (Singer 2013). Auch in der Schweiz wer-

den derartige Systeme bereits erfolgreich getestet. Die Swisscom stellte z.B. kürzlich ein Programm vor, welches anhand von Stimmdaten den Anrufer eindeutig identifizieren kann (Swisscom 2019).

Durch all diese Methoden können KI-Systeme zu Aussagen und Ergebnissen führen, die mittels traditioneller Marktforschung nicht erreicht werden können. Selbst private Merkmale und Eigenschaften, die Konsumentinnen und Konsumenten nicht aktiv mitteilen (möchten), können mit hoher Wahrscheinlichkeit geschätzt werden, oft ohne dass dies den Betroffenen bewusst ist (Kosinsiki et al. 2013). Die Vorhersagen werden hierbei umso genauer, je mehr Daten dem System vorliegen. Die Daten werden daher häufig als das neue «Öl» für Unternehmen bezeichnet (Gentsch 2019).

3.3.2.2. Vorteile für Konsumentinnen und Konsumenten

In Zeiten der Digitalisierung wird **Informationsaufbereitung und -kuration** als eine der Hauptaufgaben künstlicher Intelligenz diskutiert (Chalmers 2018). Laut Informationen des sozialen Netzwerks Facebook haben Nutzer/-innen dieser Plattform beispielsweise täglich über 1500 neue Beiträge im eigenen Netzwerk, während im Schnitt 300 davon von einer Person pro Tag tatsächlich konsumiert werden (Luckerson 2015). Um den Konsumentinnen und Konsumenten möglichst relevante und interessante Inhalte zu liefern – und damit möglichst lang beim Unternehmen zu halten –, werden daher Inhalte und Angebote von Unternehmen wie GAFA mithilfe fortgeschrittener Algorithmen personalisiert. Personalisierung endet hierbei heute nicht mehr bei einer gezielten Werbemassnahme, sondern zielt vielmehr auf ein individuelles, personalisiertes Kundenerlebnis ab.

Im Allgemeinen beschreibt Personalisierung eine Rekonfiguration der informationellen Umgebung (Lenk 2018). Bei personalisierten Onlinediensten geschieht dies durch vollautomatisierte KI-Systeme, welche die Umgebung mittels einer feingranularen Klassifikation der Nutzer/-innen auf ihre zugeordneten Persönlichkeitswerte und Bedürfnisse anpassen (Krafft & Zweig 2018). Zusätzlich zu diesem Persönlichkeitsprofil arbeitet der Dienst mit einem ergänzenden Auftrag des Nutzers (beispielsweise einer Suchanfrage) oder ohne dessen explizite Eingabe (beispielsweise der Feed auf Facebook).

Frühe Anwendungsbeispiele für Personalisierungssysteme finden sich in der digitalen Kommunikation: Spamfilter von E-Mail-Diensten werden spätestens seit der Jahrtausendwende erfolgreich auf die automatisierte Aussortierung unerwünschter E-Mails trainiert. Seit 2013 sortieren viele Anbieter für einen besseren Überblick eingehende E-Mails automatisch in verschiedene Posteingänge (Allgemein, Soziale Netzwerke, Werbung, Foren, Updates etc.). Seit 2018 profitieren Google-Nutzer/-innen von einem automatischen Vervollständiger für E-Mails, welcher personalisierte Satzbausteine zur Verfügung stellt.

Heute begegnen Konsumentinnen und Konsumenten im Internet tagtäglich Personalisierungsalgorithmen, sei es auf Netflix, Amazon, Facebook oder Google. Die angezeigten Inhalte werden auf die Präferenzen des Individuums abgestimmt und angepasst, was insbesondere auch im Bereich Medien wichtige Konsequenzen hat, die dort diskutiert werden (Abschnitt 3.4). Dies trifft natürlich auch auf Werbeinhalte zu. Wie Forscher zeigen konnten, kann durch ein solches *psychological targeting* die Nutzungs- bzw. Kaufabsicht des Konsumenten für ein persönlich zugeschnittenes Angebot erhöht werden (Matz et al. 2017). Auch wenn das Vertrauen der Nutzer/-innen in Personalisierungsalgorithmen im Zuge des Cambridge-Analytica-Skandals 2018 grundsätzlich gesunken ist, ergab eine darauffolgende Studie in Deutschland, dass knapp ein Drittel der Teilnehmenden personalisierte Werbung als legitimes Finanzierungsmittel von Social-Media-Plattformen sieht (PwC 2018). Dennoch bevorzugen Konsumentinnen und Konsumenten gemäss dieser Studie einen kostenlosen Service mit *nicht* personalisierter Werbung (74 %) oder eine pauschale Bepreisung der Dienstleistung ohne Werbung (48 %).

Analog zu Unternehmen sind KI-Systeme auch für Konsumentinnen und Konsumenten eine **Entscheidungshilfe**. Sie entlasten diese, wenn es darum geht, eine Vorauswahl an relevanten Optionen zu erstellen, und ermöglichen zunehmend die Automatisierung von einfachen Routineeinkäufen und Kaufentscheidungen (Dawar & Bendle 2018). Wachsende Erfahrungswerte und Datenmengen erzeugen Netzwerkeffekte, welche KI-Systeme für den Konsumenten immer nützlicher machen. Mit zunehmender Genauigkeit können die Systeme vorhersagen, welche Kombination aus Preis, Funktion und Leistung für die Nutzer/-innen wichtig ist und letztendlich deren Kaufentscheidungen unterstützen oder sogar übernehmen. Einer Person mit ausgeprägtem ökologischem Bewusstsein könnte so beispielsweise ein kostspieligeres Produkt vorgeschlagen werden, wenn dieses mehr Nachhaltigkeit verspricht (Curran 2018).

In der Mobilität unterstützen KI-Systeme seit Jahren Konsumentinnen und Konsumenten in der Entscheidungsfindung. Navigationsdienste wie Google Maps zeigen Personen die schnellsten Strecken zu gewohnten Zielen an, erinnern diese, wenn es Zeit ist, den Weg anzutreten, oder warnen, wenn auf dem täglichen Weg zur Arbeit mit Verspätungen zu rechnen ist (Lakshmanan 2019). Um die Wegfindung für Nutzer/-innen zu optimieren, misst Google Maps anonymisiert die Ortungsdaten und Bewegungsgeschwindigkeiten aller Nutzergeräte, die nicht vom sogenannten *traffic crowdsourcing* abgemeldet wurden (Official Google Blog 2009). Die Summe aller aufgezeichneten Daten kann Verkehrsprobleme identifizieren und damit den Nutzerinnen und Nutzern helfen, den besten Weg zu finden.

Künftig werden KI-Systeme Ziele, Präferenzen und Verhalten der Konsumentinnen und Konsumenten noch besser verstehen und ihren Alltag weiter erleichtern (Rossow 2018). Eine Umfrage in den USA zeigt, dass einige bereits heute bereit wären, wichtige Entscheidungen an eine KI zu delegieren, allem voran die Ruhestandsplanung (33 %) oder die Schulwahl der Kinder (27 %; CBS News 2016).

3.3.3. Herausforderungen von KI im Konsumbereich

Die zunehmende Verbreitung von KI im Konsumbereich kann kritisch diskutierte Aspekte der Digitalisierung wie beispielsweise Datenschutz akzentuieren. Hier folgt nun eine Zusammenstellung der wichtigsten Herausforderungen, die in der Expertenumfrage detaillierter untersucht wurden.

3.3.3.1. Erkennbarkeit der KI

Für viele Konsumentinnen und Konsumenten ist heute mehr denn je undurchsichtig, wann und wie KI-Systeme eingesetzt werden. Gemäss einer aktuellen Umfrage des sozialen Netzwerks Facebook ist beispielsweise einem Grossteil der Nutzer/-innen nicht bewusst, dass der persönliche Newsfeed durch einen (nicht einsehbaren) Algorithmus des Unternehmens gesteuert wird (Karahalios 2014). Auch in der Schweiz wird die Erkennbarkeit von KI-Systemen immer häufiger thematisiert. Im Zürcher Hauptbahnhof wurden 2018 beispielsweise einzelne Werbebildschirme mit Kameras und Detektoren ausgestattet, mit welchen das Start-up Advertima jeweils Alter, Geschlecht und Bewegungen der Passanten berechnet, um eine persönlich zugeschnittene Werbung anzuzeigen (Weinmann 2018). Nach Recherche der IG Plakat Raum wurde der Zürcher Hauptbahnhof zudem mit 840

Beacons ausgestattet, welche sich mit den Smartphones der Passanten verbinden, um Wege «möglichst lückenlos mit Aussenwerbung abzudecken» (Metzler & Siegrist 2019). Zwar wurde das aktuelle Schweizer Datenschutzrecht geprüft und eine Kommunikationsstrategie für eine langfristige Lancierung aufgegleist; dennoch wurden die Passantinnen und Passanten nicht vorgängig über diese Systeme informiert und eine Einwilligung wurde grundsätzlich vorausgesetzt.

Auch immer menschenähnlicher wirkende digitale Assistenten reduzieren die Erkennbarkeit von KI-Systemen. Ein gutes Beispiel für diesen Umstand stellt Googles Vorstellung von Duplex im Jahr 2018 dar: In der Einführung telefoniert der digitale Assistent mit einer anthropomorphen Stimme und Sprache («hm», «aha» etc.), um im Namen einer Person einen Tisch im Restaurant zu reservieren oder einen Termin für den nächsten Friseurbesuch zu vereinbaren (Yaniv et al. 2018). Die Tatsache, dass Google in den präsentierten Anrufen darauf verzichtet, den digitalen Assistenten als solchen einzuführen, hat in der Folge zu umfangreichen Diskussionen in der Öffentlichkeit geführt. Google hat nach diesen Reaktionen eingeräumt und zugesichert, den digitalen Assistenten zukünftig darauf hinweisen zu lassen, dass es sich um einen maschinellen Anrufer handelt (Siegle 2019).

3.3.3.2. KI als Blackbox für Konsumentinnen und Konsumenten

Das bereits diskutierte Blackbox-Problem für bestimmte Anwendungen des maschinellen Lernens (siehe Abschnitt 2.2.4.2) akzentuiert sich bei praktischen Anwendungen von KI im Konsumbereich unabhängig von der verwendeten KI-Technologie. Für Konsumentinnen und Konsumenten ist nicht nur schwer ersichtlich, *wann* KI-Systeme eingesetzt werden, sondern auch *wie* und *wofür*. Selbst wenn diese wissen (könnten), dass persönliche Daten aufgezeichnet werden, heisst dies nicht, dass sie wissen, welche Aussagen fortgeschrittene Algorithmen auf Basis der vorhandenen Daten mit welcher Detailgenauigkeit treffen können. Umfangreiche Medienberichte, wie beispielsweise über den US-Retailer Target, der durch einen Algorithmus die Schwangerschaft einer jungen Frau korrekt vorhergesagt und passende Coupons an die Kundin versendet hat, sind ein gutes Indiz dafür (Hill 2012).

Auch in der Schweiz gibt es vermehrt Diskussionen in der Öffentlichkeit zum zweckdienlichen Einsatz von KI. Die Swisscom hat im Februar 2019 beispielsweise bekannt gegeben, dass Anrufe, welche ohne Opt-Out des Konsumenten

«zu Schulungszwecken aufgezeichnet» wurden, zur Generierung von Stimmprofilen und damit auch der Anruferidentifikation dienen können (Franke 2019). Eine Wissenschaftsjournalistin und KI-Expertin hat den zugrunde liegenden Dienst *Precire* getestet und berichtete: «Und jetzt behauptet dieser Sprachcomputer doch glatt, ich sei sehr neugierig (8 von 9 möglichen Punkten), verträglich (8), kontaktfreudig (7), [...] sei nicht besonders ausgeglichen (4) – und eben nicht besonders gut organisiert (4). Das Beängstigende: Der Algorithmus hat in fast allem recht. Ich fühle mich ertappt» (Wolfangel 2018). Obwohl die Nutzung der Stimmdaten datenschutzkonform war und die Resultate des KI-Systems zu Persönlichkeit, Gesundheit, Gewicht, Alter und Lohn der Anrufer vorerst nicht verwendet wurden, musste Swisscom den Dienst aufgrund öffentlicher Aufregung einstellen (Swisscom 2019). Im Mai 2019 wurde bekannt, dass viele andere Firmen diesen Dienst ebenfalls ohne explizites Einverständnis der Betroffenen eingeführt haben, eine beispielhafte Petition gegen dieses Vorgehen konzentriert sich dabei auf die Schweizerische Post (Campax 2019).

Forschende zeigen, dass die algorithmische Blackbox zu einer Aversion (Dietvorst et al. 2015) führen kann, d.h. zu einem allgemeinen Misstrauen gegenüber Ergebnissen, deren Herleitungen unklar sind. Im Einklang mit dieser These haben empirische Befunde ergeben, dass die Konsumentinnen und Konsumenten eher einer suboptimalen menschlichen Empfehlung vertrauen als der – wenn auch besser passenden – Empfehlung eines KI-Systems, wenn dieses nicht (genügend) verstanden wird (Yeomans et al. 2017). Einige Studien haben daher eine erhöhte Transparenz auf Websites vorgeschlagen (Wang & Benbasat 2007) und getestet (Buell & Norton 2011), um den Nutzern einen Einblick in die dahinterliegenden Prozesse zu vermitteln und das Vertrauen zu erhöhen. Um der algorithmischen Aversion der Konsumentinnen und Konsumenten zu begegnen, schlagen Yeomans et al. (2017) vor, Massnahmen in moderne KI-Systeme zu integrieren, die eine «illusion of explanatory depth» (Rozenbilt & Keil 2002) erzeugen – also den Betroffenen das Gefühl geben, sie würden das System verstehen.

Generell wird von Konsumentinnen und Konsumenten häufig gefordert, mehr Aufmerksamkeit für das Verständnis der verwendeten Informationen aufzubringen. Doch gerade die algorithmische Blackbox macht ein tiefergehendes Verständnis von Zweck und Funktionsweise für die Betroffenen unmöglich. Die Verantwortung auf die Betroffenen abzuschieben, muss daher infrage gestellt werden.

3.3.3.3. Blindes Vertrauen in KI

Nebst einer Aversion gegen Blackbox-Algorithmen gibt es ein umgekehrtes Phänomen. Aktuelle Studien belegen, dass Konsumentinnen und Konsumenten automatisierten Empfehlungen durch KI-Systeme in gewissen Situationen zu stark vertrauen (*algorithmic appreciation*; Logg et al. 2019) – unter anderem deshalb, weil die Ergebnisse als objektiver und neutraler bewertet werden (Jargo 2017). Dabei können Empfehlungen für Filme, Musik, Partner, Jobs u.a. wirtschaftlich oder politisch motiviert und nicht nur im Interesse der Betroffenen sein. Dies kann bewusst oder auf Grundlage mangelhafter Datensätze geschehen (Larson & Angwin 2016).

Neben der technischen Korrektheit gilt es, auch die Interessen zu hinterfragen, die zwischen Nutzer/-innen und Anbietern nicht zwangsläufig übereinstimmen. «Vertrauen» wird hier im metaphorischen Sinne verwendet, vergleichbar mit funktionaler Sicherheit. Arbeitet der Algorithmus in einer Weise, die die Person erwartet (einschliesslich eines festgelegten Zwecks und einer akzeptablen Fehlerquote)? Oder folgt das KI-System vielmehr einem verdeckten Interesse? Ein Beispiel für diesen Interessenkonflikt liefert Amazon: Eine Untersuchung des digitalen Assistenten «Alexa» zeigt, dass Amazon meist zwei Optionen liefert, bei welchen die hauseigenen Produkte achtfach häufiger empfohlen werden, als Verkaufszahlen erklären könnten (Cheris et al. 2017).

Grundsätzlich erwarten Nutzer/-innen von einem KI-System Ergebnisse, welche ihre Bedürfnisse ins Zentrum stellen und nicht jene von Drittparteien oder der Plattform selbst. Häufig wird dabei übersehen, dass diese Plattformen auch weiteren Interessengruppen dienen (Dawar & Bendle 2018). Suchmaschinen wie Google werden daher rechtlich dazu verpflichtet, organische Suchergebnisse deutlich von gekauften Ergebnissen abzugrenzen. Bereits 2013 beklagte die *US Federal Trade Commission*, dass diese Vorschrift zunehmend seltener eingehalten werde, obgleich die Markierungen heute noch deutlich subtiler geworden seien (Marvin 2019). Folglich stellt sich auch hier die Frage, in welchem Interesse und mit welchen Zielvorgaben die KI-Systeme handeln sollen. Auch scheint ungeklärt, ob und wie ein Interessenabgleich zwischen Nutzerinnen und Nutzern sowie den Anbietern erzielt werden soll.

3.3.3.4. Datenmonopole und Netzwerkeffekte

Nicht alle Herausforderungen von KI im Konsumbereich betreffen die direkte Interaktion zwischen Vertreiber und Konsument. Es stellen sich auch wettbewerbsrechtliche Fragen, die aus der Tatsache folgen, dass der Zugang zu und die Kontrolle über Daten ein entscheidender Faktor bei der Entwicklung von künstlicher Intelligenz ist. Eine ungleiche Verteilung kann daher eine Konzentration an Verhandlungsmacht und Kapitalanlagen auf eine kleine Anzahl an Unternehmen zur Folge haben. Eine Untersuchung des Google Research Center bestätigt, dass die Datenmenge ein elementares Kriterium für die Effizienz und Präzision von KI-Systemen darstellt (Sun et al. 2017). Dies zeigen auch vollständige Veröffentlichungen von Open-Source-Software (Meth 2015) im Vergleich zu zurückhaltenden Veröffentlichungen von dienlichen Datensätzen (Simonite 2017). Mit immer grösser werdenden Datensätzen aufgrund der Interaktion des Individuums mit einem bestimmten System können KI-Systeme laut Dawar und Bendle (2018) so gut werden, dass der Aufwand, ein neues System mit eigenen Daten zu «trainieren», eine starke Wechselbarriere für Konsumentinnen und Konsumenten darstellen könnte, falls die Daten nicht portabel sind.

Bereits heute sind wachsende Anteile der digitalen Plattformen in ihren Marktbereichen zu beobachten. Facebook (einschliesslich Instagram, Messenger und WhatsApp) besitzt rund 66 % des Marktanteils von Social Media auf mobilen Geräten, Google hält rund 95 % des Marktanteils in der Suchmaschinenwerbung (Daten für die Schweiz; Statcounter 2019). Diese Anteile basieren zum Teil auf nachfrageseitigen Skaleneffekten. Diese beschreiben den Umstand, dass der Wert eines Systems wesentlich durch die Anzahl der anderen Personen innerhalb des Netzwerks bestimmt wird, auch wenn davon nur wenige im direkten Kontakt zum Individuum stehen (Combs 2017). Neben diesen *economies of scale* tragen laut Dawar und Bendle (2018) auch *economies of scope* zur Konzentration bei, d.h. Konsumentinnen und Konsumenten wollen Transaktionskosten reduzieren und bevorzugen deshalb eine Plattform für zahlreiche unterschiedliche Tätigkeiten.

Aus diesen Gründen haben Anbieter mit grossen Datenbeständen erhebliche Wettbewerbsvorteile. Bergemann und Bonatti (2018) geben einen detaillierten Überblick über die Informationsmärkte des 21. Jahrhunderts und unterscheiden zwischen verschiedenen Modellen von Datenmärkten. Sie unterstreichen insbesondere die Relevanz des indirekten Informationsverkaufs. Dieser beinhaltet nicht den Verkauf der Konsumentendaten selbst, sondern den gezielten Zugang zu den

Nutzerinnen und Nutzern. So werden keine persönlichen Daten an die Kundinnen und Kunden weitergegeben. Mit dieser zentralen Rolle können Plattformen beide Seiten des Marktes (Reichweitennachfrager und Nutzer/-innen) bespielen, was ihnen erhebliche Verhandlungsmacht einbringt. Darüber hinaus werden kleine verwandte Plattformen und Datensätze oft von den Marktführern gekauft oder mitfinanziert (Yang & Ji 2016). Facebook kaufte beispielsweise in den letzten Jahren Parse (App Analytics Software, 85 Mio. $), face.com (Gesichtserkennungsservice, 100 Mio. $), das Werbetool Atlas (100 Mio. $), Instagram (1 B $), Oculus VR (2 B $), Whatsapp (19 B $) und über 70 weitere Unternehmen (Ramzeen 2019).

Die Kombination von Daten verschiedenster Quellen (auch physikalische Daten oder Messwerte aus den Fabriken) bilden unerwartet hohe Mehrwerte (House of Lords 2017a). Mehrere Universitäten, Start-ups, KMUs und NGOs kritisieren den schwierigen Zugang zu grossen, qualitativen Datensätzen und streichen die Hürden im Wettbewerb mit grossen Konkurrenten heraus (House of Lords 2017a; Simonite 2017). So entstehen sogenannte *data moats*, Gräben zwischen den Daten-Oligopolen und den restlichen Wettbewerbern.

Andere Expertinnen und Experten sind jedoch weniger besorgt und sehen keinen wesentlichen Unterschied zwischen konzentrierten Datensätzen und anderen Monopole begünstigenden Skaleneffekten (Casado & Lauten 2019). Daten seien zudem nicht rivalisierend, da das Individuum die gleichen Daten mehreren Anbietern zur Verfügung stellen könnte. Das übliche Verhalten von Monopolisten (beispielsweise Marktanteile mit einer Tiefpreisstrategie zu erobern) könne weiter (noch) nicht umfangreich beobachtet werden (House of Lords 2017b). Diverse Gerichte verfolgen die Aktivitäten der Hauptakteure jedoch aufmerksam (Dunleavy 2019).

3.3.3.5. Datenschutz und Privatsphäre

Privatsphäre beschreibt das Recht einer Person, bestimmte Informationen über sich selbst oder einer Gruppe, der sie angehört, nicht zu teilen und damit unbefugten Zugriff zu verhindern (APA 2019; Merriam Webster 2019). Datenschutzverletzungen waren schon lange vor dem Einsatz von KI heikel, aber die Problematik wurde durch die neuen Möglichkeiten verschärft. Massendatenerfassung und die Möglichkeit, grosse Datenbestände kostengünstig zu analysieren, haben den Druck auf die Plattformen erhöht, mehr Informationen über die Benutzer/-innen zu sammeln. Auch deren Wunsch nach personalisierten Empfehlungen drängt die

Plattformen partiell zu diesem Schritt. Ein autorisierter Zugriff auf die Daten ist jedoch nicht zwingend gegeben, wenn der Person der Umfang und die zukünftige Nutzung der Daten nicht bewusst ist (John 2018). Die durchschnittliche Person bräuchte 76 volle Arbeitstage im Jahr, um die AGBs ihrer genutzten Dienste tatsächlich zu lesen; eine anerkannte Zustimmung zu den Datennutzungen kann daher möglicherweise nicht vorausgesetzt werden (Board 2019; McDonald & Cranor 2008). Das Unternehmen PC Pitstop hat in einem Selbstversuch den Gewinn von 1000 Dollar in ihren AGBs versteckt, die man erhält, wenn man den Text bis zum Ende liest. Es dauerte mehrere Monate und Tausende Verkäufe, bis sich eine Person gemeldet hatte (PC Pitstop 2012).

Durch die wachsende Anzahl digitaler Fussabdrücke können heute mithilfe von Algorithmen ohne aktive Befragung detaillierte und stellenweise auch sehr intime Persönlichkeits- und Empfindlichkeitsprofile der Konsumentinnen und Konsumenten erstellt werden. Insbesondere die Möglichkeit, personenbezogene Daten aus verschiedenen Quellen zu kombinieren, setzt das Potenzial leistungsfähiger Algorithmen frei. Facebook selbst veröffentlichte eine Studie mit 86 220 Volontären, welche eine maschinelle Beurteilung von persönlichen Merkmalen präsentierte, die die Charaktereigenschaften besser einschätzt als die Freunde der betroffenen Personen. Die angewandte Technik ist treffsicherer als Mitarbeiter ab zehn Likes, als Freunde ab 70 Likes, als Familienmitglieder ab 150 Likes und als Ehepartner ab 300 Likes (Youyou et al. 2015). Aber auch ein einziges Profilbild kann genügen, um beispielsweise die sexuelle Orientierung algorithmisch mit bis zu 81 % Korrektheit einzuschätzen (Wang & Kosinski 2018). Neben diesen relativ statischen Persönlichkeitsprofilen experimentieren Unternehmen aktuell zunehmend daran, mithilfe von KI die aktuelle Emotionslage der Konsumenten einzufangen und das Angebot und Kundenerlebnis dynamisch darauf anzupassen. So hat Facebook beispielsweise ein KI-System vorgestellt, welches suizidgefährdete Nutzer/-innen anhand von Facebook Posts frühzeitig erkennen soll (Constine 2017; Facebook 2018). Damit erzeugt Facebook Gesundheitsdaten, ohne sich an die hohen Datenschutzstandards von Anbietern im Gesundheitswesen halten zu müssen (Goggin 2019).

Privatsphäre und Datenschutz sind breit diskutierte Themen. So stellte beispielsweise der US-Meinungsbericht der Oxforduniversität Datenschutzthemen als vordringlichstes Anliegen der Öffentlichkeit dar (Zhang & Dafoe 2019). Die Autoren registrierten auch eine vergleichsweise grosse Medienpräsenz zu diesem Thema.

Die Herausforderung des Datenschutzes lässt sich dabei mit der Notwendigkeit beschreiben, dass die Nutzer/-innen ihre Daten sorgfältig aufbewahren und kontrollieren können (House of Lords 2017a). Diese Kontrolle über die eigenen Daten kann darüber hinaus wichtig sein, da das Recht auf Gedanken- und Meinungsfreiheit, der Wunsch nach Vertraulichkeit, die Freiheit für politische Aktivitäten und das Recht auf eine zweite Chance kompromittiert wird. Es ist jedoch auch wichtig zu beachten, dass die Debatte um den Datenschutz keine eindimensionale ist und vielmehr auf einem Kompromiss zwischen Verlangen nach Datenschutz und dem etablierten Wunsch nach Personalisierung beruht, denn Personalisierung beruht nun einmal auf persönlichen Daten.

3.3.4. Fazit: Themenauswahl für die Expertenumfrage

Verschiedene Anwendungsfelder an der Kundenschnittstelle scheinen aufgrund des Drucks internationaler Anbieter auch in der Schweiz aktuell zu sein. In der Expertenumfrage war deshalb von Interesse, welche Einsatzgebiete im aktuellen Jahrzehnt im Vordergrund stehen werden und welche Herausforderungen sie implizieren. Es wurden sämtliche mittels Literatur erarbeiteten Herausforderungen jeweils für Konsumentinnen und Konsumenten sowie für Unternehmen nach ihrer Relevanz abgefragt. Hierzu wurden theoretische Beiträge, eigene qualitative Experteninterviews und empirische Untersuchungen in vergleichbaren Märkten beigezogen. Um die Beurteilung der Herausforderungen besser nachvollziehen zu können, wurden die empfundene zeitliche, räumliche und hypothetische Distanz von diversen Szenarien befragt. Dabei standen potenzielle Wünsche und Intentionen aus Konsumenten- und Unternehmerperspektive im Fokus.

3.4. KI und Medien[66]

Spätestens seit den Enthüllungen rund um Cambridge Analytica und Facebook gibt es eine gesteigerte öffentliche Aufmerksamkeit dafür, wie soziale Medien, Onlinekommunikation und gezielte Falschmeldungen (Fake News) politische Prozesse und die öffentliche Meinungsbildung beeinflussen. Im Kontext dieser Studie

[66] Dieser Abschnitt beruht auf Arbeiten von Tarik Abou-Chadi und Hauke Licht vom Institut für Politikwissenschaften der Universität Zürich.

interessiert dabei die Frage, inwieweit KI-Systeme diese Prozesse verstärken bzw. als Gegenmassnahmen eingesetzt werden können. Wie auch im Konsumbereich können KI-Systeme als Trendbeschleuniger wirken, indem sie es ermöglichen, eine grosse Menge an Informationen zu verarbeiten, die – oft unfreiwillig oder unbewusst – von den Nutzern selbst zur Verfügung gestellt werden. Auch wenn die meisten dafür eingesetzten Technologien derzeit noch vergleichsweise simpel sind (z.B. beruhen die meisten sogenannten Social Bots auf vergleichsweise einfachen Programmen), könnten künftig KI-gestützte Anwendungen vorab zwei Phänomene verstärkt beeinflussen: Falschnachrichten (Fake News) und das Problem der «Filterblasen» aufgrund zunehmend personalisierter Informationsflüsse.

Generell lässt sich dabei feststellen, dass in der öffentlichen Berichterstattung unterschiedliche Phänomene häufig nicht differenziert werden und dass nach einer zunächst grossen Euphorie darüber, wie soziale Medien demokratische Politik verändern können, nun eher die Schreckensszenarien überwiegen. Es soll im Folgenden daher zuerst ein Überblick darüber gegeben werden, wie sich Informationsverhalten und Nachrichtenangebot im digitalen Zeitalter verändert haben, welche Herausforderungen und Probleme damit verbunden sind und welche Rolle KI-Systeme für diese Entwicklung spielen. Zu diesem Zweck wird der veränderte mediale Kontext des digitalen Zeitalters dargestellt, um danach zu zeigen, welche neuen Akteure es gibt, die Einfluss auf die Meinungsbildung nehmen und wie sich diese letztlich auswirken können. Danach werden die Phänomene Personalisierung und Filter Bubble sowie Fake News genauer analysiert.

3.4.1. KI im Kontext eines sich verändernden Informationsverhaltens

Die letzten gut zehn Jahre sind von einem radikalen Umbruch der Medienlandschaft und des Informationsverhaltens geprägt. Während über Jahrzehnte hinweg Nachrichtenanbieter dominierten, die relativ zentralisiert über Zeitung, Radio und Fernsehen Nachrichten verbreiten konnten, so hat die digitale Revolution Kommunikationswege und damit Nachrichtenangebot und -konsum fundamental verändert. Die Verbreitung des Internets hat nicht nur dazu geführt, dass allgemein der Online-Nachrichtenkonsum zu einer der wichtigsten Informationsquellen geworden ist (Newman 2017), sondern dass sich spätestens mit dem Web 2.0 die Kanäle zur Verbreitung von Nachrichten ausdifferenzierten. Soziale Medien spielen dabei eine immer wichtigere Rolle für die Information und Meinungsbildung der Bürgerinnen. So zeigt der Reuters Digital News Report für 2017, dass in der Schweiz 83 % der Bevölkerung Nachrichten online konsumieren (Newman 2017).

45 % der Bevölkerung nutzen hierfür soziale Medien. Zeigt sich zwar noch eine gewisse Dominanz von TV (69 %) und Print (59 %) gegenüber sozialen Medien, lässt sich doch zweifelsohne festhalten, dass diese zumindest eine wichtige zusätzliche Nachrichtenquelle ausmachen.

Bekannte soziale Medien wie Facebook oder Twitter stellen dabei nicht die einzige neue Form der Nachrichtenvermittlung dar. Zunehmend werden auch Nachrichtenaggregatoren (Google News, Apple News) oder auch Messenger-Dienste wie Whatsapp zur Verbreitung von Nachrichten verwendet. Eine der wichtigsten Veränderungen besteht darin, dass Nachrichten zunehmend nicht mehr intermediiert sind, sondern vielmehr direkt zwischen Konsumentinnen und Konsumenten sowie Nachrichtenverfassern ausgetauscht werden (Bessi et al. 2015). Konnten Zeitungsredaktionen oder Nachrichtensendung noch gezielt filtern, bewerten und kontextualisieren, so ist dies in der pluralisierten digitalen Nachrichtenlandschaft viel weniger möglich bzw. diese Funktionen werden zunehmend von neuen Akteuren übernommen wie z.B. Google News, Apple New oder Facebook Newsfeed.

Letztlich lassen sich daher in diesem neuen Kontext zwei zentrale Felder der Veränderung voneinander unterscheiden. Zum einen hat sich das individuelle Informationsverhalten der Bürgerinnen und Bürger verändert. Anders als beim Konsum einer Zeitung oder einer Nachrichtensendung sind diese nun einer Flut von Informationen und Nachrichten ausgesetzt, deren Ursprung weniger klar ist und die häufig nicht unmittelbar einer Quelle zugeordnet werden. Die soziale Medienwelt hat nun die einzelne Person quasi selbst zu einer Art Redaktion gemacht. Sie teilen, liken, kommentieren Beiträge und tragen so selbst zur Verbreitung und Bewertung von Inhalten bei. Gleichzeitig führt dies auch dazu, dass Freunde und Bekannte eine zunehmend wichtige Quelle für Nachrichten und Informationen werden.

Diese Entwicklung kann nun zu bewussten und unbewussten Mechanismen führen, die in einer einseitigen Information oder gar vollkommenen Fehlinformation der Bürgerinnen und Bürger resultieren (Barberá 2018). Gezielte Fehlinformationen werden meistens unter dem Stichwort Fake News diskutiert, während für verstärkt einseitige Informationen Schlagwörter wie Filterblasen oder Echokammern (Echo Chambers) verwendet werden. Beide Phänomene werden im nächsten Teil ausführlicher diskutiert.

Die zunehmende Bedeutung von sozialen Medien hat allerdings nicht nur einen direkten Einfluss auf die öffentliche Meinungsbildung. Sie führt auch zu einer veränderten Situation für die traditionellen Medienanbieter und die Funktionsweise

des Journalismus generell. Eine zentrale Rolle spielt hierbei die Aufmerksamkeitsökonomie innerhalb der digitalen Kommunikation (Marwick u. Lewis 2017). Traditionelle Medien wie beispielsweise Zeitungsverlage sehen sich einem zunehmenden finanziellen Druck ausgesetzt und haben sich über die letzten Jahre von traditionellen journalistischen Prozessen wegbewegt, um neue wirtschaftliche Modelle zu entwickeln. Online-Nachrichtenanbieter achten verstärkt darauf, wie häufig Artikel gelesen und geteilt werden, weil dies einen entscheidenden Anteil an ihrem Geschäftsmodell ausmacht (Marwick u. Lewis 2017). Die Optimierung, die mit dieser Strategie einhergeht, führt nun dazu, dass Nachrichtenanbieter selbst durch Algorithmen zur einseitigen Berichterstattung beitragen können. Gleichzeitig führt ihre veränderte Rolle im Online-Nachrichtenkonsum dazu, dass zunehmend infrage gestellt wird, ob sie der Verbreitung von Fake News entschlossen und effektiv genug entgegentreten (Siegel 2018). Letzteres führt unter Umständen wiederum zu einem Glaubwürdigkeitsverlust.

Es lässt sich also festhalten, dass sich das veränderte Informationsverhalten der Bürgerinnen und Bürger sowohl direkt als auch indirekt auf die öffentliche Meinungsbildung auswirkt. Als Konsequenz sollen hier vor allem zwei Phänomene kritisch beleuchtet werden, die im Rahmen der demokratischen Willensbildung als besonders problematisch betrachtet werden: 1) zunehmend einseitige Information und das Entstehen von Filterblasen und Echokammern; 2) gezielte Fehlinformation aufgrund von Fake News. Diese werden im nächsten Abschnitt näher diskutiert. Zugleich wird darauf eingegangen, welche Rolle KI-Systeme hierbei spezifisch einnehmen.

3.4.2. Filterblasen und Echokammern

Zwei Begriffe haben im Zusammenhang mit der (algorithmischen) Personalisierung von Onlineinhalten grosse Aufmerksamkeit erfahren: Filterblasen und Echokammern. Diese sind verwandte, aber nicht identische Konzepte. Filterblasen bezeichnen eine informative und intellektuelle Isolierung, die dadurch auftritt, dass Menschen online selektiv Informationen basierend auf angenommenen Präferenzen zur Verfügung gestellt werden (Pariser 2012). Echokammern sind Online-Kommunikationsräume, in denen ähnliche oder gleiche Ideen von Teilnehmenden geteilt werden, ohne dass diese kritisch hinterfragt oder mit Gegenpositionen konfrontiert werden (Flaxman et al. 2016). Sowohl Filterblasen als auch Echokammern stellen Probleme für die öffentliche Meinungsbildung und die demokratische Deliberation dar (Stroud 2011).

Die Personalisierung von Nachrichteninhalten mittels KI trägt nun potenziell dazu bei, dass sich verstärkt Filterblasen und Echokammern bilden (Haim et al. 2017; Möller et al. 2018; Thurman u. Schifferes 2012). Menschen beziehen zunehmend Nachrichten, die nicht mehr von Redaktion, sondern zunehmend von Algorithmen kuratiert wurden (Newman 2017). Dies geschieht zum Beispiel über die Verwendung von Nachrichtenaggregatoren wie Google News, aber auch über die Selektion auf sozialen Medien (z.B. der personalisierte Facebook Newsfeed) oder auch bei Nachrichtenseiten selbst (z.B. über vorgeschlagene Artikel). Die Entscheidung, welche Inhalte den Nutzerinnen und Nutzern gezeigt werden, basiert dann auf gesammelten Informationen, die entweder freiwillig zur Verfügung gestellt oder über beispielsweise vorheriges Online-Verhalten gesammelt wurden (Thurman u. Schifferes 2012) – in analoger Weise, wie dies auch bei der Personalisierung im Bereich Konsum geschieht (siehe Abschnitt 3.3.1.1).

Wenn nun nicht mehr Redaktionen, sondern Algorithmen Nachrichten zusammenstellen, entstehen zwei Probleme. Im Gegensatz zu Redaktionen grösserer Qualitätszeitungen zum Beispiel müssen Algorithmen keineswegs einen Anspruch auf ausgewogene Nachrichten und Inhalte legen (Messing u. Westwood 2013). Hinzu kommt das Risiko, dass eben genau das Erlernen von Nutzerpräferenzen dazu führen kann, Inhalte immer einseitiger darzustellen (Haim et al. 2017). Schliesslich wird es mit grösserer Komplexität der Algorithmen und vor allem dann, wenn das Erlernen tatsächlich KI-basiert ist, zunehmend schwieriger festzustellen, aufgrund welcher Kriterien Nachrichteninhalte personalisiert werden.

Diese Dynamik wird auf sozialen Medien potenziell noch verstärkt. Nutzerinnen und Nutzer auf sozialen Medien entscheiden sich weder direkt für einen Nachrichtenanbieter, noch ist es prinzipiell ihr primäres Ziel, auf sozialen Medien Nachrichten zu konsumieren. Dadurch verlieren Quellen der Information stark an Bedeutung und werden durch Empfehlungen in Form von Teilen oder Liken ersetzt (Messing u. Westwood 2013). Filterblasen und Echokammern können nun sowohl dadurch entstehen, dass Personen vor allem die Artikel lesen, die ihr Onlinenetzwerk empfiehlt, oder dadurch, dass Algorithmen verstärkt die Artikel vorschlagen, die in einem Netzwerk beliebt sind. Vor allem Letzteres kann nun zu einem sich verstärkenden Kreislauf führen, bei dem Netzwerke zunehmend ähnliche Inhalte sehen, die ihre Meinungen verstärken, während sie zeitgleich von gegenläufigen Informationen abgeschnitten sind.

Während dieser Mechanismus der Entstehung von Filterblasen und Echokammern viel mediale Aufmerksamkeit erhalten hat und als fundamentales Problem

für die demokratische Willensbildung identifiziert wurde, fällt das wissenschaftliche Urteil dazu durchaus differenzierter aus (Bakshy et al. 2015; Guess 2018). Die meisten Studien stellen fest, dass Onlinenetzwerke keineswegs per se homogener sind als Offlinenetzwerke (Boxell et al. 2017; Gentzkow u. Shapiro 2011). Gleiches gilt für die Nachrichten, die online konsumiert werden. Sie sind in der Regel und für eine Vielzahl von Menschen eher moderat und divers. Auch Nachrichtenaggregatoren und andere Algorithmen scheinen nicht per se zu einer stärkeren Segregation und Polarisierung von Nachrichteninhalten beizutragen (Haim et al. 2017; Möller et al. 2018).

Die wissenschaftliche Skepsis gegenüber der Prävalenz von Filterblasen und Echokammern bedeutet allerdings nicht, dass die beschriebenen Prozesse kein potenzielles Problem für die demokratische Willensbildung darstellen. Zunächst ist festzuhalten, dass die Personalisierung von Inhalten und algorithmisierte Zurverfügungstellung von Medieninhalten erst im Anfangsstadium stehen. Es ist anzunehmen, dass bei bisherigen Geschäftsmodellen die Tendenz von Anbietern zur Personalisierung steigen wird. Gleichzeitig wird bei zunehmender Komplexität der KI-Systeme das Blackbox-Problem zunehmen. Die Bestandsaufnahme, dass Filterblasen und Echokammern den Online-Meinungsraum nicht dominieren, sollte daher keineswegs so gedeutet werden, dass die beschriebenen Probleme per se nicht bestehen, sondern dass sie *noch* nicht Verbreitung gefunden haben (Tucker et al. 2018).

3.4.3. Fake News

Nebst dem Problem der einseitigen Information und potenzieller Polarisierung als Folge der veränderten Medienlandschaft gibt es ein zweites Phänomen, welches in diesem Kontext viel Aufmerksamkeit erfahren hat: die gezielte Falschinformation von Nutzerinnen und Nutzer, was häufig unter dem Label «Fake News» beschrieben wird. Bevor nun die Prävalenz und Konsequenzen von Fake News diskutiert werden, lohnt es sich, einen Überblick darüber zu geben, in welchen Formen gezielte Desinformation über vor allem soziale Medien verbreitet wird. Drei wichtige Formen sind hier Trolls, Social Bots und Fake-News-Webseiten.

Trolling ist eine Online-Kommunikationsstrategie, die es darauf anlegt, durch provokative und emotionale Kommentare weitere emotionale Reaktionen zu erzeugen (Siegel 2018). Diese Kommentare, die häufig vage oder faktisch inkorrekte Informationen enthalten, können dafür genutzt werden, um bestimmte Gruppen

über ihre gewöhnliche Reichweite hinaus bekannt zu machen. So lässt sich beispielsweise die gesteigerte Bekanntheit von bestimmten Alt-Right-Gruppen oder Persönlichkeiten auf erfolgreiches Trolling zurückführen (Marwick & Lewis 2017). Bei Trolls ist es auch wichtig, unabhängige von bezahlten Trolls zu unterscheiden (Siegel 2018). Während unabhängige Trolls häufig ideologisch motiviert sind, so handeln bezahlte Trolls im Interesse von Auftraggebern. Bezahlte Trolls oder sogar Troll Factories werden hierbei gezielt eingesetzt, um Nachrichteninhalte im Sinne von beispielsweise Regierungen zu beeinflussen (Sanovich & Stukal 2018).

Social Bots sind Computerprogramme, die geschrieben wurden, um gezielt Nachrichten auf sozialen Medien zu verbreiten (Siegel 2018). Man kann sie als eine Form von Chatbot (siehe Abschnitt 2.2.3.6) auffassen; sie ermöglichen aber meist nur einfache Interaktionen oder überhaupt keine Reaktivität.

Schliesslich gibt es **Fake-News-Webseiten**, die einzig dem Zweck dienen, Falschnachrichten zu verbreiten. Sie treten häufig in der Verbindung mit Trolls oder Social Bots auf und können dazu beitragen, dass Fehlinformationen glaubwürdiger empfunden werden. Für manche Betreiber von Fake-News-Webseiten stehen auch vor allem monetäre Anreize im Vordergrund.

Es gibt nun eine Reihe von Wegen, wie Trolls und Social Bots die öffentliche Wahrnehmung und Meinungsbildung über soziale Netzwerke beeinflussen können. Zunächst können diese dazu beitragen, einen Eindruck zu erwecken, dass bestimmte Personen oder Positionen beliebter sind, als dies eigentlich der Fall ist. So können beispielsweise Politiker Bots einkaufen, um die Zahl ihrer Followers zu erhöhen und somit beliebter oder interessanter zu erscheinen (Siegel 2018). Gleiches gilt für ausgedrückte Meinungen und Positionen, beispielsweise in Form von Facebook-Posts oder Tweets. Vor allem im Wahlkampf kann es für Politiker interessant sein, den Eindruck zu erwecken, dass ihre Botschaften und Positionen grössere Unterstützung in der Bevölkerung erfahren. Für die Bürgerinnen und Bürger bedeutet dies dann, dass sie einen stark verfälschten Eindruck der öffentlichen Meinung bekommen.

In ähnlicher Weise können Social Bots eingesetzt werden, um Suchalgorithmen oder Trends auf sozialen Medien zu beeinflussen. Wie schon zuvor beschrieben, funktionieren Algorithmen häufig so, dass sie Nutzern beliebte Nachrichten eher anbieten als andere Nachrichten. Dies gilt sowohl für Suchmaschinen, Algorithmen von sozialen Medien und Nachrichtenaggregatoren als auch direkt für Artikelempfehlungen auf Nachrichtenseiten. Um eine höhere Verbreitung von Nachrichten zu erzielen, können Social Bots eingesetzt werden, um diese gezielt

stärker über Suchalgorithmen und Empfehlungen zu verbreiten (Sanovich & Stukal 2018). Auf diesem Weg können Meinungen, Nachrichten oder sogar Verschwörungstheorien eine Reichweite erhalten, die sonst undenkbar gewesen wäre.

Schliesslich können Social Bots und Trolls auch direkt falsche und frei erfundene Nachrichten verbreiten. Dies geschieht häufig in Verbindung mit den schon zuvor erwähnten Webseiten, die sich auf falsche Nachrichten spezialisieren, und entspricht am ehesten der allgemein verbreiteten Idee von Fake News. Eine Hauptstrategie ist es, darauf abzuzielen, dass die falschen Nachrichten von seriösen Nachrichtenquellen aufgegriffen werden und sich somit über legitime Kanäle verbreiten (Siegel 2018). Für die Bürgerinnen und Bürger ist es damit zunehmend schwierig, die ursprüngliche Quelle der Nachricht zu identifizieren. Eine beliebte Taktik zur Verbreitung von Falschinformationen in den USA ist es daher, damit zunächst die Aufmerksamkeit von lokalen Nachrichtensendern zu gewinnen, da diese über weniger Möglichkeiten verfügen, Fakten zu überprüfen und Geschichten zu recherchieren (Marwick & Lewis 2017). Sind die Falschnachrichten erst einmal von den lokalen Nachrichten aufgenommen worden, ist es umso wahrscheinlicher, dass auch nationale Nachrichtenanbieter darüber berichten werden.

Die direkte Verbreitung von Falschnachrichten wird generell als eine der grössten Bedrohungen für die öffentliche Meinungsbildung angesehen (Tucker et al. 2018). Technischer Fortschritt, vor allem auch KI-Technologien, trägt dazu bei, dass Falschnachrichten immer ausgereifter werden. So beschränken sich die Fälschungen nicht mehr allein auf Textmaterial, sondern beinhalten zunehmend auch gefälschte Fotos und Videos. Unter dem Schlagwort Deepfakes werden dabei Foto- und Videofälschungen verstanden, die basierend auf KI-Lernprozessen immer besser darin werden, menschliche Muster zu imitieren. Damit wird es zunehmend leichter, beispielsweise Gesichter von Menschen in andere Videos zu implantieren oder auch Stimmen nachzuahmen.[67]

Social Bots und Fake News haben mittlerweile grosse Verbreitung gefunden. Vor allem im Zuge der US-Präsidentschaftswahlen von 2016 haben Forschende ein hohes Aufkommen von Social Bots dokumentiert. Bessi und Ferrara (2016) zeigen beispielsweise, dass während der US-Wahlkampagne Social Bots ca. 3,8 Millionen Tweets produziert haben (im Vergleich zu ca. 17 Millionen, die von Menschen abgeschickt wurden). Social Bots nehmen auch bei der Verbreitung von Falschnachrichten eine prominente Rolle ein. Nach Shao et al. (2018) wurden etwa 34 %

[67] Siehe z.B.: https://lyrebird.ai/vocal-avatar.

aller Artikel aus Quellen mit niedriger Glaubwürdigkeit von Social Bots verbreitet. Das Aufkommen von Social Bots und ihre Rolle für die Verbreitung von Falschnachrichten wurde auch für zahlreiche andere Länder wie zum Beispiel Russland, Frankreich oder Deutschland dokumentiert (Sanovich & Stukal 2018).

Generell lässt sich festhalten, dass nicht nur eine grosse Menge an Falschnachrichten produziert und von Social Bots geteilt wird, sie finden auch eine weite Verbreitung. Zwei prominente Analysen für das Nachrichtenmagazin Buzzfeed zeigen beispielsweise, dass sich Falschnachrichten in der Regel mehr und besser verbreiten als echte Nachrichten (Sanovich & Stukal 2018). Für Deutschland zeigte sich, dass sieben der zehn am weitesten verbreiteten Nachrichten über Angela Merkel als Fake News einzustufen sind.[68]

Die zuvor diskutierte Anwendung von Optimierungsalgorithmen kann dabei eine entscheidende Rolle für die Verbreitung von Fake News spielen. So finden diese vor allem über Facebook Verbreitung. Gerade die von der Aufmerksamkeitsökonomie getriebene Anreizstruktur beflügelt demnach Taktiken, die zur Verbreitung von Falschnachrichten verwendet werden. In Kombination mit zunehmender Personalisierung basierend auf KI-Lernprozessen sind potenziell immer mehr Leute Fake News ausgesetzt. Dies erzeugt den Eindruck, dass dies vielgelesene und häufig geteilte Nachrichten sind. Dies verleiht diesen falschen Berichten zunehmend an Glaubwürdigkeit.

Schliesslich ist es aber nicht nur die Verbreitung von Falschnachrichten über soziale Medien, die dieses Phänomen zum Problem für demokratische Meinungsbildung macht, sondern auch das gleichzeitige Misstrauen, welches traditionellen Medien entgegengebracht wird. Wie der Reuters Digital News Report 2017 (siehe Abbildung 9) zeigt, sind in fast allen beteiligten Ländern weniger als die Hälfte der Befragten der Meinung, dass Medien gut darin sind, Fakt von Fiktion zu unterscheiden (Newman 2017). Für soziale Medien ist die Zahl noch viel geringer.

Auch in der Schweiz geben nur 46 % der Befragten an, Nachrichten generell zu vertrauen (Newman 2017). In einem Kontext, in dem immer weniger Menschen den Medien vertrauen, wird es auch immer schwieriger, gegen Fake News vorzugehen. Mögliche Lösungsansätze müssen also notwendigerweise über technische Massnahmen hinausgehen und auch dazu beitragen, dass bestimmte mediale Akteure Vertrauen zurückgewinnen können.

[68] Siehe: https://www.buzzfeednews.com/article/karstenschmehl/top-merkel-news.

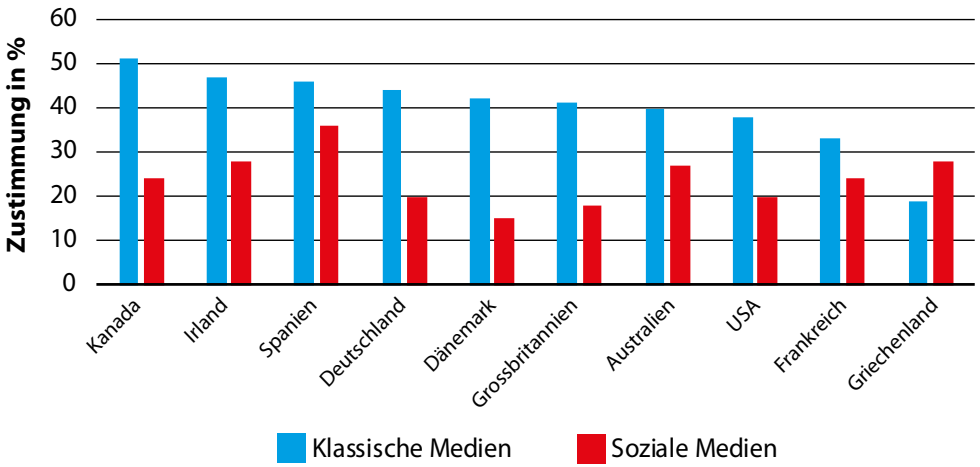

Abbildung 9: Anteil der Personen, die zustimmen, dass konventionelle Medien bzw. Social Media Fake News zuverlässig erkennen können (adaptiert aus Newman 2017).

