# Gebildeter und vernetzter Mensch – Vier Thesen zur soziotechnischen Gestaltung der Zukunft[1]

Toni Wäfler

FH Nordwestschweiz, Hochschule für Angewandte Psychologie

ZUSAMMENFASSUNG

Künstliche Intelligenz und Autonome Systeme erlauben die Automatisierung von Prozessen, die zuvor nicht automatisierbar waren. Der folgende Beitrag diskutiert, welche Konsequenzen sich daraus für die soziotechnische Systemgestaltung ergeben könnten. Ausgehend von den neuen technischen Fähigkeiten und deren Grenzen wird die Frage verfolgt, inwiefern sich Mensch und Technik noch unterscheiden und welche Formen der komplementären Integration vom Mensch und Technik sich daraus ergeben. Dabei wird zwischen informatisierender, interaktiver und kollaborativer Gestaltung des Zusammenwirkens von Mensch und Technik unterschieden. Der Beitrag schließt im Fazit mit vier Thesen in Bezug auf die soziotechnische Systemgestaltung unter Berücksichtigung der neuen technischen Fähigkeiten. (a) Menschlicher Kontrollverlust über automatisierte Prozesse wird zunehmen. (b) Die Resilienz soziotechnischer Systeme muss daher erhöht werden. (c) Entsprechend sind (neue) Formen resilienzförderlicher, soziotechnischer Systemgestaltung zu entwickeln. (d) Dies bedingt auch eine Reflexion des Menschenbildes hin zum gebildeten und vernetzten Menschen.

**Schlüsselwörter**

Künstliche Intelligenz – Autonome Systeme – Mensch-Maschine Komplementarität – soziotechnische Systemgestaltung – Resilienz – Menschenbild

ABSTRACT

Artificial intelligence and autonomous systems allow the automation of processes that could not be automated before. This paper refelcts on the consequences for sociotechnical system design. Based on the new technical capabilities and their limits, the question is discussed to what extent humans and technology still differ and which forms of complementary integration of humans and technology result from this. A distinction is made between informative, interactive and collaborative design of the interaction of humans and technology. The contribution concludes with four theses regarding socio-technical system design taking into account the new technical capabilities. (a) Loss of human control over automated processes will increase. (b) The resilience of sociotechnical systems must therefore be increased. (c) Accordingly, (new) forms of resilience-promoting, socio-technical system design are to be developed. (d) This also requires a reflection of the Menschenbild towards the educated and networked human being.

**Keywords**

Artificial intelligence – autonomous Systems – complementarity of humans and machines – sociotechnical system design – resilience – Menschenbild

[1] Die Literaturrecherche zu diesem Paper wurde unterstützt durch – in alphabetischer Reihenfolge – Adrian Campos, Sandra Schenkel und Cyrill Ziegler.

## 1 Einleitung

Hacker (2018) begründet Herausforderungen, die sich angesichts der Digitalisierung für die Arbeitspsychologie ergeben. Er zieht daraus u. a. das Fazit, dass die Arbeitspsychologie nicht nur eine beschreibende, sondern viel mehr eine gestaltungsleitende Wissenschaft werden müsse, um nicht marginalisiert zu werden. Dabei müsse sie sich auch in Richtung kognitiver Anforderungen ausweiten. Vor diesem Hintergrund beinhaltet der folgende Beitrag einen arbeits- und organisationspsychologischen Positionsbezug zu neuen Technologien, worunter hier im Wesentlichen Künstliche Intelligenz und Autonome Systeme verstanden werden. Ziel ist, zu verstehen, was an diesen Technologien tatsächlich neu ist und welche neuen technischen Fähigkeiten sich daraus ergeben (Kap. 2), welche Grenzen sie haben (Kap. 3) und inwiefern Mensch und Technik nach wie vor qualitativ unterschiedlich und komplementär sind (Kap. 4). Anlass dieses Positionsbezugs ist die Annahme, dass diese neuen Technologien erlauben, nicht nur programmierbare sondern auch nicht-algorithmisierbare, anspruchsvolle kognitive Tätigkeiten zu automatisieren. Damit stellt sich die Frage von neuem, ob und ggf. welche Konsequenzen sich daraus für die soziotechnische Systemgestaltung ergeben. Dazu werden vier Thesen formuliert (Kap. 5).

Für den Beitrag gelten folgende Einschränkungen: Bekannte Auswirkungen von konventioneller Automatisierung auf Menschen sind nicht Gegenstand. Diese werden an anderen Stellen ausführlich beschrieben (z. B. Bainbridge, 1987; Parasuraman, Mouloua & Molloy, 1996; Grote, Weik & Wäfler, 1996; Parasuraman & Riley 1997; Manzey, 2012). Ebenso wird nicht auf die Diskussion um technologische Singularität eingegangen. Damit wird der Zeitpunkt bezeichnet, zu dem die Künstliche Intelligenz die Menschliche Intelligenz übertrifft. Der Zukunftsforscher Raymond Kurzweil schätzt diesen Zeitpunkt auf 2045 (Mitchell, 2019). Ob dieser jemals eintritt, und ob Technologie den Menschen dann vollständig ersetzt, ist aktuell noch eine Glaubensfrage. Im Moment jedenfalls sind Menschen noch nicht ersetzbar. Sie haben in allen soziotechnischen Systemen noch eine wichtige Rolle, sei es im Engineering, in der Maintenance oder auch im Operating. Vor diesem Hintergrund wird hier der Standpunkt eingenommen, dass Mensch und Technik auch in Zeiten der Künstlichen Intelligenz und der Autonomen Systeme komplementär sind, und dass Wege gefunden werden müssen, Mensch und Technik synergetisch zu kombinieren. Dabei steht hier weniger die Mensch-Maschine Schnittstelle im Fokus sondern die Mensch-Maschine Kollaboration (z. B. Grote, Weik & Wäfler, 1996; Hollnagel & Woods, 2005). Eine letzte Einschränkung soll hier erwähnt werden, auch wenn sie im Verlauf des Beitrags nicht mehr explizit thematisiert wird. Technikgestaltung für Konsumenten und Laien ist von Technikgestaltung für professionell-arbeitende Experten ihres Faches zu unterscheiden. Die folgenden Überlegungen beziehen sich auf das Verhältnis von Mensch und Technik in der professionellen Arbeitswelt.

## 2 Neue Technologien und deren Fähigkeiten

Im Folgenden ist beschrieben, was in diesem Beitrag unter „neuen Technologien“ verstanden wird, und welche neuen technischen Fähigkeiten diese ermöglichen.

### *2.1 Neue Technologien*

In Anlehnung an Brynjolfsson und McAfee (2014), Floridi (2015), Mitchell (2019), Russel (2019), Kuhn und Liggesmeyer (2019), Zehnder (2019) sowie EASA (2020) wird hier Folgendes unter „neuen Technologien“ verstanden:

*Künstliche Intelligenz:* Software, die fähig ist, zu lernen und damit Aufgaben zu lösen, ohne dass der Lösungsweg von Menschen programmiert werden muss.[2] Hauptsächliche Methoden sind Neuronale Netzwerke und Reinforcement Learning. Neuronale Netzwerke werden trainiert. Dies bedeutet beispielsweise hinsichtlich Bilderkennung, dass der Software viele Bilder präsentiert werden und sie selbständig lernt, darin Muster zu erkennen, sodass sie mit der Zeit innerhalb der Bilder Objekte erkennen kann. Sie muss vom Menschen in diesem Trainingsprozess insofern unterstützt werden, als dass die präsentierten Bilder ein Label haben müssen. Dieses Label wird von Menschen dem Bild gegeben. Soll die Software etwa lernen Männer von Frauen zu unterscheiden, so muss dieses Label angeben, ob das präsentierte Bild eine Frau oder einen Mann darstellt. Demgegenüber benötigt Reinforcement Learning keine derartige menschliche Unterstützung. Reinforcement Learning beruht im Prinzip auf operantem Konditionieren. Ein Roboter beispielsweise, der derart lernt, Fußball zu spielen, wird auf ein Fußballfeld gestellt und macht zufällige Schritte und Kickbewegungen. Hat er zufällig Erfolg, d. h. trifft er bei einer Kickbewegung zufällig den Ball, so wird dieses Verhalten verstärkt, in dem sich die Wahrschein-

[2] Dies ist eine etwas engere Definition von Künstlicher Intelligenz, als sie andernorts gemacht wird, wo u.a. auch symbolische Ansätze wie Expertensysteme, die nicht lernfähig sind, zur Künstlichen Intelligenz gezählt werden. Im vorliegenden Paper werden nur subsymbolische Ansätze des Machine Learnings zur Künstlichen Intelligenz gezählt.

lichkeit erhöht, dass es wiederholt wird. Derart lernt der Roboter selbständig, sich zunehmend zielorientierter zu verhalten. Neuronale Netzwerke und Reinforcement Learning können auch kombiniert werden.