3.4.4. Fazit: Themenauswahl für die Expertenumfrage

Basierend auf diesem Forschungsstand wurden zwei Themenblöcke in die Expertenumfrage aufgenommen: Personalisierung von Nachrichten und Fake News. Für beide soll es zunächst um eine Einschätzung der Prävalenz der Probleme gehen. Für Fragen der Personalisierung von Inhalten steht folglich im Mittelpunkt, inwieweit diese mit Veränderungen des Medienbetriebs zusammenhängen und wie sehr und auf welche Weise Algorithmen den Nachrichtenkonsum von Menschen beeinflussen. Zugleich wird der Frage nachgegangen, inwieweit Onlinenetzwerke in ihrer Diversität von Offlinenetzwerken abweichen.

Für Fake News wird geprüft, inwieweit technologische Veränderungen die Möglichkeiten verändern, falsche Nachrichten zu produzieren und als glaubwürdig erscheinen zu lassen. Gleichzeitig soll eingeschätzt werden, welche Möglichkeiten KI auch für Gegenmassnahmen und vor allem das Fact Checking bietet. Generell spielt für Fake News, aber auch für den gesamten Medienbereich, die Frage eine zentrale Rolle, in welchem Verhältnis eine Regulierung zu Prinzipien der demokratischen Meinungsbildung steht und daher als problematisch zu betrachten ist.

3.5. KI in Verwaltung und Gerichtsbarkeit[69]

Beim Thema KI denkt man nicht unbedingt als Erstes an deren Nutzung durch den Staat. Dennoch zeigen jüngere Entwicklungen, dass die Bedeutung von KI-Technologien in Zukunft auch für staatliche Handlungen ansteigen könnte. Dies gilt in erster Linie für die Tätigkeiten der Verwaltung. So wird etwa der Bundesrat in einem Postulat[70] aufgefordert zu prüfen, wie die Effizienz in der Bundesverwaltung mithilfe von Prozessautomatisierung und KI optimiert werden kann. Entsprechende Empfehlungen finden sich auch im Bericht der interdepartementalen Arbeitsgruppe «Künstliche Intelligenz», die teilweise auch in die Empfehlungen dieser Studie eingeflossen sind (siehe dazu Abschnitt 2.6).

Nebst der staatlichen Verwaltung im engeren Sinn könnte KI auch in der Justiz eine Rolle spielen. Die Europäische Kommission des Europarates für die Wirksamkeit der Justiz hat aus diesem Grund im Dezember 2018 eine Charta mit Ethikgrundsätzen für die Anwendung von KI in der Justiz verabschiedet (CEPEJ 2018). Sie soll den verantwortlichen Gremien in den Mitgliedstaaten, zu denen auch die Schweiz zählt, als Leitfaden im Umgang mit KI in Justizverfahren dienen. Hingegen ist ein Einsatz von KI im Bereich der Gesetzgebung derzeit nicht absehbar. Zwar wird der Gesetzgeber früher oder später die Rahmenbedingungen eines verantwortungsvollen Einsatzes von KI festlegen müssen (siehe dazu Abschnitt 2.5). Aber eine Einbindung von KI im Rahmen der Tätigkeit der Legislative selbst ist derzeit nicht zu erkennen (Braun Binder 2018).

In dieser Studie liegt der Blickwinkel entsprechend auf der Nutzung von KI in Verwaltung und Gerichtsbarkeit. Zu diesem Zweck werden zuerst grundsätzliche Voraussetzungen des staatlichen KI-Einsatzes besprochen. Danach werden Beispiele des aktuellen oder in naher Zukunft erwartbaren Einsatzes von KI-Systemen in Verwaltung und Rechtsprechung aufgeführt. Schliesslich werden die wichtigsten in der Literatur besprochenen Probleme des staatlichen KI-Einsatzes vorgestellt.

---

[69] Dieser Abschnitt beruht auf Arbeiten von Nadja Braun Binder, bis Ende Juli 2019 Assistenzprofessorin für öffentliches Recht unter besonderer Berücksichtigung europäischer Demokratiefragen an der Universität Zürich, seit 01.08.2019 Professorin für Öffentliches Recht an der Universität Basel.

[70] Postulat FDP-Liberale Fraktion (18.3783) vom 19.09.2018: Effizienzsteigerung beim Bund durch intelligente Prozessautomatisierung in der Verwaltung.

3.5.1. Rechtliche Voraussetzungen für KI-Nutzung durch den Staat

Im Unterschied zu privaten Akteuren, die KI einsetzen, gelten für den Staat besondere Voraussetzungen und Rahmenbedingungen. So handelt der Staat in der Regel sogenannt einseitig, d.h. aus einer übergeordneten Stellung heraus, und ohne das Einverständnis der Adressaten seines Handelns. Man spricht auch von hoheitlichem Handeln. Unter gewissen Voraussetzungen kommt dem Staat eine Anordnungs- und Zwangsbefugnis gegenüber den Privaten zu. Im Gegenzug unterliegt der Staat bzw. jeder, der staatliche Aufgaben wahrnimmt, der Bindung an die Grundrechte. Diese schützen Private gegen Übergriffe der Staatsmacht. Setzt der Staat im Rahmen seiner Aufgabenwahrnehmung KI ein, sind dabei die Garantien der Grundrechte zu bedenken.

Relevant ist vor diesem Hintergrund insbesondere, dass der Staat KI regelmässig zur Automation von Entscheiden einsetzt. Teilweise wird deshalb in der internationalen Debatte auch empfohlen, anstelle von KI den Begriff «*automated decision making*» (ADM) zu verwenden (Spielkamp 2019). Die Entscheidungsautomation kann dabei unterschiedlich stark durch KI unterstützt werden. Eine grobe Unterscheidung kann danach getroffen werden, ob ein Mensch (z.B. ein Sachbearbeiter in der Verwaltung oder eine Richterin am Gericht) entscheidet und sich dabei auf KI-basierte Informationen stützt oder ob eine menschliche Entscheidung in einem konkreten Anwendungsfall komplett entfällt, weil die KI alle erforderlichen Schritte übernimmt. Im ersten Fall kommt der KI die Funktion einer Entscheidungsunterstützung zu, im zweiten Fall trifft die KI die Entscheidung anstelle eines Menschen.

Diese Zweiteilung ist allerdings nur sehr grob und zeigt nicht das volle Spektrum der möglichen Interaktion zwischen Mensch und KI bei der Automation des Entscheidens. Deshalb werden in der Literatur feinere Abstufungen vorgenommen, wie z.B. ein fünfstufiges Modell, das zwischen assistierten Entscheiden (Stufe 1), teilweisen Entscheiden (Stufe 2), geprüften Entscheiden (Stufe 3), delegierten Entscheiden (Stufe 4) und autonomen Entscheiden (Stufe 5) unterscheidet (Bitkom 2017). Je höher die Stufe, desto stärker verlagert sich der Entscheid vom Menschen auf die KI. Je nachdem, wie stark die KI in die Entscheidungsfindung involviert ist, sind damit unterschiedliche Chancen und Risiken verbunden. Im Kontext hoheitlichen Handelns ist dabei zu bedenken, dass gewisse Entscheidungen, wie z.B. eine Verfügung oder ein Gerichtsurteil, rechtsverbindliche Wirkungen entfalten. Die rechtsstaatlichen Verfahrensgarantien (z.B. Anspruch auf rechtliches Gehör, worunter auch der Anspruch auf Entscheidbegründung fällt) schaffen dabei für den staatlichen KI-Einsatz zusätzliche Herausforderungen.

Bevor auf einzelne Anwendungsbereiche eingegangen wird, ist eine letzte Differenzierung zu erwähnen: Verwaltungshandeln kann danach unterschieden werden, ob damit ein *Rechtserfolg* oder ein *Taterfolg* erzielt werden soll. Ist das Handeln auf die unmittelbare Gestaltung der Rechtslage ausgerichtet, spricht man von einem Rechtsakt. Beispiele sind etwa der vorsorgliche Entzug eines Führerausweises mittels Verfügung oder die Steuerveranlagung. Soll mit dem Handeln hingegen unmittelbar die Faktenlage gestaltet werden, handelt es sich um einen Realakt. Dazu zählen z.B. Auskünfte, Empfehlungen und Berichte. Der Einsatz von KI ist grundsätzlich sowohl beim Erlass von Rechtsakten als auch bei der Vornahme von Realakten vorstellbar. Ein Chatbot beispielsweise, der von der Verwaltung eingesetzt wird, um Bürgeranfragen zu beantworten, fällt in den Bereich der Realakte. Ein KI-basiertes Risikomanagementsystem, das im Rahmen der Steuerveranlagung Steuererklärungen auf Anzeichen von Betrug kontrolliert, würde im Vorfeld des Erlasses eines Rechtsaktes eingesetzt. Je nachdem, ob Verwaltungshandeln als Rechtsakt oder Realakt eingestuft wird, bestehen unterschiedliche Anforderungen an bzw. Auswirkungen auf die Zuständigkeit, das Verfahren, die Steuerungswirkung, die Rechtsbeständigkeit und den Rechtsschutz.

3.5.2. Bereiche für die staatliche Anwendung von KI

Anders als in der Privatwirtschaft eröffnet KI für den Staat nicht unmittelbar neue Betätigungsfelder. Die Aufgaben des Staates sind vorgegeben; KI kann demnach in erster Linie dort genutzt werden, wo sie den Staat in der Erfüllung bisheriger Aufgaben unterstützt. In der Regel wird der Staat KI einsetzen, um seine Aufgaben effizienter, kostengünstiger und/oder in materieller Hinsicht besser zu erfüllen. Im Folgenden werden – ohne Anspruch auf Vollständigkeit – einzelne Bereiche skizziert, in denen KI von staatlicher Seite getestet oder bereits eingesetzt wird.

3.5.2.1. Vorausschauende Polizeiarbeit (*predictive policing*)

Algorithmenbasierte vorausschauende Polizeiarbeit kann zweierlei bedeuten: einerseits die Erstellung einer Prognose bezüglich der Gefährlichkeit oder Gefährdung einer Person und andererseits die raumbezogene Prognose in Bezug auf die Wahrscheinlichkeit von Verbrechen. Im ersten Fall wird danach gefragt, *wer* gefährlich werden bzw. gefährdet sein könnte, im zweiten Fall danach, *wo* eine bestimmte Gefahr auftreten könnte (Leese 2018).

Die KI-basierte *personenbezogene vorausschauende Polizeiarbeit* wird vor allem in den USA eingesetzt. Zu erwähnen ist etwa die «Strategic Subject List» der Stadt Chicago.[71] Mithilfe von Daten über soziale Kontakte von Personen soll das Risiko errechnet werden, dass eine Person beispielsweise in Bandenkriminalität involviert sein könnte (Saunders et al. 2016). Computergestützte personenbezogene vorausschauende Polizeiarbeit ist auch in der Schweiz bekannt (Leese 2018). Aufmerksamkeit hat 2018 die Software DyRiAS (Dynamisches Risiko-Analyse-System) erlangt.[72] Weitere in der Schweiz eingesetzte Instrumente sind ODARA (*Ontario Domestic Assault Risk Assessment*), Patriarch (*Assessment of Risk for Honour Based Violence*) und RA-PROF (*Radicalisation Profiling*).[73] Inwieweit im Rahmen der in der Schweiz eingesetzten Systeme KI genutzt wird, ist unklar.

Die Idee hinter der *raumbezogenen vorausschauenden Polizeiarbeit* ist, dass in einem ersten Schritt auf Basis von mathematischen Prognosemethoden Zeitpunkt und Örtlichkeit von Kriminalitätsrisiken vorhergesagt werden und sodann gestützt auf diese Vorhersagen Polizeiressourcen in das betreffende Gebiet geschickt werden, um die Durchführung krimineller Handlungen zu verhindern (Mohler et al. 2011). Die im ersten Schritt zum Einsatz kommenden Algorithmen variieren erheblich. Es gibt Systeme, die auf Basis von maschinellen Lernverfahren operieren, wie zum Beispiel das System Predpol.[74] Predpol wird von der Polizei in US-amerikanischen Grossstädten (z.B. Los Angeles, Chicago, Seattle, Boston) verwendet. 2013 wurde die Software erstmals auch in Europa eingesetzt; dies in der Grafschaft Kent (England). Zwar hat Kent Ende 2018 kommuniziert, Predpol nicht mehr zu nutzen; allerdings nicht aufgrund negativer Erfahrungen, sondern weil die Polizei ein eigenes System entwickeln will.[75]

---

[71] Siehe: https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np.

[72] Siehe: https://www.srf.ch/news/schweiz/predictive-policing-polizei-software-verdaechtigt-zwei-von-drei-personen-falsch.

[73] Siehe dazu Bericht des Bundesrates vom 11.10.2017 in Erfüllung des Postulates Feri 13.3441 vom 13.06.2013: Bedrohungsmanagement, insbesondere bei häuslicher Gewalt, S. 6. Dieser ist abrufbar unter https://www.bj.admin.ch/dam/data/bj/sicherheit/gesetzgebung/gewaltschutz/ber-br-d.pdf.

[74] Siehe: https://www.predpol.com. Eine Beschreibung des Systems findet sich bei Mohler et al. 2015.

[75] Siehe: https://www.telegraph.co.uk/technology/2018/11/27/kent-police-stop-using-crime-predicting-software.

Eine weiter gehende *Predictive-policing*-Lösung ist HunchLab[76], das nicht nur Gefahrenzeitpunkte und -orte erkennen und der Polizei ermöglichen will, entsprechende Polizeipatrouillen anzuordnen. HunchLab, das ebenfalls auf maschinellen Lernverfahren basiert, will darüber hinaus die Polizeipatrouillen mit den vom betreffenden Gemeinwesen vorgegebenen Prioritäten abgleichen, Ressourcen effizient einteilen und die geeignetste Einsatztaktik ermitteln.

Das in der Schweiz in einzelnen Kantonen eingesetzte raumbezogene *Predictive policing*-System «Precobs» setzt dagegen nicht auf maschinelle Lernverfahren. Es generiert Prognosen vielmehr auf Basis von vorgängig eingegebenen «Wenn-dann-Entscheidungen» (Gerstner 2017) – es handelt sich also um eine regelbasierte Technologie. Das System wird in den Kantonen Aargau und Basel-Landschaft sowie in der Stadt Zürich regulär eingesetzt und in den Kantonen Zug und Zürich getestet.[77] Neben dem Verzicht auf KI unterliegt Precobs in der Schweiz auch einer Beschränkung auf die Prognose von Wohnungseinbruch-Diebstahlsdelikten.

3.5.2.2. Rückfallgefahr bei Straftätern

Für die Beurteilung der Rückfallgefahr von Straftätern finden teilweise KI-gestützte Systeme Anwendung. So wird im US-amerikanischen Justizsystem etwa die Beurteilungssoftware COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) eingesetzt.[78] Auf der Basis von 137 Merkmalen wird die Rückfallwahrscheinlichkeit eines Straftäters berechnet. Aufmerksamkeit wurde dem System zuteil, da dieses das Rückfallrisiko von afroamerikanischen Straftätern durchwegs höher als dasjenige von weissen Straftätern eingestuft hatte (Angwin et al. 2016; siehe dazu auch die Ausführungen in Abschnitt 2.4.2). Ausserdem war COMPAS Gegenstand eines gerichtlichen Verfahrens im Bundesstaat Wisconsin.[79] Im britischen Durham wird eine KI-basierte Risikoanalyse eingesetzt, um die Gefahr erneuter strafbarer Handlungen von Straftätern in einem Zeitraum von

[76] Siehe: https://www.hunchlab.com.

[77] Siehe: https://www.ifmpt.de.

[78] Siehe: http://www.equivant.com/solutions/case-management-for-supervision. COMPAS wurde von der Firma Northpointe entwickelt. Im Zuge von Kritik an COMPAS wurde die Firma umbenannt in Equivant. Vgl. für einen Überblick über weitere Risikobeurteilungssoftware in der US-amerikanischen Justiz Kehl et al. 2017.

[79] Vgl. Wisconsin Supreme Court, State v. Loomis, Urteil vom 13.07.2016, Az. 2015AP157-CR.

zwei Jahren einzuschätzen. Die dort verwendete Software (HART – Harm Assessment Risk Tool) musste angepasst werden, um zu verhindern, dass Personen aus weniger privilegierten Wohngegenden nicht systematisch schlechter als Personen aus wohlhabenderen Wohngegenden beurteilt werden (Burgess 2018).

In der Schweiz wird im Strafvollzug anhand des Programms ROS (Risikoorientierter Sanktionenvollzug) die Möglichkeit von Vollzugslockerungen geprüft. Ziel ist es, das Rückfallrisiko während und nach dem Vollzug zu senken (Vegh 2019). ROS wurde zwischen 2010 und 2013 in einem vom Bundesamt für Justiz unterstützten Modellversuch in den Kantonen Luzern, St. Gallen, Thurgau und Zürich getestet. Inzwischen wurde ROS in der ganzen Deutschschweiz umgesetzt.[80] Im Rahmen von ROS werden Daten zu einer Person aus dem Strafregisterauszug in das Fall-Screening-Tool (FaST) überführt. Es resultiert eine Einteilung in drei Risikokategorien, die als Basis für den Entscheid dient, ob eine weitere psychologisch-forensische Begutachtung notwendig ist. Gemäss Nachfrage basiert ROS nicht auf neueren KI-Technologien (Deep Learning).

3.5.2.3. Steuerverfahren

In Australien wird KI eingesetzt, um automatisch Geldforderungen des Staates gegenüber dem Bürger, etwa im Bereich der Einkommensteuern, zu erheben. Das Vorhaben stiess allerdings auf heftige Kritik, da das System teilweise zu nicht korrekten Schlüssen kam und es den betroffenen Bürgerinnen und Bürgern überliess, sich gegen fehlerhafte Bescheide zur Wehr zu setzen (Knaus 2017; Djeffal 2018). Dies stellte vor allem Menschen aus sozial schwächeren Schichten und benachteiligte Bevölkerungsgruppen, die sich nicht gegen den Bescheid wehren konnten, vor grosse Probleme (Commonwealth Ombudsman 2017).

Ein anderes Beispiel von KI im Besteuerungsverfahren stammt aus Deutschland. Ab 2017 sind dort gesetzliche Grundlagen für die vollautomatisierte Durchführung des Besteuerungsverfahrens in Kraft.[81] Algorithmenbasierte Risikomanagementsysteme (RMS) sollen dabei fehlerhafte Steuererklärungen erkennen und Steuerumgehungen verhindern. Angesichts der grossen Datenmengen, die ein RMS

[80] Siehe: https://www.rosnet.ch. Vgl. auch Treuthardt et al. 2018.

[81] Gesetz vom 18.7.2016 (BGBl I S. 1679). Vgl. dazu Braun Binder 2016a.

zu verarbeiten hat, liegt die Idee nahe, KI einzusetzen (Braun Binder 2019a). So hat der Bundesbeauftragte für Wirtschaftlichkeit in der Verwaltung bereits 2006 den Einsatz von «selbstlernenden» RMS empfohlen.[82] Gemäss § 88 Abs. 5 Satz 4 der Abgabenordnung dürfen Einzelheiten der RMS allerdings nicht veröffentlicht werden, soweit dadurch die Gleichmässigkeit und Gesetzmässigkeit der Besteuerung beeinträchtigt werden könnten. Deshalb bleibt unklar, ob und gegebenenfalls welche maschinellen Lernverfahren im Rahmen von RMS eingesetzt werden.[83]

In der Schweiz werden KI-basierte Besteuerungsverfahren, soweit ersichtlich, derzeit nicht eingesetzt. Zwar schreitet auch hierzulande vor allem in den Kantonen die Digitalisierung voran, insbesondere mit Blick auf die webbasierte elektronische Steuererklärung (Ferber 2018). Auf Bundesebene könnte im Kontext des Programms DaziT, in dessen Rahmen die Digitalisierung der Zollverwaltung vorangetrieben wird, mittelfristig eine automatisierte Erhebung verschiedener Verbrauchsteuern, vielleicht auch unter Abstützung auf KI, eingeführt werden.[84] Die Rechtsgrundlagen, die den Einsatz von automatisierten Verfügungen im Bereich der Zollabgaben[85], der Tabaksteuern[86], der Biersteuern[87], der Mineralölsteuern[88] sowie der Schwerverkehrsabgaben[89] ermöglichen sollen, sind im Anhang zum Entwurf des neuen Datenschutzgesetzes vorgesehen.[90]

---

[82] Vgl. Präsident des Bundesrechnungshofes in seiner Funktion als Bundesbeauftragter für Wirtschaftlichkeit in der Verwaltung (Probleme beim Vollzug der Steuergesetze, S. 165), abrufbar unter https://goo.gl/92gu6r. Vgl. zur Bedeutung lernender RMS im Steuervollzug auch Schmidt 2008.

[83] Vgl. aber den Hinweis in: Landtag Baden-Württemberg, Mitteilung der Landesregierung vom 14.12.2011, Drs. 15/1047, S. 12 und 19, wonach in einzelnen Bundesländern RMS eingesetzt werden, die auf künstlichen neuronalen Netzen basieren und die Umsatzsteuervoranmeldungen auf Anzeichen für die Beteiligung an einem sogenannten Karussellbetrug analysieren.

[84] Vgl. Botschaft des Bundesrates vom 15.02.2017 zur Finanzierung der Modernisierung und Digitalisierung der Eidgenössischen Zollverwaltung, BBl 2017, 1719, sowie die Webseite der Eidgenössischen Zollverwaltung: https://www.ezv.admin.ch/ezv/de/home/themen/projekte/dazit.html.

[85] Entwurf des Art. 38 Abs. 2 Zollgesetz, BBl 2017 7260.

[86] Entwurf des Art. 18 Abs. 4 Tabaksteuergesetz, BBl 2017 7262.

[87] Entwurf des Art. 17 Abs. 3 zweiter Satz Biersteuergesetz, BBl 2017 7263.

[88] Entwurf des Art. 21 Abs. 2^{bis} Mineralölsteuergesetz, BBl 2017 7263.

[89] Entwurf des Art. 11 Abs. 4 Schwerverkehrsabgabegesetz, BBl 2017 7263.

[90] Vgl. Entwurf des Bundesgesetzes über die Totalrevision des Bundesgesetzes über den Datenschutz und die Änderung weiterer Erlasse zum Datenschutz, BBl 2017 7193 ff.; sowie die entsprechende Botschaft des Bundesrates vom 15.09.2017 (BBl 2017 6941 ff.); vgl. Rechsteiner 2018.

3.5.2.4. Weitere Einsatzbereiche

Neben den bereits erwähnten Beispielen sind KI-Anwendungen in den unterschiedlichsten Gebieten der öffentlichen Verwaltung denkbar. Dazu zählen etwa der Einsatz von Spracherkennungssoftware, beispielsweise für die Büroautomation (Erstellung von Akten und Dokumenten mittels Spracheingabe) oder für eine stimmbiometrische Erkennung der Herkunftsregion von Flüchtlingen.[91] Sprachbasierte Assistenten können dabei helfen, Anfragen zu beantworten.[92] Im privaten und öffentlichen Verkehr kann KI zur Optimierung und Steuerung des Verkehrs oder zur Parkraumüberwachung[93] genutzt werden. Die deutsche Bundespolizei erprobte vom 01.08.2017 bis 31.07.2018 verschiedene, u.a. KI-basierte Systeme zur automatischen Gesichtserkennung am Bahnhof Berlin Südkreuz (Bundespolizeipräsidium Potsdam 2018). Hinweise auf den staatlichen Einsatz von KI finden sich auch im Zusammenhang mit der Bekämpfung von Wirtschaftskriminalität.[94] Schliesslich ist zu erwähnen, dass in den USA KI im Strafprozess verwendet wird, um Proben mit gemischten DNA-Spuren zu analysieren (Kwong 2017).

Ein weiteres Beispiel betrifft die Beurteilung von Gesuchen um unentgeltliche Rechtspflege (Rechsteiner 2018), wo ein Algorithmus die Erfolgsaussichten des betreffenden Falles aufgrund vergleichbarer Fälle ermitteln würde. Hinzuweisen ist ferner auf die im Rahmen der Dateninnovationsstrategie des Bundesamtes für Statistik (BfS 2017) entwickelten Pilotprojekte Arealstatistik Deep Learning

---

[91] Siehe: https://www.cancom.info/2019/01/spracherkennung-in-behoerden-ueber-digitale-entlastung-bis-hin-zum-datenschutz/.

[92] Vgl. das Beispiel von «Amelia» im Bezirk North London Borough. Nach dreimonatigem Training konnte Amelia 64 % aller Anfragen erfolgreich beantworten; siehe Bitkom 2017.

[93] Vgl. den Aktionsplan «Digitalisierung und Künstliche Intelligenz in der Mobilität» des deutschen Bundesministeriums für Verkehr und digitale Infrastruktur vom November 2018, abrufbar unter https://www.bmvi.de/SharedDocs/DE/Anlage/DG/aktionsplan-ki.pdf?__blob=publicationFile, sowie die Liste der KI-Projekte des BMVI unter https://www.bmvi.de/DE/Themen/Digitales/Aktionsplan-Digitalisierung-und-Kuenstliche-Intelligenz/KI-Projekte-in-der-Mobilitaet/aktionsplan.html.

[94] Vgl. die Beiträge auf der Herbsttagung des deutschen Bundeskriminalamtes zum Thema «Sicherheit in einer offenen und digitalen Gesellschaft» vom 21.–22.11.2018, abrufbar unter https://www.bka.de/DE/AktuelleInformationen/Publikationen/BKA-Herbsttagungen/2018/ProgrammUndRedebeitraege/programmUndRedebeitraege.html;jsessionid=F3875CE18706B64522A2D7F8B664B76D.live2291?nn=99088. Für die Schweiz vgl. Moser 2019.

(ADELE)[95], NOGAuto[96], ML_SoSi[97] oder Plausibilitätsprüfungen mittels Machine Learning[98]. Alle diese Projekte sind noch nicht im Einsatz.[99]

3.5.3. Kritik und Lösungsansätze

3.5.3.1. Datenschutz

Im Zusammenhang mit dem Einsatz von KI, insbesondere im Rahmen automatisierter Entscheide, werden häufig personenbezogene Daten verarbeitet. So kann KI z.B. im Rahmen von personenbezogener vorausschauender Polizeiarbeit oder bei der Beurteilung der Rückfallgefahr von Straftätern eingesetzt werden, um eine Prognose über die Gefährlichkeit einer Person abzugeben (siehe auch Abschnitte 2.9 und 2.10). In diesem Zusammenhang sind der verfassungsrechtlich im Rahmen des Schutzes der Privatsphäre (Art. 13 BV) verbürgte Schutz vor dem Missbrauch persönlicher Daten (Art. 13 Abs. 2 BV) sowie weitere Persönlichkeitsrechte relevant (Braun Binder 2019b). In der Praxis besteht bei der Umsetzung des Datenschutzrechts allerdings häufig die Schwierigkeit, dass die betroffene Person gar nicht weiss, welche Daten über sie gespeichert bzw. im Rahmen einer KI-gestützten Verarbeitung genutzt werden. Etwas Abhilfe könnte eine Überarbeitung des Datenschutzgesetzes schaffen. In seinem Entwurf für ein neues Datenschutzgesetz vom September 2017[100] sieht der Bundesrat in Art. 19 Abs. 1 E-DSG denn auch eine Informationspflicht bei automatisierten Einzelentscheidungen vor.

---

[95] Mithilfe von KI sollen Luftbildinterpretationen zur Identifizierung und Klassifizierung von Veränderungen teilweise automatisiert werden. Siehe: https://www.experimental.bfs.admin.ch/de/index.html.

[96] Mittels maschineller Lernverfahren soll die Kodierung der wirtschaftlichen Tätigkeit von Unternehmen auf Basis bereits vorhandener Daten beim BfS automatisiert werden.

[97] Bei diesem Projekt geht es um die Gruppierung typischer Verlaufsmuster bezüglich Leistungsbezügen im System der sozialen Sicherheit und Erwerbsarbeit sowie Schätzung der Gruppenzugehörigkeit durch Nutzung individueller Merkmale und retrospektiver Verlaufsdaten.

[98] Plausibilitätsprüfungen im BfS sollen anhand von maschinellen Lernverfahren erweitert und beschleunigt werden bei gleichzeitiger Steigerung der Datenqualität.

[99] Experimentelle Statistiken des BfS werden auf einer speziellen Microsite veröffentlicht: www.experimental.bfs.admin.ch.

[100] Bundesgesetz über die Totalrevision des Bundesgesetzes über den Datenschutz und die Änderung weiterer Erlasse zum Datenschutz, BBl 2017, 7193 (7215).

Abs. 4 enthält eine Sonderregelung für Bundesorgane: Diese sollen automatisierte Einzelentscheidungen (gemeint sind Verfügungen[101]) entsprechend kennzeichnen.[102]

Diese Regelung alleine würde allerdings noch keine Klarheit über die genutzten Daten schaffen. Sie würde vielmehr nur einen ersten Anhaltspunkt für betroffene Personen bieten, weitere Informationen einzuholen. Die Regelung würde sich auch lediglich auf den engen Anwendungskreis von Entscheidungen beziehen, die ausschliesslich auf einer automatisierten Bearbeitung von Personendaten beruhen. Dies ist der Fall, «wenn keine inhaltliche Bewertung und darauf gestützte Entscheidung durch eine natürliche Person stattgefunden hat. Das heisst, die inhaltliche Beurteilung des Sachverhalts, auf dem die Entscheidung beruht, erfolgte ohne Dazutun einer natürlichen Person. Darüber hinaus wird auch der Entscheid, der auf der Basis dieser Sachverhaltsbeurteilung ergeht, nicht von einer natürlichen Person getroffen.»[103] Mit anderen Worten, die vorgeschlagene Regelung in Art. 19 Abs. 4 E-DSG würde im Falle des KI-Einsatzes nur in Situationen einer durch die KI autonom getroffenen Entscheidung greifen. Aber auch in diesen Fällen würde die neue Regelung der betroffenen Person noch nicht die Möglichkeit eröffnen, umfassend Kenntnis über die von der KI verarbeiteten Daten zu erhalten. Sie würde einzig die Information erhalten, dass die sie betreffende rechtsverbindliche Entscheidung ohne menschliches Zutun gefällt wurde. Immerhin könnte diese Information für die betroffene Person Anlass sein, mithilfe des datenschutzrechtlichen Auskunftsrechts (Art. 8 f DSG) von der Bundesbehörde die Angabe weiterer zu verlangen.

Die vorgeschlagene Regelung vermag also in Fällen einer vollständig von Maschinen getroffenen Entscheidung eine gewisse Transparenz und Überprüfbarkeit herzustellen, um individuelle Rechte abzusichern. In Situationen, in denen KI lediglich zur Entscheidungsunterstützung eingesetzt wird, ist sie allerdings nicht anwendbar. Genauso wenig vermag sie, gruppen- und gesellschaftsbezogene Ziele wie Nichtdiskriminierung abzusichern.[104]

---

[101] BBl 2017, 6941 (7059).

[102] Art. 19 Abs. 4 E-DSG lautet: «Ergeht die automatisierte Einzelentscheidung durch ein Bundesorgan, so muss es die Entscheidung entsprechend kennzeichnen. Absatz 2 gilt nicht, wenn der betroffenen Person gegen die Entscheidung ein Rechtsmittel zur Verfügung steht.»

[103] BBl 2017 6941 (7056 f.).

[104] Vgl. für eine ähnliche Kritik mit Blick auf die EU-Datenschutz-Grundverordnung Dreyer & Schulz 2018. Vgl. auch den Hinweis bei Spielkamp 2019.

3.5.3.2. Datenqualität

Sowohl die Entscheidungsunterstützung als auch die Vornahme gewisser Entscheidungsschritte durch KI ist auf Daten angewiesen, die analysiert werden und auf deren Basis sodann eine Prognose oder Entscheidung getroffen wird. Es gilt die allgemeine Erkenntnis, dass die Ergebnisse von Datenverarbeitungsverfahren immer nur so gut sein können, wie es die Qualität, Aktualität und Verfügbarkeit der Daten ermöglicht (für *predicitive policing* siehe z.B. Knobloch 2018 oder Leese 2018). Dies trifft auf diejenigen Daten zu, die im Laufe des ordentlichen Betriebs in ein System eingespeist und verarbeitet werden. Im Falle von maschinellen Lernverfahren, bei denen Trainingsdaten zum Einsatz kommen, gilt dies aber auch für die genutzten Trainingsdaten. Falsche Trainingsdaten können dazu führen, dass das System unkorrekte Resultate liefert.[105] Zudem besteht das Risiko, dass historisch gewachsene Vorurteile, die sich in den Trainingsdaten niederschlagen, vom System übernommen und perpetuiert bzw. verstärkt werden (Martini 2019). Mit anderen Worten: Auch wenn die Daten an und für sich korrekt sind, können sie zu diskriminierenden Resultaten führen (vgl. Abschnitt 2.8.3.4). Bereits in der Trainingsphase ist deshalb die Qualität der verwendeten Daten sicherzustellen (Braun Binder 2019b; Martini 2019). In der Literatur findet sich daher die Empfehlung, dass im Idealfall sowohl beim Betrieb als auch im Rahmen der Trainingsphase auf Daten der öffentlichen Hand zurückgegriffen werden soll (Knobloch 2018).

Anzumerken ist hier noch, dass die Verwendung grosser Mengen von Daten dazu beitragen kann, dass die Relevanz eines einzelnen falschen Datums relativiert wird. Zudem können falsche Daten natürlich auch ausserhalb des Einsatzes von KI nachteilige Folgen haben; eventuell sogar noch mehr, weil die allfällige Relativierung falscher Daten durch die Menge der Daten entfällt.

[105] Vgl. nur etwa das Beispiel des von Microsoft-Forschern mit realen Twitter-Daten trainierten Chatbots «Tay», der sich nach kurzer Zeit rassistisch und anderweitig politisch radikal äusserte, vgl. Graff 2016, oder den am Massachusetts Institute of Technology durchgeführten Test mit Bilderkennungsprogrammen: Ein System wurde mit «normalen» Daten trainiert, das andere System mit Daten aus dem Darknet. Letztgenanntes System «erkannte» daraufhin auf Bildern, die das andere System als Vögel oder Blumen interpretierte, ausschliesslich Gewaltszenen, vgl. Wakefield 2018.

3.5.3.3. Begründung staatlicher Entscheidungen

Der staatliche Einsatz künstlicher Intelligenz steht im Vergleich zum privatwirtschaftlichen Einsatz vor zusätzlichen Herausforderungen. Die Begründung eines Rechtsaktes bzw. Gerichtsurteils gegenüber Betroffenen gehört zu den zentralen Elementen des Rechtsstaates. Die Begründungspflicht folgt aus dem Anspruch auf rechtliches Gehör (Art. 29 Abs. 2 BV) und zwingt den Staat dazu, Transparenz herzustellen bezüglich der Entscheidungsgründe. Dies soll den Betroffenen ermöglichen, den Entscheid nachzuvollziehen und allenfalls ein sachgerechtes Rechtsmittel einzureichen.

Trifft aber ein auf neueren maschinellen Lernverfahren wie Deep Learning basiertes KI-System die Entscheidung oder zumindest wesentliche vorbereitende Teile davon, können die internen Schritte, die innerhalb des Systems stattfinden, nicht mehr nachvollzogen werden. Daraus wird in der Literatur bisweilen der Schluss gezogen, eine den rechtsstaatlichen Anforderungen genügende Begründung sei nicht mehr möglich oder zumindest zweifelhaft (Rechsteiner 2018).

Andere Stimmen betonen aber, dass auch bei der Begründung einer von einem Behördenmitarbeitenden getroffenen Entscheidung die einzelnen Gedankengänge der Person nicht wiedergegeben werden müssen. Genauso wenig müssten die einzelnen maschinellen Schritte in der Begründung dargestellt werden. Auch wenn der Herstellungsprozess einer auf maschinellen Lernverfahren basierenden Entscheidung nicht im Einzelnen nachvollzogen werden könne, müsse dies nicht bedeuten, dass KI keine Entscheidungen treffen dürfe. Es stelle sich aber die Frage, ob bzw. wie KI künftig in der Lage sein werde, die Entscheidung nicht nur zu fällen, sondern auch in einer den rechtsstaatlichen Anforderungen an Begründungen genügenden Form darzustellen (Meyer 2018; Doshi-Velez & Kortz 2017).

3.5.3.4. Diskriminierungspotenzial

Das Diskriminierungspotenzial von KI wird insbesondere im Zusammenhang mit vorausschauender Polizeiarbeit und der Beurteilung der Rückfallgefahr von Straftätern in der Literatur intensiv diskutiert. Bei der vorausschauenden Polizeiarbeit bezieht sich die Diskussion sowohl auf die personen- als auch auf die ortsbezogenen Systeme.

Bei *personenbezogenem* KI-Einsatz ist neben den datenschutzrechtlichen Anforderungen insbesondere zu berücksichtigen, dass keine Vorurteile gegenüber gewissen Bevölkerungsgruppen oder Minderheiten entstehen bzw. reproduziert werden (Andrejevic 2017; Mantello 2016; Hildebrandt 2016; Martini 2019).[106]

Bei *ortsbezogenen* Systemen können ebenfalls Bedenken auftreten. Dies etwa dann, wenn der Einsatz von Polizeikräften zur Überrepräsentanz in ärmeren Gegenden mit einem hohen Anteil bestimmter Bevölkerungsgruppen führt. Durch die erhöhte Polizeipräsenz werden zusätzliche Straftaten beobachtet, wodurch es zu einer Verzerrung in der Wahrnehmung der Verteilung von Straftaten kommen kann. KI-basierte Systeme können eine erhöhte Polizeipräsenz in bestimmten Gegenden unter Umständen unabhängig von Daten wie Einkommen oder Hautfarbe empfehlen; diese Faktoren können in ethnisch stark segregierten Gebieten aber mit der Postleitzahl übereinstimmen (Knobloch 2018).

3.5.3.5. Anspruch auf rechtliches Gehör

Wird ein KI-Algorithmus vom Staat so eingesetzt, dass dieser unmittelbar selbst eine für die Rechtsunterworfenen verbindliche Entscheidung (Verfügung) trifft, dann besteht die Gefahr, dass der Anspruch auf rechtliches Gehör beeinträchtigt wird (Braun Binder 2019b). Dieser Anspruch ist Bestandteil der allgemeinen Verfahrensgarantien (Art. 29 Abs. 2 BV) und bedeutet insbesondere, dass eine Person das Recht hat, sich zu äussern, bevor eine sie nachteilig treffende Entscheidung gefällt wird. In der Literatur wird deshalb vorgeschlagen, vollautomatisierte Verfügungen grundsätzlich nur dort zuzulassen, wo den Begehren der betroffenen Person entsprochen wird (Rechsteiner 2018).

Alternativ könnte die Äusserungsmöglichkeit bei Begehren, die elektronisch gestellt werden, in Form eines sogenannten Freitextfeldes vorgesehen werden. Sobald ein Eintrag in diesem Freitextfeld gemacht wird, wäre über das Begehren nicht mehr vom Algorithmus, sondern von einem Menschen, unter Berücksichtigung der getätigten Äusserung, zu entscheiden (Braun Binder 2016a; Braun Binder 2016b).

[106] Z.B. bezüglich der Nutzung von Gesichtserkennung: Big Brother Watch, Defending Civil Liberties, Protecting Privacy, 2018, abrufbar unter: https://bigbrotherwatch.org.uk/wp-content/uploads/2018/07/Big-Brother-Watch-evidence-Policing-for-the-future-inquiry.pdf.

3.5.4. Fazit: Themenauswahl für die Expertenumfrage

Die Beispiele des staatlichen KI-Einsatzes stammen vorwiegend aus anderen Ländern. Entsprechend finden sich auch die Auseinandersetzungen damit hauptsächlich in der Literatur aus Rechtskreisen ausserhalb der Schweiz. In der Expertenumfrage war deshalb von besonderem Interesse, welche Entwicklungen in den nächsten fünf bis zehn Jahren einerseits in der Schweiz sowie andererseits in anderen Ländern erwartet werden.

Die in der Literatur diskutierten Herausforderungen unterscheiden sich massgeblich danach, in welchen Prozessen der öffentlichen Verwaltung KI eingesetzt wird und inwieweit dadurch menschliche Entscheidungen ersetzt werden. Deshalb wurden die Experten nach den erwarteten Einsatzgebieten in der öffentlichen Verwaltung gefragt. Zudem zielte die Expertenumfrage darauf ab zu erfahren, welche Vorteile, aber auch welche Risiken im Zusammenhang mit einem staatlichen KI-Einsatz erwartet werden.

4. Experten zu künstlicher Intelligenz

Der Einbezug von Fachpersonen spielte eine zentrale Rolle, um (i) die Einschätzungen des Projektteams und den Konsensus aus der Literatur kritisch zu hinterfragen sowie (ii) neue Empfehlungen zu erarbeiten. Zweck des Einbezugs von Expertinnen und Experten war dabei nicht, den Prozess der Entscheidungsfindung bezüglich Empfehlungen gewissermassen auszulagern. Vielmehr soll damit das Meinungsspektrum ausgeweitet werden – auch um blinde Flecken des Studienteams zu identifizieren. Wie in Abschnitt 1.3.1 ausgeführt, erfolgte die Expertenumfrage in zwei Runden, in denen unterschiedliche Sachverhalte erhoben wurden: In der ersten Runde ging es vorrangig um die Einschätzung der Faktenlage, in der zweiten Runde primär um die Einschätzung von möglichen Massnahmen. Ergänzt wurden die beiden Umfragen durch eine Befragung der breiteren Bevölkerung mit Schwerpunkten «Konsum» und «Ethik». Schliesslich konnten die Fachpersonen in einem Workshop zu ersten Entwürfen von Empfehlungen Stellung nehmen. Details zur Methode finden sich im Anhang.

4.1. KI-Wissen und Meinungen der Fachpersonen

In der Expertenumfrage wurde auf verschiedene Weise sowohl die generelle Meinung der Fachpersonen gegenüber KI (zwischen Skepsis und Zustimmung) als auch deren Wissen und Erfahrung mit KI erhoben. Dies erlaubt es, die später in der Umfrage gegebenen Einschätzungen auf mögliche Voreingenommenheit oder (fehlende) Erfahrung der Fachpersonen zu prüfen. Entsprechende Ergebnisse werden nur berichtet, wenn sich ein auffälliges Resultat zeigte.

4.1.1. Charakterisierungsmerkmale des Expertensamples

Die Daten der Umfrage wurden hinsichtlich dreier Merkmale aggregiert, die für die Auswertung der Resultate herangezogen wurden. So wurde erstens ein **KI-Expertise-Index** generiert, der Aufschluss darüber gibt, wie umfassend eine Person sich über KI informiert bzw. KI-Technologien im Alltag nutzt. Der Index aggregiert Antworten zu folgenden Fragen:

- Eine Frage erfasste aufgrund verschiedener Unterkategorien (Fachpresse, allgemeine Medien, Konferenzen, Weiterbildung), auf welche Weise und wie häufig sich die Experten zu KI informierten.
- Eine Frage erfasste, welche KI-Technologien die Expertinnen und Experten wie häufig in ihrem Alltag nutzen (digitale Assistenten, Übersetzungsprogramme, KI-Systeme zur Datenanalyse, KI-Systeme zur Bildanalyse).

Die damit generierten Indexwerte pro Person[107] wurden zum einen für Korrelationsanalysen, zum anderen für Gruppenvergleiche (Median-Split) genutzt. Je höher der Wert, desto höher die Expertise der jeweiligen Person.

Zweitens wurde ein **KI-Meinungs-Index** geniert, der die generelle Einstellung der Expertinnen und Experten gegenüber KI widerspiegeln sollte. Auch dieser Index wurde aus den Antworten mehrerer Fragen generiert:

- Eine Frage erfasste, inwieweit sich die Expertinnen und Experten bezüglich der durch KI eröffneten Möglichkeiten eher als «Enthusiasten» oder als «Skeptiker» bezeichnen (5 Punkte Likert-Skala)
- In einer offenen Frage wurden die Fachpersonen aufgefordert, mit Stichworten KI zu charakterisieren. Hier wurde kodiert, inwieweit dafür eher neutrale, eher positive oder eher negative Stichworte verwendet wurden.

Auch diese damit generierten Indexwerte pro Person wurden für Korrelationsanalysen und Gruppenvergleiche (Median-Split) genutzt. Je höher der Wert, desto skeptischer ist die Person bezüglich KI generell.

Drittens wurden die Personen bezüglich ihres **Ausbildungshintergrunds** in «technische» und «nicht technische Fachpersonen» unterschieden. In erstere Kategorie fallen Personen mit einer Ausbildung in Informatik, Ingenieurwesen und den Naturwissenschaften (161 Personen), in letztere alle anderen (146 Personen).

Zu jedem Bereich wurde am Schluss des Frageblocks gefragt, wie gut sich die Person im jeweiligen Thema auskennt (Skala: Experte; sehr vertraut; vertraut; wenig vertraut; gar nicht vertraut). Damit soll sichergestellt werden, dass die Resultate nicht aufgrund unterschiedlicher Expertise pro Thema verzerrt wurden.

[107] Bei der Erstellung beider Indizes wurde geprüft, wie stark die Antworten zu den einzelnen Items miteinander korrelierten, um sicherzustellen, dass alle Korrelationen signifikant sind und in die gleiche Richtung gehen; dies war der Fall. Zudem wurde geprüft, welchen Einfluss die Gewichtung von Einzel-Items auf den Gesamtindex hat, um einen möglichst robusten Index zu erhalten.

4.1.2. Zusammenhänge zwischen den Charakterisierungsmerkmalen

Aufgrund der oben definierten Indizes liessen sich Verbindungen zwischen den Merkmalen feststellen. So zeigte sich erstens ein plausibler Zusammenhang zwischen technischem Ausbildungshintergrund und Expertise-Index: Technische Experten erreichen im Schnitt einen höheren Indexwert als nicht technische Experten. Es ergab sich allerdings kein systematischer Zusammenhang zwischen technischem Ausbildungshintergrund und KI-Skeptizismus, d.h. technische Experten stehen im Schnitt KI ähnlich gegenüber wie nicht technische.

Beim Vergleich der beiden Indizes zeigte sich eine signifikante negative Korrelation zwischen KI-Expertise und KI-Meinung: Je weniger die betroffenen Personen sich über KI informieren bzw. KI nutzen, desto skeptischer sind sie ihr gegenüber eingestellt ($– 0.24$, $p < 0.001$). Eine Analyse der beiden Komponenten des KI-Expertise-Index ergab, dass die Komponente «Erfahrung mit KI» für diese negative Korrelation am stärksten verantwortlich ist. Um diesen möglichen Zusammenhang zwischen Skepsis und Wissen zu prüfen, wurden in der zweiten Umfrage die folgenden Aussagen hinzugefügt, wobei die Befragten den Grad der Ablehnung bzw. Zustimmung angeben konnten:

1. Eine skeptische Einstellung bezüglich KI ist in erster Linie ein Resultat ungenügender Information: Je mehr Personen über KI wissen, desto geringer wird die KI-Skepsis sein.
2. Personen mit technischer KI-Expertise sind gegenüber KI positiv voreingenommen und neigen dazu, Risiken zu übersehen.

Die Antworten wurden mit dem KI-Experten- und dem KI-Meinungs-Index korreliert. Hier zeigte sich nur zwischen dem Meinungs-Index und der ersten Aussage eine signifikante negative Korrelation ($– 0.37$, $< .001$). Dies bedeutet, dass Skeptiker selbst nicht der Ansicht sind, ihre Skepsis beruhe auf mangelnde Information. Weil kein Zusammenhang zwischen Expertise und der ersten Aussage gefunden wurde (also im Sinn, dass Personen mit geringerer Expertise der Aussage stärker oder weniger stark zustimmen), muss davon ausgegangen werden, dass KI-Skepsis nicht alleine als Ausdruck mangelnder Expertise gewertet werden kann.

Tabelle 5 zeigt die durchschnittliche Expertise der Fachpersonen, welche in der ersten bzw. zweiten Umfrage ihre Einschätzung zu den Fragen der jeweiligen Themenbereiche abgegeben haben. Übersetzt auf die Skalenwerte liegt diese zwi-

schen «vertraut» und «sehr vertraut».[108] Tendenziell liegen die Werte bei der zweiten Umfrage etwas höher. Dies ist darauf zurückzuführen, dass die Befragten explizit aufgefordert wurden, ihre Einschätzungen zu Empfehlungen möglichst vieler Bereiche abzugeben. Ein deutlicher Unterschied findet sich nur bei der «öffentlichen Verwaltung», wo in der ersten Umfrage generell signifikant mehr Expertise (auch im Vergleich zwischen den Bereichen) vorhanden war.

Tabelle 5: Durchschnittliche Expertise in den einzelnen Fachbereichen pro Umfrage; ein höherer Wert bedeutet eine höhere Expertise (*: signifikanter Unterschied zwischen erster und zweiter Umfrage)

	Erste Umfrage	Zweite Umfrage
Arbeitswelt	3.37	3.16
Bildung und Forschung	3.76	3.43*
Konsum	3.26	3.38
Medien	3.79	3.51
Öffentliche Verwaltung	3.92	3.40*

4.2. Beurteilungen zum Themenfeld Arbeitswelt[109]

4.2.1. Zentrale Ergebnisse der ersten Umfrage

In der ersten Umfrage haben 115 von 307 Befragten den Bereich Arbeit gewählt. Hier wurden primär Einschätzungen zu quantitativen Effekten auf makroökonomischer Ebene und zu qualitativen Veränderungen der Gestaltung von Arbeit aus Sicht von Arbeitnehmerinnen und Arbeitnehmern sowie Arbeitgebern erfragt. Ebenso wurde gefragt, ob unterschiedliche Einschätzungen bei den Auswirkungen

[108] Die Antwortmöglichkeiten waren: 1: Ich bin mit dem Thema gar nicht vertraut; 2: Ich bin mit dem Thema wenig vertraut; 3: Ich bin mit dem Thema vertraut; 4: Ich bin mit dem Thema sehr vertraut; 5: Ich bin Experte/Expertin.

[109] Dieser Abschnitt beruht auf Arbeiten von Johann Čas und Jaro Krieger-Lamina, Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften.

auf die Schweiz und auf die Europäische Union bestehen. Nachfolgend werden nur die für die Studie wichtigsten Ergebnisse aufgeführt und grafisch dargestellt.

4.2.1.1. Makroökonomische Effekte

Aus quantitativer Sicht wurde nach den Wirkungen von KI auf die Beschäftigung, die Produktivität, die Löhne, die Bedeutung des Produktionsfaktors menschliche Arbeit, die Anteile des Dienstleistungssektors sowie auf die Polarisierung bei Qualifikationen und Löhnen gefragt (Abbildung 10).

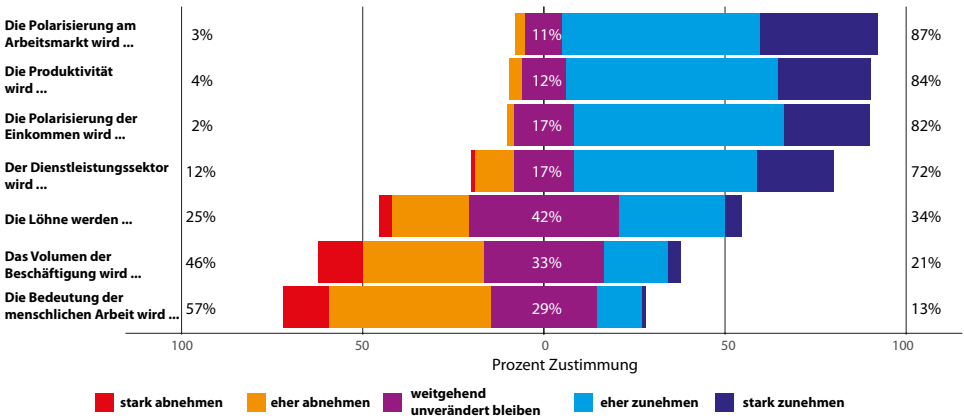

Abbildung 10: Beurteilung von quantitativen Effekten im Bereich Arbeitswelt. Die Prozentzahlen links und rechts geben jeweils den Anteil der zustimmenden bzw. ablehnenden Personen an.

Die Teilnehmenden stufen die Zunahme der Polarisierung des Arbeitsmarktes (87 % positiv), die Zunahme der Produktivität (84 %) und der Löhne (82 %) als die wahrscheinlichsten Szenarien ein. Die Bedeutung menschlicher Arbeit (57 % negativ) und das Beschäftigungsvolumen (46 %) werden als eher abnehmend eingeschätzt. Die Mehrheit stuft die Entwicklungen in der Schweiz ähnlich ein wie in der Europäischen Union. Ein Viertel bis ein Drittel der Befragten denkt, dass KI im Dienstleistungssektor, in der Produktivität, bei der Polarisierung des Arbeitsmarktes, bei den Löhnen und bei der Beschäftigung einen grösseren Einfluss haben wird im Vergleich zu Europa.

Bei den Potenzialen, menschliche Arbeit zu unterstützen, sie zu ersetzen oder zusätzliche Beschäftigung zu generieren, wurde zwischen dem Agrar-, Produktions- und Dienstleistungssektor unterschieden. Grundsätzlich stuft eine Mehrheit der Befragten die Unterstützung (53–82 % je Sektor) wie auch den Ersatz (36–64 %) von menschlicher Arbeitskraft über alle Sektoren hinweg als wahrscheinlich ein. Zusätzliche Beschäftigung wird für den Dienstleistungssektor als eher wahrscheinlich eingestuft (59 %), für den Produktionssektor sind sich die Befragten uneinig und beim Agrarsektor wird keine zusätzliche Beschäftigung erwartet.

4.2.1.2. Auswirkungen auf die Arbeitsverhältnisse

Bezüglich der Auswirkungen von KI auf die Gestaltung der Arbeit aus Sicht der **Arbeitnehmer/-innen** stufen die Befragten eine stärkere Kontrolle von Arbeitnehmer/-innen (76 %), die Zunahme instabiler Arbeitsverhältnisse (74 %) wie auch den Einsatz von KI im Zuge der Rekrutierung (73 %) als die drei wahrscheinlichsten Szenarien ein. Auch der Einsatz von KI bei der Entscheidung über Beförderungen (64 %) und für die Erhöhung der Autonomie für Arbeitnehmer (57 %) wird eher als wahrscheinlich eingestuft. Dasselbe gilt für den Erhalt sozialer Absicherungen (52 %). Eine dank KI kürzere Arbeitszeit sowie eine sinkende Arbeitsbelastung werden eher als unwahrscheinlich angesehen, wobei Letztere stark zu polarisieren scheint. Jeweils mehr als 20 % der Fachpersonen schätzen eine sinkende Arbeitsbelastung als sehr wahrscheinlich oder als sehr unwahrscheinlich ein.

Aus Sicht der **Arbeitgeber** werden eine Zunahme der Produktivität (93 %), mehr Vorteile für Grossunternehmen (90 %) sowie flexiblere Reaktionen auf Marktänderungen (88 %) als die drei am wahrscheinlichsten Szenarien eingestuft. Auch eine bessere Kontrolle von Arbeitnehmerinnen und Arbeitnehmern und flexiblere Zuteilung von Arbeitszeiten und Aufgaben werden von der Mehrheit der Befragten als wahrscheinlich eingestuft. Bei der Frage, ob KI Arbeitnehmer/-innen primär ersetzt oder lediglich unterstützt, sind sich die Befragten uneinig; beide Szenarien werden als gleich wahrscheinlich eingeschätzt (57 %).

4.2.1.3. Fazit zur ersten Umfrage

Aufgrund der Ergebnisse der ersten Umfrage wurden folgende Bereiche ausgewählt, um mögliche Massnahmen zu beurteilen: die Polarisierung von Arbeits-

märkten und von Löhnen, eine Abnahme des Volumens der Beschäftigung, stärkere Kontrolle von Arbeitnehmern, eine Zunahme prekärer Arbeitsverhältnisse sowie Vorteile insbesondere für Grossunternehmen.

4.2.2. Zentrale Ergebnisse der zweiten Umfrage

In der zweiten Umfrage haben 70 von 111 Befragten Empfehlungen zum Bereich Arbeitswelt angegeben. Die Fachpersonen wurden gebeten, zu den in der ersten Runde als wesentlich eingeschätzten Risiken eigene Massnahmen vorzuschlagen bzw. vorgegebene Massnahmen zu bewerten. Bei den vorgeschlagenen Massnahmen wurden Kommentare aus der ersten Runde der Befragung sowie die Ergebnisse der Tiefeninterviews und der Literaturanalyse berücksichtigt.

4.2.2.1. Polarisierung am Arbeitsmarkt

Wie aus Abbildung 11 ersichtlich, werden Massnahmen im Bildungsbereich als sehr effektiv zur Vermeidung von Polarisierungseffekten angesehen. Mit Werten von 90 % für eine gezielte Weiterbildung auf individueller Ebene und 82 % für Bildungsoffensiven zur generellen Förderung von KI-Kompetenzen werden diese Massnahmen eindeutig positiv bewertet.

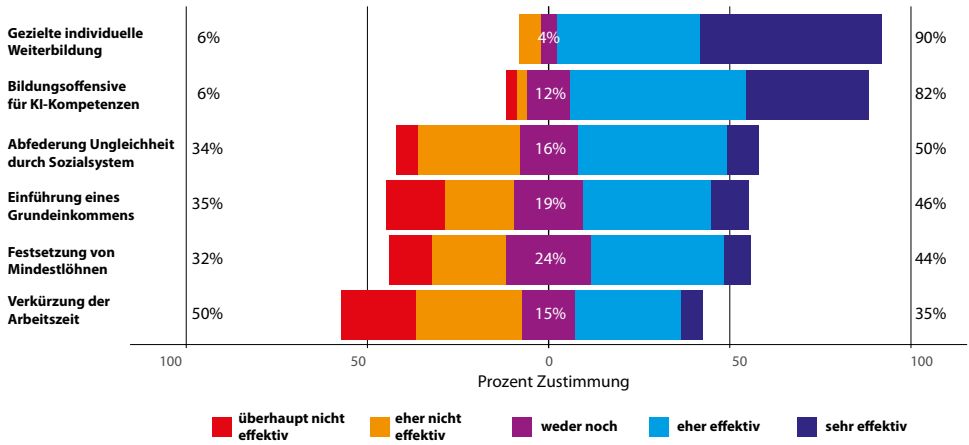

Abbildung 11: Beurteilung von Massnahmen zur Verminderung der Polarisierung am Arbeitsmarkt.

Bei einer verstärkten Abfederung sozialer Ungleichheit durch das Sozialsystem, der Einführung eines Grundeinkommens und der Festsetzung von Mindestlöhnen überwiegen die positiven Einschätzungen, die Tendenz ist aber weniger klar. Bei einer Arbeitszeitverkürzung zur Verknappung des Angebots an Arbeit überwiegen mit 50 % negative Beurteilungen.

4.2.2.2. Polarisierung der Löhne

Gegen steigende Ungleichheiten bei den Löhnen wurde die gleiche Liste an Massnahmenvorschlägen abgefragt. Wenngleich sich hier das Muster ähnelt, werden doch mit Ausnahme der Abfederung durch das Sozialsystem alle weiteren Massnahmen als weniger effektiv erachtet (Abbildung 12).

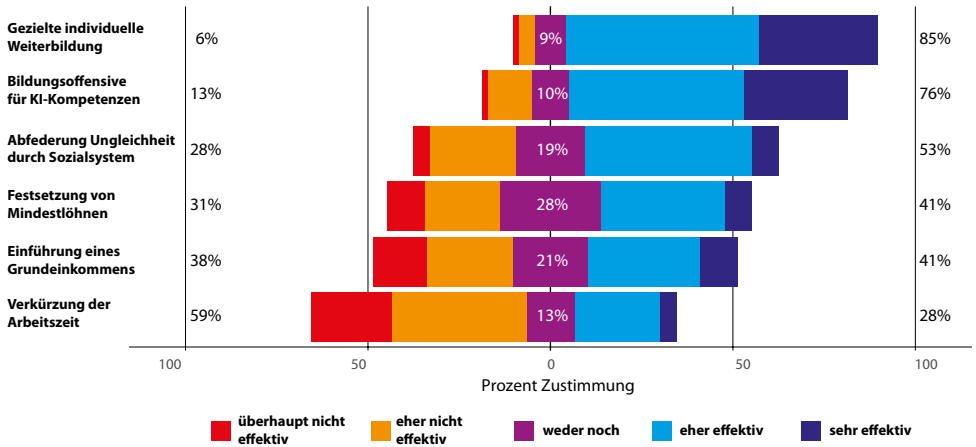

Abbildung 12: Beurteilung von Massnahmen gegen die Polarisierung von Löhnen.

Fragen der Polarisierung scheinen auch bei den Einschätzungen der Effektivität von möglichen Massnahmen zu polarisieren. Dies spiegelt sich auch in den zahlreichen, von den Fachpersonen selbst vorgeschlagenen Massnahmen wider, welche durch eine grosse Heterogenität gekennzeichnet sind.[110]

[110] Insgesamt wurden mehr als 200 eigene Vorschläge gemacht. Viele widersprechen sich und spiegeln konträre gesellschaftliche Grundhaltungen wider, indem sie etwa eine Kritik an neoliberaler Wirtschaftspolitik oder ein Misstrauen gegenüber regulativen Eingriffen des Staates ausdrücken.

4.2.2.3. Abnahme des Beschäftigungsvolumens

Bei den vorgeschlagenen Massnahmen, um einer Abnahme des Volumens der Beschäftigung entgegenzuwirken bzw. deren Folgen abzumildern, sind mit zwei Ausnahmen die Einschätzungen wenig eindeutig bzw. tendieren ins Negative (siehe Abbildung 13).

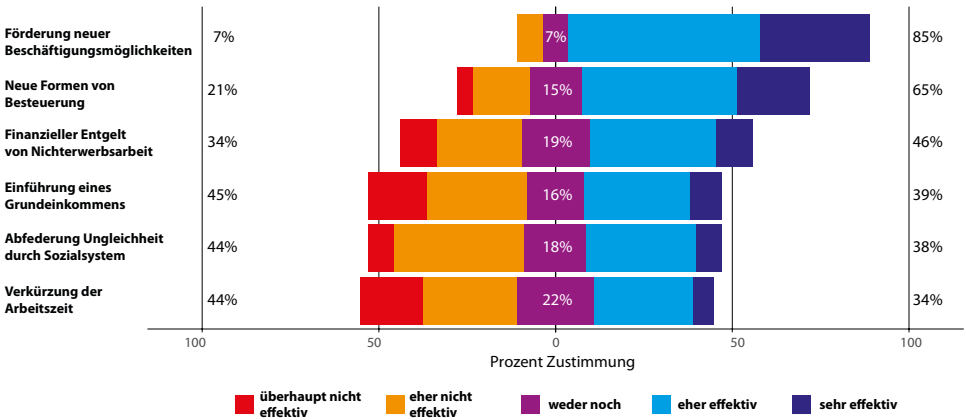

Abbildung 13: Beurteilung von Massnahmen gegen die Abnahme der Beschäftigung.

Mit einem Wert von 85 % werden die Förderung von neuen Beschäftigungsmöglichkeiten, z.B. Start-ups, gefolgt von neuen Formen der Besteuerung zur Finanzierung des Staatshaushalts/Sozialsystems (65 %) als die effektivsten Massnahmen erachtet. Bei der finanziellen Abgeltung von Nichterwerbsarbeit überwiegen mit 46 % noch die positiven Einschätzungen, bei der Einführung eines Grundeinkommens, der verstärkten Abfederung sozialer Ungleichheit durch das Sozialsystem sowie einer Arbeitszeitverkürzung zur Verknappung des Angebots an Arbeit die negativen Einschätzungen der Effektivität.

4.2.2.4. Kontrolle der Arbeitnehmerinnen und Arbeitnehmer

Um einer stärkeren Kontrolle der Arbeitnehmer/-innen entgegenzuwirken, wurden folgende Massnahmen abgefragt: Vergütung oder Vermeidung der Arbeit in der

Freizeit, Schutz der Arbeitnehmer/-innen auch in transnationalen Arbeitsverhältnissen sicherstellen, Massnahmen gegen Überwachung am Arbeitsplatz sowie eine Stärkung des Datenschutzes generell (Abbildung 14).

Mit Ausnahme einer Vergütung oder Vermeidung der Arbeit in der Freizeit, bei der die positiven und negativen Einschätzungen fast gleichauf liegen, werden die restlichen vorgeschlagenen Massnahmen überwiegend für effektiv befunden.

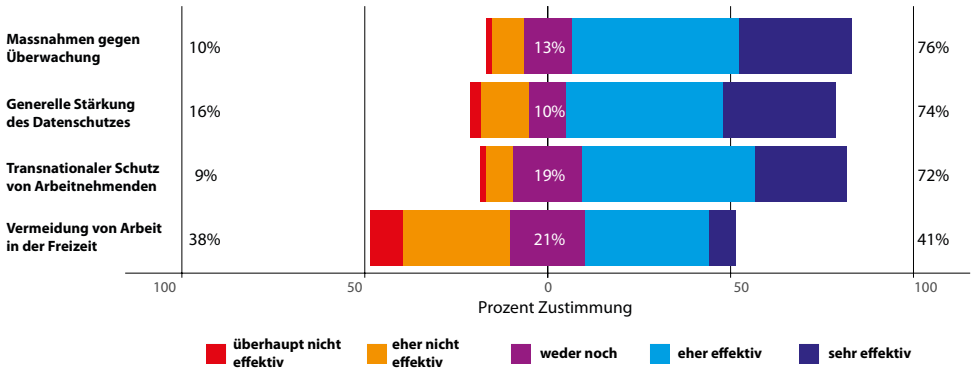

Abbildung 14: Beurteilung von Massnahmen gegen zunehmende Kontrolle der Arbeitnehmer/-innen.

4.2.2.5. Prekäre Arbeitsverhältnisse

Bei den Massnahmen, die einer Zunahme instabiler oder prekärer Arbeitsverhältnisse entgegenwirken sollen, zeigt sich ein uneinheitliches Bild (Abbildung 15).

So werden eine Sicherstellung des Arbeitnehmerschutzes auch in transnationalen Arbeitsverhältnissen (72 %) sowie Änderungen im Sozialversicherungsrecht (64 %) als überwiegend positiv beurteilt. Bei der Einschätzung einer Einführung eines Grundeinkommens und Festsetzung von Mindestlöhnen herrscht Uneinigkeit, während eine Arbeitszeitverkürzung von 60 % der Befragten als nicht effektiv angesehen wird.

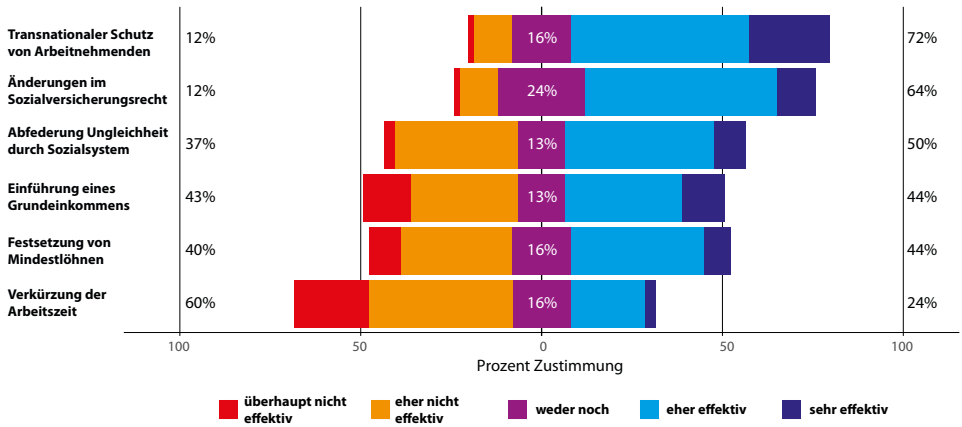

Abbildung 15: Beurteilung von Massnahmen gegen zunehmend prekäre Arbeitsverhältnisse.

4.2.2.6. Vorteile für Grossunternehmen

Als letzte Frage zur Beurteilung der Effektivität wurden Massnahmen abgefragt, um Vorteilen für Grossunternehmen und damit der Behinderung des freien Wettbewerbs entgegenzuwirken (Abbildung 16).

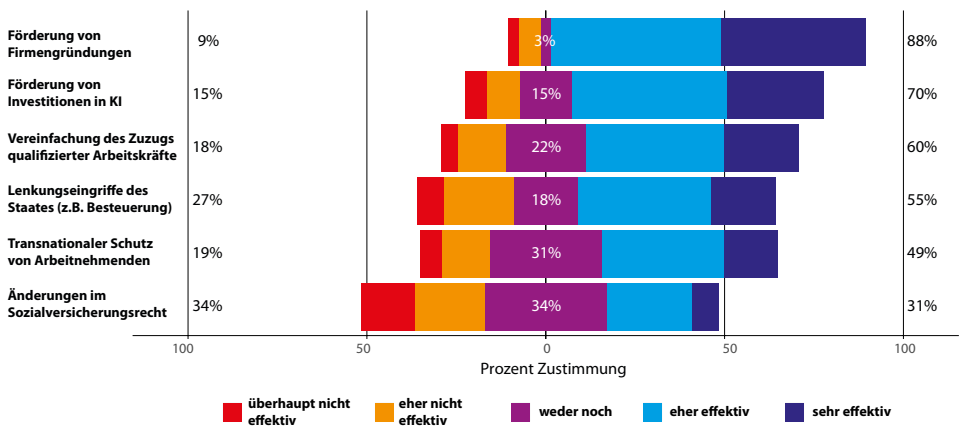

Abbildung 16: Beurteilung von Massnahmen zur Sicherung des freien Wettbewerbs.

Als eindeutig bzw. überwiegend effektiv angesehen wurden dabei die Förderung von Unternehmensgründungen (88 %), die Förderung von Investitionen in KI (70 %), eine Vereinfachung des Zuzugs qualifizierter Arbeitskräfte (60 %) sowie Lenkungseingriffe des Staates (55 %). Der Schutz von Arbeitnehmer/-innen auch in transnationalen Arbeitsverhältnissen wird eher effektiv und Änderungen im Sozialversicherungsrecht werden als eher nicht effektiv eingeschätzt, wobei bei diesen letzten beiden Handlungsoptionen der hohe Anteil an «weder noch»-Antworten darauf hindeutet, dass hier kein starker Zusammenhang zur Problematik von Wettbewerbsverzerrungen gesehen wird.

4.2.2.7. Wünschbarkeit der Massnahmen

Die abgefragten Massnahmenvorschläge wurden auch insgesamt hinsichtlich ihrer Wünschbarkeit beurteilt, da manche Massnahmen versprechen, Probleme effektiv zu lösen, während die Umsetzung dieser Massnahmen aus anderen Gründen (z.B. ethische Erwägungen oder unerwünschte sekundäre Effekte) als nicht wünschenswert erscheinen. Hinsichtlich der Tendenz decken sich die Ergebnisse weitgehend mit jenen der Wirksamkeit, wobei bei der Wünschbarkeit überwiegend höhere Zustimmungsraten erzielt wurden. Eine Ausnahme stellt etwa die Option einer Arbeitszeitverkürzung dar; je nach Zusammenhang, in dem die Effektivität dieser Massnahme abgefragt wurde, schätzen zwischen 24 % (Massnahme gegen prekäre Arbeitsverhältnisse) und 35 % (Polarisierung am Arbeitsmarkt) der Befragten eine Arbeitszeitverkürzung als effektiv ein. Damit liegt dieser Wert teilweise über und unter dem entsprechenden Wert für die Wünschbarkeit von 30 %.

Abschliessend wurde nach der Verantwortlichkeit bei der Überwindung der Herausforderungen bzw. Umsetzung der Massnahmen gefragt, wobei zwischen den Kategorien Gesetzgeber/Staat, Unternehmen und Bürger/Arbeitnehmer unterschieden wurde. Im Bereich Arbeit sehen 55 % den Staat als in erster Linie verantwortlich, 35 % die Unternehmen und 10 % die Bürgerinnen und Bürger selbst.

4.2.2.8. Zusammenfassende Beurteilung

Die Ergebnisse der zweiten Umfrage verweisen auf einen wichtigen Punkt bei der Formulierung von Empfehlungen: Insbesondere die Wahl von Massnahmen im makroökonomischen Bereich ist stark geprägt von generellen politischen Ansich-

ten über die Art und Weise, wie die Gesellschaft organisiert werden sollte. KI erweist sich dabei als ein Katalysator für die Diskussion grundlegender Themen wie Steuersystem, Ausgestaltung des Arbeitsmarktes oder Sozialpolitik.

4.2.3. Ergebnisse des Workshops zum Themenbereich Arbeit

Nach einer kurzen Vorstellungsrunde wurden die zentralen Ergebnisse der zweiten Umfrage präsentiert. In dieser wurden jene aus der ersten Runde als relevant und wahrscheinlich eingeschätzten Entwicklungen in Hinblick auf mögliche Massnahmen abgefragt. Diese umfassen eine Polarisierung der Arbeitsmärkte, steigende Ungleichheiten bei den Löhnen, eine Abnahme des Arbeitsvolumens, eine steigende Kontrolle von Arbeitnehmerinnen und Arbeitnehmern, eine Zunahme von prekären Arbeitsverhältnissen sowie Vorteile für Grossunternehmen (gegenüber KMUs). Von den Teilnehmenden der Umfrage wurde dabei eine grosse Anzahl an Vorschlägen für Massnahmen entwickelt, die sehr heterogen waren und sich teilweise widersprachen. Diese Vielfalt spiegelt offensichtlich auch eine Polarisierung bei der grundsätzlichen Ausrichtung von Massnahmen wider, die von fundamentaler Kritik an neoliberaler Wirtschaftspolitik bis zu einem Misstrauen gegenüber staatlichen Regulierungsmassnahmen reicht.

Bei den Massnahmen gegen die Polarisierung und gegen steigende Ungleichheit bei Löhnen wurde in der Umfrage eine Arbeitszeitverkürzung überwiegend als wenig effektiv angesehen, alle anderen Massnahmen (Abfederung über Sozialpolitik, Festlegung von Mindesteinkommen oder bedingungsloses Grundeinkommen) wurden überwiegend positiv eingeschätzt. Eindeutig positiv beurteilt wurden Massnahmen zur individuellen und allgemeinen Weiterbildung. Vergleichbare Muster zeigten sich bei den vorgeschlagenen Massnahmen zu den weiteren abgefragten Feldern. Mit wenigen Ausnahmen, wie etwa der Arbeitszeitverkürzung oder im Hinblick auf gewisse Fragen auch eine stärkere Unterstützung durch das Sozialsystem, wurden diese als überwiegend oder eindeutig positiv beurteilt.

Ziel der anschliessenden Diskussion war es, auf Basis der bisherigen Resultate konkrete Handlungsempfehlungen zu formulieren. Dabei wurden die folgenden Handlungsfelder thematisiert und entsprechende Vorschläge dazu entwickelt, zum Teil wurden auch spezifische Einzelmassnahmen dazu besprochen:[111]

[111] Es wurden auch explizit Fragen der beruflichen Weiterbildung in diesen Gruppen diskutiert. Diese Aspekte werden im Themenbereich Bildung (Abschnitt 4.3.3) besprochen.

Überwachung von Arbeitnehmer/-innen: Die gesetzlichen Bestimmungen wurden in diesem Bereich als ausreichend empfunden. Ein Mangel ist die Rechtsdurchsetzung. Von den Workshopteilnehmenden wird generell beobachtet (und für die Zukunft vermutet), dass Durchsetzungsmängel oder Compliance-Probleme durch den Einsatz von KI noch verstärkt werden. Woran es konkret fehlt, sind Instrumente und Verfahren, mit denen Verstösse gegen diese Regelungen aufgedeckt werden können. Des Weiteren wurde eine stärkere Beteiligung der Betroffenen beim Entscheid über den Einsatz von KI-Systemen als sinnvoll erachtet, analog zur gesetzlichen Forderung der Mitsprache bei der Auswahl der Betriebsmittel.[112] Als zentrale Voraussetzung für die mögliche Genehmigung der Überwachung der Arbeitnehmer/-innen wurden transparente Systeme genannt.

Zertifizierung von KI: Zertifizierte KI-Anwendungen stellen nach Ansicht der Teilnehmenden eine wesentliche Voraussetzung für deren Einsatz in sensiblen Bereichen dar. Durch eine Konformitätserklärung der Hersteller von KI-Systemen (analog zur CE-Kennzeichnung) könnte, so die Diskussion, das Transparenzproblem gelöst und eine rechtlich und ethisch konforme Nutzung ermöglicht werden. Dies betrifft insbesondere Aspekte des Datenschutzes (Verarbeitung von persönlichen Daten bzw. Informationen, mit denen Rückschlüsse auf persönliche Daten gezogen werden können).

Diskutiert wurden schliesslich auch **sozialpolitische Massnahmen** wie Anpassungen am Steuersystem, Verringerung der Arbeitszeit, Bekämpfung transnationaler prekärer Arbeitsverhältnisse aufgrund einer Plattformökonomie oder ein bedingungsloses Grundeinkommen. Der Tenor der ersten Runde lautete, dass hierfür die bereits bewährten Instrumente wie Vereinbarungen zwischen den Sozialpartnern über die Gesamtarbeitsverträge, bilaterale *codes of conduct* mit einzelnen Plattformen sowie die Sicherung des aktuellen Sozialversicherungssystems genutzt werden sollten. In der zweiten Runde wurde aber darauf hingewiesen, dass längerfristig eine Anpassung des Sozialversicherungssystems essenziell sein könnte. Die Aufgabe des Staates sei es, das System der Sozialversicherung dahin gehend anzupassen, dass auch alle Arbeitnehmer/-innen in neuen, möglichweise kurzfristigen und prekären Arbeitsverhältnissen (z.B. Platt-

[112] Bundesgesetz über die Information und Mitsprache der Arbeitnehmerinnen und Arbeitnehmer in den Betrieben, SR 822.14; Artikel 9 und Artikel 10 Bst. A.

formökonomien) geschützt sind. Relevant seien hier die Empfehlungen des Berichts des Bundesrates in Erfüllung der Postulate Reynard und Derder aus dem Jahr 2017.[113]

Insgesamt ergaben sich aus den Diskussionen der zweiten Runde drei Prioritäten für den Themenbereich Arbeit:

1. Förderung der beruflichen Weiterbildung (siehe dazu den Folgeabschnitt)
2. Prüfung von finanziellen Massnahmen wie z.B. Anpassungen im Sozialversicherungssystem oder Kompensationen für Lohnsteuerausfälle
3. Technische Massnahmen für Fragen, die sich aus den Auswirkungen von KI-Anwendungen auf die Gestaltung der Arbeit auswirken (Transparenz, Mitwirkung, Schutz sensibler personenbezogener Daten, Zertifizierung)

Bezüglich Massnahmen in den Bereichen 1 und 2 müsse aber festgehalten werden, dass diese den digitalen Wandel generell betreffen und nicht nur an das Thema KI geknüpft werden könnten.

4.3. Beurteilungen zum Themenfeld Bildung und Forschung[114]

4.3.1. Zentrale Ergebnisse der ersten Umfrage

In der ersten Umfrage haben 113 von 307 Befragten den Bereich «Bildung und Forschung» gewählt. Ziel der ersten Umfrage war es, von den Fachpersonen zu erfahren, für wie wahrscheinlich und wie risikobehaftet sie unterschiedliche Anwendungsmöglichkeiten von KI in Bildung und Forschung einschätzen. Hier ging es zum einen um KI-Tools für Bildung und Forschung, zum anderen um die durch das Bildungswesen zu fördernden KI-Kompetenzen.

[113] Siehe: https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-68708.html.

[114] Dieser Abschnitt beruht auf Arbeiten von Clemens Mader und Claudia Somm, Abteilung Technologie und Gesellschaft, Empa.

4.3.1.1. Einsatz von KI-Tools im Bildungswesen

Als zentraler Vorteil von KI-unterstützter Lernsoftware gilt die **Personalisierung** der Angebote auf die Stärken und Schwächen der Lernenden. Ein KI-System kann somit eine wertvolle Ergänzung zu «kollektiven» Bildungsmethoden in Klassen sein. Zum einen soll es die Lerninhalte (Schwierigkeitsgrad, Tempo, Vermittlungsart (visuell, audio etc.) für die Lernenden individuell anpassen, zum anderen werden den Lehrenden Analysen zum Lernverhalten der Schüler übermittelt, wodurch auch die Lehrenden die Lernenden gezielter fördern können. Durch die Unterstützung durch KI-Software für Lernende und Lehrende sollen Letztere gemäss der Firma Century (siehe auch Abschnitt 3.2.1.2) bis zu sechs Stunden Arbeitszeit (Gestaltung der Lehre, Administration) einsparen. Damit bleibt mehr Zeit für die individuelle Betreuung der Lernenden (Anderson 2019).

Diese Ansicht wird durch die Befragten unterstützt. In der Befragung halten 74 % der Fachpersonen die Individualisierung der Bildungsabläufe für wahrscheinlich und sehr wahrscheinlich. 77 % glauben dabei an die individuellere Förderung von Potenzialen, und 57 % halten es für wahrscheinlich bzw. sehr wahrscheinlich, dass auch das Lehrpersonal sich der individuellen Betreuung widmen kann.

Ein beispielhaftes Anwendungsfeld von KI-gestützter Personalisierung ist das Sprachenlernen. Bereits heute können Lernende über KI-unterstützte Software wie Duolingo (siehe auch Abschnitt 3.2.1.2) Sprachen lernen. Für Schulklassen gibt es eine angepasste Benutzeroberfläche für die Lehrpersonen, an der sie die Lernfortschritte der Lernenden verfolgen können. Das KI-System unterstützt die Spracherkennung und kann daher sogar Dialoge mit den Lernenden führen. Entsprechend halten es 90 % der befragten Fachpersonen für wahrscheinlich bzw. sehr wahrscheinlich, dass KI-gesteuerte Online-Lernprogramme für Sprachen umfassend in der Schule Einzug halten werden.

Ein weiteres Themenfeld im Bildungswesen ist der Einsatz von KI für **Betrugsbekämpfung**. So wird beispielsweise an der Open University in Grossbritannien für Online-Prüfungen das TeSLA-System (*An Adaptice Trust-based e-Assessment System for Learning*)[115] zur Vorbeugung von Betrug bei Prüfungen genutzt. Das Tool nutzt Nutzeridentifizierung mittels Gesichts-, Sprach- und Tastendruckerkennung sowie Anti-Plagiat-Tools. Hier sind 65 % der Befragten der Meinung, dass

[115] Open University 2019; siehe auch: https://tesla-project.eu/how-it-works/.

KI-Systeme zur Aufklärung und Vorbeugung von Prüfungsbetrug künftig einen starken oder sehr starken Einfluss haben werden.

Deutlich kritischer sind die Befragten bezüglich des Einsatzes von KI für Beurteilungen von Schüler/-innen bzw. von deren Arbeiten. Hier sind relative Mehrheiten von 35–55 % der Ansicht, dass solche Anwendungen nicht an Einfluss gewinnen werden (Abbildung 17).

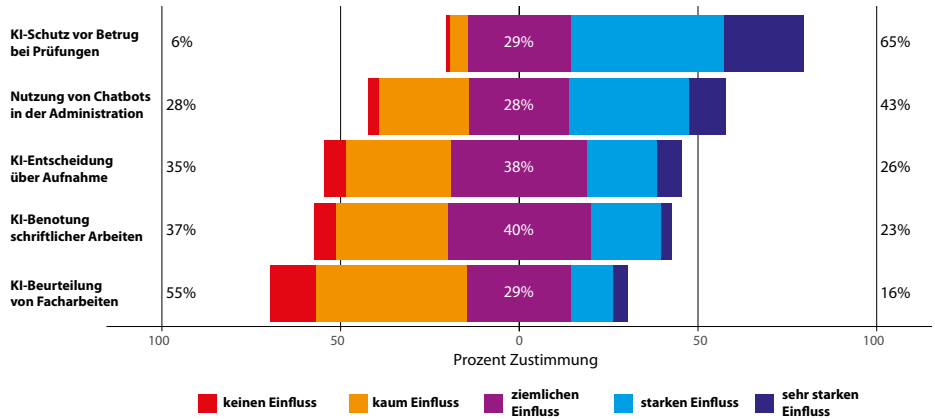

Abbildung 17: Einsatz von KI-Tools im Bildungswesen.

Derartige Einsatzformen von KI-Systemen dürfte auch den Einfluss privater Unternehmen im Bildungswesen erhöhen. In der Expertenbefragung halten es 81 % der Befragten für wahrscheinlich und sehr wahrscheinlich, dass der private Einfluss auf die Bildung durch KI-Systeme zunimmt.

4.3.1.2. Einsatz von KI-Tools in der Forschung

Bereits heute werden in **Suchmaschinen für wissenschaftliche Literatur und Daten** KI-Algorithmen genutzt (Extance 2018). Das Software-Tool Iris.ai[116] etwa liefert anhand einer 300- bis 500-Wort-Beschreibung des Problems eines Forschers oder einer Publikation eine Karte mit Tausenden von übereinstimmenden Dokumenten, die visuell nach Themen strukturiert ist. Entsprechend sehen 67 %

[116] Siehe: https://iris.ai.

der Befragten zukünftig einen starken Einfluss von KI-Anwendungen in der Literaturverwaltung und 73 % sind der Ansicht, dass diese Systeme künftig einen starken bzw. sehr starken Einfluss in der Empfehlung relevanter Literatur haben werden (Abbildung 18). Die Fachpersonen weisen aber auch darauf hin, dass es wichtig ist, sich als Forscher nicht auf ein einziges Tool zu verlassen – insbesondere wenn es um die Auswahl von Daten geht. 51 % der Befragten sehen denn auch ein grosses bis sehr grosses Risiko, wenn KI-Systeme Forschungsdaten beeinflussen, weil dadurch gewisse Daten aus dem Blickfeld verschwinden könnten.

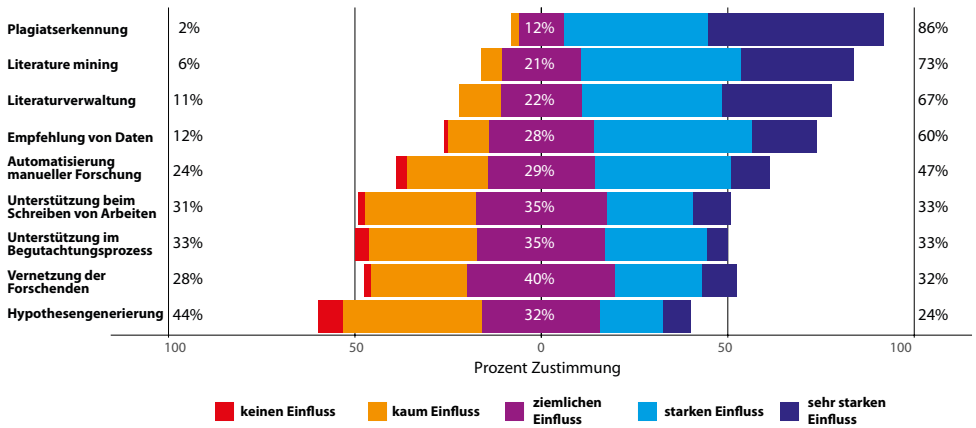

Abbildung 18: Einsatz von KI-Tools in der Forschung.

Analog wie im Bildungswesen zeigen KI-Systeme auch in der Wissenschaft bereits heute ein grosses Potenzial in der Plagiatserkennung. Gängige Tools an Hochschulen nutzen KI-Algorithmen, um eine möglichst gute Trefferquote auf Texte zu erreichen, die bereits online zu finden sind. Neu hinzu kommen Tools wie Emma[117], die zuerst mit verlässlich vom spezifischen Autor verfassten Texten «gefüttert» wurden und danach erkennen können, ob auch andere Texte von demselben Autor stammen bzw. ob aufgrund des veränderten Schreibstils der Verdacht bestehen könnte, dass das Werk abgeschrieben oder durch einen Ghostwriter verfasst wurde. 86 % der Experten sind der Meinung, dass KI-Systeme in der Plagiatserkennung einen starken bis sehr starken Einfluss haben werden (Abbildung 18).

[117] Siehe: https://emmaidentity.com.

4.3.1.3. Zu vermittelnde KI-Kompetenzen

Im Abschnitt 2.5.2 wurde bereits die Rolle der Kompetenzen im Umgang mit KI-Systemen erörtert. In der ersten Befragung wurde deshalb auch der Frage nachgegangen, wann bereits mit der Befähigung der relevanten Kompetenzen begonnen werden soll und wie die Schweiz im internationalen Vergleich gemäss der Einschätzung der Fachpersonen dasteht.

48 % der Befragten vertreten die Meinung, dass MINT-Fächer, in jeweils für die Altersstufen angepassten methodischen Zugängen, bereits ab der Grundschule vermittelt werden sollten. Knapp 7 % denken, dass dies bereits in der Vorschule geschehen soll und 25 % vertreten die Ansicht, dass die Sekundarstufe 1 hierfür eine gute Altersgruppe wäre.

43 % bestätigen das eingangs im Bildungskapitel erwähnte, gute Abschneiden der Schweiz bei der Vermittlung der MINT-Fächer und halten sie als (eher) fortschrittlich, 32 % sind bei der Frage unentschlossen und nur 25 % halten die Schweiz als eher bzw. sehr rückständig (Abbildung 19). Allerdings schätzen 58 % der Fachpersonen die Schweiz in der Vermittlung von *computational thinking* als (eher) rückständig ein. Dies sollte der Meinung von 36 % der Befragten zufolge ab der Grundschule und zufolge von 25 % der Fachpersonen ab der Sekundarstufe 1 gelehrt werden.

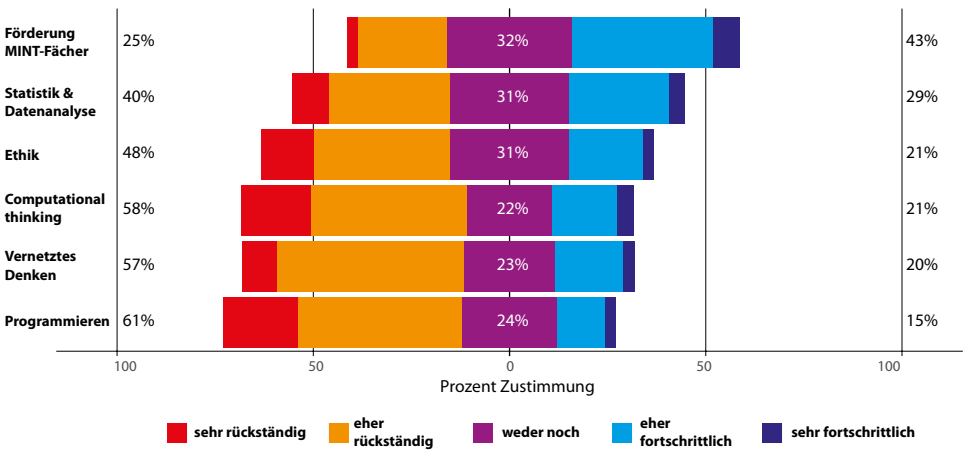

Abbildung 19: Zu vermittelnde KI-Kompetenzen.

4.3.1.4. Zusammenfassendes Fazit

Zusammenfassend kann gesagt werden, dass die Expertinnen und Experten in jenen KI-Anwendungen ein besonderes Risiko sehen, die als Beurteilung oder Entscheidung angesehen werden. Geht es um Empfehlungen oder präventive Massnahmen, werden KI-Anwendungen grosse Chancen zugesprochen.

Das Schulsystem wird sich durch KI-Anwendungen in Richtung der personalisierten, individuell geförderten Lehre entwickeln. Strukturell wird sich an den Schulen deshalb nicht viel ändern. Die Lehrpersonen haben mehr Zeit, sich um die Schüler zu kümmern, und können in der Administration Zeit einsparen. Dass die privaten Unternehmen an Einfluss zunehmen, bereitet jedoch Sorge und muss durch Massnahmen reguliert werden.

4.3.2. Zentrale Ergebnisse der zweiten Umfrage

In der zweiten Umfrage haben 56 von 111 Personen Einschätzungen zu Massnahmen im Bereich «Bildung und Forschung» abgegeben. Ziel der zweiten Umfrage war es, die Meinung zur Wirkung und Wünschbarkeit von vorgeschlagenen Massnahmen einzuholen. Hinsichtlich Massnahmen im Bereich Bildung und Forschung wurden folgende Themen abgefragt: Wie kann die Nutzung von KI-Anwendungen möglichst risikofrei und zugleich potenzialfördernd für die Lernenden genutzt werden? Wie können Weiterbildungen für Lehrende an Schulen möglichst effizient umgesetzt werden? Und wie kann der Einfluss privater Unternehmen im öffentlichen Interesse gehalten werden?

Diese Fragen führen zur Formulierung von möglichen Massnahmen im Bereich Bildung und Forschung, die zu einer Förderung der Potenziale von KI-Anwendungen und Minimierung etwaiger Risiken der Anwendung führen sollen.

4.3.2.1. Massnahmen im Bereich Bildung

Bezüglich der Thematik «Individualisierte Lehre und Speicherung personenbezogener Daten» halten 76 % der Befragten die Empfehlung, individualisierte Lernprofile zu fördern, für wirkungsvoll und 80 % für wünschenswert. Dahinter liegt das zentrale Potenzial in der Nutzung von KI-Anwendungen für die Bildung. Individualisierung ermöglicht etwa die «Reaktion auf spezifische Probleme und Herausforderungen», denen Lernende gegenüberstehen. Anwendungen für den Schul-

gebrauch (z.B. Century) stellen zugleich den Lehrenden Überblickscharts zur Verfügung, um die individuellen wie auch kollektiven Stärken und Schwächen der Klasse auszumachen. Dies ermöglicht den Lehrenden zeitliche Einsparungen und pädagogische Vorteile.

Die Umsetzung dieser Massnahme bedarf einiger Abklärungen:

- Welche Anwendung soll genutzt werden? Es gibt aktuell erst bedingt Einblicke in unabhängig erstellte Ergebnisse von Fallbeispielen der Nutzung von KI-Anwendungen. Studien werden vorwiegend von den Anbietern der KI-Anwendungen selbst veröffentlicht. Hierzu besteht noch Forschungsbedarf.
- Umgang mit den Daten: Welche Daten werden verarbeitet, wo und wie lange werden sie gespeichert und für wen zugänglich gemacht? Beispiele zeigen, dass den Schulen die Option gelassen wird zu entscheiden, welche Daten für die Auswertung der individuellen Profile genutzt werden dürfen. Einschränkungen in der Zugänglichkeit von Informationen gehen zugleich mit Einschränkungen der durch die Anwendung gemachten Aussagen einher. Hierzu muss eine Diskussion geführt werden. Die Ergebnisse aus der Befragung liefern dazu bisher kein einheitliches Bild: In der Befragung sind 65 % der Befragten der Meinung, dass die Daten nicht über die Schulzeit hinaus gespeichert werden dürfen, und 55 % meinen, dass sensitive Daten nicht erfasst werden dürfen.
- Reichweite des Einsatzes: In Belgien wird in Pilotversuchen bereits an 700 Schulklassen die Nutzung erprobt. Die grosse Anzahl der Klassen ermöglicht den breiten Test und schafft zugleich eine relevante Datenmenge zur Nutzung der KI-Anwendung.

Weitere Massnahmen betreffen die Thematik «Einfluss von Unternehmen auf Bildungsbereich». Durch die Nutzung von KI-Anwendungen an Schulen erhalten Unternehmen Zugang zu sensitiven Daten sowie auch Einfluss auf die Pädagogik und Ausbildung Minderjähriger. Zugleich besteht das Risiko, sich durch die Auswahl eines Anbieters in langfristige Abhängigkeiten zu begeben.

Dennoch sind hier 59 % der Befragten der Meinung, dass ein generelles Unterbinden von Anwendungen privater Unternehmen falsch wäre (27 % der Befragten wären dafür, 14 % sind unentschlossen). Vielmehr liegt es am Umgang mit den Daten und präzis formulierten Vereinbarungen, die zu treffen sind. So sind 71 % der Meinung, dass es eine behördliche Prüfung und Zulassung der Anwendungen braucht. Dafür müssen natürlich auch klare Kriterien festgelegt werden, und es

wäre zu vereinbaren, wie ein möglichst effizientes Prüfverfahren aussehen könnte, um einer Schulautonomie in der Nutzung nicht zu sehr entgegenzuwirken (Abbildung 20).

79 % der Befragten sind zudem der Meinung, dass der Algorithmus in der schulischen Nutzung transparent ausgeführt sein muss. Dies deutet ebenso auf den nötigen sensiblen Umgang der Nutzung von Daten im Spannungsfeld öffentlich/privater Nutzung hin und wäre auch für die behördliche Prüfung notwendig.

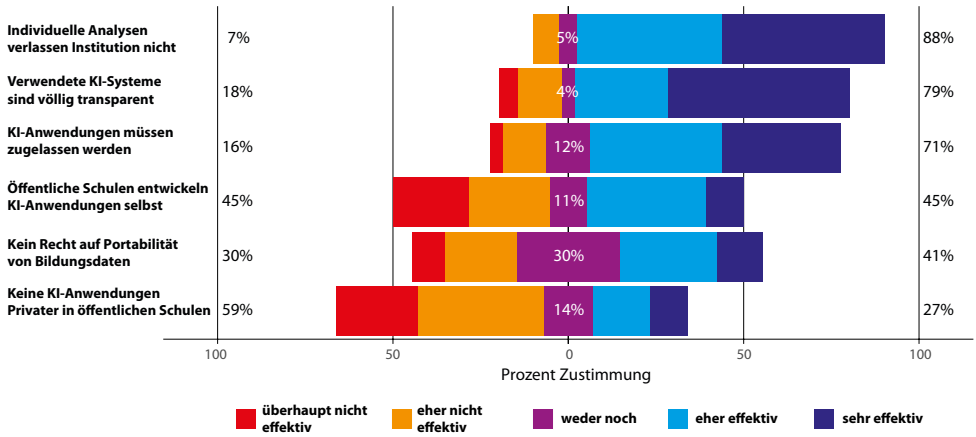

Abbildung 20: Massnahmen zur Verminderung des Einflusses von Unternehmen in der öffentlichen Bildung.

Uneinigkeit besteht in der Frage, ob Anwendungen direkt durch die Behörden entwickelt werden sollten. Je 45 % befürworten die Massnahme bzw. lehnen sie ab. Beispiele aus dem Ausland (Century, siehe Abschnitt 3.2.1.2) wie auch Expertenmeinungen im Rahmen des Workshops zeigen allerdings, dass die Einbindung der Behörden im Zuge der Entwicklung der Anwendungen von grossem Nutzen sein kann. Dies erhöht die Transparenz der Werte, die den Algorithmen zugrunde liegen, welche Daten wie genutzt werden, sowie auch die kontextspezifische Anpassung auf das Bildungssystem.

4.3.2.2. Massnahmen im Bereich Forschung

KI-Anwendungen nehmen eine wachsende Rolle in der Forschung ein. Viele innovative Leistungen werden erst durch die Nutzung von KI-Anwendungen möglich. Das Forschungsmanagement und die strategische Entwicklung der Forschung innerhalb von Institutionen muss sich dieser Entwicklung anpassen beziehungsweise kann durch gezielte Massnahmen Innovationen fördern.

In der Umfrage zeigt sich hier ein einhelliges Bild. Alle fünf vorgeschlagenen Massnahmen zur Förderung von Chancen bzw. Verminderung von Risiken von KI in der Forschung wurden mit 80–91 % Zustimmung bewertet. So sind 80 % der Befragten der Meinung, dass Forschungseinrichtungen entsprechende Anlaufstellen einrichten sollten. Diese bieten Weiterbildungen zur Nutzung von KI in der Forschung an, dienen aber auch als Knotenpunkt für Forscher unterschiedlicher Fächer für den interdisziplinären Austausch. Beispiel dafür ist etwa die Digital-Society-Initiative der Universität Zürich, die auch zu Fragestellungen zu künstlicher Intelligenz Angebote für die Mitarbeitenden in Forschung, Lehre und Betrieb anbietet.

In der Anwendung stellen sich neue Fragen des Umgangs und Eigentums mit den Daten sowie mit dem aus der Nutzung der KI-Anwendung resultierenden Wissen. Die Experten der Umfrage vertreten auch hier eine dezidierte Meinung: Ergebnisse müssen transparent und reproduzierbar sein (91 %), es braucht klare Transparenz-Guidelines, die Daten müssen im Eigentum der Forschungseinheit bleiben (83 %) oder im Open-Source/Open-Access-Format verfügbar gemacht werden (80 %).

4.3.2.3. Zusammenfassende Beurteilung

In den Umfragen wurde in Bildung und Forschung das Potenzial, das KI-Anwendungen für kontextspezifische Bedürfnisse mitbringt, sehr deutlich. KI-Anwendungen sollten gemäss den Experten für Bildung wie auch für Forschung gefördert werden. Die Potenziale liegen in der Effizienzsteigerung für Lehrende, der Personalisierung von Lehrinhalten und Erhöhung der Inklusion in der Bildung. In der Forschung werden die Innovationspotenziale deutlich, die KI-Anwendungen für spezifische Anwendungen bringen können, sowie auch die Effizienzsteigerungen, die gerade in Zeiten der stark zunehmenden globalen Menge an Publikationen und Daten vonnöten ist.

Bedenken gibt es sowohl in Bildung als auch Forschung hinsichtlich des Umgangs mit Daten, solange diese zum öffentlichen Bereich gehören und damit im öffentlichen Interesse liegen und öffentliches Gut sind. Dies darf durch KI-Anwendungen und den Einfluss privater Unternehmen nicht eingeschränkt werden. Es braucht daher in Bildung wie auch in Forschung die enge Zusammenarbeit zwischen Behörden und Unternehmen, um Datenschutz und Privatsphäre in Einklang mit den Potenzialen der Nutzung von KI-Anwendungen zu bringen.

Im Bereich **Bildung** kamen die Fachpersonen zu folgenden Schlüssen: Sie befürworten erstens die Nutzung von KI-unterstützten individuellen Lernprogrammen. Die Fachpersonen empfahlen, die Rahmenbedingungen des Fallbeispiels aus Belgien (siehe Abschnitt 4.3.2.1) genau zu betrachten und auch in der Schweiz eine Gruppe von Pilotschulen einzurichten, anhand derer die Vor- und Nachteile einer breiten Anwendung abgeschätzt werden können. In der Auswahl der Unternehmenspartner ist es wesentlich, darauf zu achten, dass eine gemeinsame Entwicklung der Lehrinhalte möglich ist und auch auf die Anforderungen des Bundes bezüglich des Schutzes der Privatsphäre eingegangen wird. Zweitens sollen die Schulbehörden über die Nutzung der Daten und den Gerichtsstand entscheiden können. Zudem muss über die Dauer der Speicherung und den Zugang zu den Daten bestimmt werden. Können etwa die Lernenden nach Beendigung ihrer schulischen Laufbahn ihre personenbezogenen Daten in einem portablen (Datei-) Format als ihr Eigentum mitnehmen? Dürfen andere Schulen beim Aufnahmeverfahren verlangen, die personenbezogenen Lerndaten der Lernenden einzusehen? Dies sind Fragen, die bewertet und entschieden werden müssen. Drittens sollen Behörden mit Unternehmen in der Entwicklung von massgeschneiderten KI-Anwendungen zusammenarbeiten. Dies passiert schon heute, wenn Unternehmen modular zusammenstellbare Angebote schaffen, die von den Bildungsbehörden für ihre spezifischen Bedürfnisse angepasst werden können.