*Internet der Dinge:* Direkte Kommunikation zwischen Dingen oder zwischen Dingen und ihrer Umwelt. Ein Beispiel solch direkter Kommunikation sind selbstfahrende Autos, die sich ohne menschliches Zutun gegenseitig koordinieren. Ein Beispiel aus der produzierenden Industrie sind Werkstücke, die ihre Bearbeitungszeit direkt mit den entsprechenden Produktionsmaschinen aushandeln.

*Big Data, Digitaler Zwilling:* Digitale Repräsentationen von Objekten, Prozessen, Situationen, Ereignissen, Wissensbeständen – kurz von allem. Big Data umfasst große Datenmengen, die aus unterschiedlichen Quellen stammen und unstrukturiert sein können. Technologien sind fähig, Big Data in großer Geschwindigkeit zu verarbeiten. Digitale Zwillinge repräsentieren Eigenschaften und Verhaltensweisen von Dingen aus der realen Welt.

*Was ist neu an diesen Technologien?* Im Zusammenhang mit diesen Technologien wird oft auch von der vierten Industriellen Revolution, bzw. von Industrie 4.0 gesprochen. Ob es sich tatsächlich um eine Revolution handelt oder um einen langjährigen Entwicklungsprozess, kann debattiert werden. Die Technologien an sich sind nicht plötzlich da, sondern sind über viele Jahre entstanden. Neu hingegen ist vor allem, dass sie heute ihre Potenziale sehr viel besser ausschöpfen können, weil die Leistungsfähigkeit der Hardware in den letzten fünf Jahrzehnten exponentiell gewachsen ist (vgl. Moore'sches Gesetz). Gleichzeitig sind auch die Kosten für die Hardware massiv gesunken. Floridi (2015) illustriert dies mit folgenden Zahlen: Ein iPad2 von 2010 hatte eine Leistungsfähigkeit zur Ausführung von 1600 Millionen Instruktionen pro Sekunde (MIPS). Setzt man den Preis für eine solche Leistung auf 100 Dollar fest, hätte eine vergleichbare Leistung in den 1950er Jahren 100 Billionen Dollar gekostet. 2013 verfügte das iPad4 bereits über eine Leistung von 17056 MIPS. Die Leistungsfähigkeit hat sich also in nur drei Jahren mehr als verzehnfacht. Verläuft die Entwicklung weiterhin exponentiell, so ist zu erwarten, dass die entsprechenden Technologien in ihren Funktionen immer besser werden und auch immer neue Funktionen übernehmen können.

Diese enorme Leistungssteigerung gilt hingegen nicht für Software: „… computer software has not shown the same exponential progress; it would be hard to argue that today's software is exponentially more sophisticated, or brain-like, than the software of fifty years ago, or that such trend has ever existed." (Mitchell, 2019, S. 108). Die enorme Entwicklung in der Leistungsfähigkeit der Technologien beruht also auf quantitativer Leistungssteigerung der Hardware und nicht auf qualitativer Verbesserung der Algorithmen. Allerdings kann auch ein quantitatives Wachstum zu qualitativ veränderter Leistungsfähigkeit führen. Ein Flugzeug beispielsweise braucht eine bestimmte Geschwindigkeit, um fliegen zu können. Die quantitative Zunahme von Geschwindigkeit ermöglicht also den qualitativen Sprung vom Fahren zum Fliegen. Vergleichbar ermöglicht das quantitative Wachstum der Hardware-Leistung qualitativ neuen Fähigkeiten, welche im folgenden Abschnitt skizziert sind.

### *2.2 Fähigkeiten der neuen Technologien*

Die oben beschriebene Leistungssteigerung der Hardware ermöglicht es, große Datenmengen zu speichern und selbstlernend zu verarbeiten. Damit kann – und das ist ein ganz wesentlicher qualitativer Sprung – das als Polanyi's Paradox bezeichnete Problem überwunden werden (Hirsch-Kreinsen & Karačić, 2019). Dieses Paradox besagt, dass Menschen mehr wissen als sie sagen können. Sagen können Menschen nur ihr explizierbares Wissen. Das implizite Wissen können sie nicht sagen. Da in der herkömmlichen Automatisierung Technik von Menschen mittels Handlungsregeln (bzw. Algorithmen) programmiert werden muss, bildet das nicht explizierbare Wissen eine Barriere dessen, was automatisierbar ist. Kann nun die Künstliche Intelligenz in Daten selbständig Muster erkennen, so werden auch Aufgaben automatisierbar, für welche Menschen ihr Wissen nicht explizieren können müssen. Damit können Aufgaben automatisiert werden, die bislang nicht automatisierbar waren. Im Sinne eines qualitativen Sprungs entstehen daraus die folgenden neuen technologischen Funktionen.

*Entscheidungsunterstützungssystemen:* Selbständige Mustererkennung durch Künstliche Intelligenz ist zum Beispiel hinsichtlich Entscheidungsunterstützungssystemen ein substantieller Vorteil (Mitchell, 2019). Herkömmliche Expertensysteme (z. B. der General Problem Solver) beruhten auf programmierten Regeln. Sie erwiesen sich jedoch zum einen als sehr fehleranfällig und zum anderen war ihre Generalisierbarkeit, d. h. ihre Anwendbarkeit auf andere Situationen sehr eingeschränkt. Grund dafür ist, dass die menschlichen Experten, welche die Regeln für die herkömmlichen Expertensysteme formulieren, über sehr viel implizites Wissen verfügen, welches in den explizierten Regeln nicht enthalten ist. Von Künstlicher Intelligenz verspricht man sich daher Entscheidungsunterstützungssysteme, die Regeln aufgrund von Daten selber erkennen und daher nicht mehr von explizierbarem menschlichem Wissen abhängig sind. Das sich daraus

ergebende neue Problem, dass nun die Künstliche Intelligenz ihrerseits implizites Wissen aufbaut, welches sie dem Menschen nicht mehr kommunizieren kann, wird weiter unten thematisiert (vgl. Kap. 4).

*Autonome Systeme:* Künstliche Intelligenz ist auch eine Schlüsseltechnologie für Autonome Systeme (Kuhn & Liggesmeyer, 2019): Ein System ist dann autonom, wenn es ohne menschliches Zutun und ohne eine detaillierte Programmierung für eine Situation ein vorgegebenes Ziel selbständig und an seine Situation angepasst erreichen kann. Autonome Systeme unterscheiden sich daher von heute existierenden eingebetteten Systemen durch ihre Fähigkeit, eigenständig Entscheidungen zu treffen – auch dann, wenn diese nicht detailliert programmiert wurden. (Kuhn & Liggesmeyer, 2019, S. 27). „Solche Systeme weisen grundsätzlich die Fähigkeit auf, komplexe Verarbeitungsketten von Daten, automatische Objektidentifikationen und Sensorfunktionen auf verschiedensten Ebenen bis hin zur Schaffung einer für die jeweilige Zielsetzung des Systems hinreichend genauen digitalen Repräsentation der Wirklichkeit realisieren und beherrschen zu können." (Hirsch-Kreinsen & Karačić, 2019, S. 9). Dies ermöglicht beispielsweise selbstfahrende Autos. Dabei besteht der Nutzen nicht nur in der Automatisierung. Vielmehr sollen Autonome Systeme auch ressourcenschonend sein. Folgendes Beispiel soll dies illustrieren. Wenn Autos autonom fahren, wenn sie also nicht von einem steuernden Menschen abhängig sind, macht es keinen Sinn, sie auf Parkplätzen herumstehen zu lassen. Die Nutzung von Mobilitätsdiensten wird dann für viele Menschen wirtschaftlicher sein, als ein Auto zu besitzen, was die Anzahl benötigter Fahrzeuge reduziert (Kuhn & Liggesmeyer, 2019).