Im Bereich **Forschung** kamen die Fachpersonen zu folgenden Schlüssen: Um das volle Potenzial der Anwendungsmöglichkeiten von KI-Systemen in der Forschung zu nutzen, wird Forschungseinrichtungen empfohlen, KI-Informations- und Innovationsstellen einzurichten. Die Stellen bündeln Weiterbildungsmöglichkeiten, schaffen interdisziplinären Austausch und informieren Forschende unterschiedlichen Fachhintergrunds über Potenziale der KI-Anwendungen in der Forschung. Die Anlaufstellen entwickeln aber auch Richtlinien für die Angehörigen der Institution, um einen achtsamen Umgang mit KI-Systemen sicherzustellen.

Es bedarf des Weiteren eines sorgfältigen Umgangs mit den Forschungsdaten. Schon die Eingabe unveröffentlichter Forschungstexte in die kostenlose Version des Übersetzungsprogramms DeepL kann ein datenschutzrechtliches Problem darstellen, da die Systeme die Texte zur Verbesserung des KI-Algorithmus speichern.[118] In der kostenpflichtigen Pro-Version von DeepL werden die Daten nur für die Dauer der Übersetzung gespeichert und somit nicht für das Training genutzt. Doch auch hier weist DeepL selbst darauf hin, dass keine personenbezogenen Daten eingegeben werden sollten.

4.3.3. Ergebnisse des Workshops zu Bildung und Forschung

Ziel der ersten Runde des Workshops war es, auf Basis der Ergebnisse der Onlinebefragung mögliche Massnahmen zu sammeln, die die Chancen der KI-Anwendungen in Bildung und Forschung für die Schweiz stärken und zugleich Risiken minimieren. Teilnehmende der ersten und zweiten Runde des Workshops waren Stakeholder aus Bereichen der Zivilgesellschaft, Wirtschaft, Verwaltung und Forschung; es mangelte aber an Teilnehmenden aus Bildungsinstitutionen oder Bildungsverwaltung. Dies hatte den Effekt, dass der Schwerpunkt der besprochenen Themen auf die öffentliche und zivilgesellschaftliche Bildung zu liegen kam und weniger auf die schulische Bildung. Für die Studie kann dies als wertvolle Ergänzung angesehen werden, da der Schwerpunkt in den beiden Befragungsrunden bisher auf schulischer Bildung gelegt wurde. Die Thematik «Berufliche Weiterbildung» wurde im Themenbereich Arbeitswelt besprochen; sie wird aber in diesem Abschnitt vorgestellt.

Ein erster, durch die Umfrage inspirierter Themenbereich betraf die zahlreichen Möglichkeiten der **Nutzung von KI im Bildungssystem**. Die Teilnehmenden des Workshops bestätigten das grosse positive Potenzial der Personalisierung der Bildung durch Auswertung des Lernverhaltens mittels KI sowohl zugunsten der Lernenden als auch der Lehrenden. Auch könnten KI-Systeme zur Inklusion von körperlich, regional oder finanziell benachteiligten Lernenden beitragen. KI-Anwendungen zur Unterbindung von Betrug (z.B. in *open education systems*) werden ebenfalls grundsätzlich positiv beurteilt. Als Hauptrisiko wird der Umgang mit den Daten gesehen, welche solche Anwendungen benötigen bzw. generieren. Es

[118] So heisst es in Art. 4 der Datenschutzerklärung von DeepL (Stand Nov. 2019): «Wir speichern Ihre Texte und die Übersetzung für einen begrenzten Zeitraum, um unseren Übersetzungsalgorithmus zu trainieren und zu verbessern.»

handelt sich dabei auch um Daten, die bisher entweder nicht erfasst wurden oder innerhalb der Bildungseinrichtung gespeichert wurden. Durch KI-Anwendungen bekommen private Unternehmen Zugang zu personenbezogenen Daten sowie inhaltlicher und methodischer Gestaltung der Lehre. Der Umgang mit diesen Daten muss persönliche Rechte schützen, und die Gestaltung der Lehre muss im öffentlichen Interesse bleiben. Sehr kritisch betrachtet werden dabei sensorunterstützte KI-Systeme, die durch die Auswertung von Bewegungs-, Ton-, Temperatur- oder Gesichtserkennungsdaten Analysen zur Anwesenheit, Aufmerksamkeit, Gesundheit und Interaktion im Klassenzimmer oder an der Bildungseinrichtung ermöglichen. Es besteht ein Konsens, dass sich solche Analysen in der Schweiz kaum mit den Werten und Rechten des Persönlichkeitsschutzes vereinbaren liessen.

Ein zweiter Themenkomplex betraf die Förderung von **KI-Kompetenzen**. Unter diesem Überbegriff wird eine Vielzahl an Kompetenzen betrachtet, die zum Umgang mit einer von KI-Anwendungen geprägten Welt befähigen sollen. Darunter fällt *digital literacy* generell, ebenso wie Systemkompetenz, interdisziplinäres Denken, als auch das Verständnis für konkrete Anwendungskontexte von KI und den dafür notwendigen Daten. Auch die Workshopteilnehmenden bekräftigten, dass KI-Kompetenzen bereits ab einem frühen Alter (Vorschulalter) vermittelt werden sollten. Deshalb sollte hier der Fokus auf die Schaffung der notwendigen Voraussetzungen gelegt werden. Diskutiert wurden hierzu insbesondere der Aufbau einer nationalen Bildungsplattform zur Sammlung von Angeboten und die Schaffung entsprechender Lehrmittel. Plädiert wurde auch für ein vermehrtes Experimentieren mit Methoden an Pilotschulen, um solche Kompetenzen zu vermitteln.

Ein im Themenfeld «Arbeitswelt» besprochener Schwerpunkt betraf die kontinuierliche, lebenslange **Weiterbildung**. Hierzu bestand ein Konsens darüber, dass die Verantwortlichkeit geteilt ist. Arbeitnehmende sind dafür ebenso verantwortlich wie die Unternehmen und auch der Staat, in dessen Interesse es ja ebenfalls ist, die Arbeitnehmer/-innen fit für den Arbeitsmarkt zu machen bzw. zu halten. Derzeit scheint sich der Staat nur für eine grundsätzliche Ausbildung bis zum Alter von 16 Jahren verantwortlich zu fühlen. Wichtig ist auch, die Arbeitgeber weiterhin in die Pflicht zu nehmen, weil ein Trend erkennbar sei, wonach Arbeitnehmer vor allem bei externer Weiterbildung keine finanzielle Unterstützung erhalten. Eng mit diesem Themenkomplex verbunden ist auch die Forderung, dass die Entwicklungen am Arbeitsmarkt genau beobachtet werden und an das Bildungssystem zurückgemeldet werden sollten, damit beispielsweise Studierende abschätzen können, wie gut die Jobaussichten in Zukunft sind. Es sollten gezielt von KI betroffene

Berufsfelder auch mit entsprechenden Bildungsangeboten in der Aus- und Weiterbildung ausgestattet werden, um den Arbeitssuchenden und der Wirtschaft die besten Entwicklungsoptionen zu ermöglichen.

Im Bereich **Forschung und Innovation** wurde schliesslich hervorgehoben, dass KI-Anwendungen neue Erkenntnismöglichkeiten eröffnen. Somit geht es nicht nur um einen Effizienzgewinn, sondern um innovative Forschungsverfahren. Gesteigerte Effizienz wird durch KI-Anwendungen im *data* und *literature mining* ermöglicht. Generell wird durch KI der Bedarf an interdisziplinärer Zusammenarbeit erhöht; eine solche Zusammenarbeit soll demnach durch entsprechende Initiativen gefördert werden. Nicht vergessen werden sollte dabei, dass durch KI-Anwendungen auch eine Reihe von Fragen rund um die Nutzung der Daten, Eigentumsrechte sowie auch den Zugang zu Daten aufgeworfen werden.

Insgesamt ergaben sich aus den Diskussionen der zweiten Runde folgende Prioritäten für den Themenbereich Bildung und Forschung:

1. Schweizweit verstärkter Aufbau von KI-Kompetenzen. Dies beinhaltet zum einen die Entwicklung neuer Lehrmittel und eine Bereitschaft für entsprechende pädagogische Experimente; zum anderen soll eine nationale Bildungsplattform Angebote zum Aufbau von Kompetenzen sichtbar machen.
2. Strukturelle Massnahmen wie z.B. Finanzierungsregeln zur Förderung der beruflichen Weiterbildung, um die Befähigung zum praktischen Umgang mit KI-Systemen in der Arbeitswelt zu sichern.
3. Investitionen in interdisziplinäre Forschung mit Schwerpunkt auf KI-Anwendungen.

Einigkeit besteht auch darin, dass Aspekte des Datenschutzes besondere Beachtung verdienen. Hierbei handelt es sich aber um ein bereichsübergreifendes Phänomen, das auf die spezifischen Bedürfnisse der einzelnen Bereiche gesondert Rücksicht nehmen muss. Die durch Personalisierung im Bildungsbereich generierten Daten sind dabei besonders schutzwürdig.

4.4. Beurteilungen zum Themenfeld Konsum[119]

4.4.1. Zentrale Ergebnisse der ersten Umfrage

72 von 307 Befragten haben den Themenbereich «Konsum» in der ersten Umfrage ausgefüllt. Dabei sind die zentralen Ausprägungen der demografischen Variablen wie Alter, Geschlecht, Ausbildung, Arbeitsverhältnis mit KI und Herkunft ausreichend vertreten.

4.4.1.1. Einsatzformen von KI im Bereich Konsum

Die befragten Fachpersonen schätzen alle beschriebenen Einsatzgebiete von KI im Bereich Konsum als mindestens «sehr wichtig» ein und stimmen damit mit der Einschätzung aus der Literatur überein. Die automatisierte Generierung von Persönlichkeitsprofilen (*consumer profiling*), Erstellung von individuellen (Kauf-) Empfehlungen (*Empfehlungssysteme*) sowie die Personalisierung von Interaktions- und Kommunikationsstrategien (*predictive personalization*) werden dabei als die Top-3-Einsatzgebiete der nächsten fünf bis zehn Jahre eingestuft (86–92 % «sehr wichtig» oder «äusserst wichtig»).

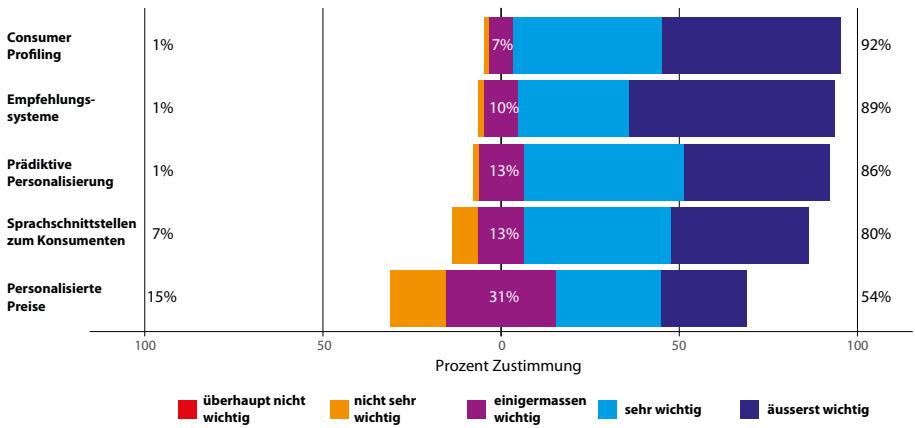

Abbildung 21: Beurteilung von Einsatzformen von KI im Bereich Konsum.

[119] Dieser Abschnitt beruht auf Arbeiten von Anne Scherer, Assistenzprofessorin am Lehrstuhl für Marketing und Marktforschung an der Universität Zürich, mit Unterstützung von Pascal Sutter.

4.4.1.2. Herausforderungen für Konsumenten und Unternehmen

Gemäss den Befragten stehen **Konsumentinnen und Konsumenten** vor wesentlichen Herausforderungen. Alle ausgewählten Themen wurden von 68–82 % der Teilnehmenden als «grosse Herausforderung» eingestuft. Die Funktionsweise des Algorithmus nachvollziehbar zu machen (*operative Transparenz*), wird dabei mit 82 % als wichtigste Herausforderung eingestuft. Aber auch das Verständnis der Kunden über den Wert der persönlichen Daten (*data-dollar awareness*, 73 %), die zunehmende Schwierigkeit für Kunden, die KI-Plattform zu wechseln (*Plattformklebrigkeit*, 72 %), oder den Einsatz von Algorithmen und KI zu erkennen (*Erkennbarkeit*, 68 %) sowie die *Entmenschlichung* von Kundenkontaktpunkten (68 %) scheinen eine kritische Rolle zu spielen.

Weiter erwarten 81 % der Befragten eine steigende Relevanz von *Datenfairness* und 68 % den *Wunsch nach Nachvollziehbarkeit* der Algorithmen. Ein hinreichendes *Datenbewusstsein* der Konsumentinnen und Konsumenten (37 %) und eine genügende *KI-Erkennbarkeit* (41 %) werden dagegen als eher unwahrscheinlich eingestuft. Die Kommentare verweisen weiter auf das Problemfeld der uninformierten Zustimmung, wonach Betroffene immer häufiger uninformiert persönliche Daten für Unternehmen freigeben, ohne genau zu wissen, zu welchem Zweck und in welcher Weise die Daten verwendet werden. Gerade in Bezug auf KI fehlt Konsumentinnen und Konsumenten hier zudem noch das Verständnis, welche Aussagen moderne Algorithmen mithilfe der Daten treffen können (*algorithm awareness*). Die Relevanz der Datenfairness (75 % der Befragten sehen höhere Relevanz für die Schweiz) und der Nachvollziehbarkeit (65 %) wird dabei für die Schweiz besonders hoch eingestuft. Die Aktualität der Herausforderungen erfordert gemäss den Befragten neben eher trägen Policy-Reformen und Bildungsinvestitionen auch agile Kontrollmechanismen, die zeitnah umgesetzt werden sollten.

Für **Unternehmen** scheint der Trade-off zwischen Personalisierung und Datenschutz die grösste Herausforderung darzustellen (88 %), eng verknüpft mit dem umkämpften Kundenvertrauen (72 %) in KI-Systeme.

Ein zweiter Schwerpunkt liegt bei der Entstehung von Oligopolen und der Konzentration von Daten und damit Marktmacht. 94 % der Befragten stufen dies als ein wahrscheinliches Szenario ein, 77 % erwarten in der Folge steigende Markteintrittsbarrieren und 80 % eine Zunahme von *metabots* im Marketing. Ent-

sprechend wird auch die *Dateninteroperabilität* von 70 % der Befragten als wesentliche Herausforderung gesehen. Die beschriebenen Szenarien werden folglich auch von über 70 % der Befragten als realistisch in fünf oder weniger Jahren eingeschätzt.

4.4.1.3. Zusammenfassendes Fazit

Zusammenfassend stellen sich im Bereich Konsum zentrale Herausforderungen auf Konsumenten- wie Anbieterseite. Im Zentrum steht dabei auf beiden Seiten ein sinnvoller, verständlicher und nachvollziehbarer Umgang mit den persönlichen Daten der Kundinnen und Kunden. Es wird erwartet, dass Konsumentinnen und Konsumenten eine Personalisierung durch KI wünschen, aber auch ihre Privatsphäre gewahrt sehen wollen. Erfolgreiche Unternehmen müssen also bei einem KI-Einsatz Datengewinnung und -verwertung mit einem transparenten und sicheren Datenumgang in eine Balance bringen.

4.4.2. Zentrale Ergebnisse der zweiten Umfrage

Durch die Literaturanalyse, die Tiefeninterviews und die Expertenumfragen konnten die wesentlichen Herausforderungen im Themenfeld Konsum identifiziert werden. Ziel der zweiten Umfragerunde war es, eine breite Auswahl an Handlungsempfehlungen zu den Schwerpunkten Kundenvertrauen, Personalisierung versus Privatsphäre und Datenkonzentration/Oligopolbildung zu diskutieren. 53 von 111 Fachpersonen haben diese Runde vollständig ausgefüllt, wobei erneut alle relevanten demografischen Ausprägungen vertreten waren. Sämtliche Handlungsempfehlungen wurden sowohl nach ihrer Effektivität als auch ihrer Wünschbarkeit abgefragt. Die Richtung der Ergebnisse deckt sich vollständig, wobei die Wünschbarkeit durchgehend höhere Zustimmung der Fachpersonen erhielt. Für eine bessere Übersicht wird in den folgenden Abschnitten nur auf die Ergebnisse zur Effektivität näher eingegangen.

Im Rahmen der Bevölkerungsumfrage von SATW und der Stiftung Risiko-Dialog wurden die Massnahmen ebenfalls integriert. Dabei zeigte sich, dass sich die Einschätzungen der Fachpersonen und der allgemeinen Bevölkerung im Wesentlichen decken; allerdings werden Oligopole von Letzterer deutlich kritischer bewertet. Diese Resultate werden deshalb separat aufgeführt.

4.4.2.1. Förderung des Kundenvertrauens

Offene Daten, Kontrolle der Konsumentinnen und Konsumenten, Offenlegung des KI-Zwecks und der Ergebnisherleitung sowie ein Recht auf Datenlöschung werden von den Teilnehmenden allesamt als effektiv (83–94 %) und wünschenswert (87–94 %) erachtet. Sämtliche Ansätze können Anbietern entsprechend nahegelegt werden, wobei die Offenlegung des Zwecks die aktuell gebräuchlichere Offenlegung der Herleitung passend ergänzen würde. Weiter wurden in den offenen Antworten eine unabhängige Prüfstelle und *privacy-by-design* gefordert. Beide Ansätze werden in den folgenden Abschnitten aufgegriffen.

4.4.2.2. Sicherung der Privatsphäre

Ein unabhängiges Aufsichtsorgan für Qualitätstests von Datensätzen und Algorithmen wird gefordert (77 % bzw. 72 %). Fachleute aus Wissenschaft, Journalismus und Regulierungsstellen sollen zudem temporäre Einsichten in Algorithmen und Daten von KI-Anwendungen erhalten (74 %). Auch die Umsetzung der EU-Datenschutz-Grundverordnung DSGVO wird gefordert (60 %), dieser Ansatz ist für rund 30 % der Teilnehmenden aber nicht effektiv genug; die Idee einer «Herkunftsangabe» von KI-Systemen stösst eher auf Ablehnung (Abbildung 22).

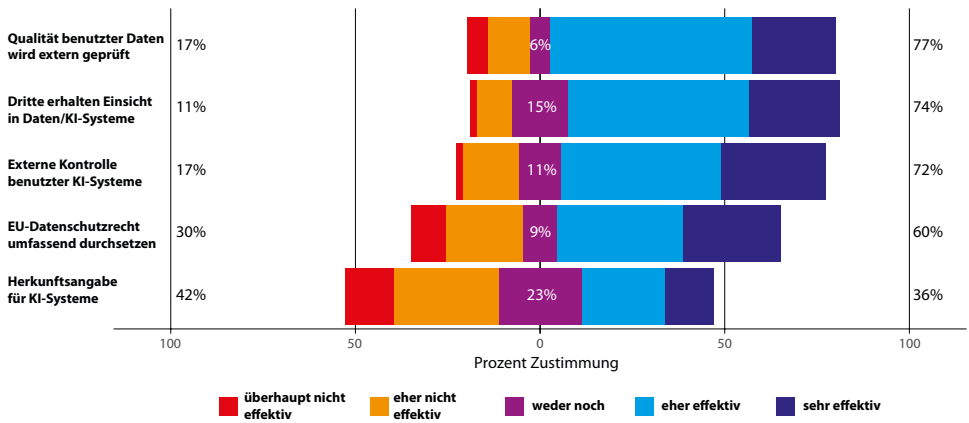

Abbildung 22: Beurteilung von Massnahmen zur Sicherung der Privatsphäre.

Bezüglich der Einstellungen der Fachpersonen ergab sich hier ein signifikanter Unterschied: «KI-Skeptiker» befürworten insbesondere einen TÜV und die Einsichten durch Dritte signifikant stärker als «KI-Enthusiasten».

4.4.2.3. Verhinderung von Konzentration der Verhandlungsmacht

Konsumentinnen und Konsumenten sollen das Recht haben, die sie betreffenden personenbezogenen Daten in einem strukturierten, gängigen und maschinenlesbaren Format zu erhalten und sie allenfalls einem anderen Anbieter übertragen zu dürfen. Dies wird von 92 % der befragten Fachpersonen gewünscht und von 77 % als effektiv eingestuft. Weiter wird ein Datenzentrum gefordert, von welchem Konsumentinnen und Konsumenten ihre personenbezogenen Daten sichern und Zugriffe kontrollieren können. Hier wird ein dezentrales System (68 %), beispielsweise mittels *distributed-ledger*-Technologie, gegenüber einem zentralisierten (58 %) Ansatz bevorzugt. Der Ansatz, die Oligopole zu akzeptieren und aktiv mitzugestalten, wird von den Fachpersonen eher kritisch gesehen (Abbildung 23).

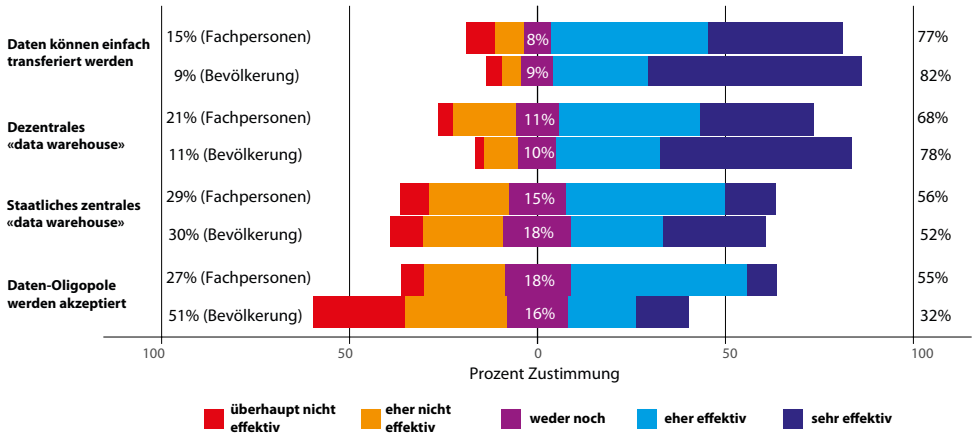

Abbildung 23: Beurteilung von Massnahmen zur Verhinderung von Oligopolen. Abgebildet werden die Ergebnisse der Umfrage unter den Fachpersonen und der allgemeinen Bevölkerung.

Im Rahmen der Bevölkerungsumfrage von SATW und der Stiftung Risiko-Dialog wurden die Massnahmen zu Oligopolen ebenfalls integriert; die Ergebnisse finden sich ebenfalls in Abbildung 23. Dabei zeigte sich, dass sich die Perspektive der allgemeinen Bevölkerung mehrheitlich mit der Einschätzung der Experten deckt, aber akzentuierter ist. Während eine erhöhte Übertragbarkeit der Daten und ein dezentrales Datencenter von über 75 % der Befragten gewünscht wird, werden oligopolartige Strukturen nur von rund 32 % der Befragten gutgeheissen.

4.4.2.4. Verantwortungsfrage

Die Fachpersonen wurden abschliessend auch zur Verantwortlichkeit bei der Überwindung der Herausforderungen befragt. Im Konsumbereich wird hier die Verantwortung wie in sonst keinem anderen Handlungsfeld beim Konsumenten selbst gesehen (für die allgemeine Übersicht siehe Abschnitt 4.7.2.1). Dies überrascht, da die Konsumentinnen und Konsumenten in einem deutlich asymmetrischeren Mächteverhältnis zu den Anbietern stehen als bei vertraglichen Strukturen wie beispielsweise in der Bildung, der Verwaltung oder dem Arbeitsverhältnis. Es ist zumindest infrage zu stellen, in welchen Formaten Konsumentinnen und Konsumenten über die Entwicklungsprozesse von KI mitentscheiden dürfen.

Weiter hat sich gezeigt, dass die Herausforderungen nirgends so zeitnah gesehen werden wie beim Konsum. Lediglich 17 % der Befragten geben an, dass die beschriebenen Veränderungen im kommenden Jahrzehnt noch nicht gravierend sein werden. Die restlichen Anwendungsgebiete lagen zwischen 21 % und 38 %.

4.4.2.5. Zusammenfassende Beurteilung der Expertenmeinungen

In den Umfragen waren sich die Experten einig, dass die Personalisierung des Angebots eine der wichtigsten Anwendungsgebiete von KI im Konsumbereich ist. Jedoch wurde auch betont, wie wichtig für Konsumentinnen und Konsumenten – gerade in der Schweiz – der Schutz der Daten und der Privatsphäre ist. Personalisierung auf der einen Seite und die Wahrung der Privatsphäre auf der anderen gut auszubalancieren, wird daher eine der wichtigen Herausforderungen für Unternehmen in der Zukunft sein. Ein dritter zentraler Punkt der Umfrage betrifft die Gefahr einer Konzentration des Angebots auf einige wenige Anbieter (Datenoligopole), insbesondere für Anwendungen im Konsum. Dies könnte nicht nur den Wettbewerb auf Unternehmensseite behindern, sondern auch das Angebot und

die Wechselmöglichkeiten auf der Seite der Konsumentinnen und Konsumenten einschränken. Entsprechend fokussieren die Empfehlungen im Workshop auf die Problemfelder Kontrolle, Datenschutz und Wettbewerb.

4.4.3. Ergebnisse des Workshops zum Themenbereich Konsum

Eine gute Balance aus Personalisierung auf der einen Seite und der Wahrung der Privatsphäre auf der anderen Seite zu finden, wurde in den Expertenumfragen als eine der zentralen Herausforderungen für Unternehmen in der Zukunft eingeschätzt. Das Fehlen einer solchen Balance könnte nicht nur zu einer Einschränkung des Wettbewerbs auf Unternehmensseite führen, sondern auch zur Einschränkung des Angebots und der Verhandlungsmacht auf der Seite der Verbraucher. Die Diskussion im Workshop fokussierte daher auf die Probleme Kontrolle, Datenschutz und Wettbewerb.

Die Umfrageteilnehmenden sehen einen wesentlichen Anteil der Verantwortung für die Risiken der KI-Nutzung bei Konsumenten, obwohl diese im Vergleich zu Abnehmern anderer Einsatzgebiete eine wesentlich geringere Verhandlungsmacht besitzen. Entsprechend wäre ein möglicher Ansatz, Konsumenten mithilfe einfacher Massnahmen zum **Selbstmanagement der Privatsphäre** zu befähigen. Im Zentrum stand dabei die Idee eines «Daten-Cockpits», das es Konsumentinnen und Konsumenten für den jeweiligen Service erlauben soll, festzulegen, 1) ob generell personalisiert werden soll (d.h. der Algorithmus muss aktiviert oder kann deaktiviert werden), 2) welche persönliche Daten nicht in den Algorithmus eingehen dürfen, 3) welche Daten nicht mit Drittanbietern geteilt werden oder einfliessen dürfen und 4) welche Daten nicht über verschiedene Services eines Anbieters hinweg geteilt werden sollen («right not to connect data»). Zur Vereinfachung könnten Datenkategorien erstellt und per Kategorie eingestellt werden (z.B. Gesundheitsdaten, Profildaten, Transaktionsdaten). Eng verbunden mit der Idee eines Daten-Cockpits ist die Forderung, KI-gestützte Angebote im Internet hätten als solche erkennbar zu sein – sei es als Interaktionspartner (ein Chatbot soll sich beispielsweise als solcher zu erkennen geben) oder als Personalisierungsalgorithmus. Als intuitive Methode wurde eine Personalisierungsampel diskutiert, welche Verbrauchern signalisiert, in welchem Ausmass ein Service personalisiert wird.

Die Idee eines Daten-Cockpits wurde in der zweiten Runde optimiert. Die Fachpersonen kritisierten, dass die Sicherung und Verwaltung der Datenmengen die Konsumenten wohl überfordere, gerade wenn die Einstellungen für jeden Service neu eingerichtet werden müssen. Da ein zentrales Datenlager als kein realistisches Szenario gesehen wird, wurde als alternativer Ansatz eine **zentrale Rechtefreigabe** diskutiert. Die Idee ist, dass Kundinnen und Kunden auf dieser Plattform zentral ihre Einstellungen zum Datenschutz und Privatsphäre speichern können. Unternehmen, die persönliche Daten des Verbrauchers nutzen wollen, müssen bei dieser Plattform die entsprechenden Einstellungen abfragen. Nur wenn eine Einwilligung der jeweiligen Person vorliegt, dürfen Daten von Unternehmen zum bewilligten Zweck erhoben und verwendet werden. Ein solches *data rights repository* sollte mittels einer dezentralen Technologie gesichert von einer staatlichen Stelle kontrolliert werden; die Rechteeinstellungen könnten dann z.B. Teil der e-ID werden.

Die Teilnehmenden sehen wie die Expertinnen und Experten der Umfrage die mögliche Bildung von Datenoligopolen als weitere zentrale Herausforderung an. Die Fachpersonen sind sich darin einig, dass eine **Portabilität der Daten**, wie in der Datenschutz-Grundverordnung der EU gefordert wird, in Zukunft unumgänglich sein wird. Es muss hierbei beachtet werden, dass der Begriff «Daten» im Fall von neuen KI-Technologien eine neue Bedeutung erhält. Die Rechte der Nutzer/-innen sollten möglichst nicht nur für aktiv eingetragene Daten (*provided data*), sondern auch für Daten gelten, welche durch KI-generiert wurden (*inferred data*).

Insgesamt ergaben sich aus den Diskussionen der zweiten Runde folgende Prioritäten für das Thema Konsum:

1. Datenhoheit für die Konsumentinnen und Konsumenten: Förderung von Transparenz und Kontrolle, beispielsweise durch ein dezentralisiertes *data-rights-repository* mit der Pflicht von Dienstleistern, bei der Nutzung persönlicher Daten diese jeweils vorher abfragen zu müssen (*opt-in*).
2. Ermächtigung der Konsumentinnen und Konsumenten durch einen möglichst internationalen Portabiliätsstandard von Daten, welcher das Profiling (d.h. auch *inferred data*) möglichst vollständig umfasst.

4.5. Beurteilungen zum Themenfeld Medien[120]

Im Themenbereich Medien wurden aus methodischen Gründen auch in der zweiten Umfrage noch einige Fragen zur Einschätzung der Faktenlage gestellt. Die entsprechenden Daten finden sich im Abschnitt 4.5.1.1.

4.5.1. Zentrale Ergebnisse der ersten Umfrage

67 von 307 Befragten haben den Fragebogen zum Thema «Medien» in der ersten Umfrage ausgefüllt.

4.5.1.1. Einschätzung der Problemlage

Anknüpfend an den Literaturüberblick, lohnt sich zunächst ein Blick auf die persönlichen Einschätzungen der befragten Fachpersonen zu einigen konkreten Sachverhalten und Entwicklungen hinsichtlich der Nutzung von (sozialen) Medien und des Einflusses, den KI darauf potenziell hat. Diese Thematik wurde in der zweiten Runde der Umfrage untersucht, bei der Daten von 55 Personen vorlagen. Grund dafür ist, dass die Wahrnehmungen und Befunde zwischen öffentlichem und wissenschaftlichem Diskurs bei Digitalisierungsthemen nicht selten weit auseinanderliegen. Vor diesem Hintergrund galt es zunächst, festzustellen, inwiefern die befragten Fachpersonen vom aktuellen Forschungsstand abweichende Problemwahrnehmungen äussern.

Anknüpfend an die Diskussion um Filterblasen und Echokammern, wurden die Teilnehmenden der zweiten Umfrage im Bereich Medien daher eingangs befragt, für wie divers sie folgende Inhalte auf sozialen Onlinenetzwerken einschätzen: Artikel, denen Nutzer/-innen in sozialen Onlinenetzwerken Aufmerksamkeit schenken; politische Meinungen, denen Nutzer/-innen in sozialen Onlinenetzwerken durch algorithmisch personalisierte Feeds ausgesetzt werden; Artikel, die Online-Nachrichtenportale den Nutzerinnen und Nutzern vorschlagen; und Artikel, die Nutzer/-innen online selber auswählen.

[120] Dieser Abschnitt beruht auf Arbeiten von Tarik Abou-Chadi und Hauke Licht vom Institut für Politikwissenschaften der Universität Zürich.

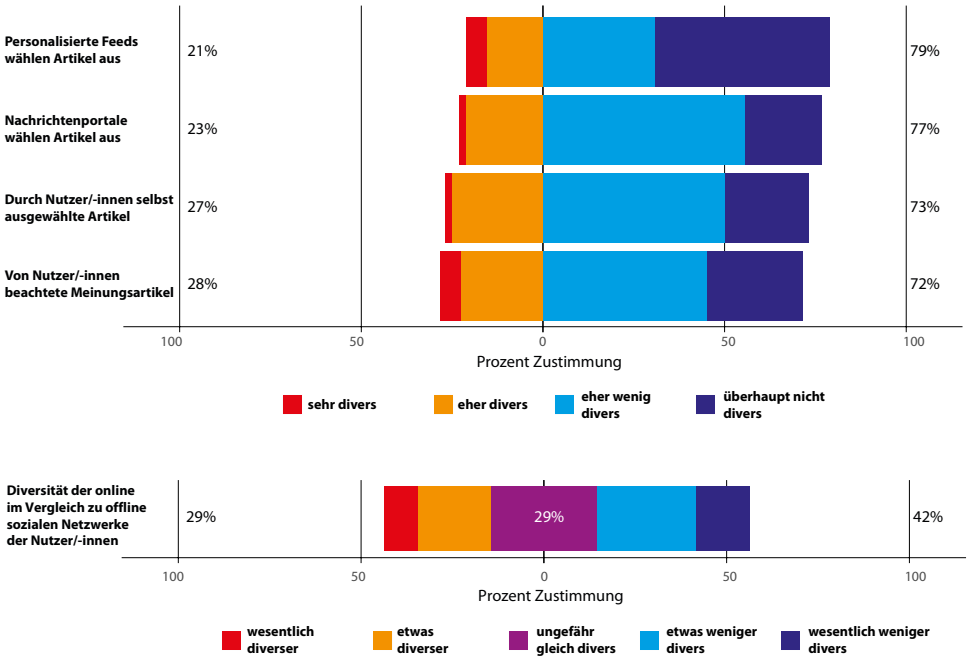

Abbildung 24: Einschätzung der Diversität von Nachrichtenquellen.

Abbildung 24 (oben) zeigt das Ergebnis der Befragung. Obwohl die Einschätzungen durchaus auseinandergehen, zeigt sich, dass die überwiegende Mehrheit dazu neigt, die Inhalte für nicht divers zu halten: Je nach Inhalt sind zwischen 72–79 % der Befragten der Ansicht, dass die Diversität gering ist. Es findet sich dabei kein Unterschied zwischen «KI-Enthusiasten» und «KI-Skeptikern» – nur im Fall der Nachrichtenauswahl durch Portale gehen Erstere signifikant stärker davon aus, dass Portale die Diversität erhöhen können.

Vergleicht man die Einschätzung der Diversität von sozialen Netzwerken der Nutzerinnen und Nutzer auf Onlineplattformen im Vergleich zu ihren persönlichen Offlinenetzwerken (z.B. Freundes- und Bekanntenkreisen), zeigt sich ein vergleichbares Bild (Abbildung 24, unten). Während 29 % der Befragten keine Unterschiede in der Diversität von sozialen Online- und Offlinenetzwerken vermuten, gaben rund 43 % an, dass Onlinenetzwerke weniger divers sind. Dem stehen rund 29 % gegenüber, die glauben, dass Onlinenetzwerke diverser sind als soziale

Offlinenetzwerke. Die Einstellung zu KI hatte dabei keinen Einfluss auf diese Einschätzungen.

4.5.1.2. Wandel des traditionellen Journalismus

Im Zuge der ersten Befragung wurden Einschätzungen eingeholt, wie sich künstliche Intelligenz und KI-basierte Systeme auf die Arbeit von Medienleuten bzw. die Erstellung und Kuration von journalistischen Inhalten auswirkt.

Unter dem Stichwort «Wandel des traditionellen Journalismus» wurden die Fachpersonen daher zunächst um ihre Einschätzung zu verschiedenen Aussagen gebeten, die mögliche Einflüsse der Nutzung von KI im Journalismus skizzierten. Anknüpfend an das Thema Fake News, fand die Aussage, dass die zunehmende Verbreitung von sogenannten *deep fakes* die Glaubwürdigkeit der Medien insgesamt untergraben wird, die stärkste Zustimmung (Abbildung 25). Insgesamt 82 % stimmten dieser Aussage zu. Ein vergleichbares Mass an Zustimmung fand die Aussage, dass die Einführung von KI die Art und Weise verändert hat, wie traditionelle Medien ihre Quellen überprüfen (61 % stimmten zu und nur 21 % lehnten die Aussage ab). Bezüglich der Frage, ob die Verwertung von Social-Media-Inhalten eine Gefährdung für den Journalismus darstellt, gingen die Meinungen etwas stärker auseinander: 31 % lehnten diese Aussage ab, während 51 % ihr zustimmten.

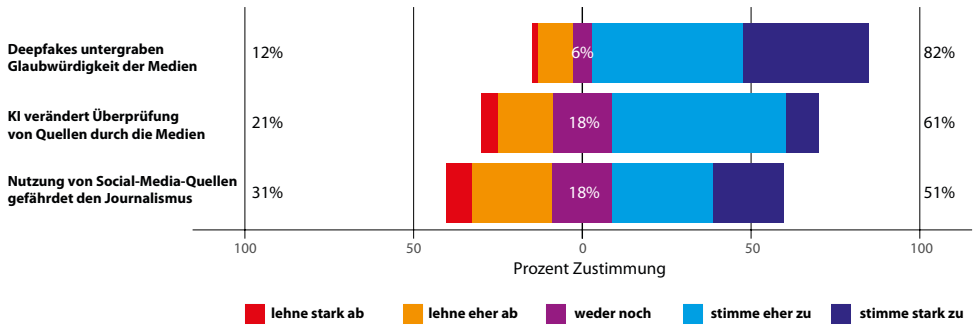

Abbildung 25: Genereller Einfluss von KI auf den Journalismus.

Bezogen auf den Einfluss auf die allgemeine Qualität des Journalismus dieser und anderer KI-bedingter Veränderungen ergab sich das folgende Meinungsbild

(Abbildung 26): Eine grosse Mehrheit (73 %) glaubt, dass das vermehrt automatische Erzeugen von kurzen Nachrichtenartikeln die Qualität des Journalismus verschlechtert. Ähnlich wird die Suche von Informationen durch Journalisten auf Social Media beurteilt: 67 % sehen hier einen negativen Einfluss und nur 24 % glauben, dass diese Entwicklung die Qualität des Journalismus verbessern werde. Weniger pessimistisch waren die Einschätzungen bezüglich der zunehmenden reinen Onlineverbreitung von Nachrichten. Die Hälfte der Befragten beurteilt diese Entwicklung weder klar negativ noch klar positiv für die Qualität des Journalismus.

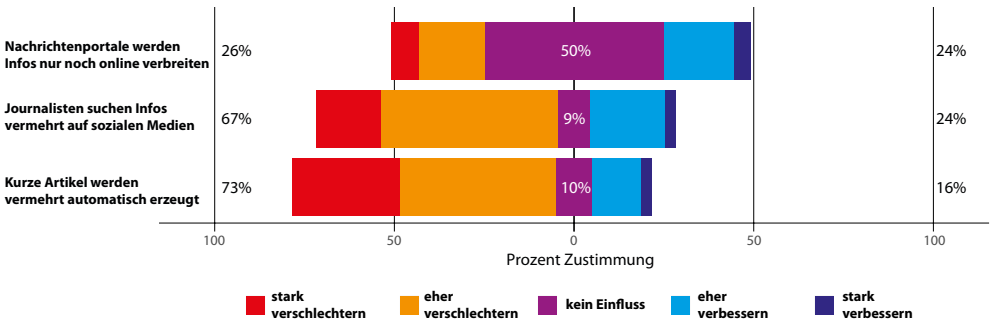

Abbildung 26: Einfluss von KI auf die Qualität des Journalismus.

Insbesondere die Einschätzungen zu dieser letzten Entwicklung unterscheiden sich hierbei stark (und statistisch signifikant) nach dem Grad an «KI-Enthusiasmus» der Befragten: Während «KI-Enthusiasten» stärker als neutrale zu einer positiven Einschätzung tendieren, ist dieser Trend für Skeptiker umgekehrt.

Die stärksten Unterschiede zwischen den Befragten zeigten sich jedoch bei der Einschätzung der Szenarien des KI-bedingten Wandels im Journalismus. Insbesondere der Zeitrahmen, in welchem sich unterschiedliche Veränderungen manifestieren werden, wurde von den Experten divers eingestuft.

Konkret wurden die folgenden Entwicklungen abgefragt:

1. Mehr als 50 % aller Kurznachrichten werden durch KI-Systeme (*robot journalists*) verfasst statt von Journalisten.
2. Recherchen werden hauptsächlich durch KI-Systeme durchgeführt.
3. Nachrichten werden praktisch nur noch online verbreitet.

4. Mehr als 50 % aller Nachrichten werden von Bürgern verfasst, nicht von Journalisten.

Die Ergebnisse (Abbildung 27) zeigen, dass sich die Fachpersonen in dieser Frage uneinig sind, ob und wenn ja, wann diese Veränderungen realisiert werden. Ein beachtlicher Anteil der Experten schätzt folglich alle vier Szenarien als unwahrscheinlich ein (zwischen 15 und 34 %). Wiederum andere glauben, dass sich diese Veränderungen bereits vollzogen haben. Entsprechend wurden die beschriebenen Szenarien als bereits zutreffend eingestuft (ebenfalls zwischen 15 und 34 %; mit Ausnahme des zweiten Szenarios). Der Rest der Befragten glaubt, dass die beschriebenen Veränderungen früher (innert der nächsten fünf Jahre) oder später (eher erst bis zu 25 Jahren) Realität werden.

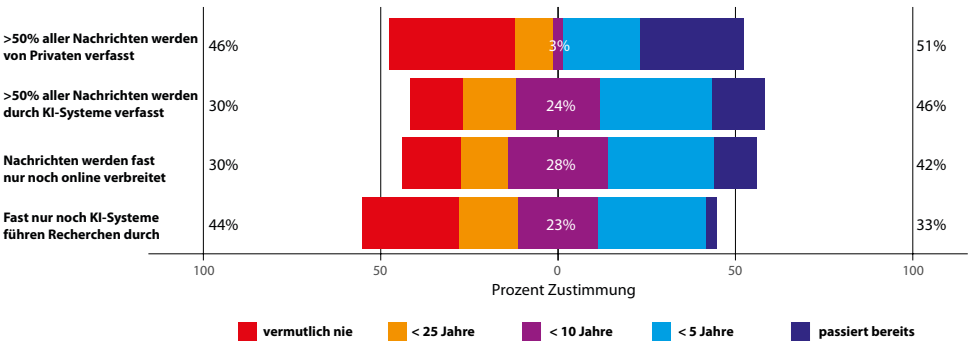

Abbildung 27: Einschätzung der Realisierung von KI-beeinflussten Veränderungen.

4.5.1.3. Personalisierung des Medienkonsums

Ein weiterer Frageblock betraf die Personalisierung des Medienkonsums. Ein erstes interessantes Ergebnis ist die Einschätzung der Fachpersonen, ab welchem Zeitpunkt Mediennutzer/-innen nur noch personalisierte Inhalte angezeigt werden. 33 % der Experten waren der Meinung, dass dies bereits der Fall sei, während 21 % glauben, dass dies wohl nie eintreten werde.

Dabei zeigte sich, dass die Fachpersonen das Thema Personalisierung durchaus differenziert betrachten: Knapp zwei Drittel (64 %) der Befragten lehnen die Aussage ab, wonach die *Personalisierung von Inhalten* auf Online-Medienportalen eine positive Entwicklung ist. Diese Ablehnung fiel allerdings schwächer aus, wenn die *Personalisierung von Empfehlungen* angesprochen wurde; konkret lehnten nur

34 % die Aussage stark oder eher ab, dass die Personalisierung von Empfehlungen ein positiver Trend ist (Abbildung 28).

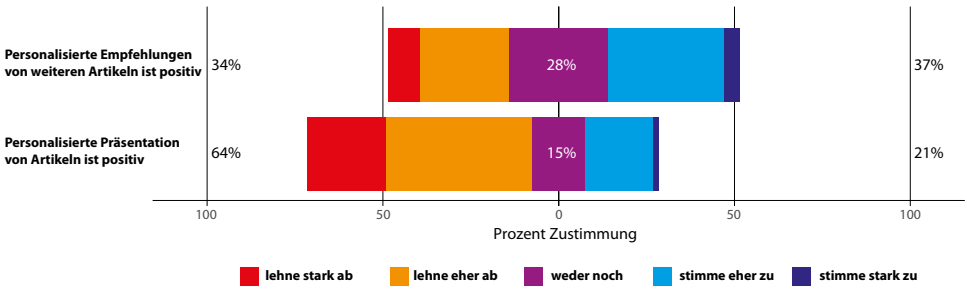

Abbildung 28: Einschätzung der Personalisierung von Medieninhalten.

Dies lässt vermuten, dass das Ausmass sowie die Art und Weise der Personalisierung einen kritischen Aspekt für die Bewertung der Fachpersonen darstellt.

4.5.1.4. Fake News

Bei der Befragung zum Themenschwerpunkt Fake News wurde in der ersten Umfrage die Meinung der Fachpersonen zu einigen kontrovers diskutierten Themen erfasst. Eine solche Kontroverse betrifft die **Rolle von KI bei der Verbreitung von Fake News**. Zum einen lässt sich vermuten, dass im Zuge einer fortschreitenden Weiterentwicklung KI dazu eingesetzt werden kann, Falschmeldungen noch effizienter zu erzeugen und zu verbreiten, gar so, dass sie kaum noch als solche erkannt werden können (*deep fakes*). 47 % der Befragten schätzen dieses Szenario als (sehr) wahrscheinlich ein. Zum anderen besteht aber auch die Möglichkeit, dass die Weiterentwicklung von Erkennungssystemen von Fortschritten in der Forschung profitiert und es daher in Zukunft einfacher sein wird, mit dem Einsatz von KI Fake News besser zu erkennen. 29 % der Befragten schätzen dieses Szenario als (sehr) wahrscheinlicher ein; 24 % äusserten keine Tendenz zwischen den Szenarien. Die Standpunkte in dieser Debatte korrespondierten stark mit dem allgemeinen «KI-Enthusiasmus» der Befragten; «KI-Enthusiasten» sehen in KI signifikant stärker ein Mittel zur Bekämpfung von Fake News.

Eine weitere Kontroverse betrifft den Einsatz von **automatischem Fact Checking** (Abbildung 29). Hier positioniert sich eine Mehrheit von 54 % gegen die Aussage, dass das automatisierte Überprüfen von Online-Inhalten auf ihren Faktengehalt abzulehnen sei; lediglich 24 % stimmten dieser Aussage zu. Ferner befürworteten die Teilnehmenden in der ersten Befragung jeweils die Einführung von Gesetzen zur Bekämpfung von Fake News (60 % stimmten zu, gegenüber nur 21 % mit ablehnender Einstellung) sowie die Entwicklung interner Kontrollmechanismen zur Bekämpfung von Fake News durch die Betreiber sozialer Medienplattformen (85 % stimmten zu, gegenüber nur 5 % mit ablehnender Einstellung).

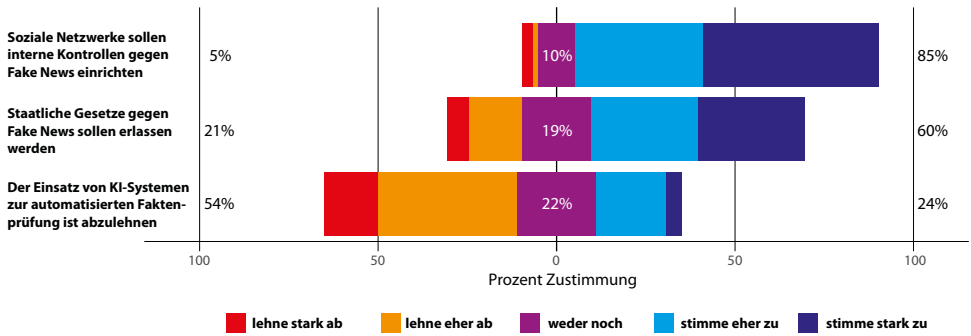

Abbildung 29: Einschätzung von Massnahmen gegen Fake News.

Ferner zeigt sich allerdings, dass diese überwiegend starke Befürwortung der verschiedenen Massnahmen gegen die Verbreitung von Fake News für viele implizit an eine Bedingung geknüpft ist: der Wahrung der freien Meinungsäusserung. So stimmten 78 % der Aussage zu, dass eine Vollmacht zur Zensur von Nachrichten für private Unternehmen (wie z.B. soziale Netzwerke) mit einer Gefährdung der Meinungsfreiheit einhergeht (nur 10 % lehnten diese Aussage ab).

In der zweiten Umfrage wurde dieser potenzielle Zielkonflikt zwischen dem effektiven Bekämpfen von Fake News bei gleichzeitiger Wahrung des Rechts der freien Meinungsäusserung nochmals genauer aufgegriffen. Konkret wurden die Befragten mit der folgenden Frage direkt vor die Wahl gestellt: *Was schätzen Sie als das grössere Problem ein: zu wenig Bekämpfung von Fake News oder eine zu starke Einschränkung der freien Meinungsäusserung?* Die Ergebnisse zeigen ein gewis-

ses Mass an Polarisierung: Während 30 % aller Befragten die zu geringe Bekämpfung von Fake News für das grössere Risiko hielten, schätzten 15 % die Einschränkung der freien Meinungsäusserung stärker bedroht ein. Jeweils 26 % tendierten zur ersten bzw. zweiten Position.

Bemerkenswerterweise hängen die Standpunkte der Befragten in dieser Kontroverse nur geringfügig und statistisch nicht signifikant mit ihren allgemeinen Einstellungen zu KI (Enthusiasmus vs. Skeptizismus) zusammen. Gleiches gilt für die KI-Expertise der Befragten. Es liegt nahe zu vermuten, dass die Einstellung zur Debatte um die Grenzen des technischen Eingriffs in die freie Meinungsäusserung bei der Bekämpfung von Fake News mit anderen politischen Grundwerten zusammenhängt. Allerdings lässt sich diese Vermutung mit den Daten, die im Rahmen beider Umfragen erhoben wurden, nicht überprüfen.

4.5.1.5. Zusammenfassendes Fazit

Die Fachpersonen sehen einen tendenziell starken Einfluss von KI-Systemen auf die Entwicklung von Medieninhalten. Sie stehen dieser Entwicklung sehr skeptisch gegenüber. Gerade auch in Bezug auf die Diversität von Inhalten und politischen Positionen innerhalb von Onlinemedien und vor allem auch sozialen Medien entwerfen die Fachpersonen ein kritisches Bild. Sie halten es für besonders wichtig, gegen Fake News vorzugehen; durchaus auch staatlich unterstützt.

4.5.2. Zentrale Ergebnisse der zweiten Umfrage

In der zweiten Umfrage haben 55 von 111 Befragten Einschätzungen zum Bereich «Medien» abgegeben. Beurteilt wurden Effektivität und Wünschbarkeit von Massnahmen, wobei nur Daten zum ersteren Punkt abgebildet werden.

4.5.2.1. Themenkomplex Personalisierung und Filterblase

Zum Themenkomplex Personalisierung wurden den Befragten verschiedene mögliche Massnahmen präsentiert (Abbildung 30).

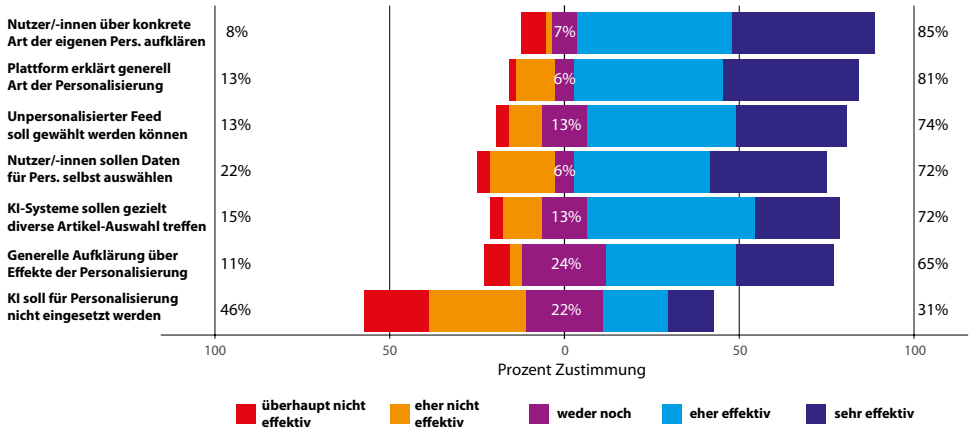

Abbildung 30: Effektivität von Massnahmen gegen Filterblasen.

Die Ergebnisse dieser Befragung zeigen, dass die Fachpersonen stark dazu tendieren, Massnahmen zu befürworten, die den Nutzenden mehr Kontrolle geben und/oder transparenter machen, welche Inhalte personalisiert werden und welche Daten zur Personalisierung genutzt werden. Die Fachpersonen sind allerdings dahin gehend gespalten, inwiefern die Verbannung algorithmisch gesteuerter Personalisierung effektiv und wünschenswert ist. Mit Ausnahme dieser letzten Massnahme zeigt sich generell eine Tendenz, die genannten Massnahmen zwar für richtig zu halten (Wünschbarkeit), aber ihre Effektivität als vergleichsweise etwas geringer einzuschätzen.

4.5.2.2. Themenkomplex «Fake News»

In der zweiten Runde der Befragung wurde den Experten eine Reihe konkreter Massnahmen zur Bekämpfung von Fake News präsentiert (Abbildung 31).

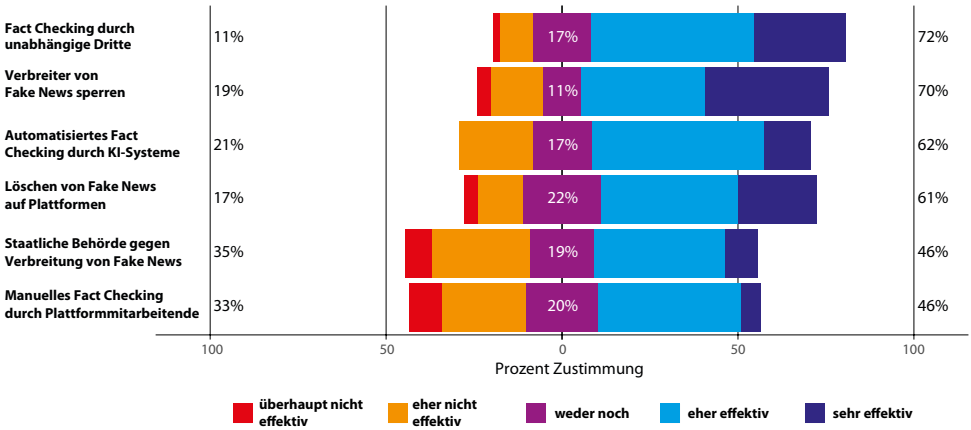

Abbildung 31: Effektivität von Massnahmen gegen Fake News.

Die zwei Massnahmen mit dem höchsten Mass an Meinungspolarisierung sind das Einrichten von staatlichen Behörden, um der Verbreitung von Fake News entgegenzuwirken, sowie das manuelle *fact checking* durch die Mitarbeitenden der Plattformen. Konkret stehen der Idee des Einrichtens einer staatlichen Behörde 25 klare Befürworter genau 22 klaren Gegnern gegenüber. Im Falle des manuellen *fact checking* durch Plattformmitarbeitern ist dieses Verhältnis 30 zu 10.

Mit Blick auf die Effektivität ist bemerkenswert, dass die Polarisierung (also Personen, die eine der äusseren Antwortkategorien wählen) hinsichtlich des Einrichtens von staatlichen Behörden zur Fake-News-Bekämpfung bei der Wünschbarkeit dieser Massnahme noch höher ist. Es gilt aber zu bemerken, dass diese Veränderung in den Verteilungen auf relativ wenige Befragte zurückzuführen ist.

4.5.2.3. Zusammenfassende Beurteilung der Expertenmeinungen

Die Fachpersonen unterstützen die Ideen für Vorkehrungen stark, die es Nutzerinnen und Nutzern erlauben, mehr Transparenz über den Einsatz von KI in Bezug auf Nachrichteninhalte zu gewinnen. Diese sollen die Möglichkeit haben, individuell darüber zu entscheiden, wie stark die eigenen Inhalte durch Algorithmen gesteuert sind und welche der persönlichen Informationen von Algorithmen genutzt werden. Sie betrachten all diese Massnahmen sowohl als effektiv wie auch als wünschbar.

In Bezug auf Fake News beurteilen die Fachpersonen das Löschen von Beiträgen und das Sperren von Nutzern als effektiv und wünschbar. Kritischer stehen sie der Einrichtung von Behörden zu diesem Zweck gegenüber. Auch *fact checking* wird nur bedingt als hilfreiches Instrument gegen Fake News betrachtet. Hilfreicher als das manuelle *fact checking* wird das automatisierte *fact checking* betrachtet.

4.5.3. Ergebnisse des Workshops zum Themenfeld Medien

Den Teilnehmenden wurde zunächst erläutert, dass das übergeordnete Thema der Gruppendiskussion der Einfluss der Anwendung von KI-basierten Systemen auf die politische Meinungsbildung ist. Der Fokus der Diskussion solle dabei auf dem Einsatz von KI in den Domänen traditioneller sowie digitaler sozialer Medien liegen. Zudem soll sich die Diskussion zunächst auf die KI-basierte Personalisierung von Inhalten durch digitale Medienplattformen und anschliessend auf das Thema Fake News richten.

Die Gruppendiskussion zum Thema **Personalisierung** wurde mit einer offenen Diskussionsrunde gestartet. Wiederkehrende Punkte in der Diskussion waren dabei der Mangel an Bewusstsein und Wissen über Selektionskriterien, die zur Personalisierung angewendet werden; ein Mangel an Möglichkeiten die Kriterien, anhand derer personalisiert wird, selbst zu beeinflussen bzw. zu kontrollieren; ein Unbehagen darüber, halb-wissentlich von Algorithmen in «Schubladen gesteckt» zu werden, sowie der negative Einfluss der Personalisierung von Medieninhalten auf den demokratischen Meinungsbildungsprozess.

In der anschliessenden Diskussion lag der Fokus zunächst auf Online-Nachrichtenplattformen. Kritisiert wurde dabei insbesondere, dass die marktwirtschaftliche Ausrichtung nicht öffentlich-rechtlicher Plattformen die Tendenz zur Aufmerksamkeitshascherei (*click baiting*) verstärke und so die journalistische Qualität abnähme. Der öffentlich-rechtliche Rundfunk stelle in dieser Hinsicht einen wichtigen Gegenpol und Qualitätsmassstab dar. Die Teilnehmenden waren zudem der Meinung, dass der Unterschied zwischen traditionell-redaktioneller und maschineller Personalisierung von Inhalten darin bestehe, dass Analyse und Produktion zusammenlaufen und es so zu einer beschleunigten Rückkopplung von Inhalt und Monetarisierungsinteressen komme. Vergleichbare Probleme seien auch in sozialen Medien zu beobachten. Die dort stattfindende Personalisierung verstärke negative Tendenzen wie Sensationalismus und die Verbreitung von Gerüchten. In beiden Fällen gelte aber, dass KI zwar verstärkend, aber nicht ursächlich negative Folgen

hervorrufen würde. Als Gegenmassnahmen wurde vorab eine Erhöhung von Transparenz und Kontrolle durch die Nutzenden gesehen. Transparenz sollte vorab dahin gehend geschaffen werden, welche (persönlichen) Daten in Personalisierungsalgorithmen einfliessen, als auch hinsichtlich des Outputs der Algorithmen (also Klassifikationen, Scorings etc.). Als positiv wurde auch die Idee erachtet, auf Online-Nachrichtenplattformen ein optionales Anwählen nicht personalisierter Seiten zu etablieren.

Beim Thema **Fake News** wurden die enorme Verbreitungsgeschwindigkeit von Fake News sowie Bedenken, dass falsche und manipulierte Inhalte in der Masse von Inhalten untergehen und ihr Einfluss somit unbemerkt auftritt, kritisch diskutiert. Betreffend die Frage, welche Massnahmen ergriffen werden könnten, um der Verbreitung von Fake News vorzubeugen, wurden folgende Aspekte diskutiert: Reputation und Rechenschaftspflicht seien wichtige Mechanismen, um Unternehmen für die Bekämpfung zu interessieren. Im Gegensatz zu den Ergebnissen der zweiten Umfrage lehnten die Teilnehmenden das Sperren von Accounts, die Fake News verbreiten, ab. Sollte dennoch auf ein solches Mittel zurückgegriffen werden, sollten Unternehmen verpflichtet werden, offenzulegen, wie beim Ermitteln solcher Accounts vorgegangen werde. Technisch gesehen könnte das automatische Entfernen von Fake News die beste Möglichkeit bieten, ihre Verbreitung zu verhindern. Allerdings könnte das Löschen von Inhalten «demokratiegefährdend» sein und Systeme, die über das Löschen von Inhalten oder Sperren von Nutzern operieren, die Gefahr bergen, hinsichtlich politischer Zwecke von Dritten manipuliert zu werden. Ferner wurde auch diskutiert, inwiefern die Bewertung von Quellen bzw. Online-Nachrichtenplattformen ein geeignetes Mittel sei, um die Vertrauenswürdigkeit von Nachrichten zu signalisieren. Die flächendeckende Umsetzung eines solchen Bewertungssystems blieb ebenfalls umstritten.

Auf die Frage, wer für die Bekämpfung der Verbreitung von Fake News verantwortlich sei, äusserten die Teilnehmenden zurückhaltende Skepsis dahin gehend, ob staatliche Regulation ein effektives Instrument darstelle. Unternehmen müssten viel mehr transparent machen, was sie gegen die Verbreitung von Fake News unternehmen, und staatliche Prüfstellen sollen dabei kontrollierend agieren. Wie auch in anderen Themenbereichen zeigte sich hierbei eine Präferenz der Teilnehmenden für marktwirtschaftliche Selbstregulierung.

Insgesamt ergaben sich aus den Diskussionen der zweiten Runde folgende Prioritäten für den Themenbereich Medien:

1. Erhöhung der Transparenz darüber, aufgrund welcher Daten personalisiert wird und welche Effekte die Personalisierung tatsächlich hat. Als Referenzmassstab für einen solchen Vergleich könnte standardmässig die Option einer «nicht personalisierten Nachrichtenauswahl» auf Nachrichtenportalen verlangt werden, wobei dann aber noch zu klären ist, was «nicht personalisiert» konkret bedeutet.
2. Algorithmen dahin gehend anwenden, dass die Diversität in den Inhalten, die Nutzerinnen und Nutzern vorgeschlagen werden, gesteigert wird.
3. Bekämpfung von Fake News sowohl durch Massnahmen, die Falschinformationen als solche kenntlich machen, als auch durch eine Bewertung hinsichtlich der (ideologischen) Einseitigkeit/Ausgewogenheit von Nachrichten. Löschen von Nachrichten oder Sperren von Nutzerinnen und Nutzern werden eher kritisch betrachtet.

4.6. Beurteilungen zum Themenfeld Verwaltung und Gerichtsbarkeit[121]

4.6.1. Zentrale Ergebnisse der ersten Umfrage

In der ersten Umfrage haben 73 von 307 Befragten den Bereich «Verwaltung» gewählt.

4.6.1.1. Wahrscheinlichkeit von KI-Einsatz in der Verwaltung

Grundsätzlich wird von den Befragten der KI-Einsatz in den nächsten fünf bis zehn Jahren in anderen Ländern als wahrscheinlicher eingestuft als in der Schweiz. Teilautomatisierung (sowohl bei einfachen als auch bei komplexen Sachverhalten) wird als wahrscheinlicher eingestuft als Vollautomatisierung.

[121] Dieser Abschnitt beruht auf Arbeiten von Nadja Braun Binder, bis Ende Juli 2019 Assistenzprofessorin für öffentliches Recht unter besonderer Berücksichtigung europäischer Demokratiefragen an der Universität Zürich, seit 01.08.2019 Professorin für Öffentliches Recht an der Universität Basel.

Eine deutliche Mehrheit der Befragten betrachtet den Einsatz von KI bei der Teilautomatisierung sowohl einfacher als auch komplexer Sachverhalte als wahrscheinlich in den nächsten fünf bis zehn Jahren. Dies gilt sowohl für die Schweiz als auch für andere Länder. Allerdings sind die entsprechenden Werte für die Schweiz etwas zurückhaltender.

Mit Blick auf die Vollautomatisierung (also eine Bearbeitung durch KI ohne jegliche menschliche Bearbeitung) zeigt sich ein deutlicherer Unterschied zwischen der Einschätzung der Entwicklungen in den nächsten fünf bis zehn Jahren in der Schweiz gegenüber jener in anderen Ländern. Die Befragten sind gespalten bezüglich der Erwartbarkeit vollautomatisierter Bearbeitung einfacher Sachverhalte (51 %) in der Schweiz und stufen die vollautomatisierte Bearbeitung komplexer Sachverhalte in der Schweiz deutlich als unwahrscheinlich ein (21 %). Für andere Länder wird die Vollautomatisierung von einfachen Sachverhalten in den nächsten fünf bis zehn Jahren von einer Mehrheit der Befragten als ein wahrscheinliches Szenario eingestuft; die Wahrscheinlichkeit einer Vollautomatisierung komplexer Sachverhalte wird dagegen zurückhaltend eingeschätzt.

4.6.1.2. Wahrscheinliche Einsatzgebiete von KI durch den Staat

Der Einsatz von KI im Rahmen von Kontrollen (z.B. Abgleich aller in einem Antrag enthaltenen Daten mit Daten aus vorhandenen Datenbanken), der Betrugsbekämpfung (z.B. Steuerbetrugsbekämpfung durch Musterabweichungserkennung) und der Erkennung von drohenden Gefahren für Polizeigüter (z.B. für die öffentliche Sicherheit und Ordnung) werden von einer Mehrheit der Befragten als wahrscheinlich in den nächsten fünf bis zehn Jahren eingestuft. Dies gilt sowohl für die Schweiz als auch für andere Länder. Allerdings sind die entsprechenden Werte für die Schweiz etwas zurückhaltender. Dies betrifft insbesondere die Wahrscheinlichkeit eines Einsatzes im Rahmen von Kontrollen; sie wird für die Schweiz signifikant geringer eingestuft als für das Ausland.

Die Wahrscheinlichkeit des Einsatzes von KI zur Eruierung des Risikopotenzials von Personen (z.B. die Rückfallgefahr von Straftätern) wird für die Schweiz in den nächsten fünf bis zehn Jahren ebenfalls signifikant geringer eingestuft als für das Ausland.

4.6.1.3. Risiken von KI für Bürgerinnen und Bürger

Die Befragten schätzen folgende Risiken für Bürgerinnen und Bürger tendenziell als wahrscheinlich ein:

- Einschränkungen des Rechts auf Datenschutz (76 %)
- Das Risiko, unverschuldet ins Visier der Behörden zu gelangen (69 %)
- Nicht nachvollziehbare (intransparente) Verfahrensabläufe (61 %)

4.6.1.4. Risiken von KI für die Verwaltung

Die Befragten schätzen folgende Risiken für die Verwaltung tendenziell als wahrscheinlich ein:

- Abhängigkeit von IT-Anbietern ausserhalb des eigenen Rechtsraums (90 %)
- Maschinenhörigkeit (z.B. mangelnde Motivation, eine von einem KI-System abweichende Position zu begründen) (76 %)
- Schwierigkeit, Datenschutz zu gewährleisten (76 %)
- Schwierigkeit, Datensicherheit zu gewährleisten (73 %)
- Mangelnde Kontrollierbarkeit der automatisierten Verfahrensabläufe (61 %)

4.6.1.5. Zusammenfassendes Fazit

Insgesamt fällt die Experteneinschätzung mit Blick auf die Schweiz konservativer aus als für andere Länder. So wird der KI-Einsatz in der schweizerischen öffentlichen Verwaltung als weniger wahrscheinlich eingeschätzt als in anderen Ländern. Ein Einsatz im Rahmen von vollautomatisierten komplexen Sachverhalten in der Schweiz wird klar als unwahrscheinlich beurteilt. Die (insgesamt zurückhaltende) Erwartung ist, dass KI in der öffentlichen Verwaltung in der Schweiz eher bei einfachen Sachverhalten und in teilautomatisierten Prozessen eingesetzt wird.

4.6.2. Zentrale Ergebnisse der zweiten Umfrage

In der zweiten Umfrage haben 58 von 111 Befragten Einschätzungen zum Bereich «Verwaltung» angegeben. Die Teilnehmenden konnten zu fünf Risiken (Datenschutz, Unschuldsvermutung, Intransparente Verfahren, Maschinenhörigkeit, Diskriminierungspotenzial) Massnahmen vorschlagen bzw. die vorgeschlagenen Massnahmen hinsichtlich Effektivität und Wünschbarkeit bewerten.

Die ersten vier Risiken wurden von den Ergebnissen der ersten Umfrage abgeleitet. Darüber hinaus ergab die Literaturanalyse, dass KI-Systeme zu systematisch falschen Resultaten kommen können, die bestimmte Gruppen von Menschen benachteiligen, ohne dass dies sachlich gerechtfertigt ist. Ein solcher Bias kann seinen Ursprung in fehler- oder lückenhaften Daten haben, er kann aber auch aufgrund von fehlerhafter KI entstehen. Ein Bias ist aber auch bei korrekten Daten und fehlerfreiem KI-System möglich, nämlich wenn die genutzte Datengrundlage selbst einseitig geprägt ist. Diesem letztgenannten Umstand ist am schwierigsten zu begegnen. Obwohl dieses Risiko in der ersten Expertenumfrage nicht hervorgehoben wurde, wurde es aufgrund der Relevanz in der Literatur in die zweite Umfrage aufgenommen.

Unter den Antwortenden sind die Männer deutlich in der Überzahl (45 von 58), dasselbe gilt für Deutschsprachige und Schweizer (je 46 von 58).

4.6.2.1. Datenschutz

Die im Bereich des Datenschutzes vorgeschlagenen Massnahmen umfassen verschieden ausgeprägte Informationspflichten und Zustimmungserfordernisse sowie Verbote, personenbezogene Daten weiterzugeben. Fast alle davon wurden klar mehrheitlich als effektiv (77–82 %) und wünschenswert (73–88 %) betrachtet. Dies betrifft die Information der betroffenen Person, wenn Daten über sie genutzt bzw. neue Daten über sie generiert werden; ein Verbot, personenbezogene Daten Dritten zur Verfügung zu stellen, sowie die Erhebung bzw. Verwendung personenbezogener Daten nur nach Zustimmung durch die betroffene Person. Lediglich eine Minderheit beurteilte das Verbot, personenbezogene Daten anderen Verwaltungsstellen zur Verfügung zu stellen, als effektiv (47 %) bzw. wünschenswert (43 %).

In den Freitextfeldern ergänzten die Teilnehmenden verschiedene Massnahmen. Diese umfassen Abstufungen an Transparenz- bzw. Offenlegungserfordernissen

(Transparenz von Datenflüssen, Transparenz von Datensammlungen, Offenlegungspflicht für Quellcode und Architektur von Soft- und Hardware); den Grundsatz «privacy by design»; die Einführung einer Secure ID und der Möglichkeit, dass jede betroffene Person die Autonomie über ihre Daten und die Vergabe von Berechtigungen zu deren Einsicht hat bzw. transparent über jegliche Zugriffe auf die Daten informiert wird; sowie allgemeine Vorgaben (Datensparsamkeit, gesetzliche Grundlagen, klare Zuständigkeiten bzw. Stärkung des EDÖB bzw. Durchführung von internen/externen Audits und Kontrollen, mit der EU und anderen Ländern abgestimmte Rechtsgrundlagen).

4.6.2.2. Unschuldsvermutung

Als Mittel gegen das Risiko, unverschuldet ins Visier der Behörden zu gelangen, wurden als Massnahmen einerseits verschieden strikte Testphasen und andererseits Einschränkungen mit Blick auf Verfahrensarten bzw. Verfahrensstadien, in denen KI eingesetzt werden darf, vorgeschlagen. Die Ergebnisse sind in Abbildung 32 aufgeführt.

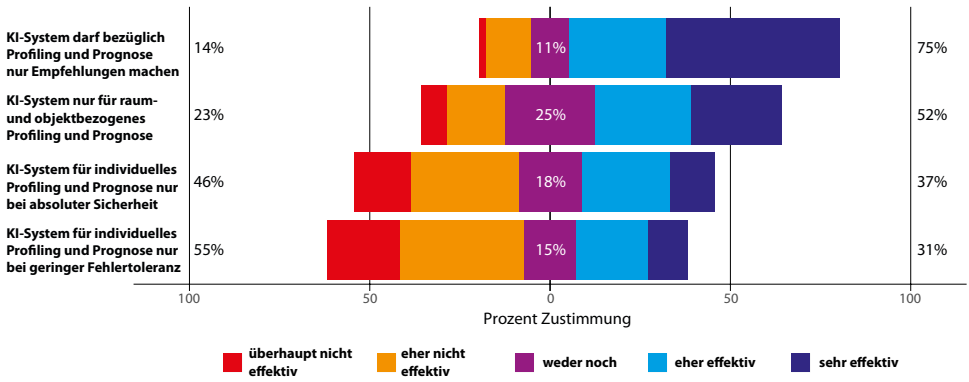

Abbildung 32: Erwartete Effektivität von Massnahmen im Bereich Unschuldsvermutung.

Die einzige Massnahme, die von einer deutlichen Mehrheit der Antwortenden als effektiv (76 %) und wünschenswert (84 %) eingestuft wurde, ist die Vorgabe, dass KI im Rahmen von Profiling oder vorausschauenden Analysen nur eine Entscheidungsempfehlung abgeben, aber niemals selbst entscheiden darf. Eine knappe

Mehrheit der Antwortenden beurteilt die Massnahme, dass KI lediglich für die Vorbereitung von raum- bzw. objektbezogenen Entscheidungen (nicht aber für die Vorbereitung personenbezogener Entscheidungen) eingesetzt werden darf, als effektiv (51 %) und wünschenswert (59 %). Nur eine Minderheit hält die Vorgabe, dass KI erst eingesetzt werden darf, wenn in Tests *keine* falschen Verdächtigungen bzw. lediglich 1 % falsche Verdächtigungen vorkommen, für effektiv (36 % bzw. 31 %). Interessanterweise hält eine knappe Mehrheit allerdings die Nulltoleranz von falschen Verdächtigungen für wünschenswert (54 %). Die 1%-Toleranz wird nur von einer Minderheit für wünschenswert (43 %) befunden.

In den Freitextfeldern haben die Antwortenden wiederum verschiedene Massnahmen ergänzt. Dazu zählen verschiedene Restriktionen des Einsatzes von KI (z.B. KI nur als Entscheidungsempfehlung – nicht aber als Ersatz für menschliche Entscheidungen); eine Deklarationspflicht für KI-gestützte Entscheide; die Stärkung der Rechtsschutzmöglichkeiten für Betroffene (z.B. durch die Möglichkeit, einen «digitalen» Anwalt beizuziehen, der das Unwissen einer Person in diesem Bereich ausgleichen kann) bzw. eine unkomplizierte Schadenersatzregelung (z.B. persönliche Haftung von Entscheidungsträgern und/oder Dritten); eine Zertifizierung der genutzten KI-Systeme durch externe Prüfstellen (TÜV); die Offenlegung statistischer Daten über den KI-Einsatz (*false positives*, Trainingseffizienz); öffentliche Tests von KI-Systemen sowie keine bzw. eingeschränkte Weitergabe von Daten (allgemein oder im Bereich der Strafverfolgung).

4.6.2.3. Intransparente Verfahren

Die vorgeschlagenen Massnahmen zur Sicherstellung der Transparenz umfassen Informations-, Erklärungs- und Begründungspflichten sowie unterschiedliche Kontrollvorgaben. Alle Massnahmen wurden von einer deutlichen Mehrheit der Antwortenden als effektiv (67–87 %) und wünschenswert (78–95 %) beurteilt. Im Einzelnen handelt es sich um die folgenden Massnahmen:

- Verpflichtung des Staates, allgemein darüber zu informieren, wenn er KI in bestimmten Verfahren einsetzt;
- Verpflichtung des Staates, die Grundzüge des Ablaufs des Verfahrens, in dem KI eingesetzt wird, in allgemeiner Weise zu erklären;
- Verpflichtung des Staates, jede Entscheidung, die sich auf KI abstützt, individuell zu begründen;

- Konfigurierung der KI-Systeme so, dass sie selbst die zentralen Punkte, die sie bei der Vorbereitung ihres Resultats berücksichtigen, erklären können;
- Verpflichtung der Verwaltungseinheit, die KI einsetzt, die KI-Systeme regelmässig zu überprüfen.

Technische Fachpersonen stufen den Bedarf nach einer Erklärung jeder staatlichen Entscheidung als deutlich effektiver und wünschenswerter ein als Laien. Dies gilt auch für Personen mit höherer KI-Skepsis im Vergleich zu jenen mit tiefer Skepsis.

Als weitere Massnahmen schlugen die Antwortenden in den Freitextfeldern vor die Förderung bzw. Finanzierung der Forschung zu *explainable AI*; eine restriktive Anwendung von KI-Systemen (nur zur Entscheidungsunterstützung), eine Deklarationspflicht, Offenlegungspflichten sowie verschiedene Kontrollen sowie die Schaffung einer Instanz, bei der als falsch erachtete Entscheidungen rapportiert werden können.

4.6.2.4. Maschinenhörigkeit

Die drei vorgeschlagenen Massnahmen im Bereich des Risikos der Maschinenhörigkeit (z.B. mangelnde Motivation, eine von einem KI-System abweichende Position zu begründen) wurden jeweils von einer Mehrheit als effektiv (66–75 %) und von einer sehr deutlichen Mehrheit (89–99 %) als wünschenswert eingestuft. Im Einzelnen handelt es sich um die folgenden Massnahmen:

- Schulung der Personen, die KI-Systeme nutzen (hinsichtlich deren Möglichkeiten und Grenzen).
- Regelmässige Sensibilisierung der Personen, die KI-Systeme nutzen.
- Wichtige Entscheidungen, die sich auf Empfehlungen von KI-Systemen stützen, müssen von einer zweiten Person überprüft werden.

In den Freitextfeldern wurden ergänzend als Massnahmen vorgeschlagen: Ausbildung bzw. Aufklärung zu mündigen Bürgerinnen und Bürgern; eine restriktive Anwendung von KI-Systemen (nur zur Entscheidungsunterstützung) sowie verschiedene Kontrollen und Offenlegungspflichten sowie eine Pflicht zur Einzelfallbegründung.

4.6.2.5. Diskriminierungspotenzial

Zwei der im Bereich Diskriminierungspotenzial vorgeschlagenen Massnahmen wurden von einer deutlichen Mehrheit als effektiv (69–88 %) und wünschbar (79–93 %) gewertet. Es handelt sich um die beiden folgenden Massnahmen:

- Ein KI-System darf erst eingesetzt werden, wenn in mehreren Trainingszyklen diskriminierende Resultate ausgeschlossen werden konnten.
- Der Staat muss beim Einsatz von KI transparent machen, welche Datenbasis genutzt wird und welche zentralen Kriterien die KI-Systeme aus den genutzten Datensätzen entwickeln.

Die dritte vorgeschlagene Massnahme – dass der Staat nur Daten nutzen darf, die er selber gesammelt hat – wurde lediglich von einer Minderheit als effektiv (23 %) bzw. wünschbar (45 %) eingestuft (Abbildung 33).

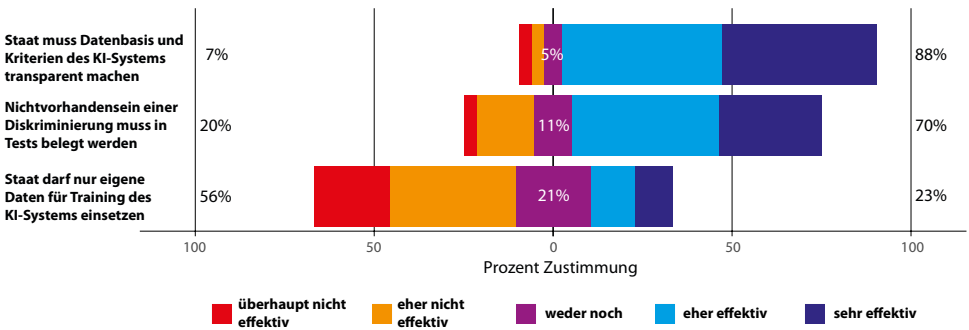

Abbildung 33: Erwartete Effektivität von Massnahmen im Bereich Diskriminierung.

In den Freitextfeldern wurden als Massnahmen ergänzend angeführt die Notwendigkeit einer grossen Datenbasis und intensiver Trainings; die Einführung von Qualitätsstandards für Daten; eine restriktive Anwendung von KI-Systemen (nur zur Entscheidungsunterstützung) sowie verschiedene Kontrollen (insbesondere genügend grosse statistisch signifikante Stichproben), Zertifizierung von KI-Systemen und Offenlegungspflichten sowie die Angabe eines *bias-levels* (Wahrscheinlichkeit, dass ein Bias bezüglich Geschlecht, Alter, Religion, Nationalität etc. vorliegt).

4.6.2.6. Zusammenfassende Beurteilung der Expertenmeinungen

Die Resultate aus beiden Umfragen ergeben mit Blick auf die Entwicklungen in der schweizerischen öffentlichen Verwaltung ein kohärentes Bild. Auf der einen Seite wird der Einsatz von KI als weniger wahrscheinlich als in anderen Ländern eingeschätzt und in eher unkritischen Abläufen (im Rahmen von teilautomatisierten Prozessen bei einfachen Sachverhalten) erwartet. Auf der anderen Seite findet sich unter den Massnahmen zur Eingrenzung verschiedener Risiken wiederholt die Empfehlung, KI-Systeme nur restriktiv (zur Unterstützung und nicht zum Fällen von Entscheidungen) einzusetzen.

Interessanterweise befand sich das in der Literatur vielfach beschriebene Diskriminierungspotenzial nicht unter den Risiken, die sich aufgrund der ersten Umfrage als besonders relevant herauskristallisiert hatten. Dies könnte darauf hindeuten, dass in diesem Bereich keine hohe Risikoerwartung mit Blick auf die öffentliche Verwaltung (in der Schweiz) besteht. Ein Teil der entsprechenden Massnahmen zur Verhinderung von Diskriminierung wurde dann allerdings in der zweiten Umfrage von einer Mehrheit als wünschbar beurteilt sowie um weitere Massnahmen ergänzt. Als Gegenmassnahme für verschiedene Risiken haben sich im Rahmen der zweiten Umfrage sodann Offenlegungs- bzw. Transparenzpflichten herauskristallisiert sowie die Notwendigkeit präventiver bzw. nachträglicher Kontrollen.

4.6.3. Ergebnisse des Workshops zum Themenfeld öffentliche Verwaltung

Nach einer kurzen Vorstellungsrunde wurden die zentralen Resultate der beiden Umfragen zusammengefasst. Insgesamt lassen sich unter den Antworten wenige Widersprüche erkennen. Aus der ersten Umfrage geht hervor, dass der staatliche KI-Einsatz in den nächsten fünf bis zehn Jahren in anderen Ländern als wahrscheinlicher eingestuft wird als in der Schweiz. Teilautomatisierung (sowohl bei einfachen als auch bei komplexen Sachverhalten) wird als wahrscheinlicher eingestuft als Vollautomatisierung. Der KI-Einsatz wird im Rahmen der Betrugsbekämpfung (z.B. Steuerbetrugsbekämpfung durch Musterabweichungserkennung) und Erkennung von drohenden Gefahren für Polizeigüter (z.B. für die öffentliche Sicherheit und Ordnung) als wahrscheinlich betrachtet. Der Einsatz von KI zur Eruierung des Risikopotenzials von Personen (z.B. die Rückfallgefahr von Straftäterinnen oder Straftätern) wird für die Schweiz als signifikant unwahrscheinlicher eingestuft als für andere Länder. In der zweiten Umfrage konnten die Teil-

nehmenden zu fünf Themenfeldern (Datenschutz, Unschuldsvermutung, Intransparente Verfahren, Maschinenhörigkeit, Diskriminierungspotenzial) Massnahmen vorschlagen bzw. die vorgeschlagenen Massnahmen bewerten. Basierend darauf entwickelte sich eine Diskussion um Empfehlungen bezüglich vier Kategorien: Risikodifferenzierung, Beschaffung, Datenqualität und Transparenz.

Ein wichtiger Faktor für die Bewertung des KI-Einsatzes durch den Staat ist eine **Risikodifferenzierung** basierend auf der Tatsache, dass nicht alle Verwaltungshandlungen gleich sensibel sind. Entsprechend ist der KI-Einsatz auch nicht in jedem Bereich mit denselben Risiken behaftet. Es empfiehlt sich deshalb, grundsätzliche Kategorien mit unterschiedlichen Risiken zu identifizieren. Als Differenzierungskriterien könnten insbesondere zwei Aspekte dienen: 1) Unterscheidung danach, gegenüber wem die Verwaltung KI einsetzt, z.B. verwaltungsintern, Verwaltung – Verwaltung, Verwaltung – Unternehmen, Verwaltung – Bürger; und 2) Unterscheidung danach, ob mit der entsprechenden KI-basierten Anwendung in Grundrechte von Betroffenen oder von Dritten eingegriffen werden könnte.

Grundlegende Weichenstellungen für einen verantwortungsvollen KI-Einsatz durch den Staat müssen im Rahmen der **Beschaffung** getroffen werden. Die Verwaltung muss dabei sicherstellen, dass sie die technischen Systeme versteht, die sie beschafft. Zudem muss die Verwaltung bei der Beschaffung auf eigenes, aber auch auf vom Hersteller unabhängiges Know-how zurückgreifen können. Entsprechendes Know-how muss aufgebaut und gefördert werden.

Die Korrektheit, Vollständigkeit und Eignung der Daten, die von der Verwaltung im Rahmen von KI-Anwendungen genutzt werden, sind von zentraler Bedeutung. Die Struktur der Verwaltungslandschaft in der Schweiz kann die Sicherstellung der **Datenqualität** erschweren, da Daten häufig dezentral (Föderalismus, ausgeprägtes «Silo-Denken» zwischen einzelnen Departementen) gesammelt, gepflegt und in unterschiedlichen Formaten gespeichert werden. Dies birgt mit Blick auf den KI-Einsatz gewisse Risiken. Auch wenn die einzelnen Gemeinwesen jeweils über korrekte Daten verfügen, kann die Qualität der Daten z.B. bei Datenaustausch oder deren organisationsübergreifender Nutzung beeinträchtigt werden. Bei der Definition von entsprechenden Massnahmen gilt es, das Rad nicht neu zu erfinden, sondern es kann (und sollte) auf bisherigen Arbeiten – vgl. z.B. den Ausbau einer gemeinsamen Stammdatenverwaltung des Bundes[122] – aufgebaut werden.

[122] Siehe: https://www.isb.admin.ch/isb/de/home/ikt-vorgaben/strategien-teilstrategien/sb018-ikt-teilstrategie_stammdatenverwaltung.html.

Die Herstellung von **Transparenz** wird schliesslich als mögliches Korrektiv für verschiedene Risiken betrachtet. Die Offenlegung der genutzten Daten (siehe Datenqualität) kann dabei helfen, einen Bias bzw. Diskriminierung zu verhindern (oder zumindest sichtbar zu machen), die Betroffenenrechte zu stärken, Fairness herzustellen und die Nachvollziehbarkeit von Verfahren zu erhöhen. Die Offenlegung der genutzten Algorithmen kann unter Umständen dazu beitragen, Schwachstellen zu erkennen (durch externe Überprüfung). Transparenz kann aber auch kontraproduktiv sein, da heute gegenüber KI eine gewisse Skepsis besteht. Insofern sollte eine Verwaltungseinheit sich nicht vorschnell damit brüsten, dass sie KI einsetzt, wenn es sich dabei um harmlose Anwendungen handelt (z.B. die Nutzung der online frei zugänglichen Übersetzungssoftware DeepL zur Übersetzung verwaltungsinterner, nicht rechtsverbindlicher Texte).

Teil der Transparenz ist dabei auch, dass der Erfolg (etwa im Sinn von Effizienz oder Effektivität) einer KI-Nutzung messbar gemacht werden sollte. Die Messbarkeit des Erfolgs enthält nämlich die Pflicht, dass man sich die Ziele, welche man mit KI verfolgt, bewusst machen und darüber Rechenschaft ablegen muss. Nur so kann man auch den Erfolg messen und nach aussen begründen. Dazu sollten – eventuell in Anlehnung an einzelne Sensibilitätsstufen (siehe Risikodifferenzierung) – geeignete Messkriterien entwickelt werden. Man könnte dies auch mit einer Begründungspflicht verbinden, wonach die Verwaltung begründen muss, wann und warum sie KI einsetzen will.

Aus den Diskussionen des zweiten Teils des Workshops ergaben sich folgende Prioritäten für den Themenbereich öffentliche Verwaltung:

1. Gewinnen eines Verständnisses darüber, welche KI-Anwendungen durch den Staat mehr oder weniger risikobehaftet sind.
2. Sicherstellung der Qualität der (staatlichen) Daten, die für KI-Anwendungen verwendet werden.
3. Sicherstellen der Transparenz des KI-Einsatzes dahin gehend, dass die Begründungspflicht, die dem Staat im Rahmen seines Verwaltungshandelns obliegt, weiterhin eingehalten werden kann.

4.7. Beurteilungen zum Themenfeld Ethik und Recht[123]

Einschätzungen zu ethischen und rechtlichen Aspekten wurden sowohl in der ersten als auch zweiten Umfrage erfasst. Diese Einschätzungen waren meist genereller Natur und bezogen sich nicht auf bestimmte KI-Technologien. Sie dienen also mehr zur Erfassung eines Stimmungsbildes bezüglich grundlegender ethischer und rechtlicher Aspekte von KI. Der Aspekt der «Verantwortung» hingegen wurde anhand einer spezifischen Fallvignette untersucht – ein Aspekt, der in der SATW-Bevölkerungsumfrage erneut aufgenommen wurde.

4.7.1. Zentrale Ergebnisse der ersten Umfrage

In der ersten Umfrage wurde zunächst geprüft, welche generellen Entwicklungsziele (orientiert an den 17 Zielen für nachhaltige Entwicklung der Agenda 2030 der Vereinten Nationen) die befragten Experten für wichtig erachten und inwiefern sie denken, dass KI die Erreichung dieser Ziele in der Tendenz eher fördert oder eher behindert. Danach wurden Einschätzungen zu einigen der wichtigsten ethischen und rechtlichen Bedenken bezüglich des Einsatzes von KI in sensitiven Entscheidungen abgefragt. Ein Fokus wurde dann auf den Bereich der Transparenz gelegt – zum einen hinsichtlich der Frage, wie viel Verständnis über die Funktionsweise man für bestimmte KI-Anwendungen haben sollte, und zum anderen, inwieweit Betroffene informiert werden sollten. Schliesslich wurden anhand eines Einsatzszenarios von KI in der Medizin Fragen von Vertrauen und Verantwortung untersucht – ein Aspekt, zu dem auch die in der Bevölkerungsumfrage gesammelten Daten hier vorgestellt werden.

4.7.1.1. Globale Entwicklungsziele und KI

Den Befragten wurden insgesamt zehn Entwicklungsziele vorgelegt: Armutsbekämpfung, Freiheit, Frieden, gesunde Umwelt, globale Zusammenarbeit, Rechtsstaatlichkeit, Sicherheit, Sicherung der Grundbedürfnisse (wie z.B. Essen oder Wohnen), soziale Beziehungen zu anderen Menschen und Zugang zu

[123] Dieser Abschnitt beruht auf Arbeiten von Markus Christen und Markus Kneer von der Digital Society Initiative der Universität Zürich.

Bildung und Information. Aus diesen konnten sich die Befragten die nach ihrer Ansicht fünf wichtigsten auswählen und dann dahin gehend beurteilen, ob KI das Erreichen dieses Ziels eher erleichtert oder eher behindert.

Das von den Fachpersonen am häufigsten genannte Entwicklungsziel war «Zugang zu Bildung und Information»; also jenes, zu dem KI generell der positivste Einfluss zugeschrieben wurde (Abbildung 34). Ein klar begünstigender Einfluss wird KI im Hinblick auf die Ziele «globale Zusammenarbeit» und «gesunde Umwelt» zugeschrieben. Ein negativer Effekt von KI wird vorab für die Bereiche «soziale Beziehungen» und «Freiheit» vermutet, was unter anderem Ausdruck der Angst einer durch KI verbesserten Überwachung von Menschen sein könnte.

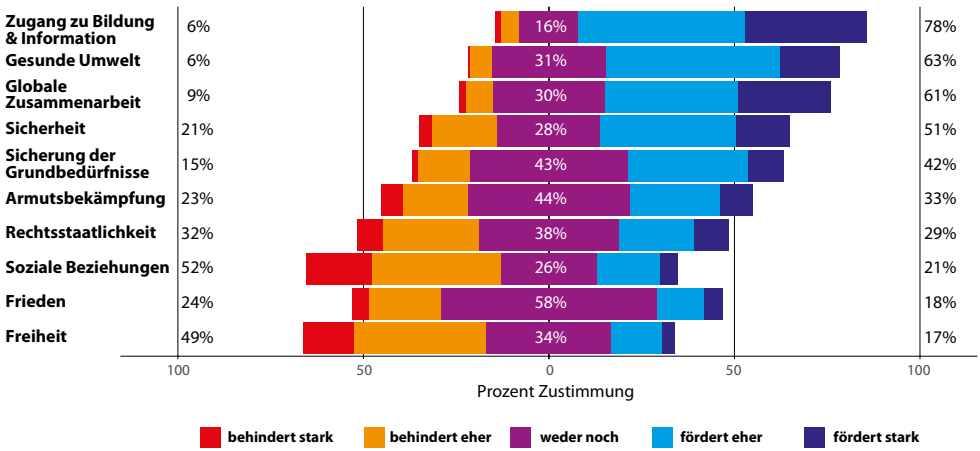

Abbildung 34: Einschätzung des Einflusses von KI auf die globalen Entwicklungsziele.

Signifikante Unterschiede zwischen den Gruppen finden sich nur selten. Bemerkenswert ist, dass nicht technische Experten KI eher als Bedrohung der Rechtsstaatlichkeit ansehen, während technische Experten hier eher eine Chance sehen.

4.7.1.2. Beurteilung ethischer und rechtlicher Bedenken

Die Experten nahmen dann Stellung zu einer Reihe von Bedenken, die in der ethischen (siehe Abschnitt 2.4) und rechtlichen (siehe Abschnitt 2.5) Debatte regelmässig geäussert werden. Es handelt sich dabei um Folgendes:

- Die mangelnde Nachvollziehbarkeit einer KI-Entscheidung (Blackbox)
- Die mögliche Unfairness der KI-generierten Entscheidungen (Fairness)
- Das Potenzial für Voreingenommenheit der KI aufgrund von Verzerrungen in den Trainingsdaten hinsichtlich z.B. Herkunft oder Geschlecht (Bias)
- Die mögliche Verschleierung der Tatsache, dass die Beurteilung durch KI und nicht durch einen Menschen vorgenommen wird (Intransparenz)
- Die Schwierigkeit, KI-Systeme zur Rechenschaft zu ziehen (Verantwortung)
- Die Möglichkeit, dass KI-Systeme Menschen täuschen und irreführen (Täuschung)

Die Ergebnisse (Abbildung 35) zeigen, dass im Schnitt bezüglich sämtlicher Aspekte Bedenken geäussert werden. Die grössten Bedenken bestehen bezüglich des Daten-Bias, der Verantwortung und des Blackbox-Problems.

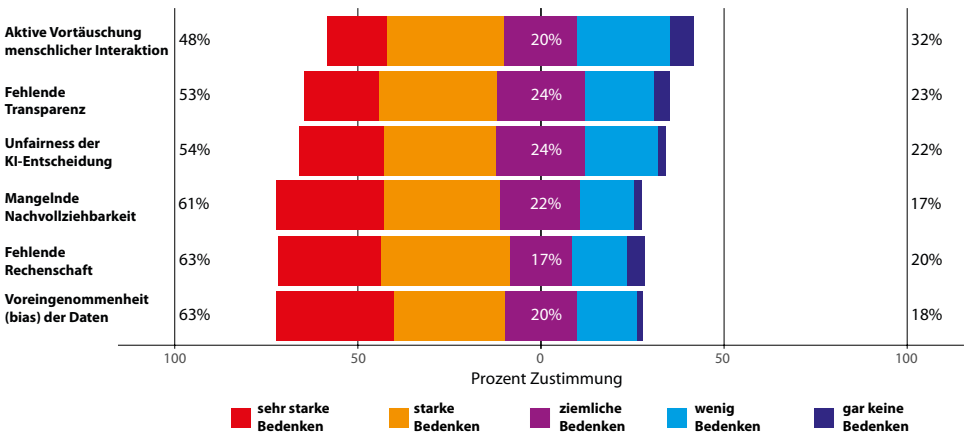

Abbildung 35: Einschätzung ethischer und rechtlicher Bedenken gegenüber KI.

4.7.1.3. Beurteilung der Transparenzfrage

Hinsichtlich der Transparenz des KI-Einsatzes nahmen die Befragten zu einer Reihe von Szenarien Stellung, bei denen KI für spezifische Zwecke verwendet wird. Gefragt wurde, inwiefern der Anbieter der Dienstleistung wissen sollte, wie genau das von ihm eingesetzte KI-System zu einer Lösung gekommen ist und ob die vom System betroffene Person informiert werden sollte, dass KI im Entscheidungsprozess involviert war. Das Ergebnis zeigt (Abbildung 36), dass Anbieter – mit der erwarteten Ausnahme von KI-Übersetzungen – alle Anwendungen mindestens in ihrem statistischen Verhalten verstehen sollten (Werte um 2). Bei den mutmasslich sensibelsten Anwendungen wie z.B. im Strafverfahren oder Gesundheitswesen sollte eine Erklärung der Entscheidungen im Einzelfall (Werte nahe 3) möglich sein. Analog präsentiert sich der Anteil der Personen, die verlangen, dass die betroffenen Menschen jeweils über den Einbezug von KI informiert werden sollten. Mit Ausnahme der Übersetzung liegen die Werte deutlich über 50 % und steigen mit der Sensitivität der Anwendung.

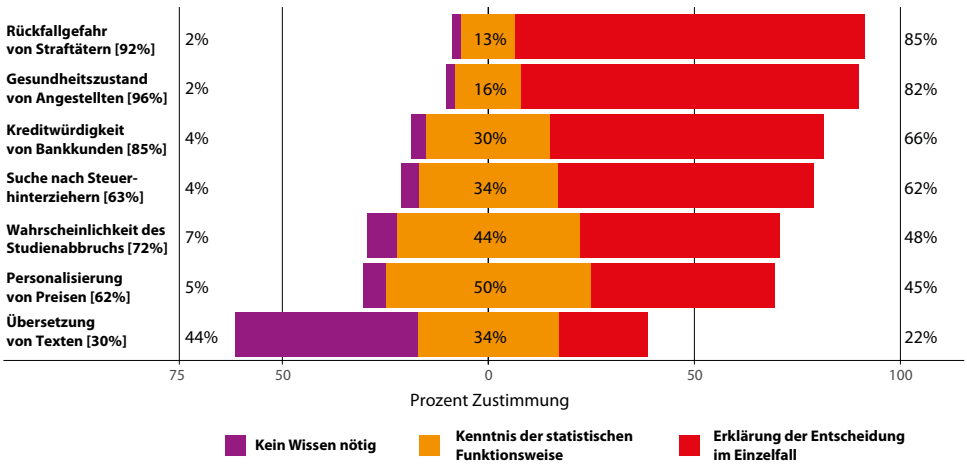

Abbildung 36: Einschätzung der Transparenzerfordernisse für KI-Anwendungen. Die Prozentzahlen in eckigen Klammern geben den Anteil an Zustimmung an, dass die Betroffenen über den KI-Einsatz informiert werden sollen.

4.7.1.4. Beurteilung von Vertrauen und Verantwortung

Am Schluss der ersten Umfrage beurteilten die Expertinnen und Experten schliesslich ein fiktives Szenario der Anwendung von KI im Gesundheitswesen. Das Szenario liest sich wie folgt:

> *Dr. Schmitt ist der leitende Arzt in einer Hautklinik.*
>
> *Die Klinik hat kürzlich die Software CANCERFIX gekauft, die sich künstlicher Intelligenz auf Basis neuronaler Netzwerke bedient, um Hautkrebs zu diagnostizieren. Die Software wurde mit über 400 000 Bildern von Hautläsionen trainiert.*
>
> *Die Rate der korrekten Diagnose liegt bei 95 %, der genau gleichen Rate wie jener des Hautarztes Dr. Franke.*

Die Befragten wurden zuerst gefragt, welcher Diagnose sie mehr vertrauen würden – jener des Arztes, jener der KI oder beiden gleich (was gemäss der Beschreibung die rationale Antwort wäre). In einem zweiten Schritt wurde das Szenario so ergänzt, dass entweder die KI oder der Arzt (die Befragten erhielten zufällig das eine oder andere Szenario) einen Diagnosefehler begehen, dem eine Patientin zum Opfer fällt. Danach wurde gefragt, welchen Grad an moralischer Verantwortung die involvierten Personen (je nach Szenario: leitender Arzt, diagnostizierender Arzt, diagnostizierende KI, KI-Trainer, KI-Programmierer) für den Fehler tragen.

In der Bevölkerungsumfrage wurde das medizinische Szenario durch zwei weitere, strukturell gleiche Szenarien ergänzt. Im ersten Fall geht es um Personalselektion (durch einen Menschen bzw. eine KI), in zweiten Fall um das Löschen von Fake News (durch einen Menschen bzw. eine KI). Auch in diesen Szenarien geschehen dann Fehler mit gravierenden Konsequenzen und die Verantwortlichkeiten werden erfragt.

Die Ergebnisse zeigen eine deutliche Kontextabhängigkeit des Vertrauens in die Entscheidungsgüte von KI-Anwendungen. Im medizinischen Szenario unterscheiden sich Fachpersonen und allgemeine Bevölkerung kaum – Letztere haben eine Tendenz, stärker dem Menschen zu vertrauen. Interessant ist, dass der KI in der Anwendung «Personalselektion» klar weniger vertraut wird, während in der Anwendung «Fake-News-Entdeckung» der KI deutlich mehr zugetraut wird.

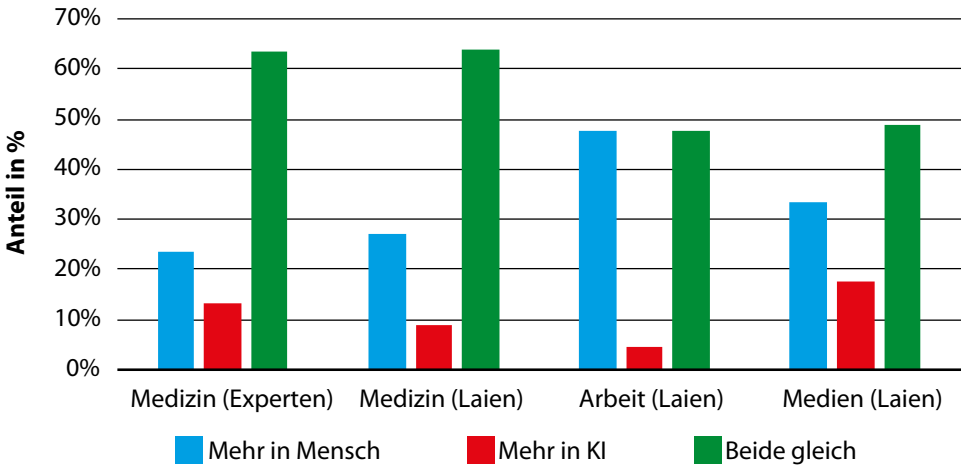

Abbildung 37: Einschätzung des Vertrauens in KI-Anwendungen der Fachpersonen und der Bevölkerung.