*Vernetzte Autonome Systeme:* Ergänzt man die Künstliche Intelligenz Autonomer Systeme mit Kommunikationsfähigkeit (Internet der Dinge) und Datenverfügbarkeit (Big Data, Digitale Zwillinge), so entstehen vernetzt agierende Autonome Systeme, die sich nicht nur gegenseitig koordinieren können, sondern die beispielsweise auch von verteilt gemachten „Erfahrungen" profitieren. Entsprechend können sich selbstfahrende Autos nicht nur an Kreuzungen abstimmen oder den Verkehr gleichmäßig über eine Stadt verteilen, sodass weniger Staus entstehen und der Verkehr flüssiger wird. Ein einzelnes Auto kann zudem auch Routen „kennen", die es selber noch gar nie gefahren ist, weil ihm andere Autos oder auch andere Datenquellen Daten dazu bereitstellen (Kuhn & Liggesmeyer, 2019). Die Vernetzung Autonomer Systeme erlaubt diesen also, sich gegenseitig zu koordinieren und vom gegenseitigen Erfahrungsschatz zu profitieren, womit sich Effizienz und vielleicht auch Effektivität erhöhen.

*Gott Perspektive:* Zehnder (2019) spricht zudem von der „Gott-Perspektive"[5]. Damit meint er, dass ein digitalisiertes System viel besseren Überblick haben kann, als es Menschen möglich ist. Als Beispiel gibt er das Rail Control System (RCS) der Schweizerischen Bundesbahn (SBB) an. Auf dem Schienennetz der SBB verkehren täglich über 11000 Züge. Diese werden vom RCS überwacht und koordiniert: „Alle zwei Sekunden berechnet es den Verkehr auf dem gesamten Bahnnetz der Schweiz mit einem Prognosehorizont von zwei Stunden. … Auf diese Weise kann das RCS Probleme auf den Schienen rasch und exakt vorausberechnen und bei Bedarf frühzeitig Massnahmen in die Wege leiten." (Zehnder, 2019, S. 228).

Diese qualitativ neuen Fähigkeiten der Technik werden auch die Arbeitswelt verändern. Wie diese Veränderung aussehen könnte und welche Konsequenzen sich daraus für die Arbeitsgestaltung ergeben könnten, dazu werden in Kapitel 5 vier Thesen aufgestellt.

### *2.3 Exkurs 1: Mensch*

Menschen zeigen viele Fähigkeiten der neuen Technologien ebenfalls: Sie können autonom und zielgerichtet handeln, sind lernfähig, sowohl als Individuen als auch als soziale Systeme, sie nehmen Informationen aus der Umwelt auf und verfügen individuell über mentale Modelle und Situation Awareness. Als soziale Systeme können sie Prozesse steuern durch Distributed Situation Awareness wie auch formale und informale Zusammenarbeit. Sie kommunizieren miteinander, können sich koordinieren und verfügen über kulturelle Regeln des Zusammenlebens. Einzig über die „Gott Perspektive" verfügen Menschen nicht. Und natürlich unterliegen sie in all diesen kognitiven Prozessen Biases, Irrtümern, Missverständnissen, etc.

## 3 Grenzen der neuen Technologien

Die neuen Technologien haben auch ihre Grenzen. Bezüglich Künstlicher Intelligenz im Sinne subsymbolischer, selbstlernender Software gibt Mitchell (2019) dazu einen Überblick. Die Grenzen Autonomer Systeme beschreiben zusammenfassend Kuhn und Liggesmeyer (2019). Einen besonderen Fokus auf die Problematik der Ziele legt Russel (2019) und auf die Frage von Ethik und gesetzlicher Konformität Chen (2019). Im Folgenden werden die hauptsächlichen Grenzen zusammenfassend beschrieben.

[5] Zehnder (2019) nennt dies „Gott-Perspektive", weil im Buch Hiob (34:21) beispielsweise der Zürcher Bibel stehe: „Denn seine Augen wachen über den Wegen des Menschen, und er sieht alle seine Schritte".

### *3.1 Grenzen Neuronaler Netzwerke*

Nach Mitchell (2019) sind Neuronale Netzwerke nicht wirklich selbstlernend. In der Regel ist ihr Lernen von Menschen „supervised". Zwar gibt es auch Ansätze des „unsupervised Learnings", doch diese sind zumindest aktuell noch nicht wirklich erfolgreich. „Supervised Learning" bedeutet – wie oben beschrieben – dass Trainingsdaten von Menschen mit Labels versehen werden müssen. Es braucht also menschliche Intelligenz als Voraussetzung für künstlich intelligentes Lernen. Erschwerend kommt hinzu, dass die Künstliche Intelligenz auch Fehler lernt, die Menschen beim Labeling machen. Auch abgesehen vom Labeling, ist menschliche Intelligenz im Spiel, wenn Neuronale Netze selbständig lernen. Dies, weil Menschen die Software-Architektur Neuronaler Netze maßgeblich beeinflussen und damit die Feineinstellung der Parameter des Netzwerkes. Demis Hassabis, Mitgründer von Google Deep-Mind meinte dazu: „It is almost like an art form to get the best out of these systems. … There's only a few hundred people in the world that can do that really well." (zit. nach Mitchell, 2019, S. 166).

Mitchell (2019) beschreibt noch weitere Grenzen selbstlernender Neuronaler Netzwerke:

- Overfitting to training set: Das Entscheidungsmodell, das ein Neuronales Netzwerk selbstlernend aufbaut, kann den Trainingsdaten überangepasst sein. Sollen in der Bilderkennung beispielsweise Schäferhunde von Huskies unterschieden werden, kann es vorkommen, dass im Trainingsdatensatz die Huskies immer im Schnee und die Schäferhunde nie im Schnee stehen. Die Künstliche Intelligenz lernt dann möglicherweise nicht, die beiden Hunderassen zu unterscheiden, sondern Bilder mit Schnee von solchen ohne Schnee.
- Biases in trainings set: Steckt in den Trainingsdaten eine Verzerrung, so lernt die Künstliche Intelligenz diese mit. Soll sie beispielsweise Vorhersagen über beruflichen Erfolg machen, und wird sie dazu mit Daten aus der Vergangenheit trainiert, in der z. B. Frauen gegenüber Männern benachteiligt waren, so wird sie die Erfolgschance von Männern höher einschätzen als jene von Frauen.
- Impenetrability, dark secret: Die beschriebenen Effekte von Overfitting und Biases sind für Menschen nur schwer erkennbar, da das selbsterlernte Entscheidungsmodell für Menschen unverständlich ist. Für den Menschen ist damit per Definition nicht nachvollziehbar, worauf genau die Künstliche Intelligenz ihre Entscheide begründet (man spricht in diesem Zusammenhang auch von „Black Box Artificial Intelligence", vgl. unten).
- Theory of mind: Der Mensch verfügt nicht über eine theory of mind bezüglich der Künstlichen Intelligenz und die Künstliche Intelligenz auch nicht bezüglich des Menschen. Das heißt, dass keine gegenseitigen Gefühle, Bedürfnisse, Ideen, Absichten, Erwartungen und Meinungen vermutet werden, was die Interaktion behindert.
- Easy to cheat: Künstliche Intelligenz ist (noch) einfach zu täuschen. Experimente in der Bilderkennung zeigen, dass bereits die Änderungen einzelner Pixel dazu führen, dass die Künstliche Intelligenz Bilder nicht mehr erkennt. Dies kann sogar durch Veränderungen ausgelöst werden, welche für das menschliche Auge nicht wahrnehmbar sind. Gute Abwehr gegen böswillige Attacken ist noch nicht gegeben.