In der Verantwortungsverteilung im Fall eines Fehlers finden sich dann deutlichere Unterschiede zwischen allgemeiner Bevölkerung und den Fachpersonen. Wenn die KI einen Fehler macht, hält Erstere sowohl das System selbst als auch die involvierten technischen Experten (Trainer, Programmierer) tendenziell für stärker verantwortlich. Interessanterweise ist für die allgemeine Bevölkerung die Verantwortungslast beim Menschen im Fall eines menschlichen Diagnosefehlers gleich gross wie beim Diagnosefehler der KI. Unabhängig davon liegt aber bei allen drei Szenarien die grösste Last der Verantwortung im Fall eines KI-Fehlers bei jener Person, die entschieden hat, die KI einzusetzen.

Vertrauensunterschiede scheinen teilweise auch Effekte zu haben. So sind im Fall der Medien, wo das Vertrauen in die KI am grössten ist, auch die Verantwortungswerte der KI bzw. in die involvierten Techniker am grössten.

4.7.1.5. Zusammenfassendes Fazit

Die Antworten der Fachpersonen zu den ethischen und rechtlichen Aspekten sind in zweierlei Hinsicht relevant: Zum einen besteht durchaus eine wohlwollende Grundstimmung gegenüber KI im Hinblick auf die zentralen Entwicklungsziele der Vereinten Nationen. Für die meisten dieser Ziele rechnet man generell mit einem eher positiven Einfluss von KI. Zum anderen bestätigen die Ansichten der Fachpersonen die in den diversen ethischen Richtlinien als problematisch identifizierten Aspekte von KI: So werden vorab ein möglicher Daten-Bias sowie Intransparenz in der Art und Weise, wie eine KI zu einer Entscheidung kommt, als problematisch beurteilt. Gerade bei sensitiven Entscheidungen werden entsprechende Transparenzforderungen laut. Vertrauen wie auch die Zuschreibung der Verantwortlichkeiten bei der Nutzung von KI dürften stark vom jeweiligen Kontext abhängen.

4.7.2. Zentrale Ergebnisse der zweiten Umfrage

In der zweiten Umfrage waren generelle ethische und rechtliche Aspekte kaum mehr ein Thema. Hier wurde zum einen in Abhängigkeit von den fünf Themenbereichen gefragt, wer die Hauptverantwortung der Umsetzung von Massnahmen zur Risikokontrolle bzw. Förderung von KI-Anwendungen trägt. Zum anderen wurde detaillierter nachgefragt, wo im Fall von KI-Fehlern die Hauptlast der moralischen Verantwortung liegen soll. Alle 111 Personen haben hierzu ihre Einschätzungen abgegeben.

4.7.2.1. Verantwortung für die Umsetzung von Massnahmen

Es wurden für jeden der fünf Anwendungsbereiche drei Hauptgruppen als möglicher Hauptadressat für die Umsetzung von Massnahmen benannt: der Staat bzw. Gesetzgeber, die involvierten Unternehmen oder die Nutzer/-innen (Bürger, Konsumentinnen etc.). Bei den Ergebnissen zeigen sich deutliche Unterschiede zwischen den Anwendungsfeldern. Erwartungsgemäss ist der Staat der Hauptadressat für Massnahmen im Bereich Verwaltung. Im Bereich Konsum wird dem Konsumenten die grösste Zuständigkeit zugewiesen, während in den Medien die involvierten Unternehmen die grösste Verantwortung tragen sollen (Abbildung 38).

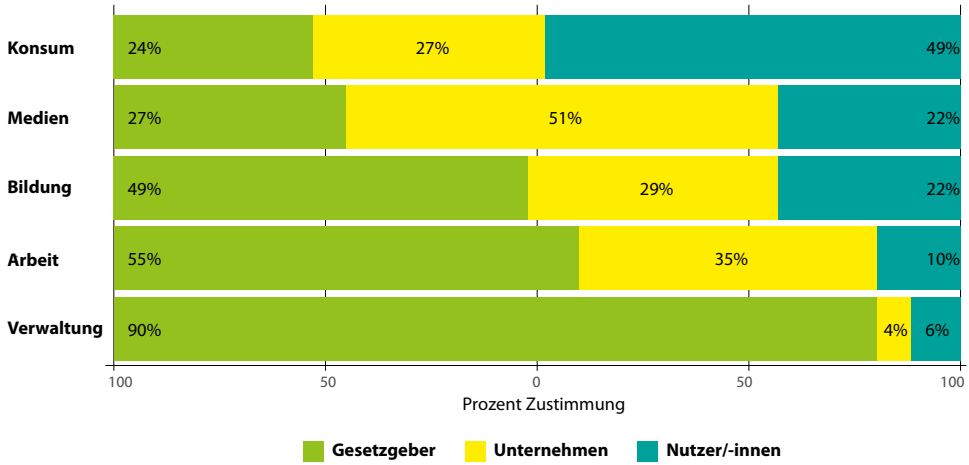

Abbildung 38: Einschätzung der Zuständigkeit für die Umsetzung von Massnahmen pro Themenbereich.

4.7.2.2. Verantwortung für die Umsetzung von Massnahmen

Schliesslich wurden aufgrund der teilweise überraschenden Antworten bei den vorgelegten Entscheidungsszenarien in der zweiten Umfrage noch Einschätzungen zur moralischen Verantwortung bei KI-Fehlern erhoben, wobei folgende Optionen zur Auswahl standen:

1. Die Beurteilungen des KI-Systems sind immer nur als Empfehlung zu verstehen, die von einem Menschen geprüft werden müssen. Die moralische Verantwortung für die Entscheidung liegt allein beim Nutzer.
2. In gewissen Bereichen werden KI-Systeme zur Entscheidungshilfe mindestens so hohe Erfolgsraten wie Experten haben. Nutzt eine Person diese Beratungssysteme nicht, muss sie dafür Rechenschaft ablegen.
3. Den Entwickler/Hersteller eines KI-Systems trifft keine moralische Verantwortung, wenn das System durch den Gebrauch des Nutzers mit unzureichenden Daten trainiert wurde und daher falsche Entscheidungen trifft.

Die Befragten konnten ihre Ablehnung bzw. Zustimmung mit einer 5-Punkte-Likert-Skala (von «lehne stark ab» bis zu «stimme stark zu») kenntlich machen. Die Ergebnisse sehen wie folgt aus (Abbildung 39):

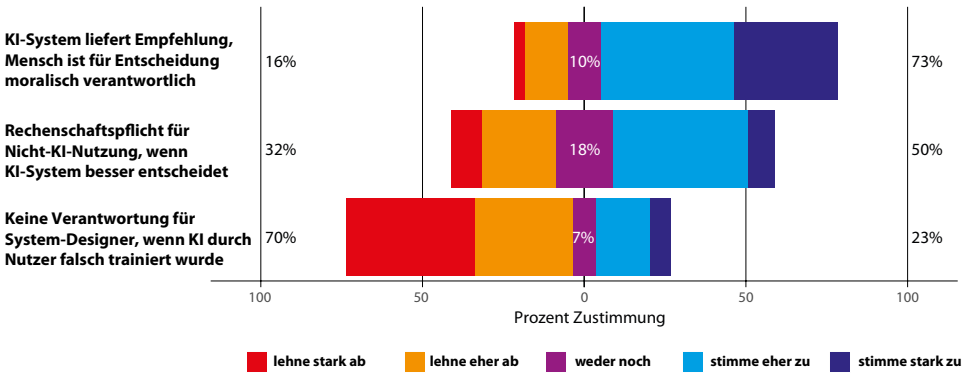

Abbildung 39: Einschätzung von Verantwortlichkeitsfragen.

Es zeigt sich auch hier der klare Wunsch, dass die Verantwortlichkeit für KI-Entscheidungen beim Mensch liegen muss, wobei aber Unterschiede bezüglich KI-Expertise und KI-Skepsis ersichtlich sind: Je höher die Expertise, desto stärker wird eine Rechenschaftspflicht gewünscht (Korrelation 0.26, $p < 0.01$), während höhere KI-Skepsis mit einer Ablehnung der Rechenschaftspflicht korreliert (– 0.23, $p = 0.02$). KI-Skeptiker sehen auch eine höhere Verantwortung beim Hersteller (– 0.22, $p = 0.03$).

4.7.3. Zusammenfassende Beurteilung der Expertenmeinungen

Die in der Umfrage zutage getretenen ethischen Einschätzungen resultieren nicht in Empfehlungen, sondern bestimmen gewissermassen deren Tonlage. So stösst eine insgesamt negative Einschätzung von KI auf Ablehnung, d.h. Empfehlungen sollten Chancen wie Risiken gleichermassen betreffen. Generell sollten die Empfehlungen Transparenz in mehrfacher Hinsicht betonen, sobald KI in Entscheidungen eingebunden ist, die Menschen betreffen: Nutzerinnen und Nutzer sollten wissen, wenn sie mit KI interagieren, und sollten in der Lage sein, die Entscheidungsfindung von KI-Systemen ausreichend zu verstehen.

5. Empfehlungen

Nachfolgend werden Empfehlungen zuhanden des Gesetzgebers bzw. der Bundesbehörden sowie anderer Stakeholder (z.B. Unternehmen oder Öffentlichkeit) vorgestellt. Angesichts der Vielfalt an Themen, die durch diese Studie angesprochen wird, und der zahlreichen offenen Fragen gilt es, eine Balance zwischen dem Vorsorgeprinzip zur Vermeidung von Risiken und der Gefahr der regulatorischen Verhinderung vielversprechender technischer Entwicklungen zu finden. Zu beachten ist dabei, dass in einem gesellschaftlichen Prozess geklärt werden muss, welche Risiken als gravierend einzustufen sind, und dass das Fehlen von Regulation auch Innovationen verhindern kann; etwa wenn Ängste vor dem Verlust von Arbeitsplätzen Akzeptanzprobleme verursachen. Deshalb beschränken sich die Empfehlungen nicht nur auf regulatorische Aspekte, sondern sie benennen auch Diskussionsbedarf, verweisen auf die Förderung technischer Lösungen für die Unterstützung von bestehendem Recht und schlagen Massnahmen vor, wie einzelne Akteure befähigt werden, ihre Aufgaben angesichts der Herausforderungen von KI besser wahrnehmen zu können. Zu jeder Empfehlung wird eine kurze Erläuterung gegeben und der Adressat spezifiziert.

5.1. Bereichsübergreifende Empfehlungen

Die bereichsübergreifenden Empfehlungen sind nach sachlogischen Gesichtspunkten geordnet und widerspiegeln keine Priorisierung.

Empfehlung 1	**Der Gesetzgeber soll im Bereich der KI eine technologieneutrale und bereichsspezifische Herangehensweise verfolgen: Anstelle eines allgemeinen «KI-Gesetzes» sollen bereichsbezogen konkrete Probleme und Fehlentwicklungen in regelmässigen Abständen identifiziert, evaluiert und gegebenenfalls mittels geeigneter Rechtsnormen gelöst werden. Dabei sollen insbesondere auch die Entwicklungen in der Europäischen Union Beachtung finden.**

Erläuterung: Bei KI handelt es sich um eine Basistechnologie, die über grosses wirtschaftliches und gesellschaftliches Potenzial verfügt, aber auch Herausforderungen und Risiken mit sich bringt. Die Anwendungsformen von KI sowohl durch staatliche als auch private Akteure sind äusserst vielfältig und lassen sich, was zukünftige Einsatzmöglichkeiten anbelangt, aus heutiger Sicht schwer abschätzen, zumal viele experimentelle Anwendungen noch weit von marktfähigen Produkten entfernt sind. Chancen und Risiken der Basistechnologie KI lassen sich nicht generell beurteilen, weil es entscheidend vom Anwendungskontext (d.h., wer nutzt ein KI-System für was genau, unter Nutzung welcher Daten und unter Berücksichtigung welcher rechtlicher Vorgaben?), aber auch von der konkreten KI-Technologie abhängt, inwieweit beispielsweise negative Folgen zu erwarten sind. Entsprechend werfen die Entwicklung und Anwendung von KI unterschiedliche Rechtsfragen auf, die keiner einheitlichen Regelung zugeführt werden können. Der Erlass eines allgemeinen «KI-Gesetzes» wäre daher nicht zweckdienlich. Vielmehr müssen die Folgen spezifischer Anwendungen in regelmässigen Abständen bereichsbezogen untersucht, konkrete Probleme und Fehlentwicklungen identifiziert und diese – soweit sinnvoll – unter Berücksichtigung der gesamtgesellschaftlichen und volkswirtschaftlichen Interessen einer Regulierung zugeführt werden.

Eine entsprechende regelmässige Überprüfung sollte im Rahmen der «Strategie Digitale Schweiz» eingebunden werden. Zu beachten ist dabei auch die internationale Entwicklung, insbesondere in der Europäischen Union. So hat die neue Präsidentin der EU-Kommission Ursula von der Leyen angekündigt, sie wolle binnen der ersten 100 Amtstage der neuen Kommission (Start: Dezember 2019) eine Gesetzesinitiative für einen «koordinierten Ansatz für die menschlichen und ethischen Auswirkungen der Künstlichen Intelligenz» (KI) auf den Weg bringen.

Adressat: Gesetzgeber.

Empfehlung 2	**Der Gesetzgeber soll den Einsatz von KI-Systemen nicht als datenschutzrechtliches Problem auffassen. Zwar greift der Datenschutz, wenn Personendaten zur Entwicklung und Anwendung von KI genutzt werden. Risiken, die durch die Nutzung von Sachdaten entstehen, oder Diskriminierung als Folge der Datenbearbeitung brauchen aber neue bzw. andere Ansätze, die entworfen und weiterentwickelt werden sollten.**

Erläuterung: Die Vielfalt der möglichen Anwendungsfelder und die Unterschiedlichkeit der durch den KI-Einsatz aufgeworfenen Rechtsfragen sprechen auch gegen einen umfassenden Regelungsansatz im Rahmen des Datenschutzgesetzes. Unbestritten ist, dass im Rahmen der Entwicklung und Anwendung von KI auch Personendaten genutzt werden können. Selbstverständlich sind die Vorgaben des Datenschutzrechts entsprechend bei der Entwicklung und Nutzung von KI einzuhalten. Insbesondere ist der Umstand zu berücksichtigen, dass im Rahmen von KI vermehrt sogenannte *inferred data* anfallen können (Abschnitt 2.5.3). Können diese Daten auf konkrete Personen bezogen werden, gilt das einschlägige Datenschutzrecht.

Das breite Anwendungsfeld des Datenschutzrechts darf den Gesetzgeber aber nicht dazu verleiten, den KI-Einsatz als primär datenschutzrechtliches Problem zu verstehen und entsprechend regulatorische Fragen der KI allein durch diesen Ansatz lösen zu wollen. Der heutige Ansatz des Datenschutzrechts ist auf den Vorgang der Bearbeitung von Personendaten ausgerichtet und weitgehend blind für die Folgen dieser Datenbearbeitungen. Gerade diese sind aber für die betroffenen Personen relevant – im Bereich der Nutzung von KI ebenso wie anderswo.

Auch die im Rahmen der Revision des DSG vorgesehene Bestimmung betreffend automatisierte Einzelentscheidungen (Art. 19 E-DSG) ist im Kontext von KI-Systemen nur in einem sehr eng zugeschnittenen Anwendungskreis von Bedeutung. Der Regelungsentwurf bezieht sich auf Entscheide, die ausschliesslich auf einer automatisierten Bearbeitung von Personendaten beruhen. Dies ist der Fall, «wenn keine inhaltliche Bewertung und darauf gestützte Entscheidung durch eine natürliche Person stattgefunden hat. Das heisst, die inhaltliche Beurteilung des Sachverhalts, auf dem die Entscheidung beruht, erfolgte ohne Dazutun einer natürlichen Person. Darüber hinaus wird auch der Entscheid, der auf der Basis dieser Sachverhaltsbeurteilung ergeht, nicht von einer natürlichen Person getroffen.»[124] Mit anderen Worten: Es geht in Art. 19 E-DSG um Vorgänge, bei denen keinerlei menschliche Intervention erfolgt. Dies wird angesichts der in der Praxis eingesetzten oder diskutierten KI-Anwendungen relativ selten der Fall sein (vgl. z.B. Abschnitt 3.5.2).

Zudem ist das Datenschutzrecht nicht geeignet, um alle mit dem Einsatz von KI verbundenen Herausforderungen und Probleme zu erfassen, weil sich diese auch

[124] BBl 2017 6941 (7056 f.). Vgl. auch Abschnitt 3.5.3.1.

aus der Bearbeitung von Sachdaten ergeben können, die nicht in den Anwendungsbereich des Datenschutzgesetzes fallen. Aus diesen Gründen ist der Datenschutz ein unzureichender Ansatz, um KI-Risiken umfassend zu kontrollieren; er kann lediglich einen partiellen Bereich abdecken. Mit Blick auf mögliche künftige Risiken von KI-Anwendungen ist dabei insbesondere der Aspekt der Diskriminierung zu beachten. Entsprechend soll der Gesetzgeber prüfen, inwieweit die geltenden Regulierungen im Bereich Diskriminierung angepasst bzw. ergänzt werden sollten.

Adressat: Gesetzgeber.

Empfehlung 3	**Gesetz- und Verordnungsgeber sollen sicherstellen, dass für staatliche Akteure (z.B. Gerichte, Polizei und Verwaltung) höhere Anforderungen an die Nutzung von KI gelten als für Private, wenn hoheitliche KI-Nutzung Menschen in relevanter Weise betrifft: In diesem Fall muss stets gewährleistet sein, dass die Betroffenen die Rechtmässigkeit des staatlichen Handelns beurteilen können.**

Erläuterung: Wie in den Abschnitten 2.8.1 und 2.8.2 ausgeführt, unterscheidet sich der staatliche KI-Einsatz in fundamentaler Art und Weise von jenem durch private Akteure. Dies schlägt sich auch in den Anforderungen an den staatlichen KI-Einsatz nieder. Handelt der Staat hoheitlich, also aus einer gegenüber den Bürgerinnen und Bürgern übergeordneten Stellung heraus, so sind den Betroffenen die notwendigen Anhaltspunkte dafür zu gewähren, dass sie beurteilen können, ob der Staat unrechtmässig handelt. Nur so können sie bei Bedarf ein Rechtsmittel ergreifen. Über die existierenden Verfahrensvorgaben[125] hinaus hat der Staat deshalb transparent zu machen, *ob* er KI-Systeme einsetzt, *welche Personendaten* er nutzt und *wie* diese Systeme funktionieren. Denkbar ist, dass er dies an zentraler Stelle, etwa über ein öffentlich zugängliches Register, macht. Je nachdem, welches KI-System zum Einsatz kommt, sind zusätzliche Transparenzvorgaben zu berücksichtigen. Zudem gilt sinngemäss Empfehlung 4 auch für staatliche Stellen. Die Nutzung von KI im alltäglichen Verwaltungshandeln hingegen, die Menschen (wenn überhaupt) nur mittelbar betrifft (z.B. für Übersetzungen), unterliegt keinen

[125] Vgl. nur das Bundesgesetz vom 20.12.1968 über das Verwaltungsverfahren (VwVG), SR 172.021.

zusätzlichen Transparenzpflichten. Mit Blick auf Realakte wird zudem auf Empfehlung V-1 verwiesen. Insgesamt ist festzuhalten, dass die Bindung an die Grundrechte und an das Legalitätsprinzip selbstverständlich auch für den staatlichen KI-Einsatz gelten.

Adressat: Staatliche Verwaltungsstellen, Gesetzgeber.

Empfehlung 4	**Setzen Private (Unternehmen und andere Organisationen) KI für Entscheidungen ein, die Menschen in relevanter Weise betreffen, sollen sie neben der aktiven Information über den Umstand des KI-Einsatzes auch eine Transparenz auf Nachfrage sicherstellen. Über die Massgabe des Datenschutzrechtes hinaus sollen beispielsweise Institutionen der Zivilgesellschaft auf Nachfrage alle Informationen erhalten, die eine Einschätzung möglicher Fehlentwicklungen erlauben. Der Schutz der Geschäftsgeheimnisse der Unternehmen und anderen Organisationen ist dabei in angemessener Weise zu gewährleisten.**

Erläuterung: Transparenz darüber, dass und wie genau ein KI-System in eine Entscheidung involviert ist (im Sinn einer Entscheidungsunterstützung oder gar -automatisierung), wird sowohl in der Literatur als auch von den in der Studie befragten Fachpersonen gefordert. Die Forderung nach Transparenz wird umso höher gewichtet, je mehr die Entscheidungen relevant sind, d.h. wichtige Lebensvollzüge von Menschen betreffen (z.B. im Bereich Gesundheit, Kreditvergabe oder Versicherungen). Eine auf die betroffene Person ausgerichtete Transparenz ist für automatisierte Einzelfallentscheide im Datenschutzrecht der EU verankert; eine vergleichbare Bestimmung findet sich im DSG-Entwurf. Nebst dem Recht auf Information darüber, dass ein Entscheid automatisiert getroffen wurde, soll die betroffene Person gemäss DSG-E im Falle von automatisierten Einzelentscheidungen auch einen Anspruch auf Stellungnahme und auf Beurteilung durch eine natürliche Person haben. Diese Vorgaben verlangen aber nicht, dass darüber informiert wird, ob ein KI-System zum Einsatz kam oder welche Art der KI-Technologie eingesetzt wurde.

Hier setzt die Empfehlung 4 an. Die Art der eingesetzten KI-Technologie erschwert die Umsetzung von Transparenz (siehe Abschnitt 2.2.4), Menschen unterscheiden

sich bezüglich Expertise über KI und auch der Wunsch nach Transparenz hängt stark von der Art der Anwendung ab. Zudem ist aus der Forschung bekannt, dass ein Übermass an Information oft den gegenteiligen Effekt bei den Betroffenen hat (Informationsüberflutung). Deshalb ist es zwar begrüssenswert, dass Nutzerinnen und Nutzer von KI-Systemen in geeigneter Weise informiert werden, wenn sie mit KI interagieren. Wie genau diese Information aber erfolgen soll, wird sich je nach Kontext unterscheiden und muss sich an typischen Nutzergruppen und gängigen Erwartungen orientieren. Deshalb ist der Fokus weniger auf eine generelle Information der Betroffenen zu legen, sondern auf die *Sicherstellung einer Transparenz auf Nachfrage*, sodass insbesondere Akteure wie Konsumentenschutzorganisationen, Vertreter der Zivilgesellschaft, Journalisten, Wissenschaftler etc. in die Lage versetzt werden, Indizien für mögliche Fehlentwicklungen bei der KI-Nutzung zu überprüfen. Eine solche Transparenz auf Nachfrage impliziert, dass die Nutzer/-innen von KI-Systemen interne Kontrollstrukturen aufbauen müssen, damit eine solche Transparenz dann im konkreten Fall auch erreicht werden kann.

Werden KI-Systeme in kritischen Anwendungen eingesetzt, in denen z.B. Menschen zu Schaden kommen können, sollten die Systeme so aufgesetzt sein, dass Ursachen und Verantwortlichkeiten dieser Fehlfunktion retrospektiv derart bestimmt werden können, um allfällige Haftungsfragen (notfalls gerichtlich) klären und technische Verbesserungen für künftige Anwendungen entwickeln zu können.

Adressat: Regulator, Unternehmen und andere Nutzer/-innen von KI-Systemen.

Empfehlung 5	**Organisationen (beispielsweise im Bereich Konsumentenschutz) sollen durch staatliche Unterstützung besser befähigt werden, private KI-Zertifizierungen zu prüfen. Private Initiativen für KI-Zertifizierung und die Vergabe entsprechender Labels sind zu begrüssen. Die Nutzung von KI-Systemen soll aber nicht generell von einer Marktzulassung abhängig gemacht werden.**

Erläuterung: Zunehmend werden branchenübergreifende KI-Anwendungen von Unternehmungen aus unterschiedlichen ökonomischen Sektoren und Ländern entwickelt. Einige dieser Anwendungen – ein aktuelles Beispiel ist die Entwicklung von autonomen Fahrzeugen – werden eine Form von Marktzulassung benötigen. Die Frage aber, ob eine solche Marktzulassung nötig ist, sollte nicht von der Tatsache abhängen, ob eine bestimmte Anwendung KI-Systeme involviert oder nicht.

Die Frage des Erfordernisses einer Marktzulassung stellt sich bei einer besonders hohen Gefährdung von Menschen durch konkrete Systeme (z.B. Autos oder Drohnen) und nicht aufgrund der dahinterstehenden Technologie. Zu beachten ist dabei allerdings, dass gewisse Eigenschaften von KI-Systemen das Gefährdungspotenzial erhöhen oder die Prüfung der Sicherheit von Produkten und damit eine Marktzulassung einer konkreten Anwendung erschweren können (z.B. das Blackbox-Problem). Dennoch ist klar, dass eine Abschwächung der Kriterien für eine Zulassung aufgrund solcher KI-Spezifika abzulehnen ist – KI-Systeme haben sich den Kriterien der Marktzulassung anzupassen, nicht umgekehrt.

Des Weiteren ist die (staatlich definierte) Marktzulassung von einer allfälligen Zertifizierung von KI-Systemen zu unterscheiden, und Letztere sollte keine Bedingung für die Nutzung von KI-Systemen sein. Hingegen dürfte es durchaus Anwendungsbereiche geben (z.B. im Bereich Bildung), in denen (künftig) Zertifikate von den Nutzern von KI-Systemen erwünscht oder gar gefordert werden. In solchen Bereichen kann die öffentliche Verwaltung im Beschaffungswesen durch entsprechende Vorgaben die Schaffung von Zertifikaten fördern.

Aufgrund der Vielzahl an Anwendungsbereichen und Faktoren, die bei einer Zertifizierung Berücksichtigung finden können, sollte die Entwicklung von Zertifizierungen – wie in anderen Bereichen auch – Privaten überlassen werden. Solche Zertifizierungen können z.B. erfassen, welche Daten zu welchem Zweck verarbeitet werden und wie die einzelnen Faktoren gewichtet werden, wie hoch die statistische Fehlerwahrscheinlichkeit ist, wie das statistische Input-Output-Verhalten aussieht und ob ein möglicher Bias zu erwarten ist. Auch ethische Aspekte können hier eine Rolle spielen, wobei die Vielzahl bereits bestehender Richtlinien (siehe Abschnitt 2.3) als Orientierungspunkt gelten können. Die konkrete Ausgestaltung wird von der verwendeten KI-Technologie und der Anwendung abhängen.

Private KI-Zertifikate können unter Umständen falsche Vorstellungen wecken und die Konsumenten in die Irre führen. Das Wettbewerbsrecht (UWG) hält zwar an sich ausreichende Möglichkeiten bereit, um gegen solche Zertifikate vorzugehen, es besteht aber ein gewisses Vollzugsproblem, weil es Kontrollorganisationen, wie z.B. dem Konsumentenschutz, oft an Mitteln und Kompetenz mangelt, um solche (und andere) Prüfungen durchzuführen. Deshalb sollten diese Organisationen mit geeigneten Mitteln ausgestattet und besser befähigt werden, um (künftige) KI-Zertifikate für definierte Anwendungskontexte prüfen zu können.

Adressat: Regulator, Unternehmen, Konsumentenschutz-Organisationen.

Empfehlung 6	**Hochschulen und weitere Bildungsinstitutionen, welche KI-Fachleute ausbilden, sollen auch nicht technische Kompetenzen fördern: Fachleute, welche KI-Systeme entwickeln, implementieren oder über deren Einsatz entscheiden, sollen sich Kenntnisse über rechtliche, ethische und soziale Aspekte der Nutzung von KI aneignen.**

Erläuterung: Die Beurteilung der meisten möglichen Risiken, welche die Nutzung von KI-Systemen in bestimmten Anwendungen haben können, erfordert ein vertieftes Verständnis der technischen Grundlagen. Deshalb tragen die Entwickler und Designer von KI-Technologien wie auch Entscheidungsträger eine besondere Verantwortung, die sich in der Ausbildung dieser Fachleute niederschlagen sollte. So sollte beispielsweise die Prüfung von Trainingsdaten bezüglich eines möglichen Daten-Bias zur «good practice» im Bereich des Maschinenlernens werden und fester Bestandteil in der Ausbildung von KI-Fachleuten werden. Auch das Thema KI-Entscheide und Fairness sollte Gegenstand der Ausbildung von Computerfachleuten wie auch Technologiemanagern sein, weil es über Aspekte wie *computational thinking* und Programmieren hinausgeht. Es bedarf in der Ausbildung dieser Fachpersonen deshalb auch eines ethischen Verständnisses, interdisziplinären Denkens und Handelns sowie grundlegender rechtlicher Kenntnisse.

Adressat: Hochschulen, Institutionen der beruflichen Aus- und Weiterbildung.

Empfehlung 7	**Bund, Hochschulen, Unternehmen und zivilgesellschaftliche Organisationen sollen gemeinsam den gesellschaftlichen Dialog über Chancen und Risiken der KI fördern. In Bereichen mit unklarer Risikolage muss dabei auch die Forschung zur Erkennung solcher Risiken intensiviert werden, was durch entsprechende Massnahmen der Hochschulen und Institutionen der Drittmittelförderung unterstützt werden soll.**

Erläuterung: Die grosse Zahl an KI-Anwendungen, welche derzeit erforscht, experimentell eingesetzt oder bereits breit angewendet werden, führt zu einer unübersichtlichen öffentlichen Debatte, die oft von übermässigen Ängsten oder übertriebenen Hoffnungen geprägt ist. Die Einschätzung der Chancen und Risiken

einer Basistechnologie wie KI ist damit ein Unterfangen, das nicht nur die involvierten Fachleute betrifft, sondern eine Aufgabe der ganzen Gesellschaft ist. Dies bedarf eines zielgruppenorientierten, strukturierten Dialogs, an dem sich alle massgeblichen Stakeholder (Forschung, Unternehmen, Verbände, Verwaltung und zivilgesellschaftliche Organisationen) beteiligen sollten. Die bereits etablierten Foren (z.B. die «Nationale Konferenz Digitale Schweiz») sollten demnach genutzt werden, um diesen Dialog fortzusetzen. Für Gebiete mit noch unklarer Risikolage (z.B. Nutzung in den Medien; siehe bereichsspezifische Empfehlungen) sollte zudem auch die Forschung intensiviert werden. Gleichzeitig sollte der Bund im Interesse des Technologie- und Forschungsstandorts Schweiz Bemühungen unternehmen, um die Ausschöpfung des Potenzials der Entwicklung und Nutzung von KI durch den Staat und die Privatwirtschaft bestmöglich sicherzustellen.

Adressat: Bund, Forschung, Unternehmen, zivilgesellschaftliche Organisationen.

5.2. Bereichsspezifische Empfehlungen

5.2.1. Arbeitswelt

Empfehlung A-1	**Der Bund soll mögliche makroökonomische Auswirkungen der digitalen Transformation im Allgemeinen und von KI im Speziellen verstärkt zum Anlass nehmen, gesellschaftliche Debatten über Anpassungsprozesse anzustossen: Dies betrifft insbesondere die Bereiche Arbeitszeitverkürzung, Flexibilisierung der Arbeit hinsichtlich Zeit, Ort und Mittel, ökonomische Polarisierung und Weiterbildung.**

Erläuterung: Gerade hochentwickelte Länder wie die Schweiz können durch die Nutzung von KI-Systemen Produktivitätsgewinne erzielen. Dies bedingt aber auch eine soziale und wirtschaftliche Abfederung negativer Folgen dieser Transformationsprozesse sowie die Schaffung von Anreizen für eine breite Nutzung von KI bei der Produktion von Sachgütern und Dienstleistungen. Wie genau die Folgen aussehen, lässt sich schwer vorhersagen, und welche Massnahmen ergriffen werden sollten, wird von der politischen Ausrichtung der Diskussionspartner abhän-

gen. Entsprechend ist es wichtig, dass der Diskurs über solche Aspekte in aktuellen Initiativen wie der «Strategie Digitale Schweiz» angemessen berücksichtigt wird. Mögliche Diskussionsfelder sind beispielsweise, ob eine flexible Gestaltung und Reduktion von Arbeitszeiten zu einer Vermeidung von höheren Arbeitslosenraten beitragen kann, inwieweit möglicherweise fehlende Sozialversicherungsbeiträge durch neue Formen der Besteuerung kompensiert werden können oder wie die Weiterbildung von Arbeitnehmern gefördert werden kann. Gewiss bestehen viele dieser Themen unabhängig von KI – doch die künstliche Intelligenz kann als ein Faktor verstanden werden, der bestehende Trends verstärkt und beschleunigt. Entsprechend ist es zielführend, solche Debatten auch im Sinn von Empfehlung 7 am Beispiel der KI zu führen.

Adressat: Verantwortliche «Strategie Digitale Schweiz», Gesetzgeber.

Empfehlung A-2	**Der Gesetz- und Verordnungsgeber soll das Mitspracherecht der Mitarbeitenden sicherstellen, wenn Unternehmen KI-Systeme für deren Überwachung und Kontrolle einsetzen: Die Arbeitsinspektorate müssen angemessen ausgestattet werden, um die Einhaltung der gesetzlichen Bestimmungen effektiv kontrollieren zu können.**

Erläuterung: Um auf individueller Ebene Arbeitnehmer vor möglichen negativen Konsequenzen des Einsatzes von KI zu schützen und eine Teilhabe an den positiven Wirkungen zu eröffnen, ist sicherzustellen, dass die geltenden Mitwirkungsrechte der Arbeitnehmer bezüglich der Nutzung von Systemen für eine technische Überwachung[126] durchgesetzt werden. Unternehmen sollten hierbei auch Fachexperten, Arbeitsinspektorate, Gewerkschaften und die Personalvertretungen einbeziehen.

Adressat: Unternehmen, Arbeitsinspektorate.

[126] Art. 26 «Überwachung der Arbeitnehmer» der Verordnung 3 zum Arbeitsgesetz, SR 322.113

5.2.2. Bildung[127]

Empfehlung B-1	**Kantonale Gesetz- und Verordnungsgeber und die Erziehungsdirektoren sollen Leitlinien formulieren, wie mit Daten über Leistung und Verhalten von Lernenden und den daraus mittels KI-Systemen gewonnenen Schlüssen umgegangen werden soll: Insbesondere ist zu prüfen, ob und welche über den aktuellen Datenschutz hinausgehenden Mechanismen zu schaffen sind, um Lernende vor negativen Folgen der Nutzung und Bekanntgabe ihrer Lern- und Leistungsdaten an Dritte zu schützen.**

Erläuterung: Im Zuge der Nutzung von KI-Anwendungen zum Zweck des personalisierten Lernens werden grosse Mengen personenbezogener Daten generiert, gesammelt und gespeichert, und es werden daraus Schlüsse über mögliche Bildungspfade und Erfolgschancen aller Art gewonnen (*inferred data*). Diese Daten werden in einer sensitiven Phase des Lebens von Menschen gewonnen und haben das Potenzial, die Zukunft der Betroffenen weitreichend zu prägen, insbesondere bezüglich Berufschancen. Relevante Fragen, die hier diskutiert werden müssen, sind:

- Welche Daten werden erhoben?
- Aufgrund welcher Algorithmen und Kriterien werden Potenziale, Schwächen und weitere personenbezogene Auswertungen vorgenommen?
- Wo werden die Daten gespeichert (insbesondere wenn Dienstleistungen privater Unternehmen in Schulen verwendet werden)?
- Wie lange sollen solche Daten gespeichert werden?
- Wer soll Zugang zu solchen Daten haben (Schule, Lernende, Behörden [z.B. regionale Arbeitsvermittlungszentren], Unternehmen als potenzielle Arbeitgeber etc.)?

[127] Empfehlungen zum Themenbereich Forschung werden im Abschnitt 5.3 vorgestellt. Die berufliche Weiterbildung wird in Empfehlung A-1 aufgenommen.

- Sollen die Daten so aufbereitet werden, dass sie zwischen Bildungseinrichtungen portiert werden können? Welche Rolle kann dabei das durch die EDK im November 2019 lancierte edulog-System einnehmen?
- Welche Rolle sollen solche Daten im Bewerbungsprozess spielen dürfen? Soll es Unternehmen erlaubt sein, im Zuge von Stellenbewerbungen Einsicht in solche Daten zu erhalten oder gar zu verlangen?

Beispielsweise der letzte Punkt könnte auch im Konflikt mit geltenden datenschutzrechtlichen Bestimmungen geraten. Unter Nutzung der Datenportabilität könnte es Lernenden erlaubt sein, umfassende Profile einem Unternehmen selbst vorlegen zu dürfen, was einen generellen sozialen Druck zur Offenlegung dieser Daten erzeugen könnte. Mit Blick auf die Sicherung einer Zukunftsoffenheit für die Laufbahn junger Menschen könnte demnach eine Einschränkung des Grundsatzes der informationellen Selbstbestimmung diskutiert werden.

Adressat: Gesetzgeber und Erziehungsdirektoren der Kantone.

Empfehlung B-2	**Die Bildungsinstitutionen und insbesondere die pädagogischen Hochschulen sollen untersuchen, welche spezifischen Kompetenzen vermittelt werden müssen, um ein allgemeines Verständnis von Fähigkeiten und Grenzen von KI-Systemen zu erhalten: Entsprechende Erkenntnisse sollen in Lehrmittel einfliessen und unter Nutzung bestehender Plattformen für Lehrkräfte und Lernende verfügbar gemacht werden.**

Erläuterung: Kenntnisse über KI bilden Teil eines breiten Sets an *digital skills*, die zunehmend in den Fokus der schulischen Ausbildung kommen. Welche Kenntnisse und Kompetenzen genau dies sind und wie diese vermittelt werden können, ist aber noch weitgehend unklar. Deshalb sollen durch gezielt geförderte Pilotprojekte Erfahrungswerte geschaffen werden, um darauf beruhend die weitere Implementierung von KI-Lösungen im öffentlichen Bildungswesen voranzutreiben. Nationale KI-Bildungsangebote sollen zielgruppenorientierte Angebote darstellen, wofür bestehende Plattformen wie educa.ch genutzt werden können. Durch die Zusammenarbeit zwischen Schulen, KI-Forschung, Unternehmen der KI-Bildungsbranche und pädagogischen Hochschulen sollen einerseits KI-Themen in

Lehrmitteln fachübergreifend dargestellt werden, zum anderen auch KI-Anwendungen via Lernplattformen in der Pädagogik vermehrt zum Einsatz kommen. Dieser Einsatz muss durch pädagogische Hochschulen wissenschaftlich begleitet und evaluiert werden. Schliesslich sollen auch Lehrende aller Schulstufen regelmässig Weiterbildungen erhalten, die ihnen aufzeigen, welche Rolle KI-Anwendungen für ihr Fachgebiet spielen und wie sie den Umgang mit und die Funktionsweise von KI-Anwendungen den Lernenden näherbringen können. Die Weiterbildung sollte, wenn immer möglich, von Unternehmen oder Vereinen angeboten werden, die ihrerseits selbst keine KI-Anwendungen entwickeln/vertreiben, um Interessenkonflikte zu vermeiden.

Adressat: Erziehungsdirektoren der Kantone, pädagogische Hochschulen.

5.2.3. Konsum

Empfehlung K-1	**Unternehmen, welche KI-Systeme im Konsumbereich nutzen und dafür Personendaten erheben, sollen die Transparenz des KI-Einsatzes (Empfehlung 4) und die sonstigen Anforderungen an den Datenschutz möglichst einfach vermitteln. Entsprechende Forschung und Best Practices sind zu fördern.**

Erläuterung: Aktuell beruhen die meisten Ansätze zur Kontrolle allfälliger Risiken des KI-Einsatzes im Konsumbereich auf dem Datenschutzrecht, das zurzeit in der Schweiz revidiert wird und sich weitgehend an der EU-Datenschutz-Grundverordnung orientieren wird. Dortige Ansätze wie z.B. eine *privacy by default* (Art. 25 DSGVO) sind zwar begrüssenswert, doch es zeigen sich klare Grenzen des Datenschutzrechts für den Umgang mit KI-Risiken (siehe Empfehlung 2). Für den Konsumbereich ist vorab die Schaffung von Transparenz ein zentrales Erfordernis. Transparenz soll hierbei über eine Transparenz auf Nachfrage (vgl. Empfehlung 5) hinausgehen. Grundsätzlich sollen Konsumentinnen und Konsumenten einfach einsehen können, welche Aspekte des Service (wie zum Beispiel Preis, Werbung oder Inhalt) personalisiert werden. Weiter soll auf der Benutzeroberfläche rasch zugänglich sein (z.B. innerhalb von drei Klicks), wer welche Daten sammelt, zu welchem Zweck sie verwendet werden und auf welche Weise diese in die Personalisierung einfliessen. Auch die Benutzung von Daten von Drittparteien soll auf einfache Weise ersichtlich sein (dies gilt auch für Daten, die innerhalb eines An-

bieters geteilt werden; z.B. YouTube mit GoogleMaps etc.). Je nach Kontext sollten auch Korrektur bzw. Löschen von Falschinformationen unproblematisch umgesetzt werden können.

Eine vielversprechende Möglichkeit zur Verbesserung der Transparenz ist der sogenannte *layered approach*, bei welchem den Konsumentinnen und Konsumenten Informationen auf drei Stufen kommuniziert werden: (1) in Form von Piktogrammen, welche bestimmte Arten von Datenbearbeitungen symbolisieren, (2) durch Kurztexte, welche die Nutzung knapp und präzise erklären, und (3) in Form der heute gängigen, umfassenden Datenschutzerklärungen.

Des Weiteren sollte geprüft werden, welche technischen Lösungen auf einfache Weise informationelle Selbstbestimmung erlauben – etwa wenn Konsumenten eine Erlaubnis geben sollen, dass bestimmte Services Daten aufzeichnen (*opt-in*). Dies könnte beispielsweise über Ansätze einer zentralen Rechtefreigabe (*data rights repository* mit abgespeicherten Privacy-Wünschen) geschehen, wobei folgende Punkte zu beachten sind:

- Die Idee sollte als benutzerfreundliche Plattform mittels dezentralisierter Technologie umgesetzt werden, welche von Anwendern bei Bedarf an Daten angefragt werden.
- Es könnten spezifische Bereiche (z.B. Gesundheit, Kaufverhalten oder der Aufenthaltsort) sowie vordefinierte Regeln für diese festgelegt werden. Zum Beispiel könnte so im Notfall automatisch auf kritische Gesundheits- und Ortungsdaten zugegriffen werden, ohne dass Verbraucher im Moment ihre Zustimmung geben müssten.
- Nennenswerte Erfahrungen aus vergleichbaren Pilotprojekten sollten beigezogen werden. Unter anderen könnten HIT Foundation, HubOfAllThings, Mesinfos, MiData, Mydata, Open Humans, Pryv und Vetri nützlich sein.

Adressat: Serviceanbieter.

Empfehlung K-2	**Der Gesetzgeber soll prüfen, wie Datenportabilität im Bereich von KI-Systemen umgesetzt werden kann, insbesondere um Konsumentinnen und Konsumenten den Wechsel zu einem anderen Anbieter zu erleichtern.**

Erläuterung: KI-Anwendungen im Konsum wie z.B. personalisierte Bots werden umso besser, je mehr Daten sie von der jeweiligen Person sammeln. Dies führt zum Phänomen der «Plattformklebrigkeit» (oder auch dem sogenannten *locked-in effect*), d.h. Konsumentinnen und Konsumenten werden kaum mehr in der Lage sein, eine einmal gewählte Plattform zu verlassen. Um dies im Sinne der Förderung des Wettbewerbs zu vermeiden, sollten Standards zur Datenportabilität entwickelt werden. Generell gilt es, folgende Punkte zu beachten (Näheres dazu siehe Weber & Thouvenin 2017):

- Das Recht auf Datenportabilität ist ein Instrument, das es dem «Dateninhaber» erlaubt, bei einem Datenbearbeiter gespeicherte Daten mit geringem Aufwand auf einen anderen Datenbearbeiter zu übertragen. Die Einführung eines solchen Rechts, das für die EU mit der DSGVO geschaffen worden ist, sollte auch in der Schweiz in Betracht gezogen werden.
- Ein auf dem datenschutzrechtlichen Ansatz basierendes Recht auf Datenportabilität lässt sich auf das Grundrecht der informationellen Selbstbestimmung stützen und auf der Grundlage des bestehenden datenschutzrechtlichen Auskunftsrechts als dessen Weiterentwicklung ausgestalten. Dieser Ansatz erlaubt es auch, den Interessen der Datenbearbeiter angemessen Rechnung zu tragen, indem ein solches Recht im Einzelfall gewissen Einschränkungen unterworfen werden kann.
- Zur Einführung eines Rechts auf Datenportabilität müsste das DSG im Zusammenhang mit der Regelung des Auskunftsrechts durch einige Elemente präzisiert und ergänzt werden.

Aus Konsumentensicht wird dabei zentral sein, dass der Kunde genügend über das Recht zur Datenportabilität informiert wird. Anbieter könnten hier einen Anreiz haben zu informieren, um möglichen Neukunden einen leichten Wechsel zu ermöglichen.

Zu prüfen ist schliesslich auch, inwieweit neben den Rohdaten (*provided data*) auch Daten bzw. Aussagen portabel sein sollen, welche durch ein KI-System erarbeitet wurden (*inferred data*). Die Analyse dieser Frage ist nicht einfach, weil keine einheitliche Definition von *inferred data* besteht (insbesondere nicht im Recht; siehe dazu Abschnitt 2.5.3) und sich auch das Portabilitätsrecht in der DSGVO und dem E-DSG auf «data provided by the user» beschränkt. Generiert ein KI-System aus solchen Daten Klassifikationen (z.B. bezüglich des Zivilstands einer Person), die ohne Weiteres von der Person selbst erhoben werden können,

stellt sich die Frage der Portabilität vermutlich nicht. Möglich wäre aber, dass die Nutzung von KI-Systemen dereinst zu gewissen standardmässig verwendeten Kategorisierungen führen (z.B. bezüglich Bonität einer Person), die als *inferred data* verstanden werden und einer Portabilität unterworfen werden könnten. Entsprechend wird dem Gesetzgeber empfohlen, solche Entwicklungen zu beobachten und deren Bedeutung für das Recht auf Datenportabilität abzuklären.

Adressat: Gesetzgeber.

5.2.4. Medien

Empfehlung M-1	**Die Betreiber von Medienplattformen sollen ihren Nutzerinnen und Nutzern auf einfache Weise erkennbar machen, wie die Personalisierung von Medieninhalten mittels KI die Auswahl angezeigter Inhalte beeinflusst.**

Erläuterung: Aufgrund der Bedeutung der Medien für die Meinungsbildung und damit für den politischen Prozess sind Effekte von Personalisierung kritischer anzusehen als in anderen Bereichen. Auch hier ist die Schaffung von Transparenz das primäre Ziel, wobei folgende Instrumente eingesetzt werden können:

- *Bildung und Aufklärung:* Medienunternehmen sollten dazu verpflichtet werden, in einem einheitlichen Disclaimer auf ihren Plattformen ersichtlich zu machen, wie sie KI zur Personalisierung von Inhalten einsetzen. Zu beachten ist dabei, dass der Detaillierungsgrad dieser Erläuterungen den durchschnittlichen Erwartungen der Nutzerinnen und Nutzern angepasst ist.
- *Nicht personalisierter Feed:* Medienunternehmen sollten dazu verpflichtet werden, die Option eines nicht personalisierten Feeds auf ihren Plattformen zu integrieren und deutlich sichtbar im Nutzerinterface zu präsentieren. Im Fall von Online-Nachrichtenportalen würde ein nicht personalisierter Feed redaktionell ausgewählte oder mittels aggregierter Präferenzen festgelegte Inhalte enthalten; es muss aber ausgeschlossen werden, dass diese Inhalte mittels individueller Nutzerdaten auf die (vermeintlichen) Charakteristiken und Präferenzen der individuellen Person angepasst werden. In sozialen Medien könnten Inhalte eines nicht personalisierten Feeds schlicht chronologisch präsentiert werden.

- *Meinungsdiversifikation:* Es sollten Systeme entwickelt werden, die es Nutzer/-innen erlauben, Inhalte hinsichtlich ihrer Ausgewogenheit bzw. Einseitigkeit in der Präsentation von Sachverhalten zu bewerten. Die gesammelten Daten könnten auch dazu verwendet werden, Algorithmen darauf zu trainieren, Nutzenden Inhalte vorzuschlagen, die eine balanciertere Perspektive auf einen Sachverhalt einnehmen oder aber eine konträre Position beziehen.

Adressat: Medienunternehmen, Forschung.

Empfehlung M-2	**Der Bund soll in Zusammenarbeit mit Medienunternehmen und zivilgesellschaftlichen Akteuren die gesellschaftliche Diskussion über den Umgang mit Fake News, Filter Bubbles und Echokammern intensivieren. Sicherheitsbehörden (Polizei, Nachrichtendienst und Armee) sollen unter der Voraussetzung der parlamentarischen Kontrolle Fähigkeiten entwickeln, systematische Fake-News-Kampagnen mit dem Ziel politischer Manipulation rascher zu identifizieren und die Öffentlichkeit entsprechend zu informieren.**

Erläuterung: Die Debatte um Fake News rührt an Grundrechte wie die Wahl- und Abstimmungsfreiheit, die Meinungs- und Informationsfreiheit (was ein Recht auf Irrtum miteinschliesst) und die Rolle von Zensur in politischen Diskursen, in denen Propaganda, Fehlinformationen und dergleichen immer schon eine gewisse Rolle gespielt haben. Da KI-Technologien die Möglichkeiten der manipulativen Nutzung von (Falsch-)Informationen erhöhen, ist der gesellschaftliche Diskurs über den Umgang mit diesen Risiken unumgänglich. Eine Möglichkeit, um der Verbreitung von Fake News entgegenzuwirken, ist die Bewertung von Inhalten – sowohl nach den Kategorien «richtig» oder «falsch» als auch hinsichtlich der (ideologischen) Einseitigkeit/Ausgewogenheit. Eine Einbindung dieser Bewertungen mittels eines Browser-Plug-ins könnte beispielsweise Nutzer/-innen informieren, wie ausgewogen bzw. einseitig ein Inhalt ist oder ob dieser (vermutlich) falsch ist. Hier ist aber festzuhalten, dass solche Massnahmen immer eine Autorität voraussetzen, welche vorgibt, was «ausgewogen» bzw. «wahr» ist. Zentrale Fragen sind dann, aufgrund welcher Kriterien beispielsweise «Ausgewogenheit» bestimmt werden soll

und wer solche Bewertungen vornehmen soll. Diese Fragen sollten nicht von einzelnen Akteuren (z.B. Plattformbetreibern) im Alleingang entschieden werden, sondern Teil einer gesellschaftlichen Debatte sein, welche insbesondere auch die Medien, die politischen Parteien und zivilgesellschaftliche Organisationen einbeziehen soll. Ziel der Debatte soll nicht sein, endgültig über «wahr» oder «falsch» zu richten, sondern eine breite Debatte zu Meinungs- und Informationsfreiheit und Zensur in einer Zeit zu führen, in der KI-Systeme zunehmend zu Instrumenten der Beeinflussung der Meinungsbildung werden können.

Darüber hinaus sollen die Sicherheitsbehörden (insbesondere Cyber Defence und Nachrichtendienst) unter Massgabe politischer Neutralität und entsprechender parlamentarischer Kontrolle ihre Fähigkeiten zur Erkennung systematischer Kampagnen zur Meinungsmanipulation ausbauen und regelmässig über allfällige Aktivitäten Bericht erstatten.

Empfehlung M-1 enthält bereits Vorschläge, wie man Filter Bubbles und Echokammern entgegenwirken kann. Darüber hinaus sollten auch diese Themen in der gesellschaftlichen Debatte eingebunden werden.

Adressat: Bund, Medienunternehmen, Parteien, zivilgesellschaftliche Organisationen.