Die beschriebenen Grenzen selbstlernender Neuronaler Netzwerke können dazu führen, dass diese falsche Ergebnisse hervorbringen, welche für den Menschen – falls er sie überhaupt erkennen kann – oft völlig unerwartet sind.

### *3.2 Grenzen des Reinforcement Learnings*

Reinforcement Learning funktioniert im Prinzip wie Operantes Konditionieren (s. o.). Mitchell (2019) zählt auch dazu einige Grenzen auf:

- In complex tasks states are unclear: Reinforcement Learning beruht darauf, dass zufällig erfolgreiche Aktionen verstärkt und damit die Wahrscheinlichkeit ihrer Wiederholung erhöht wird. Was genau der Zustand ist und was genau eine erfolgreiche Aktion, ist in der VUCA-Welt jedoch unklar.
- Balancing exploring and exploiting: Die Balance zwischen der Wiederholung erfolgreicher Aktionen und dem Ausprobieren neuer Aktionen zu finden, ist nicht trivial. Für das weitere Lernen ist das Ausprobieren neuer Situationen wichtig.
- Ökologische Validität: Da Reinforcement Learning nicht immer (bzw. selten) mittels Robotern im realen Umfeld möglich ist, werden oft Simulationen benutzt, innerhalb derer die Künstliche Intelligenz lernt. Dies wirft die Frage der ökologischen Validität des Gelernten auf. Je weniger valide die Simulation die Realität repräsentiert, desto geringer auch die Generalisierbarkeit des Gelernten.

Aufgrund der erwähnten Grenzen des Reinforcement Learnings hält es Mitchell (2019) nicht für erstaunlich, dass Reinforcement Learning aktuell vor allem bei Spielen wie Go, wo Reinforcement Learning und

Deep learning mitels Neuronaler Netzwerke kombiniert werden, erfolgreich ist. Go mag komplexe Situationen hervorbringen, beruht aber – wie alle Spiele – auf wenigen klar definierten Regeln. Dies ist in der realen VUCA-Welt nicht gegeben. Im Übrigen – so Mitchell (2019) – hat auch Alpha Go Zero nicht wirklich alles aus den Daten und ohne Zutun menschlicher Intelligenz gelernt. Auch hier waren Menschen bei der Gestaltung der Software-Architektur, der Monte Carlo Simulationen und der Festlegung von Hyperparametern beteiligt.

### *3.3 Grenzen Autonomer Systeme*

Den Autonomen Systemen setzen folgende Aspekte grundsätzliche Grenzen. Zum einen können sie ihre Ziele nicht in übergeordnete Wertesysteme einordnen. Zum anderen ist ihr Entscheiden und Verhalten durch die Möglichkeiten der Data Analytics eingeschränkt. Im Folgenden werden diese beiden Sachverhalte diskutiert. Darüber hinaus bestehen noch einige praktische Probleme, die am Ende dieses Abschnittes beschrieben sind.

#### *3.3.1 Ziele und Wertesysteme*

Verschiedene Autorinnen und Autoren weisen auf die Ziel-Problematik Autonomer Systeme hin (Behymer & Flach, 2016; Kuhn & Liggesmeyer, 2019; Russel, 2019). Das Problem liegt vor allem darin, dass Autonome Systeme über programmierte Ziele verfügen, welche sie mit aller Konsequenz verfolgen, ohne sie vor dem Hintergrund eines übergeordneten Wertesystems zu hinterfragen oder zu priorisieren. Zwar sind Autonome Systeme fähig, zwischen eigenen Zielen und übergeordneten Systemzielen abzuwägen: „Ähnlich wie menschliche Fahrer/-innen wird ein autonomes Fahrzeug entscheiden können, ob es in einer Stausituation einem Fahrzeug einer nicht vorfahrtsberechtigten Straße Vorfahrt gewährt und so das Systemziel, Fortkommen für alle Fahrzeuge dem eigenen Ziel der möglichst schnellen Zielerreichung überordnet.“ (Kuhn & Liggesmeyer, 2019, S. 27). Es kann dabei aber nur programmierte Ziele berücksichtigen. Die Problematik, die sich daraus ergibt, brachte Stuart Russel anlässlich eines Vortrags (Zürich, 31.10.2019) folgendermaßen auf den Punkt: Schickt man seinen autonomen Roboter bei Starbuck’s um die Ecke einen Kaffee besorgen, so zögert dieser keine Sekunde, die anderen Gäste bei Starbuck’s tot zu machen, falls der Kaffee grad einmal knapp sein sollte. Das Autonome System ist nicht fähig, Ziele und Werte abzuwägen, die nicht programmiert sind. Andererseits scheint es unmöglich, ein vollständiges Wertesystem zu programmieren. Auch ein „fokussiertes“ Wertesystem zu programmieren, ist eine große Herausforderung. Dies weil ein Überblick über alle möglichen Nebenwirkungen einer konsequenten Zielverfolgung außerhalb des Vorstellungsvermögens der Programmierer liegt. Stuart Russel bringt im selben Vortrag dazu folgendes Beispiel: Die homogenen Blasen in den sozialen Systemen sind als Nebeneffekt einer Künstlichen Intelligenz entstanden, die das Ziel verfolgte, Werbung besser zu verkaufen. Da sich Werbung besser an homogene als an heterogene Gruppen verkaufen lässt, hat die Künstliche Intelligenz Möglichkeiten gefunden, die Gruppierung der Nutzer in homogene Blasen zu fördern. Die Meinungs-Polarisierung, die daraus entstand, war Nebeneffekt und nicht Ziel der Aktionen der Künstlichen Intelligenz. Ihr Ziel war einzig, Werbung besser zu verkaufen. Stuart Russel ist eine der großen Koryphäen der Künstlichen Intelligenz. Seine Lehrbücher zählen zu den Wichtigsten. Vor diesem Hintergrund schlägt er vor, das Ziel-Problem mittels einer Künstlichen Intelligenz zu lösen, welche das Wertesystem der Menschheit lernt (Russel, 2019).

Auch Chen (2019) diskutiert das Problem, dass Autonome Systeme nicht über ein programmiertes Wissen zu Gesetzen und Ethik verfügen. Solches Wissen ist seiner Ansicht nach auch nicht programmierbar. Dies, weil legale und ethische Beurteilungen immer Resultat von Aushandlungsprozessen sind, die auch hinterfragt werden können bzw. sollen. Zu solchen Prozessen des Argumentierens und Aushandelns seien Maschinen nicht fähig. Konkrete Anwendungen von Wissen zu Gesetzen und Ethik ist seiner Ansicht nach daher nicht trivial: „… law and ethics belong to the human domain.“ (Chen, 2019, S. 73). Chen (2019) sieht drei hauptsächliche Ursachen dafür:

- Emotion: Maschinen haben keine Emotionen. Menschen hingegen schon: Sie können sich schuldig (guilty) fühlen, entehrt (dishonoured) und / oder schändlich (disgraceful), wenn sie etwas Falsches tun.
- Strafe: Maschinen können nicht bestraft werden.
- Verständnis: Maschinen können zwar entscheiden, wie eine Aktion auszuführen ist. Sie haben jedoch kein Verständnis des Zwecks oder der strategischen Bedeutung einer Aktion. Dies weil sie nicht verstehen, was sie tun, und entsprechend auch nicht zu einer kritischen Reflektion fähig sind.

Zwar gibt es Ansätze, Maschinen Wissen über Gesetze und Ethik beizubringen. Entweder indem abstraktes Wissen in die Maschine programmiert wird, welches sie in konkreten Situationen anwenden soll (top-down approach) oder indem aufgrund angemessenen konkreten Verhaltens in konkreten Situationen mittels Künstlicher Intelligenz moralische Heuristiken hergeleitet werden sollen (bottom-up approach). Vorstellbar

sind auch Kombinationen dieser beiden Ansätze. Aus den oben erwähnten Gründen hält Chen die beiden Ansätze für nicht anwendbar, weswegen er ein alternatives Konzept vorschlägt (vgl. Kap. 4.3). Hinsichtlich der Grenzen Autonomer Systeme ist v. a. entscheidend, dass Autonome Systeme ohne menschlichen Beitrag nicht fähig sind, sicherzustellen, dass ihr Verhalten im Rahmen eines ethischen und / oder legalen Wertesystems bleibt.

### *3.3.2 Limiten der Data Analytics: Rechnen ist nicht Denken*

Zwar können Autonome Systeme ohne menschliches Zutun Entscheide fällen und sich zielorientiert verhalten. Entscheidungen und Verhalten finden jedoch immer innerhalb eines Kontextes statt (Kuhn & Liggesmeyer, 2019). Entscheide der Künstlichen Intelligenz können durchaus komplex sein. Sie basieren aber immer auf definierten Regeln, die „… auf Variablen und numerischen Werten beruhen." (Kuhn & Liggesmeyer, 2019, S. 28). Dabei können diese Regeln programmiert oder durch eine Künstliche Intelligenz selbstlernend eruiert worden sein. Die Einsatzfähigkeit Autonomer Systeme ist damit jedoch beschränkt auf Situationen, über die Informationen in maschinenlesbarer Form zur Verfügung stehen (Chen, 2019; Gerst, 2019), und in welchen Entscheide anhand messbarer Kriterien getroffen werden können (Kuhn & Liggesmeyer, 2019). Kuhn und Liggesmeyer (2019) gehen davon aus, dass Autonome Systeme einfache Aufgaben mit beschränktem Entscheidungsspielraum bald übernehmen werden. Komplexere Aufgaben werden dann übernommen, wenn relevante Information numerisch vorliegt. Das Verhalten Autonomer Systeme ist jedoch immer kontextgebunden. Sie können sich nur im bekannten Raum bewegen. Dies „… definiert eine Grenze der Einsatzpotenziale Autonomer Systeme. Kreative Entscheidungen, also das sinnvolle Reagieren auf vollkommen neue Situationen, werden auch in Zukunft dem Menschen vorbehalten bleiben." (Kuhn & Liggesmeyer, 2019,S. 29).

Andere Autorinnen und Autoren gehen hier noch einen Schritt weiter, indem sie grundsätzlich in Frage stellen, ob Maschinen überhaupt Entscheide treffen können (Abbass, 2019; Franken & Wattenberg, 2019; Mitchell, 2019). Für sie produzieren Data Analytics nicht Entscheide an sich, sondern nur die Grundlage von Entscheiden. Diese Grundlage erlaubt es, in einem Sense-Making-Prozess, eine Situation (besser) zu verstehen, Vorhersagen zu treffen und Konsequenzen unterschiedlicher Optionen zu antizipieren. Daher müssen Autonome Systeme immer sozial integriert sein: „Even the most autonomous and clever AI will exist within a social system in which it needs to interact with humans and other AI systems. AI must become socially integrated." (Abbass, 2019, S. 170). Gerst (2019) meint dazu, „… KI wird immer zwingend auf die Zusammenarbeit mit Menschen angewiesen sein; umgekehrt gilt das nicht." (Gerst, 2019, S.107). Hinsichtlich des Sense-Makings fehlen den Autonomen Systemen Hintergrundwissen (background knowledge) und Alltagswissen (common sense). Für beispielsweise Clinical Decision Support Systems (CDS) beschreiben Bezemer, de Groot, Blasse, ten Berg, Kappen & Bredenoord (2019), dass deren output in der Regel einer Interpretation bedarf, die den Kontext miteinbezieht. Dabei muss nicht nur kontextuales Wissen von Fachpersonen in die Entscheidung integriert werden, sondern beispielsweise auch Bedürfnisse der betreffenden Patientinnen und Patienten. Floridi (2015) weist darauf hin, dass Maschinen zwar sehr viel besser rechnen können als Menschen. Menschen hingegen können denken. Rechnen und denken sei nicht dasselbe.

### *3.3.3 Praktische Begrenzungen*

Autonome Systeme, als deren Kerntechnologie Kuhn und Liggesmeyer (2019) die Künstliche Intelligenz bezeichnen, unterliegen ebenfalls den oben erwähnten Grenzen der Neuronalen Netze und des Reinforcement Learnings. Darüber hinaus nennen Kuhn und Liggesmeyer (2019) aber noch folgende weitere Grenzen:

- Interoperabilität von Diensten: Der eigentliche Nutzen Autonomer Systeme liegt nicht im einzelnen System, sondern in deren Vernetzung (s. o.). Damit setzt die Interoperabilität Grenzen. Diese ergeben sich weniger aus den uneinheitlichen Kommunikationsprotokollen, was technisch zwar noch ungelöst ist, aber lösbar wäre: „Es wäre jedoch zu kurz gegriffen, nur die Kommunikation und die ausgetauschten Daten zu standardisieren. Die komplexeren Herausforderungen bei der Kopplung von Autonomen Systemen liegen in der Integration von Diensten." (Kuhn & Liggesmeyer, 2019, S. 38). Damit ist die Funktionalität der einzelnen Autonomen Systeme gemeint, die integriert werden müssten, um einen Nutzen aus der Vernetzung Autonomer Systeme zu bekommen.
- Testung von Sicherheit und Zuverlässigkeit: Die Ingenieurwissenschaften kennen Methoden der Testung und Validierung technischer Systeme, die in der Softwareentwicklung wie auch in der Entwicklung physischer Produkte eingesetzt werden. Bei Autonomen Systemen und noch mehr bei der Vernetzung Autonomer Systeme stoßen diese Methoden an ihre Grenzen. Durch die Fähigkeit der Autonomen Systeme, selbständig Entscheide zu treffen, sind die möglichen Verhaltensweisen und damit auch die Fehlerquellen derart zahlreich, dass sie sich nicht mehr ermitteln lassen.

Es müssen daher neue Methoden der Testung und Validierung erarbeitet werden.
- Digitale Zwillinge: Autonome Systeme sind auf adäquate Digitale Zwillinge relevanter Umweltaspekte angewiesen, um die Auswirkungen ihrer Entscheidungen situationsgerecht vorhersagen und bewerten zu können. Valide und reliable Digitale Zwillinge bereitzustellen ist eine große Herausforderung.
- Security und geistiges Eigentum: Ebenfalls (noch) ungelöst sind Fragen des Schutzes Autonomer Systeme vor Cyberkriminalität sowie auch Fragen des Schutzes geistigen Eigentums, wenn sich Autonome Systeme vernetzen.

Darüber hinaus zweifeln Hirsch-Kreinsen und Karačić (2019) die Machbarkeit und die Sinnhaftigkeit des Einsatzes Autonomer Systeme an. Aus ihrer Sicht weisen die Systeme, die aktuell im Einsatz sind, einen nur sehr eingeschränkten Autonomiegrad auf, sodass fraglich ist, ob man sie überhaupt als Autonome Systeme bezeichnen kann. Eine Weiterentwicklung der Autonomie ist sehr aufwändig und mit hohen Investitions- und Implementierungskosten verbunden. Demgegenüber ist der Nutzen aus Sicht der betrieblichen Praxis fraglich, da Rentabilitätsaussichten unklar sind. Zudem wird aus Sicht der Praxis auch in Frage gestellt, ob Autonome Systeme noch beherrschbar sind.

### *3.4 Exkurs 2: Mensch*

Es sei hier erwähnt, dass auch Menschen nicht immer perfekte Entscheide fällen oder Handlungen zeigen. So unterliegt individuelles menschliches Entscheiden einer Vielzahl kognitiver Biases und auch soziale Systeme erreichen ihre Ziele keineswegs immer erfolgreich. Mitchell (2019) weist jedoch darauf hin, dass Mensch und Künstliche Intelligenz beim Entscheiden ganz unterschiedliche Fehler machen. Künstliche Intelligenz macht Fehler, weil sie Objekte nicht versteht und kein Kontextwissen dazu hat. Sie kennt die Rolle eines Objekts im Kontext nicht, sie hat keine Erinnerung des Objekts in anderen Kontexten, sie verfügt nicht über verschiedene Perspektiven auf das Objekt, sie nimmt auch nicht Informationen zum Objekt über verschiedene Sinneskanäle (z. B. riechen) auf. Kurz: Der Künstlichen Intelligenz fehlt das Hintergrundwissen zum Objekt. Deswegen kommt es zu Fehlern, wie sie in Kapitel 3.1 beschrieben sind. Der Mensch hingegen macht viele Fehler aufgrund von kognitiven Biases, die gerade wegen des Kontextes und Hintergrundwissens entstehen.