5.2.5. Öffentliche Verwaltung

Empfehlung V-1	**Die öffentliche Verwaltung soll Kriterien definieren, anhand derer bestimmt werden kann, wie eine verantwortliche staatliche KI-Nutzung (siehe auch Empfehlung 3) konkret umgesetzt werden kann.**

Erläuterung: Verwaltungen sollten Kriterien für eine Risikodifferenzierung der staatlichen Nutzung von KI entwickeln. Dies soll anhand folgender Faktoren geschehen: a) Input (welche Daten werden genutzt?); b) Output, wobei das Schadenspotenzial anhand der verwaltungsrechtlichen Einteilung von Verwaltungshandeln differenziert werden soll, d.h. es wird unterschieden zwischen schlichtem Verwaltungshandeln (Realakt) und belastenden bzw. begünstigenden Rechtsakten; c) evidenzbasierte Herangehensweise (wo gibt es in der aktuellen Verwaltungspraxis viele Probleme mit Verwaltungshandeln?).

Auch bei einem zurückhaltenden Einsatz ist von Anfang an darauf zu achten, dass Entscheidungsträger die spezifischen Herausforderungen kennen, die mit KI-

gestützten Empfehlungen einhergehen. Eine entsprechende Schulung bzw. Sensibilisierung – beispielsweise von Richterinnen und Richtern – ist deshalb von zentraler Bedeutung.

Zu beachten ist dabei auch, dass grundlegende Weichenstellungen für einen verantwortlichen KI-Einsatz durch den Staat bereits im Rahmen der Beschaffung getroffen werden. Schon vor der Ausschreibung sollte die Verwaltung deshalb auf unabhängiges Know-how zurückgreifen können bzw. sich eigenes Know-how aneignen, um im Rahmen des Beschaffungswesens die Einhaltung vorgängig definierter Standards prüfen zu können. D.h. entsprechendes Know-how muss frühzeitig aufgebaut und gefördert werden. Gerade im Rahmen der Beschaffung, aber auch darüber hinaus, ist ausserdem sicherzustellen, dass die Kantone und Gemeinden ebenfalls auf entsprechendes Know-how zurückgreifen können.[128]

Adressat: Öffentliche Verwaltungen von Bund, Kantonen und Gemeinden.

Empfehlung V-2	**Die öffentliche Verwaltung soll sicherstellen, dass Daten, welche für die staatliche KI-Nutzung genutzt werden, eine ausreichende Qualität haben.**

Begründung: Die Förderung der Datenqualität soll durch eine Reihe von Massnahmen sichergestellt werden, wobei Datenqualität durch drei Merkmale charakterisiert wird: a) möglichst korrekte Daten; b) möglichst geeignete Daten; c) Erhebung von Trainingsdaten nach ethischen Standards. Diese Massnahmen sind:

- Schaffung verbindlicher Standards sowohl für die Erhebung von Trainingsdaten als auch mit Blick auf eine effiziente Zusammenarbeit zwischen verschiedenen Behörden bzw. Behörden und Privaten bzw. Unternehmen;[129]

[128] Vgl. in diesem Kontext auch den Schlussbericht des Eidgenössischen Finanzdepartements und der Konferenz der Kantonsregierungen vom Oktober 2019: «Digitale Verwaltung: Projekt zur Optimierung der bundesstaatlichen Steuerung und Koordination», abrufbar unter https://www.newsd.admin.ch/newsd/message/attachments/58761.pdf.

[129] Der Bedarf für einheitliche Standards im eGovernment wurde längst erkannt und wird beispielsweise vom Verein eCH abgedeckt.

- Einheitliche und genaue Beschreibung der Daten (Metadaten) sowie Schaffung einheitlicher Strukturen und Inhalte der Stammdaten – jeweils unter Berücksichtigung bestehender Vorgaben;[130]
- Offenlegung der genutzten Daten gegenüber der Allgemeinheit (nicht in Form von Einzeldaten, sondern als Benennung von Merkmalen in Form von Kategorien wie «Beruf» etc.);
- Offenlegung der genutzten Personendaten gegenüber der betroffenen Person;
- Verantwortlichkeit für Datenqualität festlegen (es darf nicht alles auf die «IT» abgewälzt werden);
- Schaffen der Möglichkeiten, adäquate Datenqualität *organisatorisch* sicherzustellen, z.B. indem der Hersteller einer KI-basierten Anwendung die Datenqualität garantiert (Selbstzertifizierung), diese von einer unabhängigen Stelle zertifiziert wird oder eine Selbstkontrolle durch betroffene Personen ermöglicht wird.

Adressat: Öffentliche Verwaltungen von Bund, Kantonen und Gemeinden.

5.3. Forschungsbedarf

Viele der genannten Empfehlungen verlangen nach vertieften Kenntnissen sowohl bezüglich technischer Ausgestaltung als auch ethischer, rechtlicher und sozialer Aspekte von KI. Der daraus resultierende Forschungsbedarf ist im Rahmen bestehender Instrumente der Forschungsförderung (z.B. nationale Forschungsschwerpunkte) anzugehen. Dabei sollten folgende Gesichtspunkte besondere Beachtung finden:

- **Erklärbare KI:** Es werden signifikante Forschungsanstrengungen geleistet werden müssen, um ein besseres Verständnis darüber zu erlangen, *wie die*

[130] Zu erwähnen sind etwa die Strategie für den Ausbau einer gemeinsamen Stammdatenverwaltung des Bundes, vgl. https://www.isb.admin.ch/isb/de/home/themen/bundesarchitektur/schwerpunkte/stammdatenverwaltung.html, oder die Dateninnovationsstrategie des Bundesamtes für Statistik; siehe: https://www.bfs.admin.ch/bfs/de/home/aktuell/neue-veroeffentlichungen.gnpdetail.2017-0673.html.

neueren KI-Systeme zu Entscheidungen kommen und wie die Entscheidungsfindung Menschen gegenüber *transparent* gemacht werden kann. Ohne dieses Verständnis ist nicht zu erwarten, dass es KI-Systemen erlaubt sein wird, insbesondere Entscheidungen in Bereichen zu fällen, die einer rechtsstaatlich robusten Beurteilung standhalten können müssen – also beispielsweise in der ärztlichen Entscheidungsfindung. Es wird deshalb eine Kernaufgabe der Ingenieur- wie auch der Sozialwissenschaften werden, neue Methoden, Metriken, Kriterien und vertrauenswürdige Überprüfungsmechanismen zu entwickeln, damit Nutzerinnen und Nutzer solcher Systeme wie auch der Regulator die Validität, Transparenz und Fairness automatisierter Entscheidungssysteme bewerten können. Dem Aspekt der Mensch-Maschine-Interaktion ist dabei besondere Beachtung zu schenken.

- **Regulation autonomer Systeme:** Ausgehend von etablierten Nutzungen von Automatisierung (z.B. im *algorithmic trading*) wird ML zu einer *graduellen Zunahme der Autonomiefähigkeit von technischen Systemen* führen, die bald einmal nach *neuen rechtlichen Regeln* verlangen wird. Diese «Grauzone» dürfte dadurch charakterisiert sein, dass Menschen entweder nur mittelbar von den Entscheidungen der Systeme betroffen sind oder die Tragweite der Entscheidungen (z.B. im Fall der personalisierten Werbung für Produkte) als ethisch weniger bedeutsam angesehen wird. Die Ausgestaltung des Regulierungsrahmens in diesen Zonen erweiterter Autonomiefähigkeit technischer Systeme wird zu einer zentralen Herausforderung für die rechtswissenschaftliche Forschung werden.

- **Sicherstellung menschlicher Kontrolle:** Es besteht in ethischer und rechtlicher Hinsicht ein breiter Konsens darüber, dass die Nutzung von KI in relevanten Entscheidungskontexten einer *menschlichen Kontrolle* unterliegen sollte. Allerdings dürfte zunehmend unklar werden, was genau «menschliche Kontrolle» beinhaltet. Automatisierte Entscheidungsprozesse sind natürlich immer in grössere soziale Kontexte eingebunden, wo menschliche Entscheidungskompetenz auf einer höheren Ebene noch vorhanden ist – allerdings stellt sich die Frage, inwieweit dieser Entscheidungsspielraum durch die Automatisierung überhaupt noch als solcher empfunden und auch tatsächlich wahrgenommen wird bzw. ob die erlebte Autonomie nur eine scheinbare ist. Hier Klärung zu finden, dürfte eine wichtige Aufgabe der human- und sozialwissenschaftlichen Forschung werden, welche Grundlagen für eine allfällige rechtliche Regulierung bilden sollte.

- **Bedingungen für Vertrauen in KI:** Nutzungsumfang und Akzeptanz automatisierter Entscheidungen werden von *psychologischen Faktoren* beeinflusst, die derzeit noch kaum erforscht sind. Es wird zunehmend wichtiger werden, KI-Systeme so zu designen, dass sie zum einen angenommen werden können, zum anderen ihre Fähigkeiten aber nicht überschätzt werden (*design for appropriate reliance*). Hier wird sich ein wichtiges Tätigkeitsfeld für Psychologie und Verhaltenswissenschaften ergeben. Zu beachten ist dabei, dass das Vertrauen in KI-Systeme nicht nur von deren Design, sondern auch von der Gestaltung und Regulierung ihrer Nutzung abhängen wird.
- **Technologien für Sicherung der Privatsphäre:** Neuere KI-Anwendungen brauchen eine grosse Menge an Daten. Das hohe Bewusstsein zum Persönlichkeits- und Datenschutz, das in der Schweiz und Europa (im Gegensatz zu z.B. China) besteht, sollte dafür genutzt werden, KI-Anwendungen zu entwickeln, die sich global gerade dadurch auszeichnen, auf den Schutz im Umgang mit persönlichen Daten besonders zu achten. Daten-Bias sollten möglichst ausgeschlossen, Transparenz geschaffen und algorithmische Fairness gefördert werden. Zugleich sollte durch ausgebaute Forschungsprivilegien sichergestellt werden, dass der Datenschutz die Entwicklung von KI-Systemen in Europa und der Schweiz nicht unverhältnismässig behindert.
- **Förderung von interdisziplinärer Forschung:** Sowohl Entwicklung als auch Einsatz von KI-Systemen bedürfen eines interdisziplinären Forschungszugangs, in dem die fachspezifischen Forschenden und die KI-Fachpersonen gegenseitiges Verständnis über die Möglichkeiten und Bedürfnisse des Faches und der KI-Anwendungen aufbringen. Diese Forschung bedarf zielgerichteter öffentlicher Förderung. Zu beachten ist hierbei auch, dass KI selbst ein interdisziplinäres Gebiet ist. Dazu gehören z.B. kognitive Psychologen, Linguisten (Spracherkennung) und Maschinenbauer (Robotik). Deshalb sollte auch für die Weiterentwicklung der KI selbst die interdisziplinäre Zusammenarbeit von Fächern wie Linguistik, Informatik, Psychologie etc. gefördert werden.

Um das Potenzial von KI-Systemen in der Forschung und Innovation zu nutzen, wird Forschungseinrichtungen empfohlen, KI-Zentren einzurichten, welche innovative Forschungsverfahren aufzeigen, die Forschenden in der Anwendung von KI unterstützen und den interdisziplinären Austausch fördern. Diese Anlaufstellen sollen auch Richtlinien für die Angehörigen der Institution entwickeln, um einen achtsamen Umgang mit KI-Systemen sicherzustellen.

Annex

Zur Methodik der Umfrage

Der Einbezug von Fachpersonen in die Studie spielte eine zentrale Rolle, um (i) die Einschätzungen des Projektteams und den Konsensus aus der Literatur kritisch zu hinterfragen sowie (ii) neue Empfehlungen zu erarbeiten. Zweck des Einbezugs von Expertinnen und Experten war dabei nicht, den Prozess der Entscheidungsfindung bezüglich Empfehlungen gewissermassen auszulagern. Vielmehr sollte damit das Meinungsspektrum ausgeweitet werden – auch um allfällige «blinde Flecken» des Studienteams zu identifizieren. Wichtig war dabei zu prüfen, inwieweit die Experteneinschätzung der Technologie mit dem Kenntnisgrad und der Haltung gegenüber KI (KI-Optimismus v. KI-Pessimismus) korrelieren. Wie in Abschnitt 1.3.1 ausgeführt, erfolgte die Expertenumfrage in zwei Runden, in denen jeweils unterschiedliche Sachverhalte erhoben wurden: In der ersten Runde ging es vorrangig um die Einschätzung der Faktenlage, in der zweiten Runde primär um die Einschätzung von möglichen Massnahmen. Ergänzt wurden die beiden Umfragen durch eine Befragung der allgemeinen Bevölkerung; fokussiert wurden dabei die Bereiche «Konsum» und «Ethik».

Rekrutierung der Teilnehmenden

Ziel der Expertenumfrage war es zum einen, technische Fachleute im Bereich KI anzusprechen, und zum anderen, Experten aus den fünf Themenbereichen Arbeit, Bildung und Forschung, Konsum, Medien und öffentliche Verwaltung, die Interesse am Thema KI haben, zu konsultieren. Die Umfragen wurden online durchgeführt, die Teilnehmerinnen und Teilnehmer wurden per E-Mail rekrutiert, wobei insgesamt zwei Erinnerungen pro Verteiler versandt wurden. Es wurden sowohl Experten in der Schweiz als auch international kontaktiert, wobei der Fokus auf Fachpersonen aus der Schweiz lag.

Die Expertinnen und Experten wurden über folgende Kanäle kontaktiert:

- Expertenpool TA-SWISS[131] (~ 4000 Personen)

[131] Siehe: https://www.ta-swiss.ch/.

- Expertenpool Digitale Schweiz[132] (~ 1600 Personen)
- Expertenpool SwissCognitive[133] (~ 380 Personen)
- Expertenpool KI der Schweizerischen Akademie der Technischen Wissenschaften[134] (~ 350 Personen)
- Konsolidierte Kontaktliste des Studienteams (~ 200 Personen)
- Mitglieder der Société Suisse d'Informatique[135] (~ 150 Personen)

Das Kontaktschreiben enthielt die Bitte, den Umfragelink allfälligen weiteren Experten zukommen zu lassen. Deshalb – sowie aufgrund der Tatsache, dass Überschneidungen zwischen den Vertriebskanälen sicher bestanden, aber nicht quantifiziert werden konnten – kann keine genaue Zahl der insgesamt kontaktierten Personen eruiert werden; es dürfte sich aber um mehrere Tausend handeln. Die erste Umfrage wurde im Verlauf des Oktobers und Novembers 2018 einem intensiven *pretesting* unterworfen (zuerst intern im Projektteam, danach mit Dritten) und am Montag 3. Dezember 2018 gestartet. Sie dauerte bis zum 15. Januar 2019.

Für die zweite Umfrage wurden nur jene Personen kontaktiert, die Antworten für die erste Umfrage geliefert und sich explizit für die Teilnahme an der zweiten Runde bereit erklärt hatten. Dies waren 258 Personen (6 Personen hatten sich für eine Teilnahme ausgesprochen, gaben aber eine ungültige E-Mail-Adresse an). Auch diese Personen erhielten zwei Einladungs-E-Mails. Die Umfrage wurde vor dem offiziellen Start getestet; sie dauerte vom 24. April bis zum 16. Mai 2019.

Aufbau des ersten Fragebogens

Ziel des ersten Fragebogens war es, Einschätzungen zu faktischen Aspekten von KI zu erhalten – einerseits in genereller Hinsicht und andererseits bezogen auf die fünf thematischen Bereiche. Der Fragebogen war so aufgebaut, dass er in ca. 20 Minuten beantwortet werden konnte, wobei die Fachpersonen aber die Möglichkeit

[132] Siehe: https://www.bakom.admin.ch/bakom/de/home/digital-und-internet/strategie-digitale-schweiz.html.

[133] Siehe: https://swisscognitive.ch/.

[134] Siehe: https://www.satw.ch/de/ueber-satw/themenplattformen/themenplattform-kuenstliche-intelligenz/.

[135] Siehe: https://sisr.ch/.

hatten, zu mehr als einem Bereich Antworten zu geben. Die Medianzeit betrug rund 30 Minuten. Der Fragebogen war folgendermassen strukturiert:

- Zu Beginn erfolgten generelle Informationen über Zweck, Inhalt und Auftraggeber der Studie sowie über Datenschutzaspekte, sodass die Befragten eine informierte Zustimmung zur Teilnahme an der Studie geben konnten. Die Teilnehmenden gaben auch ihre Kontaktangaben an.
- Danach wurden generelle demografische Faktoren erhoben (Geschlecht, Alter, geografische Herkunft, Ausbildung und berufliche Tätigkeit).
- In einem ersten Teil wurden danach Fragen zur generellen Einschätzung und persönlichen Nutzung von diversen KI-Technologien gestellt. Zu diesem Zweck wurde auch eine Arbeitsdefinition von KI gegeben, um sicherzustellen, dass die Teilnehmenden mit einem ähnlichen Konzept von KI an den Fragebogen herangingen. Insbesondere wurde klargestellt, dass es bei der Studie nicht um eine «starke KI» im Sinn einer «Superintelligenz» geht.
- Danach konnten die Personen eines der fünf studienspezifischen Themengebiete auswählen. Die konkreten Fragen zu den jeweiligen Gebieten werden in den Abschnitten 3.3 bis 3.7 vorgestellt. Am Ende der Themenblöcke konnten die Befragten weitere Themengebiete auswählen. Im Schnitt bearbeiteten die Fachpersonen ca. 1,5 Themengebiete.
- Schliesslich folgten Fragen zum Themenkomplex «Ethik und Recht», die allen Befragten vorgelegt wurden (zu den konkreten Fragen siehe Abschnitt 3.8).

Die Fachpersonen konnten die Fragebögen auf Deutsch, Englisch oder Französisch beantworten.

Aufbau des zweiten Fragebogens

Im zweiten Fragebogen wurden primär Einschätzungen zu den Massnahmen in den fünf Themenbereichen erhoben. Dabei wurde zum einen gefragt, für wie effektiv eine Reihe möglicher Massnahmen erachtet wird, zum anderen wurde die Wünschbarkeit der jeweiligen Massnahmen erhoben – dies aus der Überlegung, dass es zwar effektive, aber nicht wünschbare Massnahmen geben könnte (und umgekehrt), z.B. bestimmte Zensurmassnahmen im Bereich Medien.

Auch hier konnten die Personen ihre Einschätzungen zu mehr als einem Themengebiet abgeben. Im Schnitt wurden pro Person Einschätzungen zu 2,6 Themengebieten abgegeben, die Medianzeit für die Beantwortung des Fragebogens betrug rund 37 Minuten.

Um eine Verknüpfung zu den Daten der ersten Umfrage zu ermöglichen, mussten die Teilnehmenden erneut ihren Namen und ihre Heiminstitution angeben. Zudem wurde die gleiche Arbeitsdefinition von KI gegeben wie in der ersten Umfrage, um sicherzustellen, dass die Teilnehmenden mit einem ähnlichen Konzept von KI an den Fragebogen herangingen. Die Fachpersonen konnten schliesslich auch angeben, ob sie Interesse haben, am Abschlussworkshop teilzunehmen.

Der zweite Fragebogen wurde auf Deutsch und Englisch angeboten, weil die Anzahl der französischsprechenden Personen klein war und erwartet werden konnte, dass diese die Fragen entweder auf Deutsch oder Englisch beantworten konnten.

Datenqualität und Demografie des Expertensamples

Wie bereits erwähnt, kann keine präzise Schätzung der Zahl der kontaktierten Personen angegeben werden. Insgesamt wurde der Link der ersten Umfrage 851 Mal angeklickt. 388 Personen lieferten Antworten (Konversionsrate: 46 %). Nach der Datenaufbereitung (Hauptkriterium war, dass eine Person mindestens zu einem Themenbereich vollständige Antworten gegeben hat) blieben die Daten von 307 Personen für die Auswertung zur Verfügung (*drop-out*-Rate 21 %).

In der ersten Umfrage gaben 265 Personen (86 %) ihre Bereitschaft zur Teilnahme an der zweiten Umfrage an. Auf die zweite Umfrage wurde insgesamt 223 Mal zugegriffen und 158 Personen lieferten Antworten (Konversionsrate: 71 %). Davon wurden 47 Personen gestrichen, weil sie nicht mindestens zu einem Bereich vollständige Antworten gegeben hatten (41) oder weil sie in der ersten Umfrage nicht teilgenommen hatten bzw. die Zuordnung nicht möglich war (6). Es verblieben die Daten von 111 Personen für die Auswertung (*drop-out*-Rate: 30 %). Die *drop-out*-Raten entsprechen den Erwartungen von Onlineumfragen, während die Konversionsrate hoch war, was auf ein überdurchschnittliches Engagement der angefragten Personen schliessen lässt.

Die wichtigsten demografischen Angaben zum Expertensample finden sich in Tabelle 6. Hier zeigt sich, dass die zweite Umfrage tendenziell mehr von Männern, Personen schweizerischer Herkunft und mit technischem Hintergrund ausgefüllt

wurden. Das Durchschnittsalter lässt generell auf erfahrene Experten schliessen; der Frauenanteil in der ersten Umfrage dürfte leicht höher sein, als in technischen Bereichen zu erwarten ist. Fast alle Fachpersonen haben eine universitäre oder eine Fachhochschulausbildung; die Diversität bezüglich der Arbeitsfelder Universität, Wirtschaft und Verwaltung ist gut.

Tabelle 6: Demografische Beschreibung des Expertensample; Angaben in Prozent.

		Erste Umfrage	**Zweite Umfrage**
Geschlecht	Männer	74 %	80 %
	Frauen	24 %	16 %
	Divers, k.A.	2 %	4 %
Alter (Schnitt)		48,1	48,4
Herkunft	Schweiz	71 %	78 %
	Europa	20 %	17 %
	Andere Länder	9 %	5 %
Sprache (basierend auf Sprachwahl in der Umfrage)	Deutsch	59 %	71 %
	Englisch	26 %	29 %
	Französisch	15 %	–
Ausbildung	Universität/ETH	84 %	86 %
	Fachhochschule	11 %	10 %
	Sonstiges	5 %	4 %
Art der Ausbildung	Technisch	52 %	59 %
	Nicht technisch	48 %	41 %
Arbeitsfeld	Universitär	40 %	35 %
	Verwaltung	18 %	20 %
	Wirtschaft	30 %	34 %
	NGO etc.	12 %	11 %
Anzahl antwortende Personen pro Themenbereich	Arbeit	115	70
	Bildung und Forschung	113	56
	Konsum	72	53
	Medien	67	55
	Öffentliche Verwaltung	73	58

Zusatzumfrage

Im Kontext der Hauptstudie wurde eine Zusatzumfrage im Rahmen eines Projekts der SATW und der Stiftung Risiko-Dialog durchgeführt, bei der es primär um konsumbezogene und ethische Fragen ging.

Die Zusatzumfrage richtete sich an die allgemeine Bevölkerung, wobei aber aufgrund des Verteilers (Kontakte der Stiftung und der SATW) angenommen werden darf, dass Personen mit einem akademischen Hintergrund überrepräsentiert sind. Die Umfrage wurde im Zeitraum vom 30. April bis 28. Mai 2019 durchgeführt. Es wurde insgesamt 482 Mal auf die Umfrage zugegriffen; die Daten von 269 Personen wurden in die Analyse aufgenommen. Unvollständig ausgefüllte Fragebögen wurden gestrichen. Die Medianzeit für das Ausfüllen betrug knapp 10 Minuten. Die Umfrage wurde auf Deutsch und Französisch zur Verfügung gestellt.

Der Aufbau des Fragebogens war wie folgt:

- Zu Beginn erfolgten Informationen über Zweck, Inhalt und Auftraggeber der Studie sowie Datenschutz, sodass die Befragten eine informierte Zustimmung zur Teilnahme an der Studie geben konnten.
- Danach wurden demografische Faktoren erhoben (Geschlecht, Alter, Ausbildung, Wohnkanton).
- Nach einer Kurzdefinition von «Künstlicher Intelligenz» folgten Fragen über die generelle Nutzung und Einschätzung von KI-Technologien.
- Danach folgte die Beurteilung von Massnahmen zur Förderung des Kundenvertrauens in KI und gegen die Bildung von Oligopolen in der Datenwirtschaft.
- Hauptteil der Umfrage war die Beurteilung von Szenarien aus drei Bereichen (Arbeitswelt, Medien und Medizin), in denen entweder ein Mensch oder ein KI-System für eine Entscheidung zuständig ist. Beurteilt wurde das Vertrauen in Mensch bzw. KI sowie die Verantwortlichkeit bei einem Fehler. Jede Person bekam ein zufällig ausgewähltes Szenario zur Beurteilung vorgelegt.
- Schliesslich konnten die Personen noch angeben, ob sie an einer von der SATW und der Stiftung Risiko-Dialog organisierten Veranstaltung teilnehmen wollten.

Tabelle 7 zeigt einen Überblick über die demografischen Angaben zur Zusatzumfrage im Vergleich zur Expertenumfrage. Es zeigt sich im Vergleich zur Expertenumfrage eine deutlich paritätischere Verteilung der Geschlechter; das Sample ist aber, wie erwartet, stark «akademikerlastig».

Tabelle 7: Demografische Beschreibung der Samples der Zusatzumfrage.

		Zusatzumfrage	**Expertenumfrage**
Geschlecht	Männer Frauen Divers, k.A.	53 % 45 % 2 %	74 % 24 % 2 %
Alter (Schnitt)		52,4	48,1
Sprache (basierend auf Sprachwahl in der Umfrage)	Deutsch Französisch Englisch	84 % 16 % –	59 % 15 % 26 %
Ausbildung	Universität/ETH Fachhochschule Sonstiges	78 % 14 % 8 %	84 % 11 % 5 %

Statistische Berechnungen und grafische Darstellung

Auswertungen der Resultate wurden unter Verwendung der Statistikprogramme Mathematica, R und SPSS vorgenommen; einfache deskriptive Untersuchungen auch mittels Excel. Für Signifikanztests wurde in der Regel ein Signifikanzniveau von $p = 0.05$ angenommen. Je nach Problemstellung wurden T-Tests, nicht parametrische Tests (Mann-Whitney) oder der Chi-Quadrat-Test verwendet. Die grafischen Darstellungen wurden in den jeweiligen Statistikprogrammen erstellt und im Grafikprogramm Illustrator bezüglich Stil und Schriftarten vereinheitlicht.

Zur Methodik der Expertenworkshops

Es wurden zwei Workshops durchgeführt. Der erste Workshop fand am 22. Mai in Zürich statt und umfasste eine ausgewählte Gruppe von Fachpersonen, die sich zum grossen Teil an den beiden Umfragen beteiligt hatten und dadurch bereits Einblicke in die Studie erhalten hatten. Der zweite Workshop fand am 22. August mit der Begleitgruppe der TA-SWISS-Studie statt.

Erster Workshop

Insgesamt hatten 36 Teilnehmende der zweiten Runde ihr Interesse bekundet, am Workshop teilzunehmen. Davon wurden nur jene Personen in Betracht gezogen, die deutsche oder französische Muttersprachler waren und in der Schweiz oder dem grenznahen Ausland wohnten. Des Weiteren wurden 13 Fachpersonen, die nicht an den Umfragen teilgenommen hatten, auf die Einladungsliste genommen. Von den 40 eingeladenen Personen haben 15 abgesagt und 10 konnten jeweils nur am Morgen bzw. Nachmittag des ganztägigen Workshops teilnehmen.

Am Workshop wurden vertiefend mögliche Empfehlungen zur Nutzung von Chancen und Kontrolle von Risiken von künstlicher Intelligenz in den Anwendungsbereichen Arbeit, Bildung, Konsum, Medien und Verwaltung in zwei Runden diskutiert. Dazu wurden gemäss den zuvor erfassten Präferenzen der Teilnehmenden Kleingruppen gebildet. Die Gruppen umfassten in der Regel 4–5 Personen (jeweils verschiedene Personen am Morgen und Nachmittag) sowie je einen Moderator und Protokollant des Studienteams.

Die Ergebnisse wurden in einem Bericht zusammengefasst und allen Teilnehmenden des Workshops geschickt. Acht Personen nahmen die Gelegenheit wahr und lieferten weitere Kommentare.

Zweiter Workshop

Der zweite Workshop wurde mit den Mitgliedern der Begleitgruppe und dem Studienteam durchgeführt. Dazu erhielten alle Beteiligten eine erste Fassung des Berichts der Studie, der Entwürfe von ausformulierten Empfehlungen enthielt. Diese

wurden in der Runde weiter diskutiert und die Anregungen der Begleitgruppe wurden aufgenommen. Dabei wurden vorab folgende Punkte diskutiert:

- Es wurde – in Anlehnung an den Bericht der Arbeitsgruppe des Bundesrates – von allen Beteiligten bestätigt, dass eine technologiespezifische Regulierung (ein generelles «KI-Gesetz») der falsche Ansatz ist, um den Herausforderungen von KI zu begegnen.
- Es wurde festgehalten, dass zwischen dem staatlichen und dem privaten Einsatz von KI unterschieden werden muss, zumal staatliches Handeln bereits regulatorisch höheren Anforderungen zu genügen hat, was sich auch im staatlichen KI-Einsatz niederschlagen sollte.
- Generell muss bei allen Empfehlungen darauf geachtet werden, dass ein klarer Adressat definiert wird, der für die Umsetzung allfälliger Massnahmen zuständig wäre.
- Aufgrund der Tatsache, dass viele Entwicklungen im Bereich KI eine internationale Dimension haben, sollen die Empfehlungen auf einer realistischen Beurteilung des Schweizer Handlungsrahmens beruhen.
- Die Nutzung von KI in zahlreichen Anwendungsbereichen hat offenbar den Effekt, dass seit Längerem diskutierte kontroverse Themen, z.B. zu Diskriminierung, Fairness oder Privatsphäre, eine neue Schärfe gewinnen, weil der Einsatz der Technologie Entscheidungen verlangt (z.B. welche Art von Fairness soll in einen Entscheidungsalgorithmus technisch implementiert werden), die vorher pluralistisch angegangen und verdrängt werden konnten. Es besteht ein Konsens, dass die Beantwortung solcher Fragen nicht den technischen Experten überlassen werden sollte.

Diese Anregungen flossen in die Ausformulierung der Empfehlungen ein.

Prozess der Finalisierung und Priorisierung

Basierend auf den Ergebnissen des zweiten Workshops wurden in mehreren internen Treffen und sonstigem Austausch der Begleitgruppe die Empfehlungen finalisiert. Es ergaben sich dabei zwei Arten von Empfehlungen:

1. Empfehlungen, die übergreifenden Charakter haben, bilden die prioritären Empfehlungen dieser Studie. Zu jeder Empfehlung folgt eine Erläuterung, bei der auch die Adressaten spezifiziert werden.

2. Empfehlungen, welche die einzelnen Bereiche betreffen, sind primär das Ergebnis der einzelnen Projektteams. Auch sie werden erläutert, wobei auch auf allfällige Umsetzungsschwierigkeiten hingewiesen wird.

Literatur

Vorbemerkung: Wenn immer möglich (im Fall von *Open-Access*-Publikationen), wird ein direkter Link zur Quelle angegeben; das Datum in Klammern bezeichnet den Zeitpunkt des letzten Zugriffs.

Abbott, R. (2016). I Think, Therefore I Invent: Creative Computers and the Future of Patent Law. B.C. Law Review, 57(4): 1079–1126

Access Now, Amnesty International (2018). The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems. Zugang: https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/ (18.12.2018)

Ackerman, E. (2017). Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms. IEEE Spectrum. Zugang: https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms (12.01.2018)

Adobe (2018). Digital Intelligence Briefing. Digitale Trends 2018. Zugang: https://www.adobe.com/content/dam/acom/de/modal-offers/pdf/econsultancy-2018-digital-trends_de.pdf (26.06.2019)

Albus, J. S. (1991). Outline for a theory of intelligence. IEEE Transactions on Systems, Man, and Cybernetics, 21(3): 473–509

Ali, A. (2018). The Rise of Virtual Digital Assistant Usage – Statistics and Trends. Mobiteam. Zugang: https://mobiteam.de/en/the-rise-of-virtual-digital-assistant-usage-statistics-and-trends/ (29.06.2019)

Altmann, J., Sauer, F. (2017). Autonomous Weapon Systems and Strategic Stability. Survival, 59(5): 117–142

Altschool (2017). Zugang: https://www.altschool.com/about/about (12.01.2018)

Alvarez-Melis, D., Jaakkola, T. S. (2017). A causal framework for explaining the predictions of black-box sequence-to-sequence models. Zugang: https://arxiv.org/abs/1707.01943 (10.11.2019)

Anderson, J. (2019). A British Startup will put AI into 700 schools in Belgium, Quartz. Zugang: https://qz.com/1577451/century-tech-signs-deal-to-put-ai-in-700-classrooms-in-belgium/ (21.03.2019)

Andrejevic, M. (2017). To Preempt a Thief. International Journal of Communication, 11: 879–896

Angwin, J., Larson, J. (2016). Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say. ProPublica. Zugang: https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say (30.12.2016)

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016). Machine Bias – There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. Zugang: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (10.11.2019)

Antoninis, M., Montoya, S. (2018). A global framework to measure digital literacy. Data for Sustainable Development Blog, 19 March 2018. Montreal, UIS

APA Dictionary of Psychology. (2019). Privacy. Zugang: https://dictionary.apa.org/privacy (08.06.2019)

Apt, W., Bovenschulte, M., Hartmann, E. A., Wischmann, S. (2016). Foresight-Studie «Digitale Arbeitswelt». (Forschungsbericht / Bundesministerium für Arbeit und Soziales, FB463). Berlin: Bundesministerium für Arbeit und Soziales. Zugang: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-47039-5 (16.12.2019)

Arntz, M., Gregory, T., Zierahn, U. (2016). The risk of automation for jobs in OECD countries: A comparative analysis. OECD Social, Employment, and Migration Working Papers (189). Paris: OECD Publishing

Azucar, D., Marengo, D., Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. Personality and Individual Differences, 124: 150–159

Baggi, D. (1989). NeurSwing: An intelligent workbench for the investigation of swing in jazz. In: Baggi D.: Computer-Generated Music, IEEE Computer Society Press, 79–93

Bakshy, E., Messing, S., Adamic, L. A. (2015). Political science. Exposure to ideologically diverse news and opinion on Facebook. Science, 348: 1130–1132

Barberá, P. (2018). The Consequences of Exposure to Disinformation and Propaganda in Online Settings. In: Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., Nyhan, B. (eds.): Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. New York: Hewelett Foundation, 15–21

Barocas, S., Selbst, A. D. (2016). Big Data's Disparate Impact, California Law Review, 104: 671–732

Barrelet, D., Egloff, W. (2008). Das neue Urheberrecht. Kommentar zum Bundesgesetz über das Urheberrecht und verwandte Schutzrechte. 3. Auflage, Bern: Stämpfli Verlag

Bartha, P. (2013). Analogy and Analogical Reasoning. Stanford Encyclopedia of Philosophy. Zugang: https://plato.stanford.edu/entries/reasoning-analogy/ (16.12.2019)

BCG (2018). AI in the Factory of the Future. The Ghost in the Machine. Zugang: https://www.bcg.com/publications/2018/artificial-intelligence-factory-future.aspx (26.06.2019)

Beck, S. (2017). Der rechtliche Status autonomer Maschinen. Allgemeine Juristische Praxis, 2: 183–191

Bergemann, D., Bonatti, A. (2018). Markets for information: An introduction. Annual Review of Economics, 11. doi: 10.1146/annurev-economics-080315-015439

Berger, M. (2004). Schutz von Software – Überblick über die Rechtslage in der Schweiz. In: Trüeb, H.-R. (Hrsg.). Softwareverträge. Zürich: Schulthess, 25–60

Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. Zugang: https://arxiv.org/abs/1703.09207v2 (12.01.2018)

Bertschinger, C. (2002). Patentfähige Erfindung. In: Bertschinger, C., Geiser, P., Münch, G. (Hrsg.). Schweizerisches und europäisches Patentrecht. Basel: Helbing Lichtenhahn Verlag, 87–162

Bessi, A., Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. First Monday 21. doi: 10.5210/fm.v21i11.7090

Bessi, A., Zollo, F., Del Vicario, M., Scala, A., Caldarelli, G., Quattrociocchi. W. (2015). Trend of Narratives in the Age of Misinformation. PloS one 10: e0134641

Betschon, S. (2019). Computer und künstliche Intelligenz machen Arbeit, viel Arbeit. Sie ist schlecht bezahlt – und das KI-Prekariat bleibt unsichtbar. NZZ, 08.10.2019

Beuth, P. (2017). Die rätselhafte Gedankenwelt eines Computers. Zeit Online. Zugang: https://www.zeit.de/digital/internet/2017-03/kuenstliche-intelligenz-black-box-transparenz-fraunhofer-hhi-darpa (24.03.2019)

BfS (2017). Dateninnovationsstrategie des Bundesamtes für Statistik vom 21.11.2017. Zugang: https://www.bfs.admin.ch/bfs/de/home/aktuell/neue-veroeffentlichungen.gnpdetail.2017-0673.html (29.12.2019)

Binder, A., Bach, S., Montavon, G., Müller, K.-R., Samek, W. (2016). Layer-wise relevance propagation for deep neural network architectures. In: Kim, K. J., Joukov, N. (eds.). Information Science and Applications (ICISA). Singapore: Springer, 913–922

Binns, R. (2018). Algorithmic Accountability and Public Reason. Philosophy & Technology, 31: 543–556

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N. (2019). «It's Reducing a Human Being to a Percentage». Perceptions of Justice in Algorithmic Decisions. Zugang: https://arxiv.org/abs/1801.10408 (12.02.2019)

Bitkom e.V./DFKI (Hrsg.) (2017). Entscheidungsunterstützung mit Künstlicher Intelligenz. Berlin. Zugang: https://www.uni-kassel.de/fb07/fileadmin/datas/fb07/5-Institute/IWR/Hornung/170901-KI-Gipfelpapier-online.pdf (15.12.2019)

Block, P. (2017). The Inventor's New Tool: Artificial Intelligence – How Does it Fit in the European Patent System? European Intellectual Property Review, 39(2): 69–73

Bloom, B. S. (1984). The Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational Researcher, 13(6): 4–16

Board, T. E. (2019). Opinion. How Silicon Valley Puts the «Con» in Consent. The New York Times. Zugang: https://www.nytimes.com/2019/02/02/opinion/internet-facebook-google-consent.html (15.12.2019)

Boeing, N. (2018). Dein Freund und Lauscher. Technology Review. Zugang: https://www.heise.de/tr/artikel/Dein-Freund-und-Lauscher-4050426.html (29.06.2019)

Bonin, H., Gregory, T., Zierahn, U. (2015). Übertragung der Studie von Frey/Osborne (2013) auf Deutschland: ZEW Kurzexpertise. Zugang: https://www.econstor.eu/bitstream/10419/123310/1/82873271X.pdf (15.12.2019)

Bostrom, N., Yudkowsky, E. (2014). The ethics of artificial intelligence. In: Frankish, K., Ramsey, W. M. (eds.). The Cambridge Handbook of Artificial Intelligence. Cambridge: Cambridge University Press

Bowles, J. (2014). The computerisation of European jobs. Breughel, Brussels. Zugang: http://bruegel.org/2014/07/thecomputerisation-of-european-jobs/ (15.12.2019)

Boxell, L., Gentzkow, M., Shapiro. J. M. (2017). Greater Internet use is not associated with faster growth in political polarization among US demographic groups. Proceedings of the National Academy of Sciences of the United States of America, 114: 10612–10617

Braun Binder, N. (2016a). Ausschliesslich automationsgestützt erlassene Steuerbescheide und Bekanntgabe durch Bereitstellung zum Datenabruf. DStZ, 14: 526–535

Braun Binder, N. (2016b). Vollständig automatisierter Erlass eines Verwaltungsaktes und Bekanntgabe über Behördenportale. DÖV, 21: 891–898

Braun Binder, N. (2018). Algorithmic Regulation – Der Einsatz algorithmischer Verfahren im staatlichen Steuerungskontext. In: Hill, H., Wieland, J. (Hrsg.). Zukunft der Parlamente. Berlin: Duncker & Humblot, 107–120

Braun Binder, N. (2019a). Algorithmisch gesteuertes Risikomanagement in digitalisierten Besteuerungsverfahren. In: Unger, S., von Ungern-Sternberg, A. (Hrsg.). Demokratie und künstliche Intelligenz. Tübingen: Mohr Siebeck, 161–181

Braun Binder, N. (2019b). Künstliche Intelligenz und automatisierte Entscheidungen in der öffentlichen Verwaltung. SJZ, 15: 467–476

Briner, A. (2006). Patentierungsvoraussetzungen. In: von Büren, R., David, L. (Hrsg.). Schweizerisches Immaterialgüter- und Wettbewerbsrecht, Bd. IV, Basel, 47–169

Brockman, G., Sutskever, I. (2015). Introducing Open AI. Zugang: https://blog.openai.com/introducing-openai/ (15.12.2019)

Brynjolfsson, E., McAfee, A. (2011). Race against the machine. Lexington, Massachusetts: Digital Frontier Press

Brynjolfsson, E., McAfee, A. (2014). The Second Machine Age. Work, Progress, and Prosperity in a time of Brilliant Technologies. New York: W. W. Norton & Company

Bucher, R. (2019). Anwendungen von Künstlicher Intelligenz in der Bildung – Chancen und Risiken, Bachelorarbeit im Fach Wirtschaftsinformatik, Universität Zürich

Buell, R. W., Norton, M. I. (2011). The labor illusion: How operational transparency increases perceived value. Management Science, 57(9): 1564–1579

Bundespolizeipräsidium Potsdam (2018). Abschlussbericht zum Teilprojekt 1 «Biometrische Gesichtserkennung» vom 18.09.2018. Zugang: https://www.bundespolizei.de/Web/DE/04Aktuelles/01Meldungen/2018/10/181011_abschlussbricht_gesichtserkennung_down.pdf?__blob=publicationFile (29.12.2019)

Bundesrat (2017). Auswirkungen der Digitalisierung auf Beschäftigung und Arbeitsbedingungen – Chancen und Risiken; Bericht des Bundesrates in Erfüllung der Postulate 15.3854 Reynard vom 16.09.2015 und 17.3222 Derder vom 17.03.2017. Zugang: https://www.newsd.admin.ch/newsd/message/attachments/50248.pdf (15.12.2019)

Bundesverband Digitale Wirtschaft (2017). BVDW-Studie: Mehrheit nutzt digitale Sprachassistenten. Zugang: https://www.bvdw.org/presse/detail/artikel/bvdw-studie-mehrheit-nutzt-digitale-sprachassistenten/ (26.06.2019)

Burgess, M. (2018). UK police are using AI to inform custodial decisions – but it could be discriminating against the poor, WIRED, 01.03.2018. Zugang: http://www.wired.co.uk/article/police-ai-uk-durham-hart-checkpoint-algorithm-edit (15.12.2019)

Bürgi-Schmelz, A., Bürgisser, M., Schwarzenbach, F.-H., von Arb, C. (1990). Künstliche Intelligenz im menschlichen Umfeld. Forschungspolitische Früherkennung. Bern: Schweizerischer Wissenschaftsrat

Burrus, L. (2016). Technologie et avocature: ROSS et Big Data. Revue de l'avocat, 8, 325–329

Buttarelli, G. (2018). Keynote speech on privacy, data protection and cyber security in the era of AI, Telecommunications and Media Forum: Artificial Intelligence

and the future Digital Single Market. Zugang: https://edps.europa.eu/sites/edp/files/publication/18-04-24_giovanni_buttarelli_keynote_speech_telecoms_forum_en.pdf (15.12.2019)

Calame, T. (2006). Besonderheiten von computerimplementierten Erfindungen. In: von Büren, R., David, L. (Hrsg.). Schweizerisches Immaterialgüter- und Wettbewerbsrecht, Bd. IV, Basel, 651–680

Caliskan, A., Bryson, J. J., Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356: 183–186

Campax. (2019). PostFinance: Keine Stimmprofile ohne Einwilligung! Zugang: https://act.campax.org/petitions/postfinance-keine-stimmprofile-ohne-einwilligung (05.06.2019)

Campbell, M., Campbell, A. J. H. Jr, Hsu, F. (2002). Deep Blue. Artificial Intelligence, 134(1-2), 57–83

Canellas, M. C., Haga, R. A. (2015). Toward Meaningful Human Control of Autonomous Weapons Systems through Function Allocation. IEEE International Symposium on Technology in Society (ISTAS) Proceedings. Zugang: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2927702

Čas, J., Rose, G., Schüttler, L. (2017). Robotik in Österreich, Kurzstudie. Entwicklungsperspektiven und politische Herausforderungen. Vienna: ITA. Zugang: http://epub.oeaw.ac.at/ita/ita-projektberichte/2017-03.pdf (09.05.2018)

Casey, B., Farhan-Gi, A., Vogl, R. (2019). Rethinking explainable machines: The GDPR's «right to explanation» debate and the rise of algorithmic audits in enterprise. Berkeley Technology Law Journal, 34. Zugang: https://ssrn.com/abstract=3143325 (12.02.2019)

Castellano, O. (2018). Will the Next Picasso Be a Robot? Medium, 03.09.2018. Zugang: https://medium.com/s/story/will-the-next-picasso-be-a-robot-9438482b4208 (11.02.2019)

CBS News (2016). 60 minutes vanity fair poll artificial intelligence. Zugang: https://www.cbsnews.com/news/60-minutes-vanity-fair-poll-artificial-intelligence/ (26.06.2019)

CEPEJ (2018). European Commission for the efficiency of Justice (CEPEJ), European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and

their environment, 03./04.12.2018. Zugang: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c (15.12.2019)

Cerka, P., Grigiene, J., Sirbikyte, G. (2015). Liability for damages caused by artificial intelligence. Computer Law & Security Review, 31(3): 376–389

Cerka, P., Grigiene, J., Sirbikyte, G. (2017). Is it possible to grant legal personality to artificial intelligence software systems? Computer Law & Security Review, 33(5): 685–699

Chalmers, T. (2018). 4 Reasons AI Content Curation Is The Next Key Marketing Tool. Forbes. Zugang: https://www.forbes.com/sites/theyec/2018/06/28/4-reasons-ai-content-curation-is-the-next-key-marketing-tool/ (05.06.2019)

Chang, J.-H., Huynh, P. (2016). ASEAN in transformation the future of jobs at risk of automation: International Labour Organization. Zugang: www.ilo.org/wcmsp5/groups/public/---ed_dialogue/---act_emp/documents/publication/wcms_579554.pdf (19.12.2019)

Chang, M., Ventura, M., Ahn, J., Foltz, P. et al. (2018). Dialogue Based Tutoring at Scale: Design and Challenges, CEUR Workshop Proceedings, 2128. Zugang: http://ceur-ws.org/Vol-2128/industrial1.pdf (19.12.2019)

Cheris, A., Rigby, D., Tager, S. (2017). Dreaming of an Amazon Christmas? Bain: https://www.bain.com/insights/retail-holiday-newsletter-2017-issue-2/ (05.06.2019)

Chouldechova, A. (2016). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Zugang: http://arxiv.org/abs/1610.07524 (10.11.2019)

Christen, M., Guillaume, M., Jablonowski, M., Lenhart, P., Moll, K. (2018). Zivile Drohnen – Herausforderungen und Perspektiven. Zürich: vdf Hochschulverlag

Clark, A., Fox, C. C., Lappin, A. (2010). The Handbook of Computational Linguistics and Natural Language Processing. Hoboken: Wiley

Combs, T. D. (2017). «Data is like a Moat!» and Other Bad Ways to Talk About Data and AI. Medium. Zugang: https://medium.com/boundlessai/data-is-like-a-moat-and-other-bad-ways-to-talk-about-data-and-ai-3aa9268db605 (06.06.2019)

Commonwealth Ombudsman (2017). Centrelink's automated debt raising and recovery system. Report No. 2/2017. Zugang: https://www.ombudsman.gov.au/

__data/assets/pdf_file/0022/43528/Report-Centrelinks-automated-debt-raising-and-recovery-system-April-2017.pdf (15.12.2019)

Constine, J. (2017). Facebook rolls out AI to detect suicidal posts before they're reported. Zugang: https://techcrunch.com/2017/11/27/facebook-ai-suicide-prevention/ (15.12.2019)

Contratto, F. (2014). Hochfrequenzhandel und systemische Risiken, Schweizerische Zeitschrift für Gesellschafts- und Kapitalmarktrecht sowie Umstrukturierungen, 143–160

Cooke, A. D. J., Sujan, H., Sujan, M., Weitz, B. A. (2002). Marketing the Unfamiliar: The Role of Context and Item-Specific Information in Electronic Agent Recommendations. Journal of Marketing Research, 39(4): 488–497

Council of Europe (2018a). Details zum Vertrag-Nr. 108. Zugang: https://www.coe.int/de/web/conventions/full-list/-/conventions/treaty/108?_coeconventions_WAR_coeconventionsportlet_languageId=en_GB (10.12.2018)

Council of Europe (2018b). Modernisation of the Data Protection «Convention 108». Zugang: https://www.coe.int/de/web/portal/28-january-data-protection-day-factsheet (10.12.2018)

Council of Europe (2019). Guidelines on Artificial Intelligence and Data Protection, Nr. T-PD(2019)01 25.01. Strassburg. Zugang: https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8 (15.12.2019)

Crootof, R. (2016). A Meaningful Floor for «Meaningful Human Control». Temple International and Comparative Law Journal, 30: 52–62

Curran, D. (2018). Are you ready? Here is all the data Facebook and Google have on you. The Guardian. Zugang: https://www.theguardian.com/commentisfree/2018/mar/28/all-the-data-facebook-google-has-on-you-privacy (26.06.2019)

Dähler, M. (2018). Haftungsrecht. In: Dähler, M., Schaffhauser, R. (Hrsg.). Handbuch Strassenverkehrsrecht, Basel, 1–96

Dale, R. (2016). The return of the chatbots. Natural Language Engineering, 22(5): 811–817

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. Philosophy & Technology, 29: 1–24

Davis, R. Buchanan, B. G., Shortliffe, E. H. (1977). Production Rules as a Representation for a Knowledge-Based Consultation Program. Artificial Intelligence, 8(1): 15–45

Dawar, N., Bendle, N. (2018). Marketing in the age of Alexa. Harvard Business Review, 96(3): 80–86

De Montjoye, Y. A., Quoidbach, J., Robic, F., Pentland, A. S. (2013). Predicting personality using novel mobile phone-based metrics. In: International conference on social computing, behavioral-cultural modeling, and prediction. Berlin/Heidelberg: Springer, 48–55

Decker, M., Fischer, M., Ott, I. (2017). Service Robotics and Human Labor: A first technology assessment of substitution and cooperation, Robotics and Autonomous Systems, 87: 348–354

Dietvorst, B. J., Simmons, J. P., Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1): 114–126

Dinges, M., Leitner, K.-H., Dachs, B., Rhomberg, W., Wepner, B., Bock-Schappelwein, J., Fuchs, S., Horvath, T., Hold, P., Schmid, A. (2017). Beschäftigung und Industrie 4.0. Technologischer Wandel und die Zukunft des Arbeitsmarkts, im Auftrag von: BMVIT, Wien: AIT Austrian Institute of Technology, WIFO, Fraunhofer Austria Research

Djeffal, C. (2018). Künstliche Intelligenz in der öffentlichen Verwaltung, Berichte des NEGZ Nr. 3, 12. Zugang: https://negz.org/wp-content/uploads/2018/11/NEGZ-Kurzstudie-3-KuenstlIntelligenz-20181113-digital.pdf (15.12.2019)

Döbel, I., Leis, M., Vogelsang, M. M., Neustoev, D., Petzka, H., Riemer, A., Rüping, St., Voss, A., Wegele, M., Welz, J. (2018). Maschinelles Lernen, eine Analyse zu Kompetenzen, Forschung und Anwendung, Fraunhofer Gesellschaft. Zugang: https://www.bigdata.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf (15.12.2019)

Doshi-Velez, F., Mason, K. (2017). Accountability of AI Under the Law: The Role of Explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper. Zugang: https://dash.harvard.edu/handle/1/34372584 (15.12.2019)

Dovas, M.-U. (2017). Automatisierte Einzelentscheidungen. Digma, 2: 98–103

Dreyer, S., Schulz, W. (2018). Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?, Bertelsmann Stiftung. Zugang: https://www.hans-bredow-institut.de/uploads/media/Publikationen/cms/media/p4ymg73_BSt_DSGVOundADM_dt.pdf (15.12.2019)