Auch soziale Systeme unterliegen kognitiven Biases (wie Groupthink oder Risky Shift) oder verhalten sich beispielsweise wegen Kommunikationsschwierigkeiten unkoordiniert, sodass sie ihre Ziele nicht erreichen. Andererseits zeigen sie aber eine erstaunliche Fähigkeit zur Resilienz, gerade auch in unerwarteten Situationen. Mensch und Maschine sind also sehr unterschiedlich.

## 4 Komplementarität von Mensch und Technik

### *4.1 Automatisieren oder Informatisieren*

Zuboff (1988) beschrieb, wie Technik in Bezug auf Menschen eingesetzt werden kann. Dabei unterschied sie grundsätzlich zwischen einem Technikeinsatz zwecks Automatisierung (Automate) und einem solchen zwecks Informatisierung (Informate). Während bei ersterem das Ziel ist, menschliche Anstrengung (Effort) und menschliche Fähigkeiten (Skills) zu ersetzen, wird bei zweiterem Technik so eingesetzt, dass sie für den Menschen und sein Handeln Informationen generiert. Für beides können auch die oben beschriebenen neuen Technologien eingesetzt werden.

Herkömmliche Automatisierung ist nicht Gegenstand dieses Papers. Auswirkungen von Automatisierung auf den Menschen sind an anderer Stelle ausführlich beschrieben (z. B. Bainbridge, 1987; Parasuraman, Mouloua & Molloy, 1996; Grote, Weik & Wäfler, 1996; Grote, 1997; Wäfler, Windischer, Ryser, Weik & Grote, 1999; Sheridan & Parasuraman, 2005; Manzey, 2012). Wichtige negative Auswirkungen der Automatisierung auf den Menschen sind u. a.: Überforderung des Menschen bei der Überwachung automatisierter Prozesse, Übervertrauen und Untervertrauen des Menschen in die Technik, Fehleinschätzung von Prozesszuständen infolge der Automatisierung, Verlust von Fähigkeiten und Erfahrungswissen infolge der Automatisierung, Demotivation infolge der Automatisierung. All dies kann sich auch ergeben, wenn Künstliche Intelligenz und Autonome Systeme eingesetzt werden. Es wird hier jedoch darauf nicht weiter eingegangen.

Hinsichtlich Künstlicher Intelligenz fordern verschiedene Autorinnen und Autoren, diese zum Zwecke der Informatisierung einzusetzen. Damit soll die menschliche Leistungsfähigkeit erhöht werden, was auch als augmented intelligence oder augmented cognition bezeichnet wird (Crowe, LaPierre & Kebritchi, 2017; Kirste, 2019). Derart soll der Mensch befähigt werden, exaktere Entscheide zu fällen (Scherk, Pöchhacker-Tröscher & Wagner, 2017), oder durch die überlegenere Maschine hinsichtlich seiner Fehleranfälligkeit unterstützt werden (Both & Weber, 2014). Dazu soll die Maschine proaktiv mit dem Menschen kommunizieren, ihm also Informationen auch dann zur Verfügung stellen, wenn er nicht oder noch nicht danach sucht (Ittermann, Niehaus, Hirsch-Kreinsen, Dregger & ten Hompel, 2016). Dies bedeutet auch, dass die Maschine den Menschen bzw. seine aktuellen

Intentionen und Bedürfnisse verstehen können soll (Heim, 2011; Ludwig, 2015).

Eine zentrale Forderung, die sich daraus ergibt, ist, dass die Kommunikation zwischen Mensch und Künstlicher Intelligenz optimiert werden muss. Dies betrifft einerseits die Sprache, in der Mensch und Künstliche Intelligenz kommunizieren (Hager, Bryant, Horvitz, Matarić & Honavar, 2017; Crandall, Oudah, Chenlinangjia, Ishowo-Oloko, Abdallah, Bonnefon, Cebrian, Shariff, Goodrich & Rahwan, 2018). Es betrifft aber auch die Transparenz von Entscheidungen / Entscheidungsvorschlägen, die von der Künstlichen Intelligenz generiert werden (Samek, Wiegand & Müller, 2017). Diese Transparenz wird auch als Explainable Artificial Intelligence (XAI) bezeichnet. Ein hauptsächliches Problem hinsichtlich dieser Transparenz ist, dass die selbstlernende Künstliche Intelligenz – also sub-symbolische Ansätze des Machine Learnings – per Definition bezüglich ihrer Entscheidungsmodelle intransparent ist. Dies wird als „Black Box Artificial Intelligence" bezeichnet. Die Entscheidungsmodelle entstehen und verfeinern sich bei Neuronalen Netzwerken im Verlaufe des Trainings und bei Reinforcement Learning im Verlaufe des operanten Konditionierens. Dabei wird eine Vielzahl von Parametern selbstlernend justiert, was für Menschen nicht mehr transparent ist. Gewissermaßen lernt die Künstliche Intelligenz derart implizites Wissen, welches sie dem Menschen nicht kommunizieren kann. Für den Menschen sind in der Folge Entscheidungen / Entscheidungsvorschläge der Künstlichen Intelligenz intransparent und dementsprechend nicht nachvollziehbar.

Im Sinne der Informatisierung können zwei Formen der Kombination von Mensch und Maschine identifiziert werden, welche im Folgenden dargestellt sind.

### *4.2 Interaktive Systeme: Gegenseitiges Lernen*

Als Erweiterung der Forderung nach Explainable Artificial Intelligence (XAI) steht die Forderung nach interaktiven Systemen, in denen die Künstliche Intelligenz dem Menschen nicht nur transparent Informationen bereitstellt, sondern in denen Mensch und Künstliche Intelligenz interaktiv zusammenarbeiten. Dies hat zum einen den Zweck, die Entscheidungsmodelle der Künstlichen Intelligenz zu verbessern, indem der Mensch der Künstlichen Intelligenz Feedback geben kann, welches die Künstliche Intelligenz in ihren Entscheidungsmodellen berücksichtigt (Hager, Bryant, Horvitz, Matarić & Honavar, 2017). Zum anderen geht es aber auch um ein gegenseitiges Lernen, d. h. die Künstliche Intelligenz lernt vom Menschen und der Mensch lernt von der Künstlichen Intelligenz (Samek, Wiegand & Müller, 2017). Konkrete Möglichkeiten dies zu erreichen, beschreibt Kirste (2019). Dabei unterscheidet er zwischen (a) Interaktivem Lernen, in dem Menschen die Modellbildung vor oder nach der Lernphase beeinflussen können, (b) Visual analytics, in denen große Datenmengen durch Methoden der Künstlichen Intelligenz so visualisiert werden, dass der Mensch daraus Erkenntnisse gewinnen kann und der Künstlichen Intelligenz beispielsweise zeigen kann, welche Merkmale wichtiger oder unwichtiger sind, sowie (c) die Dimensionsreduktion, in der der Mensch die Parameter reduziert, die die Künstliche Intelligenz selbstlernend beispielsweise bei der Bilderkennung identifiziert hat. Diese reduzierten Parameter sollen dann auch mit menschlicher Sprache beschreibbar sein.

Eine andere Form des interaktiven, gegenseitigen Lernens beschreiben Schmid und Finzel (2020), indem sie sub-symbolische Black Box Methoden der Künstlichen Intelligenz mit symbolischen White Box Methoden, die für den Menschen verständlich sind, kombinieren. Dies soll zudem eine kooperative Entscheidungsfindung ermöglichen.

Kirste (2019) weist zudem darauf hin, dass interaktives Zusammenarbeiten von Mensch und Künstlicher Intelligenz voraussetzt, dass die Künstliche Intelligenz vom Menschen akzeptiert wird. Voraussetzung dazu ist Vertrauen in die Künstliche Intelligenz. Dieses soll allerdings nicht blind sein, sondern entsteht seinerseits durch Verständnis und Erklärbarkeit, bzw. wiederum durch Explainable Artificial Intelligence (XAI). Rodriguez, Schaffer, O'Donovan und Höllerer (2019) benutzen den Begriff „Automation complacency". Diese entsteht, wenn Menschen infolge von Übervertrauen Vorschläge von Entscheidungsunterstützungssystemen annehmen, ohne diese zu prüfen oder nach zusätzlichen Informationen zu suchen. Sie empfehlen vor diesem Hintergrund, bei der Gestaltung solcher Systeme auf Features zu verzichten, die sie als kompetenter oder überzeugender erscheinen lassen, da dies zu Übervertrauen führen könnte. Auch Banker und Khetani (2019) finden negative Effekte von Übervertrauen, allerdings im Consumer-Bereich.

Parasuraman und Manzey (2010) benutzen im Zusammenhang mit Automatisierung – unabhängig davon, ob Künstliche Intelligenz eingesetzt wird – ebenfalls den Begriff „Complacency". Diese führt dazu, dass Menschen den Output eines Entscheidungs- oder Alarmsystems nicht ausreichend prüfen, in der Annahme, dass „alles in Ordnung sei". Sie beschreiben dafür drei Ursachen: (a) „Error of omission" also Unterlassung der eigenen Suche nach Informationen, da der Output als Ersatz für eigenes aufmerksames (vigilantes) Suchen und Verarbeiten von Informationen genommen wird. (b) Übervertrauen in das automatisierte Unterstützungssystem, das als mächtiges Tool mit überlegener Analysefähigkeit wahrgenommen wird. Und (c) Verantwortungsdiffusion, die – wie auch bei sozialer Interaktion – entstehen kann, wenn der

Entscheidungsprozess mit einer automatisierten Unterstützung geteilt wird. Aufgrund eines Reviews entsprechender empirischer Studien finden Parasuraman und Manzey (2010), dass Complacency bei multiplen Aufgabenanforderungen auftritt, Novizen wie auch Experten betrifft, nicht einfach überwunden werden kann, auch nicht mittels Trainings oder Instruktion, Individuen und Teams betrifft, davon beeinflusst ist, wie die Betreffenden ihre Fähigkeiten wahrnehmen und eher bei zuverlässigeren Entscheidungsunterstützungssystemen auftritt, bzw. bei zunehmender Fehlerrate des Systems abnimmt.

### *4.3 Kollaborative Systeme: Mensch-Maschine Teaming*

Dass Mensch und Maschine unterschiedlich sind, betonen viele Autorinnen und Autoren. Maschinen können große Datenmengen schnell und reproduzierbar verarbeiten (Fraunhofer, 2017; Gerst, 2019), können regelbasiert Suchtätigkeiten unterstützen (Lunze, 2016) und so den Menschen entlasten (Spath, Ganschar, Gerlach, Hämmerle, Krause & Schlund, 2013). Demgegenüber haben Menschen Intuition, Motive, Erfahrungs- und Kontextwissen, Lebenserfahrung, Common Sense, Kreativität, Imagination, Inspiration und verstehen Sinn (Schmidt & Herrmann, 2017; Chen, 2019; Gerst, 2019, Mitchell, 2019). Auch haben sie Willen: „Menschen lernen im Unterschied zur Technik, weil sie etwas lernen wollen und weil Lerninhalte in einem sinnvollen Zusammenhang mit ihren Erfahrungen und Bedürfnissen stehen.“ (Gerst, 2019, S. 106) und können Ziele setzen sowie Prioritäten verändern. Die Autorinnen und Autoren folgern, dass Künstliche Intelligenz Menschen nicht ersetzen kann in ihrer Rolle als Entscheidende, Gestaltende, Optimierende oder Kontrollierende.

Vor dem Hintergrund dieser Unterschiedlichkeit von Mensch und Maschine wird angeregt, den Fokus weniger auf die Mensch-Maschine Schnittstelle sondern auf die Teamarbeit von Mensch und Maschine (Norman, 2017) bzw. auf kollaborative Systeme (Behymer & Flach, 2016) oder synergetisches Zusammenwirken (Jarrahi, 2018; Kirste, 2019) zu legen. Als Team sollen Mensch und Künstliche Intelligenz nicht nur interagieren, sondern kollaborieren. Dazu sind erste Ansätze zu finden, die im Folgenden beschrieben werden.

Schulte und Donath (2018) haben eine Methode entwickelt, mit der Human-Autonomy Teaming (HAT) in Systemen beschrieben werden kann, in denen Menschen mit autonomer bzw. intelligenter Technik zusammenarbeiten. Die Autoren kritisieren, dass herkömmliche Methoden in der Regel Anforderungen (requirements) und technische Funktionen beschreiben, wobei der Mensch zwar als Akteur, jedoch als außerhalb der Systemgrenze betrachtet wird. Dies sei für einfache Automatisierung ausreichend, bei der dem Menschen die Rolle der Überwachung (supervisory control) zugeordnet wird. Aufgrund folgender Entwicklungen reiche dies jedoch nicht mehr aus: (a) Die Automatisierung kann zunehmend kognitive Aufgaben von höherem Niveau übernehmen (able to perform higher cognitive tasks). (b) Die Aufgabenverteilung zwischen Mensch und Maschine wird viel weniger statisch sein, sodass eine adaptive Automatisierung notwendig wird. (c) Mensch und Maschine werden in der Aufgabenausführung auf kognitiver Ebene hochgradig voneinander abhängig sein (highly dependent on a cognitive level). Daher ist es nach Schulte und Donath (2018) notwendig, den Zweck des Mensch-Maschine Systems (bzw. des Human-Autonomy Teamings) zu beschreiben, bevor mit der Gestaltung begonnen wird. Dabei ist der Mensch als Teil des Systems zu betrachten. Die Methode, die Schulte und Donath (2018) dazu entwickelt haben, hat drei Schritte: Schritt 1: Beschreibung des Arbeitsprozesses (WProc), der Arbeitsumgebung (WEnv), des gewünschten Outputs (WPOut) und des Arbeitsobjekts (WObj), welches den eigentlichen Zweck des Arbeitssystems beschreibt. Schritt 2: Beschreibung des Arbeitssystems (WSys), das aus Menschen (worker) und Maschinen (tools) besteht. Dabei werden dem Menschen und der Maschine elementare Rollen zugeordnet, wobei der Mensch die Maschine immer dominiert: „The main characteristic of the role of the worker is to know, understand, and pursue the WObj by own initiative. Without this initiative, the WProc would not be carried out. In principle, the worker, and only the worker, might as well self-assign a WObj. The Tools, on the other hand, will receive tasks from the worker and will only perform them when told to do so. Hence, the worker and the tools are always in a hierarchical relationship“ (Schulte und Donath, 2018, S. 5). Schritt 3: Erst jetzt wird Autonomie im Sinne autonomer bzw. intelligenter Technik ins System eingeführt. Diese können verschiedene Rollen und Beziehungen zum Menschen, zur konventionellen Automation und zur nicht automatisierten Technik einnehmen. Dabei seien zwei Trends erkennbar. Zum einen sind dies Autonome Systeme, die „… mostly serve the design goals to increase the human’s effectiveness, to increase the human’s span of control, to reduce the human’s taskload, and others.“ (Schulte und Donath, 2018, S. 5). Zum anderen sind dies entscheidungsunterstützende oder andersweitig den Menschen assistierende Systeme, die „… will mostly serve the design goals to avoid or correct human erroneous action, to moderate or modulate human mental workload, to increase the human’s situation awareness, and others“ (Schulte und Donath, 2018, S. 6). Die Methode hat sowohl deskriptive als auch normative Aspekte. Deskriptiv führt sie eine Sprache ein, mit der kollaborative Mensch-Ma-

schine Systeme beschrieben werden können. Normativ gibt sie dem Menschen eine über die Technik dominierende Rolle: „Since per definition there is always a HiR between worker and tool“ (Schulte und Donath, 2018, S. 7), wobei „HiR“ für hierarchische Beziehung steht, bei der der Mensch über dem Tool steht.