Dunleavy, J. (2019). Feds prepare investigations of Apple, Google, Amazon, and Facebook for anticompetitive activities. Washington Examiner. Zugang: https://www.washingtonexaminer.com/news/feds-prepare-investigations-of-apple-google-amazon-and-facebook-for-anti-competitive-activities (07.06.2019)

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2011). Fairness Through Awareness. Zugang: http://arxiv.org/abs/1104.3913 (15.12.2019)

Ebneter, M. (2010). Warum Tiere nicht Teil unserer Rechtsordnung sein können, Roboter es aber werden könnten. Jusletter vom 10. Mai 2010

EC (2018). Europäische Kommission. Artificial Intelligence for Europe, Brüssel. Zugang: https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe (15.12.2019)

EC (2019). Europäische Kommission, HEG-KI (2019): Ethik-Leitlinien für eine vertrauenswürdige KI. Zugang: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60425 (08.07.2019)

EDK (2014) Deutschschweizer Erziehungsdirektoren-Konferenz. Lehrplan 21. Zugang: https://v-ef.lehrplan.ch/index.php?code=b%7C10%7C0&la=yes (13.01.2020)

EDK (2018). Digitalisierungsstrategie. Zugang: http://www.edk.ch/dyn/31425.php (13.01.2020)

EDK (2019). Massnahmen zur Digitalisierungsstrategie der EDK. Zugang: http://www.edk.ch/dyn/12277.php (13.01.2020)

Edulog (2019). Organisation, Architektur, Prozesse. Zugang: https://www.edulog.ch/de/service/organisation-architektur-und-prozesse (13.01.2020)

Edwards, L., Veale, M. (2018). Enslaving the Algorithm: From a «right to an explanation» to a «right to better decisions»? IEEE Security & Privacy, 16(3): 46–54

EGE Europäische Gruppe für Ethik der Naturwissenschaften und der neuen Technologien (2018). Erklärung zu künstlicher Intelligenz, Robotik und «autonomen» Systemen. In: RTD.01 – Mechanismus für wissenschaftliche Beratung und Ge-

neraldirektion Forschung und Innovation, Brüssel: Europäische Kommission. Zugang: https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018_de.pdf (15.12.2019)

Elsevier (2018). Artificial Intelligence: How knowledge is created, transferred, and used: Trends in China, Europe and the United States. Zugang: https://www.elsevier.com/connect/resource-center/artificial-intelligence (15.12.2019)

Enago Academy (2018). Artificial Intelligence in Research and Publishing, 4. Juni 2018. Zugang: https://www.enago.com/academy/artificial-intelligence-research-publishing/ (15.12.2019)

EPA (2018). Richtlinien für die Prüfung im Europäischen Patentamt, Version vom 1. November 2018. Zugang: http://documents.epo.org/projects/babylon/eponet.nsf/0/2A358516CE34385CC125833700498332/$File/guidelines_for_examination_2018_hyperlinked_de.pdf (29.01.2019)

Epley, N., Waytz, A., Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. Psychological Review, 114(4): 864–886

Extance, A., (2018). How AI technology can tame the scientific literature. Nature, 561: 273–274

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D. (2018). Robust Physical-World Attacks on Deep Learning Models. Zugang: https://arxiv.org/abs/1707.08945 (29.12.2019)

Facebook (2018). How Facebook AI Helps Suicide Prevention. Zugang: https://newsroom.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/ (26.06.2019)

Fellmann, W., Werro, F. (2013). Kommentierung von Art. 60 OR 2020. In: Huguenin, C., Hilty, W. (Hrsg.). Schweizer Obligationenrecht 2020: Entwurf für einen neuen allgemeinen Teil. Zürich, 187–188

Ferber, M. (2018). Bei den Steuern bringt Digitalisierung mehr Effizienz, aber auch das Risiko gläserner Bürger. NZZ, 11.05.2018. Zugang: https://www.nzz.ch/finanzen/bei-den-steuern-bringt-die-digitalisierung-mehr-effizienz-aber-auch-das-risiko-glaeserner-buerger-ld.1384807 (15.12.2019)

Ferguson, A. G. (2017). The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement. New York: NYU Press

Fichter, A. (2019). Der Spion im Klassenzimmer. Republik, 02.07.2019 Zugang: https://www.republik.ch/2019/07/02/der-spion-im-schulzimmer (15.12.2019)

Field, J., Hilligoss, H., Achten, N., Levy, D. M., Feldman, J., Kagay, S. (2019). A Map of Ethical and Rights-Based Approaches; The Principled Artificial Intelligence Project. Harvard University. Zugang: https://ai-hr.cyber.harvard.edu/primp-viz.html (06.07.2019)

Fischer, J. M., Ravizza, M. (2000). Responsibility and control: A theory of moral responsibility. Cambridge: Cambridge University Press

Flach, P. (2012). Machine Learning: The Art and Science of Algorithms That Make Sense of Data. Cambridge: Cambridge University Press

Flaxman, S., Goel, S., Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. Public Opinion Quarterly, 80: 298–320

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E. (2018). AI 4 People's Ethical Framwork for a good Society: Opportunities, Risks, Principles, and Recommendations. Brussels: Atomium. Zugang: http://www.eismd.eu/wp-content/uploads/2018/11/Ethical-Framework-for-a-good-AI-Society.pdf (15.12.2019)

Følstad, A., Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. interactions, 24(4): 38–42

Franke, S. (2019). Biometrisches Stimmprofil bei Swisscom Support – intransparentes Vorgehen. Corporate Dialog Website. Zugang: https://corporate-dialog.ch/2019/02/28/biometrisches-stimmprofil-bei-swisscom-support/ (05.06.2019)

Fraser, E. (2016). Computers as Inventors – Legal and Policy Implications of Artificial Intelligence on Patent Law. SCRIPTed, 13(3): 306–333

Frenzel, E. M. (2018). Kommentierung zu Art. 5 DSGVO. in: Paal, B. P., Pauly, D. A. (Hrsg.). DSGVO Kommentar, 2. Aufl., München

Frey, C. B., Osborne, M., Holmes, C., Rahbari, E., Garlick, R., Friedlander, G., McDonald, G., Curmi, E., Chua, J., Chalif, P. (2016). Technology at work v2.0: The future is not what it used to be. CityGroup and University of Oxford

Frey, C. B., Osborne, M. A. (2013). The future of employment: How susceptible are jobs to computerisation? Zugang: https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf (12.01.2018)

Freytag, U. (2016). Sicherheitsrechtliche Aspekte der Robotik. Sicherheit & Recht, 2: 111–121

Frontier Economics (2018). The Impact of Artificial Intelligence on Work, An evidence review prepared for the Royal Society and the British Academy, UK. Zugang: https://royalsociety.org/~/media/policy/projects/ai-and-work/frontier-review-the-impact-of-AI-on-work.pdf (15.12.2019)

Future Customer (2018). Sentiment Analysis in Customer Service: Understanding Human Emotions. Zugang: https://www.future-customer.com/sentiment-analysis-customer-service-understanding-human-emotions/ (29.06.2019)

GAML (2018). Global Alliance to Monitor Learning. Pathway Mapping Methodology. Montreal: UIS

Gentsch, P. (2018). AI in Marketing, Sales and Service: How Marketers Without a Data Science Degree Can Use AI, Big Data and Bots. Heidelberg: Springer

Gentzkow, M., Shapiro, J. M. (2011). Ideological Segregation Online and Offline. The Quarterly Journal of Economics, 126: 1799–1839

George, D., Reutimann, K., Tamò-Larrieux, A. (2018). GDPR Bypass by Design? Transient Processing of Data Under the GDPR. Zugang: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3243389 (12.02.2019)

Gerstner, D. (2017). Predictive Policing als Instrument zur Prävention von Wohnungseinbruchdiebstahl. Freiburg i. Br.: Max-Planck-Institut für ausländisches und internationales Strafrecht

Ghosh, S., Ganguly, N., Mitra, B., De, P. (2017). Evaluating effectiveness of smartphone typing as an indicator of user emotion. Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), 146–151. https://doi.org/10.1109/ACII.2017.8273592

Goggin, B. (2019). Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts. Zugang: https://www.businessinsider.com/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12?r=US&IR=T (26.06.2019)

Gomez-Uribe, C. A., Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems, 6(4): 13–19

Goodman, B., Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a «right to explanation». AI Magazine, 38(3). Zugang https://arxiv.org/abs/1606.08813 (12.02.2019)

Graff, B. (2016). Rassistischer Chat-Roboter: Mit falschen Werten bombardiert, Süddeutsche Zeitung, 03.04.2016. Zugang: https://www.sueddeutsche.de/digital/microsoft-programm-tay-rassistischer-chat-roboter-mit-falschen-werten-bombardiert-1.2928421 (15.12.2019)

Graff, B. (2018). Robo-Journalismus. Süddeutsche Zeitung, 29.03.2018. Zugang: https://www.sueddeutsche.de/kultur/kuenstliche-intelligenz-robo-journalismus-1.3921660 (11.02.2019)

Greengard, S. (2015). The Internet of Things. Cambridge: MIT Press

Grigore, E. C., Pereira, A., Zhou, I., Wang, D., Scassellati, B. (2016). Talk to me: Verbal communication improves perceptions of friendship and social presence in human-robot interaction. Proceedings of the International conference on intelligent virtual agents. Springer, 51–63

Guess, A. (2018). Online Political Conversations. In: Tucker, J. A. et al. (eds.). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature, 9–14. Zugang: https://hewlett.org/library/social-media-political-polarization-political-disinformation-review-scientific-literature/ (15.12.2019)

Häberli, S., Müller, T. (2018). Autonomes Fahren ist noch ein Luftschloss. NZZ, 22.04.2018. Zugang: www.nzz.ch/wirtschaft/autonomes-fahren-ist-noch-ein-luftschloss-ld.1379422 (07.02.2019)

Haim, M., Graefe, A., Brosius, H.-B. (2017). Burst of the Filter Bubble? Digital Journalism, 6: 330–343

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors: The Journal of the Human Factors and Ergonomics Society, 53(5): 517–527

Hänsenberger, S. (2017). Wenn Drohnen vom Himmel fallen – luftrechtliche Haftungsfragen. Allgemeine Juristische Praxis, 2: 163–170

Hänsenberger, S. (2018). Die Haftung für Produkte mit lernfähigen Algorithmen. Jusletter, 26.11.2018

Hardt, M., Price, E., Srebro, N. (2016). Equality of Opportunity in Supervised Learning. Zugang: https://arxiv.org/abs/1610.02413 (12.01.2018)

Hartmann, F., Prinz, M. (2018). Immaterialgüterrechtlicher Schutz von Systemen Künstlicher Intelligenz. Wettbewerb in Recht und Praxis, 12: 1431–1438

Heinrich, P. (2018). PatG/EPÜ – Schweizerisches Patentgesetz/Europäisches Patentübereinkommen: Kommentar in synoptischer Darstellung. 3. Aufl., Bern

Henning, K., Süthoff, M., Mai, M. (Hrsg.) (1990). Mensch und Automatisierung. Eine Bestandesaufnahme. Opladen: Westdeutscher Verlag

Hermann, M., Pentek, T., Otto, B. (2016). Design Principles for Industry 4.0 Scenarios. HICSS, 3928–3937

Hess, H. J. (2016). Stämpflis Handkommentar PrHG. 3. Auflage, Bern

Hetmank, S., Lauber-Rönsberg, A. (2018). Künstliche Intelligenz – Herausforderungen für das Immaterialgüterrecht. GRUR, 6: 574–582

Hildebrandt, M. (2016). New Animism in Policing: Reanimating the Rule of Law? in: Bradford, B. et al. (Hrsg.). SAGE Handbook of Global Policing. London: SAGE, 406–428

Hill, K. (2012). How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did. Zugang: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#3f2b01f96668 (19.12.2019)

Hoff, K. A., Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3): 407–434

Holmes, W., Huang, R., Miao, F. (2019). Policy Guidelines for Artificial Intelligence in Education (Draft for comments), as of 6 March 2019, UNESCO

Horowitz, M. C., Scharre, P. (2015). Meaningful human control in weapon systems. Zugang: https://www.files.ethz.ch/isn/189786/Ethical_Autonomy_Working_Paper_031315.pdf (15.12.2019)

House of Lords (2017a). AI in the UK: ready, willing and able? – Artificial Intelligence Committee. Zugang: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm (07.06.2019)

House of Lords (2017b). Written evidence – Digital Catapult. Zugang: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69688.html (07.06.2019)

House of Lords (2018). AI in the UK: ready, willing and able? HL Paper 100. Zugang: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf (15.12.2019)

Huguenin, C. (2014). Obligationenrecht, 2. Auflage, Zürich

Huguenin, C., Hilty, R. (Hrsg.) (2013). Schweizer Obligationenrecht 2020 / Code des obligations Suisse 2020. Zürich

IEEE (2017). Ethically Aligned Design – Version 2. A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems (AI/AS). The Institute of Electrical and Electronics Engineers, Incorporated (IEEE) Global Initiative. Zugang: https://ieeexplore.ieee.org/document/8058187 (19.12.2019)

ILO (2017). Inception Report for the Global Commission on the Future of Work. Geneva. Zugang: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---cabinet/documents/publication/wcms_591502.pdf (15.12.2019)

ILO (2019). Work for a brighter future. Global Commission on the Future of Work. Geneva. Zugang: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---cabinet/documents/publication/wcms_662410.pdf (15.12.2019)

ISTA, CSTA (2011). Operational Definition of Computational Thinking for K-12 Education. Zugang: https://id.iste.org/docs/ct-documents/computational-thinking-operational-definition-flyer.pdf?sfvrsn=2 (16.01.2020)

ITU (2018). United Nations Activities on Artificial Intelligence (AI). Zugang: http://handle.itu.int/11.1002/pub/8120d5d5-en (15.12.2019)

Jago, A. S., Laurin, K. (2017). Technology and (in)discrimination. Talk presented at Academy of Management Annual Meeting 2017, Atlanta, GA, August 4–8th 2017

Jahn, T. (2008). Transdisziplinarität in der Forschungspraxis. In: Bergmann, M., Schramm, E. (Hrsg.). Transdisziplinäre Forschung. Integrative Forschungsprozesse verstehen und bewerten. Frankfurt/NewYork: Campus Verlag, 21–37

James, E. A., Milekiewicz, M. T. Bucknam, A. (2008). Participatory action research for educational leadership: Using data-driven decision making to improve schools. Sage

Jaquemart, C., Kobler, E. (2017). Riesiger Bedarf an Weiterbildung der Lehrer. SRF. Zugang: www.srf.ch/news/schweiz/riesiger-bedarf-an-weiterbildung-der-lehrer (15.12.2019)

Jobin, A., Ienca, M., Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1: 389–399

John, L. K. (2018). How Far Can the Surveillance Economy Go? Harvard Business Review. Zugang: https://hbr.org/2018/09/uninformed-consent (15.12.2019)

Johnson, M. et al. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. TACL, 5: 339–351

Jovanovic, B., Rousseau, P. L. (2005). General Purpose Technologies. In: Aghion, P., Durlauf, S. (Hrsg.). Handbook of Economic Growth. Amsterdam: Elsevier, 1181–1224

Jürgens, K., Hoffmann, R., Schildmann, C. (2017). Arbeit transformieren! Denkanstösse der Kommission «Arbeit der Zukunft»: Bielefeld: transcript Verlag

Just, N., Latzer, M. (2017). Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet. Media, Culture & Society, 39(2): 238–258

Kaminski, M. E. (2019). The right to explanation, explained. Berkeley Technology Law Journal, 34(1). Zugang: https://osf.io/preprints/lawarxiv/rgeus (12.02.2019)

Kamlah, W. (2018). Kommentierung zu Art. 22 DSGVO. In: Plath, K.-U. (Hrsg.). DSGVO/BDSG. Kommentar, 3. Aufl., Köln

Karahalios, K. (2014). Algorithm Awareness. How the news feed on Facebook decides what you get to see. MIT Technology Review. Zugang: https://www.technologyreview.com/s/531676/algorithm-awareness/ (15.12.2019)

Kehl, D., Guo, P., Kessler, S. (2017). Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Zugang: https://dash.harvard.edu/handle/1/33746041 (19.12.2019)

Kelly, K. (2017). The Myth of a superhuman AI. Wired. Zugang: https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/ (12.01.2018)

Kiener, R., Kälin, W., Wyttenbach, J. (2018). Grundrechte. Bern: Stämpfli Verlag

Kleinberg, J., Mullainathan, S., Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. Zugang: http://arxiv.org/abs/1609.05807 (15.12.2019)

Kleinberg, J., Mullainathan, S., Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. Zugang: https://arxiv.org/abs/1609.05807v2 (12.01.2018)

Klose, R. (2018). Harte Brocken, intelligent geknackt, Empa Quarterly, 61: 14–15

Knaus, C. (2017). Internal Centrelink records reveal flaws behind debt recovery system. The Guardian, 13.01.2017. Zugang: https://www.theguardian.com/australia-news/2017/jan/13/internal-centrelink-records-reveal-flaws-behind-debt-recove-ry-system (19.12.2019)

Knight, W. (2017). The Dark Secret at the Heart of AI. MIT Technology Review, 11.04.2017. Zugang: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (12.01.2018)

Knobloch, T. (2018). Vor die Lage kommen: predictive policing in Deutschland, Stiftung Neue Verantwortung/Bertelsmann Stiftung. Zugang: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/predictive.policing.pdf (19.12.2019)

Korinek, A., Stiglitz, J. (2017). Artificial Intelligence and Its Implications for Income Distribution and Unemployment: National Bureau of Economic Research. Zugang: https://dx.doi.org/10.3386/w24174 (19.12.2019)

Kosinski, M., Stillwell, D., Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15): 5802–5805

Krafft, T. D., Zweig, K. A. (2018). Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann. In: R. Mohabbat Kar, B. E. P. Thapa, P. Parycek (Hrsg.). (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft. Berlin: Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-57621-7 (19,12,19), 471–492

Krasser, R., Ann, C. (2016). Patentrecht, 7. Aufl., München

Kuner, C., Cate, F. H., Millard, C., Svantesson, D. J. B. (2012). Challenge of «big data» for data protection. International Data Privacy Law, 2(2): 47–49

Kuner, C., Svantesson, D. J. B., Cate, F. H., Lynskey, O., Millard, C. (2017). Machine learning with personal data: is data protection law smart enough to meet the challenge? International Data Privacy Law, 7(1): 1–2

Kwong, K. (2017). The Algorithm Says You Did It: The Use of Black Box Algorithms to Analyze Complex DNA Evidence. Harvard Journal of Law & Technology, 31(1): 275–301

Lakshmanan, R. (2019). Google Maps' new features for India just made my commute a lot less painful. Zugang: https://thenextweb.com/apps/2019/06/04/google-maps-new-features-for-india-just-made-my-commute-a-lot-less-painful/ (26.06.2019)

Landtag Baden-Württemberg (2011). Mitteilung der Landesregierung vom 14.12.2011, Drucksache 15/1047. Zugang: http://www.landtag-bw.de/files/live/sites/LTBW/files/dokumente/WP15/Drucksachen/1000/15_1047_D.pdf (12.01.2018)

Larson, J., Angwin, J. (2016). Machine Bias. ProPublica. Zugang: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (29.06.2019)

Larus, J., Hankin, C., Carson, S. G., Christen, M., Crafa, S., Grau, O., Kirchner, C., Knowles, B., McGettrick, A., Tamburri, D. A., Werthner, H. (2018): When Computers Decide. European Recommendations on Machine-Learned Automated Decision Making. Informatics Europe & EUACM. Zugang: https://dl.acm.org/citation.cfm?id=3185595 (19.12.2019)

Lawrence, M., Roberts, C., King, L. (2017). Managing automation: Employment, inequality and ethics in the digital age. London: Institute for Public Policy Research. Zugang: https://www.ippr.org/files/2018-01/cej-managing-automation-december2017.pdf (19.12.2019)

Lebsanft, E. W., Gill, U. (1987). Expertensysteme in der Praxis – Kriterien für die Verwendung von Expertensystemen zur Problemlösung. In: Savory, S. E. (Hrsg.). Expertensysteme: Nutzen für Ihr Unternehmen. Ein Leitfaden für Entscheidungsträger. München/Wien, 135–149.

LeCun, Y., Bengio, Y., Hinton, G. E. (2015). Deep learning. Nature, 521: 436–444

Lee, F. (2017). Die AAA-Bürger. Zeit Online. Zugang: http://www.zeit.de/digital/datenschutz/2017-11/china-social-credit-system-buergerbewertung (12.01.2018)

Leese, M. (2018). Predictive Policing in der Schweiz: Chancen, Herausforderungen, Risiken. In: Nünlist, C. (Hrsg.). Bulletin 2018 zur schweizerischen Sicherheitspolitik, 57–71

Lenk, K. (2018). Formen und Folgen algorithmischer Public Governance. In: R. Mohabbat Kar, B. E. P. Thapa, P. Parycek (Hrsg.). (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft. Berlin: Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT). https://nbn- resolving.org/urn:nbn:de:0168-ssoar-57541-2 (19.12.2019), 228–267

Leveringhaus, A. (2016). What's So Bad About Killer Robots? Journal of Applied Philosophy, 35(2): 341–358

Levinson, R., Hsu, F., Schaeffer, J., Marsland, T. A., Wilkins, D. E. (1991). The Role of Chess in Artificial Intelligence Research. ICGA Journal, 14(3): 153–161

Logg, J. M., Minson, J. A., Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151: 90–103

Lohmann, M. F. (2016). Automatisierte Fahrzeuge im Lichte des Schweizer Zulassungs- und Haftungsrechts. Robotik und Recht, Bd. 7, Baden-Baden

Lohmann, M. F., (2017). Roboter als Wundertüten – eine zivilrechtliche Haftungsanalyse. Allgemeine Juristische Praxis, 2: 152–162

Lohmann, M. F., Müller-Chen, M. (2017). Selbstlernende Fahrzeuge – Eine Haftungsanalyse. Schweizerische Zeitschrift für Wirtschafts- und Finanzmarktrecht, 1: 48–58

Loi, M. (2018). Können Algorithmen fair sein? InsideIT, 12.11.2018. Zugang: https://www.inside-it.ch/articles/52806 (19.12.2019)

Lopez, A. (2008). Statistical machine translation. ACM Computing Surveys, 40(3): 1–8

Lovergine, S., Pellero, A. (2018). This Time it Might be Different: Analysis of the Impact of Digitalization on the Labour Market. European Scientific Journal, 14(36): 68. Zugang: http://eujournal.org/index.php/esj/article/download/11600/11067 (19.12.2019)

Luckerson, V. (2015). Here's How Facebook's News Feed Actually Works. Time, 09.07.2015. Zugang: http://time.com/collection-post/3950525/facebook-news-feed-algorithm/ (12.01.2018)

Mankiw, G. (2016). The Economy Is Rigged, and Other Presidential Campaign Myths. The New York Times, 06.05.2016. Zugang: https://www.nytimes.com/2016/05/08/upshot/the-economy-is-rigged-and-other-presidential-campaign-myths.html?_r=0 (15.12.2019)

Mantello, P. (2016). The Machine That Ate Bad People: The Ontopolitics of the Precrime Assemblage. Big Data & Society, 3(2): 1–11

Manyika, J., Chui, M., Miremadi, M., Bughin, J., George, K., Willmott, P., Dewhurst, M. (2017). A Future that Works: Automation, Employment, and Productivity. McKinsey Global Institute. Zugang: https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Digital%20Disruption/Harnessing%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works_Full-report.ashx (19.12.2019)

Marauhn, T. (2018). Meaningful Human Control – and the Politics of International Law. In: von Heinegg H. et al. (eds.). The Dehumanization of Warfare – Legal Implications of New Weapon Technologies, Cham: Springer, 207–218

Mari, A. (2019). The Rise of Machine Learning in Marketing: Goal, Process, and Benefit of AI-Driven Marketing. Research Report, University of Zurich. DOI: 10.13140 /RG.2.2.16328.16649

Marly, J. (2018). Praxishandbuch Softwarerecht, 7. Auflage, München

Marsh, J. A., Pane, J. F., Hamilton, L. S. (2006). Making sense of data-driven decision making in education: Evidence from Recent RAND Research, RAND Education Occasional Paper, RAND Corporation

Martin, A., Medigan, D. (2006). Digital literacies for learning. London: Facet Publishing

Martini, M. (2019). Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz. Wiesbaden: Springer

Marvin, G. (2019). Updated: A visual history of Google ad labeling in search results. Zugang: https://searchengineland.com/search-ad-labeling-history-google-bing-254332 (26.06.2019)

Marwick, A., Lewis, R. (2017). Media Manipulation and Disinformation Online. Data and Society Institute. Zugang: https://datasociety.net/output/media-manipulation-and-disinfo-online/ (19.12.2019)

Matz, S. C., Kosinski, M., Nave, G., Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the National Academy of Sciences, 114(48): 12714–12719

McDonald, A. M., Cranor, L. F. (2008). The Cost of Reading Privacy Policies. Zugang: https://kb.osu.edu/handle/1811/72839 (19.12.2019)

Mead, R. (2016). Learn Different: Silicon Valley disrupts education. The New Yorker, 07.03.2016. Zugang: https://www.newyorker.com/magazine/2016/03/07/altschools- disrupted-education (12.01.2018)

Melullis, K.-J. (2015). Kommentierung zu § 6 PatG/DE. In: Benkard, G. (Hrsg.). Kommentar. Patentgesetz. Gebrauchsmustergesetz. Patentkostengesetz, 11. Auflage, München

Mendoza, I., Bygrave, L. A. (2017). The Right Not to be Subject to Automated Decisions Based on Profiling. In: Synodinou et al. (eds.). EU Internet Law – Regulation and Enforcement. Cham: Springer, 77–98

Merriam Webster Dictionary (2019). Privacy. Zugang: https://www.merriam-webster.com/dictionary/privacy (08.06.2019)

Messing, S., Westwood, S. J. (2013). Selective Exposure in the Age of Social Media. Communication Research, 41: 1042–1063

Metzler, B., Siegrist, P. (2019). Am HB buhlen 840 Beacons um die Smartphones der Passanten. Tages-Anzeiger. Zugang: https://www.tagesanzeiger.ch/zuerich/stadt/am-hauptbahnhof-buhlen-840-beacons-um-die-smartphones-der-passanten/story/10989198 (19.12.2019)

Meuldijk, M., Wattenhofer, T. (2017). Auswirkungen der Digitalisierung auf den Beruf des Wirtschaftsprüfers. Expert Focus, 11: 766–772

Meyer, S. (2018). Künstliche Intelligenz und die Rolle des Rechts für Innovation. ZRP, 233–238

Mitchell, T. M., Caruana, R., Freitag, D., McDermott, J., Zabowski, D. (1994). Experience with a learning personal assistant. Communications of the ACM, 37(7): 80–91

Mohamed, S., Abdelmoty, A. I. (2017). Spatio-semantic user profiles in location-based social networks. International Journal of Data Science and Analytics, 4(2): 127–142

Mohammadyari, S., Singh, H. (2015). Understanding the effect of e-learning on individual performance: The role of digital literacy. Computers & Education, 82: 11–52

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. (2011). Self-Exciting Point Process Modeling of Crime. Journal of the American Statistical Association, 106: 100–108

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., Brantingham, P. J. (2015). Ramdomized Controlled Field Trials of Predictive Policing. Journal of the American Statistical Association, 110: 1399–1411

Möller, J., Trilling, D., Helberger, N., van Es, B. (2018). Do not blame it on the algorithm. An empirical assessment of multiple recommender systems and their impact on content diversity. Information, Communication & Society, 21: 959–977

Monetary Authority of Singapore (2018). Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector. Zugang: http://www.mas.gov.sg/News-and-Publications/Monographs-and-Information-Papers/2018/FEAT.aspx (19.12.2019)

Monsch, M. (2018). Hochfrequenzhandel. Schweizer Schriften zum Finanzmarktrecht. Zürich: Schulthess Verlag

Morse, S. J. (1994). Culpability and control. U Penn Law Review, 142: 1587–1621

Moser, A. (2019). Künstliche Intelligenz wird zum Superhelden der Polizei, SRF, 3.8.2019. Zugang: https://www.srf.ch/news/schweiz/digitalisierung-bei-behoerden-kuenstliche-intelligenz-wird-zum-superhelden-der-polizei (19.12.2019)

Müller, M. F. (2014). Roboter und Recht. Aktuelle Juristische Praxis, 5: 595–608

Münch, P., Herzog, N. (2002). Berechtigung an der Erfindung. In: Bertschinger, C., Geiser, T., Münch, P. (Hrsg.). Handbücher für die Anwaltspraxis, Bd. VI: Schweizerisches und europäisches Patentrecht. Basel, 163–187

Murphy, R. R. (2000). Introduction to AI Robotics. Cambridge: MIT Press

Murray, K. B., Häubl, G. (2009). Personalization without Interrogation: Towards more Effective Interactions between Consumers and Feature-Based Recommendation Agents. Journal of Interactive Marketing, 23: 138–146

Nagl, W., Titelbach, G., Valkova, K. (2017). Digitalisierung der Arbeit: Substituierbarkeit von Berufen im Zuge der Automatisierung durch Industrie 4.0; Endbericht. Zugang: https://irihs.ihs.ac.at/id/eprint/4231/1/200800.pdf (19.12.2019)

National Science Foundation (2019). Astronomers capture first image of black hole. Zugang: https://www.nsf.gov/news/news_summ.jsp?cntn_id=298276 (13.01.2020)

Neapolitan, R. E. (1990). Probabilistic reasoning in expert systems – theory and algorithms. Hoboken: Wiley

Neff, E. F., Arn, M. (1998). Urheberrechtlicher Schutz der Software. In: von Büren, R., David, L. (Hrsg.). Schweizerisches Immaterialgüter- und Wettbewerbsrecht, Bd. II/2. Basel, 1–348

Newman, N. (2017). Reuters Institute Digital News Report 2017. Reuters Institute for the Study of Journalism. Zugang: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf (19.12.2019)

Noto La Diega, G. (2018). Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information. JIPITEC. Zugang: https://www.jipitec.eu/issues/jipitec-9-1-2018/4677

Official Google Blog (2009). The bright side of sitting in traffic: Crowdsourcing road congestion data. Zugang: https://googleblog.blogspot.com/2009/08/bright-side-of-sitting-in-traffic.html (05.06.2019)

Osterrieth, C. (2015). Patentrecht, 5. Aufl., München

Ovum (2017). Virtual digital assistants to overtake world population by 2021. Zugang: https://ovum.informa.com/resources/product-content/virtual-digital-assistants-to-overtake-world-population-by-2021 (04.06.2019)

Pagallo, U. (2018). Algo-Rhythms and the Beat of the Legal Drum, Philosophy & Technology, 31(4): 507–524

Pariser, E. (2012). The filter bubble. What the Internet is hiding from you. London: Penguin Books

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Cambridge: Harvard University Press

PC Pitstop (2012). $ 1000 for Reading a EULA (7 Years Later). Zugang: https://techtalk.pcpitstop.com/2012/06/12/it-pays-to-read-license-agreements-7-years-later/ (11.06.2019)

Pedrazzini, M. M., Hilti, C. (2008). Europäisches und schweizerisches Patent- und Patentprozessrecht, 3. Auflage, Bern

Pedreschi, D., Ruggieri, S., Turini, F. (2008). Discrimination-Aware Data Mining. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 560–568

Pemberton Levy, H. (2016). Gartner Predicts a Virtual World of Exponential Change. Gartner. Zugang: https://www.gartner.com/smarterwithgartner/gartner-predicts-a-virtual-world-of-exponential-change/ (12.01.2018)

Peneder, M., Bock-Schappelwein, J., Firgo, M., Fritz, O., Streicher, G. (2016). Österreich im Wandel der Digitalisierung. Österreichisches Institut für Wirtschaftsforschung. Zugang: https://www.wifo.ac.at/jart/prj3/wifo/resources/person_dokument/person_dokument.jart?publikationsid=58979&mime_type=application/pdf (19.12.2019)

Pradeep, A. K., Appel, A., Sthanunathan, S. (2018). AI for Marketing and Product Innovation: Powerful New Tools for Predicting Trends, Connecting with Customers, and Closing Sales. Hoboken: Wiley

Price, S., Flach, P. A., Spiegler, S., Bailey, C., Rogers, N. (2013). SubSift web services and workflows for profiling and comparing scientists and their published works, Future Generations Computer Systems, 29(2): 569–581

PwC (2018). Vertrauen in Medien. PwC Deutschland. Zugang: https://www.pwc.de/de/technologie-medien-und-telekommunikation/studie-vertrauen-in-medien.html (05.06.2019)

Ramzeen, A. V. (2019). 72 Facebook Acquisitions – The Complete List. Zugang: https://www.techwyse.com/blog/infographics/facebook-acquisitions-the-complete-list-infographic/ (29.06.2019)

Rechsteiner, D. (2018). Der Algorithmus verfügt. Jusletter, 26. November 2018

Ricci, F., Rokach, L., Shapira, B. (eds.) (2015). Recommender Systems Handbook. Springer

Rid, T. (2016). Maschinendämmerung. Eine kurze Geschichte der Kybernetik. Berlin: Propyläen Verlag

Roff, H. M. (2013). Responsibility, liability, and lethal autonomous robots. Routledge Handbook of Ethics and War: Just War Theory in the 21st Century. London: Routledge, 352–364

Roll, I., Wylie, R. (2016). Evolution and Revolution in Artificial Intelligence in Education. International Journal of Artificial Intelligence in Education, 26(2): 582–599

Rosenfeld, A., Sina, S., Sarne, D., Avidov, O., Kraus, S. (2018). A Study of WhatsApp Usage Patterns and Prediction Models without Message Content. Zugang: http://arxiv.org/abs/1802.03393 (19.12.2019)

Rosenthal, D. (2008). Autonome Informatiksysteme: Wie steht es mit der Haftung? In: Kündig, A., Bütschi, D. (Hrsg.). Die Verselbstständigung des Computers, TA-SWISS 51/2008, Zürich, 131–144

Rosenthal, D. (2017). Vorentwurf für ein neues Datenschutzgesetz: Was er bedeutet, Jusletter, 20. Februar 2017

Rossow, A. (2018). Artificial Intelligence and Smart Technology Is Bringing Convenience To A Whole New Level. Zugang: https://www.forbes.com/sites/andrewrossow/2018/05/24/artificial-intelligence-taking-convenience-to-a-whole-new-level/#28f5f8b34504 (19.12.2019)

Rozenblit, L., Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive science, 26(5): 521–562

Russell, S. J., Norvig, P. (2010). Artificial Intelligence – A Modern Approach. London: Pearson Education

Samore, W. (2013). Artificial Intelligence and The Patent System: Can a New Tool Render a Once Patentable Idea Obvious? Syracuse Journal of Science and Technology Law, 29: 113–142

Sanovich, S., Stukal, D. (2018). Strategies and Tactics of Spreading Disinformation through Online Platforms. In: Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., Nyhan, B. (eds.). Social Media, Political

Polarization, and Political Disinformation: A Review of the Scientific Literature. New York: Hewelett Foundation, 30–39

Sassoli, M. 2014). Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified. International Law Studies / Naval War College. Zugang: https://archive-ouverte.unige.ch/unige:37976 (19.12.2019)

Saunders, J., Hunt, P., Hollywood, J. S. (2016). Predictions Put Into Practice: A Quasi-experimental Evaluation of Chicago's Predictive Policing Pilot. Journal of Experimental Criminology, 12(3): 347–371

SBFI (2016). F&I Bericht, Internationaler Vergleich. Zugang: https://www.sbfi.admin.ch/sbfi/de/home/forschung-und-innovation/forschung-und-innovation-in-der-schweiz/f-und-i-bericht/internationaler-vergleich/12-informations--und-kommunikationstechnologien.html (19.12.2019)

SBFI (2017). Herausforderungen der Digitalisierung für Bildung und Forschung in der Schweiz. Zugang: https://www.digitale21.ch/wp-content/uploads/Herausforderungen-der-Digitalisierung-fuer-Bildung-und-Forschung-in-der-Schweiz.pdf (19.12.2019)

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 58(3): 377–400

Schafer, B., Komuves, D., Zatarain, J. M. N. et al. (2015). A fourth law of robotics? Copyright and the law and ethics of machine co-production. Artificial Intelligence and Law, 23(3): 217–240

Schafer, J. B., Konstan, J. A., Riedl, J. (2002). Meta-recommendation systems: user-controlled integration of diverse recommendations. Proceedings of the eleventh international conference on Information and knowledge management, 43–51

Schmidt, E. (2008). Moderne Steuerungssysteme im Steuervollzug. DStJG Band 31, Köln, 37–57

Schneider, G. (2019). Wie KI bei der Medikamententwicklung hilft. Zukunftsblog – ETH Zürich. Zugang: https://ethz.ch/de/news-und-veranstaltungen/eth-news/news/2019/03/blog-schneider-ai-medikamentenentwicklung.html (19.12.2019)

Schönberger, D. (2018). Deep Copyright: Up- and Downstream Questions Related to Artificial Intelligence and Machine Learning. In: de Werra (Hrsg.). Droit d'auteur 4.0 / Copyright 4.0, PI – P®opriété intelle©tuelle. Genève, 145–173

Schulze, G. (2018). Geschützte Werke. In: Dreier, T., Schulze, G. (Hrsg.). Urheberrechtsgesetz. München, 79–162

Scott, B., Heumann, S., Lorenz, P. (2017). Artificial Intelligence and Foreign Policy. Stiftung neue Verantwortung. Zugang: https://www.stiftung-nv.de/sites/default/files/ai_foreign_policy.pdf (12.01.2018)

SECO (2016). 12. Bericht des Observatoriums zum Freizügigkeitsabkommen Schweiz – EU: Auswirkungen der Personenfreizügigkeit auf den Schweizer Arbeitsmarkt. Zugang: https://www.seco.admin.ch/seco/de/home/Publikationen_Dienstleistungen/Publikationen_und_Formulare/Arbeit/Personenfreizuegigkeit_und_Arbeitsbeziehungen/observatoriumsberichte/12_Bericht_Observatorium.html

Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., Lepri, B. (2017). What your Facebook profile picture reveals about your personality. Proceedings of the 2017 ACM on Multimedia Conference, 460–468

Selbst, A. D., Powles, J. (2017). Meaningful information and the right to explanation. International Data Privacy Law, 7(4): 233–242

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., Menczer, F. (2018). The spread of low-credibility content by social bots. Nature Communications, 9: 1146–1154

Shawar, B. A., Atwell, E. (2005). Using corpora in machine-learning chatbot systems. International Journal of Corpus Linguistics, 10(4): 489–516

Sieber, M. (2017). Informatik in der Schule: Umsetzung bereitet Kopfzerbrechen. Der Informatikunterricht im Lehrplan 21 steht vor Hürden. Lehrmittel sind knapp und nicht alle Lehrer gut ausgebildet. Zugang: www.srf.ch/news/schweiz/informatik-in-der-schule-umsetzung-bereitet-kopfzerbrechen (19.12.2019)

Siegel, A. (2018). Producers of Disinformation. In: Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., Nyhan, B. (eds.). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature, New York: Hewelett Foundation, 22–29

Siegle, J. (2019). Google Duplex: Wenn der Roboter ins Stottern gerät. NZZ. Zugang: https://www.nzz.ch/digital/wenn-der-roboter-ins-stottern-geraet-ld.1484262 (19.12.2019)

Silver, D. et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529: 484–489

Simonite, T. (2017). AI and «Enormous Data» Could Make Tech Giants Like Google Harder to Topple. Wired. Zugang: https://www.wired.com/story/ai-and-enormous-data-could-make-tech-giants-harder-to-topple/ (19.12.2019)

Singer, N. (2013). In a Mood? Call Center Agents Can Tell. The New York Times. Zugang: https://www.nytimes.com/2013/10/13/business/in-a-mood-call-center-agents-can-tell.html (19.12.2019)

Spangler, S., Angela, D., Wilkins, B. J. et al. (2014). Automated hypothesis generation based on mining scientific literature. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1877–1886

Sparrow, R. (2007). Killer robots. Journal of applied philosophy, 24(1): 62–77

Spielkamp, M. (Hrsg.) (2019). Automating Society, Bericht von AlgorithmWatch/Bertelsmann Stiftung. Zugang: www.algorithmwatch.org/automating-society (19.12.2019)

Statcounter (2019). StatCounter Global Stats – Browser, OS, Search Engine including Mobile Usage Share. StatCounter Global Stats Website. Zugang: http://gs.statcounter.com/ (07.06.2019)

Stroud, N. J. (2011). Niche news. The politics of news choice. Oxford: Oxford University Press

Sun, C., Shrivastava, A., Singh, S., Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. Proceedings of the IEEE international conference on computer vision, 843–852

Swisscom (2019). Voiceprint. Zugang https://www.swisscom.ch/de/about/rechtliches/datenschutz/voiceprint.html (05.06.2019)

Thouvenin F., Früh, A., George, D. (2018). Datenschutz und automatisierte Entscheidungen, Jusletter vom 26. November 2018

Thouvenin, F. (2014). Erkennbarkeit und Zweckbindung: Grundprinzipien des Datenschutzrechts auf dem Prüfstand. In: Weber, R., Thouvenin, F. (Hrsg.). Big Data und Datenschutz – Gegenseitige Herausforderungen, Zürich, 61–83

Thurman, N., Schifferes, S. (2012). The future of personalization at news websites. Journalism Studies, 13: 775–790

Tichy, G. (2016). Geht der Arbeitsgesellschaft die Arbeit aus? WIFO Monatsberichte, 89(12): 853–871

Torcasso, D. (2018). Wie ein Algorithmus die Werbestrategie für Volkswagen plant. Handelszeitung. Zugang: https://www.handelszeitung.ch/unternehmen/wie-ein-algorithmus-die-werbung-von-vw-plant (26.06.2019)

Treuthardt, D., Loewe-Baur, M., Kröger, M. (2017). Der Risikoorientierte Sanktionenvollzug (ROS) – aktuelle Entwicklungen. SZK, 2: 24–32

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, A., Stukal, D., Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature: Hewlett Foundation

UNESCO (2015). Qingdao Declaration. Seize digital opportunities, lead education transformation. Zugang: https://unesdoc.unesco.org/ark:/48223/pf0000233352 (19.12.2019)

UNESCO (2018). UNESCO ICT Competency Framework for Teachers. Zugang: https://unesdoc.unesco.org/ark:/48223/pf0000265721 (19.12.2019)

UNESCO (2019). Beijing Consensus on Artificial Intelligence and Education. Zugang: https://unesdoc.unesco.org/ark:/48223/pf0000368303 (19.12.2019)

UNESCO (2019b). Institute for Information Technologies in Education. Zugang: https://iite.unesco.org/publications/qingdao-declaration-seize-digital-opportunities-lead-education-transformation/ (19.12.2019)

UNESCO (2019c). I'd blush if I could: closing gender divides in digital skills through education. Zugang: https://unesdoc.unesco.org/ark:/48223/pf0000367416 (16.01.2020)

Vedder, A., Naudt, L. (2017). Accountability for the use of algorithms in a big data environment. International Review of Law, Computers and Technology, 31(2): 206–224

Vegh, A. (2019). Künstliche Intelligenz in der Strafzumessung. In: Dal Molin-Kränzli, A., Schneuwly, A. M., Stojanovic, J. (Hrsg.). Digitalisierung – Gesellschaft – Recht, APARIUZ Band 20, Zürich/St. Gallen: Dike, 359–376

Verl, A., Schraft, R. D., Kaun, R. (1998). Automatisierung der Produktion. Berlin: Springer

Villanueva, C. C. (2003). Education Management Information Systems (EMIS) and the Formulation of Education for All Plan of Action, 2002–2015, Cooperation with UNESCO Almaty Cluster Office and the Ministry of Edcuation of Tajikistan. Zugang: https://unesdoc.unesco.org/ark:/48223/pf0000156818, (16.01.2020)

Vokinger, K. N., Mühlematter, U. J., Becker, A., Boss A., Reutter, M. A., Szucs T. D. (2017). Artificial Intelligence und Machine Learning in der Medizin. Jusletter, 28.08.2017

Von Büren, R., Meer, M. A. (2014). Der Urheber. In: von Büren, R., David, L. (Hrsg.). Schweizerisches Immaterialgüter- und Wettbewerbsrecht, Bd. II/1, 3. Aufl., Basel: 146–174

Von Büren, R., Meer, M. A. (2014). Der Werkbegriff. In: von Büren, R., David, L. (Hrsg.). Schweizerisches Immaterialgüter- und Wettbewerbsrecht, Bd. II/1, 3. Aufl., Basel, 58–145

Vuorikari, R., Punie, Y., Carretero Gomez, S., Van den Brade, G. (2016). DigComp 2.0: The Digital Competence Framework for Citizens. Publications Office of the European Union. DOI: 10.2791/11517

Wachter, S., Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Columbia Business Law Review. Zugang: https://osf.io/preprints/lawarxiv/mu2kf/ (12.02.2019)

Wachter, S., Mittelstadt, B., Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 7(2): 76–99

Wakefield, J. (2018). Are you scared yet? Meet Norman, the psychopathic AI, BBC News, 02.06.2018. Zugang: https://www.bbc.com/news/technology-44040008 (19.12.2019)

Wang, W., Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. Journal of Management Information Systems, 23(4): 217–246

Wang, Y., Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology, 114(2): 246

Weber, R. H., Thouvenin, F. (Hrsg.) (2014). Big Data und Datenschutz – Gegenseitige Herausforderungen, Zürich

WEF (2016). The Future of Jobs. Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution. Geneva. Zugang: www3.weforum.org/docs/WEF_Future_of_Jobs.pdf (19.12.2019)

WEF (2018). Artificial Intelligence Collides with Patent Law. Cologny/Genf. Zugang: http://www3.weforum.org/docs/WEF_48540_WP_End_of_Innovation_Protecting_Patent_Law.pdf (11.02.2019)

WEF (2018b). Insight Report. Towards a Reskilling Revolution. A Future of Jobs for All. Zugang: http://www3.weforum.org/docs/WEF_FOW_Reskilling_Revolution.pdf (19.12.2019)

Weinmann, B. (2018). Neue Plakate analysieren per Kamera die Kunden – und liefern so personalisierte Werbung. Zugang: https://www.watson.ch/!192655562 (05.06.2019)

Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1): 36–45

Widmer, P., Wessner, P. (1999). Vorentwurf zum Bundesgesetz über die Revision und Vereinheitlichung des Haftpflichtrechts. Zugang: https://www.bj.admin.ch/dam/data/bj/wirtschaft/gesetzgebung/archiv/haftpflicht/vn-ve-d.pdf (30.01.2018)

Wolfangel, E. (2018). Unsere Stimme haben sie. Republik. Zugang: https://www.republik.ch/2018/11/09/unsere-stimme-haben-sie (05.06.2019)

Wolter, M. I., Mönnig, A., Hummel, M., Schneemann, C., Weber, E., Zika, G., Helmrich, R., Maier, T., Neuber-Pohl, C. (2015). Industrie 4.0 und die Folgen für Arbeitsmarkt und Wirtschaft: Szenario-Rechnungen im Rahmen der BIBB-IAB-Qualifikations-und Berufsfeldprojektionen: IAB-Forschungsbericht

World Bank Group (2016). World Development Report 2016: Digital Dividends: World Bank Publications. Zugang: https://books.google.at/books?id=MquKCwAAQBAJ (19.12.2019)

Yang, Q., Ji, Y. J. (2016). The Platform Economy and Natural Monopoly: Regulating or laissez-faire? Zugang: https://www.law.uchicago.edu/files/file/the_platform_economy_and_natural_monopoly_regulating_or_laissez-faire_qingyang.pdf (19.12.2019)

Yaniv, L., Yossi, M. (2018). Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. Google AI Blog Website. Zugang: http://ai.google-blog.com/2018/05/duplex-ai-system-for-natural-conversation.html (04.06.2019)

Yeomans, M., Shah, A., Mullainathan, S., Kleinberg, J. (2017). Making sense of recommendations. Journal of Behavioral Decision Making, 32(4): 403–414

Youyou, W., Kosinski, M., Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4): 1036–1040

Zerilli, J., Knott, A., MacLaurin, J., Gavaghan, C. (2018). Transparency in Algorithmic and Human Decision-Making: Is there a Double Standard? Philosophy & Technology, 32: 661–683

Zhang, B., Dafoe, A. (2019). Artificial Intelligence: American Attitudes and Trends. University of Oxford. Zugang: https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/ (15.12.2019)

Zurkinden, N. (2017). AI and driverless cars: from international law to test runs in Switzerland to criminial liability risks. In: Jacquemin, H., de Streel, A. (Hrsg.). L'intelligence artificielle et le droit. Bruxelles, 341–356

Mitglieder der Begleitgruppe

Prof. Dr. Jean Hennebert, Leitungsauschuss TA-SWISS, Département d'informatique de l'Université de Fribourg, Präsident der Begleitgruppe

Benjamin Bosshard, Eidgenössische Kommission für Kinder- und Jugendfragen

Sabine Brenner, Geschäftsstelle Digitale Schweiz, Bundesamt für Kommunikation (BAKOM)

Dr. Christian Busch, Staatssekretariat für Bildung, Forschung und Innovation (SBFI)

Dr. Christine Clavien, Institut Ethique Histoire Humanités, Université de Genève

Daniel Egloff, Staatssekretariat für Bildung, Forschung und Innovation (SBFI)

Andy Fitze, SwissCognitive – The Global AI Hub

Matthias Holenstein, Stiftung Risiko-Dialog

Dr. Marjory Hunt, Fonds national suisse de la recherche scientifique (FNS)

Manuel Kugler, Schweizerische Akademie der Technischen Wissenschaften (SATW)

Thomas Müller, TA-SWISS Leitungsausschuss, Redaktor Schweizer Radio SRF

Katharina Prelicz-Huber, TA-SWISS Leitungsausschuss (bis 2019), Präsidentin Gewerkschaft VPOD/SSP, Nationalrätin

Prof. Ursula Sury, Rechtsanwältin und Professorin, Hochschule Luzern (HSLU)

Dr. Stefan Vannoni, TA-SWISS Leitungsausschuss, cemsuisse

Projektmanagement TA-SWISS

Dr. rer. soc. Elisabeth Ehrensperger, Geschäftsführerin

Dr. Catherine Pugin, Projektleiterin

www.ta-swiss.ch

mitglied der
akademien der
wissenschaften schweiz

Computer werden leistungsfähiger und können komplizierte Probleme immer schneller lösen. Gleichzeitig stehen, dank Internet und Smartphones, grosse Mengen an Daten zur Verfügung. Beides fördert die Entwicklung von künstlicher Intelligenz (KI).
Anspruchsvolle Aufgaben, an denen bisherige Computerprogramme gescheitert sind, löst künstliche Intelligenz scheinbar mühelos. Bekannte Beispiele sind KI-Systeme, die Sprachen übersetzen oder menschliche Gegner in Spielen aller Art bezwingen. Stetig wird die künstliche Intelligenz verbessert und übernimmt Tätigkeiten, die bisher Menschen vorbehalten waren, etwa Steuerbetrug identifizieren oder Krankheiten diagnostizieren. Künstliche Intelligenz gilt daher als wichtiger Treiber des digitalen Wandels.
Die Studie von TA-SWISS beschäftigt sich eingehend mit den Chancen und Risiken dieser Technologie in den Anwendungsbereichen Arbeit, Bildung und Forschung, Konsum, Medien und Verwaltung. Zur Sprache kommen insbesondere auch allgemeine ethische und rechtliche Aspekte. Das Hauptaugenmerk liegt auf Anwendungen, bei denen KI Entscheidungsprozesse unterstützt – Prozesse, die zu Entscheidungen mit direkten Auswirkungen auf Bürgerinnen und Bürger sowie auf unsere Gesellschaft als Ganzes führen.

v/d|f
TA-SWISS 72/2020
ISBN 978-3-7281-4001-2 (Printversion)
ISBN 978-3-7281-4002-9 (E-Book)
DOI-Nr.: 10.3218/4002-9