Schmidt und Herrmann (2017) gehen ebenfalls davon aus, dass ein kollaborierendes Mensch-Maschine System bessere Leistung erbringen kannm, als Mensch oder Maschine je alleine fähig sind. Entsprechend wird die Künstliche Intelligenz den Menschen nicht ersetzen. Vor diesem Hintergrund haben sie das „Intervention User Interface“ entwickelt, das sie als neues Paradigma bezeichnen. Da reine Überwachung automatisierter Prozesse eine nicht menschengerechte Aufgabe ist, weil sie u.a. Monotonie, Fatigue und Deskilling erzeugt, soll der Mensch (a) Möglichkeiten der Interaktion haben und (b) sich diesen Möglichkeiten auch bewusst sein. Diese Interaktionen verändern den vordefinierten Verlauf eines Prozesses und haben folgende drei Charakteristiken: (a) Sie sind ungeplant und treten ausnahmsweise auf. (b) Sie können schnell initialisiert werden und haben eine unmittelbare Auswirkung. (c) Sie haben eine Feedback-Funktion und helfen das autonome Verhalten des technischen Systems zu verbessern. Unter autonomem Verhalten verstehen die Autoren eine vordefinierte Verhaltensweise, die bestimmt ist durch Softwareentwickler, Algorithmen, Implementierungs-Konfigurationen, settings von Service Providern, Usern, Machine Learning oder einer Kombination von alledem. Eine Intervention ändert das vordefinierte autonome Verhalten und erfolgt aufgrund neu aufkommender (emerging) Bedürfnisse des Users oder Aspekte der Situation. Solche Interventionen zu ermöglichen erleichtert nach den Autoren die Implementierung und die Konfiguration Autonomer Systeme, da „… the developer, designer, or user can pay less at-tention to all kinds of exceptions that might occur.“ (Schmidt & Herrmann, 2017, S. 42). Allerdings sind auch Probleme mit Interventionen verbunden, die in spezifischen Implementierungen jeweils gelöst werden müssen. Dazu stellen die Autoren zum einen die Frage, wie Optionen für Interventionen sichtbar und ihre Konsequenzen verständlich gemacht werden können. Zum anderen ist aus ihrer Sicht auch zu klären, wann und wie die Kontrolle nach einer Intervention wieder an das Autonome System zurückgegeben wird. In Anlehnung an Ben Shneiderman’s Golden Rules[4] der Gestaltung von Mensch-Maschine Interaktionen geben Schmidt und Herrmann (2017) folgende Hinweise für die Gestaltung des „Intervention User Interface“:

- Expectability and predictability: Ensure that users are not surprised by automated behavior and that they understand how it develops.
- Communicate options for interventions: Make options for interventions that may be context-aware visible and understandable for users in an unobtrusive way.
- Exploration of interventions: Allow the safe and enjoyable exploration of interventions and their potential impacts, e.g., by simulation or previews on future system statuses.
- Easy reversal of automated and intervention actions: Offer a simple means to reverse the impact of the system’s automated behavior or of the results of interventions.
- Minimize required attention: Minimize the user attention required to operate the system by implicitly controlled feedback.
- Communicate how control is shared: Clearly communicate the distribution of responsibilities, as well as the actual control between the human and the machine (Schmidt & Herrmann, 2017, S. 45).

Ausgehend davon, dass Autonome Systeme nicht über ausreichend Wissen zu Ethik und Gesetzeskonformität verfügen können (s. o.), diskutiert Chen (2019) ebenfalls ein Konzept für die Rollenverteilung zwischen Mensch und Autonomem System. Dabei unterscheidet er drei Funktionen, die zueinander in einer Checks-and-Balances Beziehung stehen: (a) Eine Rechts- / Ethik- / Regelungsfunktion, die Regeln vorgibt. (b) Eine Ausführungsfunktion, die Aktionen regelkonform ausführen soll. (c) Eine richterliche Funktion (judicial), die die Regelkonformität von Aktionen beurteilt. Nach Chen (2019) sind an allen drei Funktionen Menschen beteiligt, womit das Problem der Haftung gelöst sei: „As humans are within all these three components, the liability can be imposed. The issue of holding machines liable is thus dissolved.“ (Chen, 2019, S. 75). Sowohl bezüglich der Rechts- / Ethik- / Regelungsfunktion wie auch bezüglich der richterlichen Funktion liegt die Verantwortlichkeit nach Chen (2019) vollständig beim Menschen. Bei der Ausführungsfunktion liegt der Lead zwar beim Autonomen System, die Verantwortung liegt jedoch trotzdem beim Menschen. Dies, weil „… machines merely carry out human instructions, humans who give instructions are held accountable for consequences.“ (Chen, 2019, S. 76). Checks-and-Balances sind bei Chen (2019) auf verschiedenen Ebenen gegeben. Bei der Ausführungsfunktion können Mensch und Maschine jederzeit gegenseitig Entscheide bewilligen oder blockieren. Sind sie sich uneinig, entscheidet auf höherer Ebene die richterliche Funk-

[4] vgl.: https://www.cs.umd.edu/users/ben/goldenrules.html

tion. Diese richterliche Funktion kann auch Entscheide der Rechts- / Ethik- / Regelungsfunktion überprüfen. Chen (2019) differenziert zudem zwischen front-end und back-end Prozessen. Die Ausführungsfunktion gehört zu den front-end Prozessen, wo Schnelligkeit (speed) und Angemessenheit (accuracy) wichtig sind. Deswegen sollen die Maschinen hier im Lead sein. Die Rechts- / Ethik- / Regelungsfunktion hingegen gehört zu den back-end Prozessen. Hier sind teilweise andere Kriterien wichtig: „In the back-end process, fairness, accuracy, completeness, verification, reliability, and speed are all required. Putting humans in charge there satisfies these requirements." (Chen, 2019, S. 76).

Das „Intervention User Interface" von Schmidt und Herrmann (2017), das „Human-Autonomy Teaming (HAT)" von Schulte und Donath (2018) wie auch das Rahmenkonzept von Chen (2019) stellen im Prinzip alle den Menschen über die Technik, indem sie ihm die Möglichkeit geben, jederzeit die Kontrolle zumindest über die Zielvorgaben für das Autonome System zu übernehmen. Alle drei Ansätze lassen aber die Frage offen, wie der Mensch (a) befähigt und (b) motiviert werden soll, die Kontrolle zu übernehmen.

Hinsichtlich der Befähigung werden Schmid und Finzel (2020) konkreter. Am Beispiel medizinischer Entscheidungsfindung (Tumor-Klassifikation) auf der Basis von Bilddaten, schlagen sie vor, dass Mensch und Künstliche Intelligenz kooperativ entscheiden, indem sie Entscheidungen gewissermassen aushandeln. Diese Aushandlung erfolgt mittels gegenseitiger Erklärungen, was durch eine Kombination von Black Box und White Box Ansätze der Künstlichen Intelligenz ermöglicht wird: „Focus of the project is to combine deep learning black box approaches with interpretable machine learning for classification of different types of medical images to combine the predictive accuracy of deep learning and the transparency and comprehensibility of interpretable models." (Schmid & Finzel, 2020, S. 1). Im „Explanation Interface", das die Autorinnen in ihrem Projekt erproben, präsentiert die Künstliche Intelligenz ihre Erklärungen in einer für den Menschen verständlichen Weise. Derart kann der Mensch verstehen, welche Argumente hinter dem Vorschlag der Künstlichen Intelligenz stehen. Das soll ihn in die Lage versetzen, den Vorschlag zu interpretieren und zu plausibilisieren. Der Mensch kann aber auch seine Argumente der Künstlichen Intelligenz kommunizieren, was deren Entscheidungsmodell beeinflusst. Dieser gegenseitige Austausch von Argumenten ermöglicht ein kooperatives Entscheiden.

### *4.4 Exkurs 3: Mensch*

Wie die oben beschriebenen Beispiele zeigen, gibt es hinsichtlich des Einsatzes neuer Technologien die Einsicht, dass der Mensch auch in Zukunft eine für die Systemleistung wichtige Rolle hat, und dass die Systeme daher informatisierend, interaktiv oder kollaborativ zu gestalten sind. Begründet wird diese wichtige Rolle mit den spezifisch menschlichen Fähigkeiten oder Eigenschaften, über welche die neuen Technologien nicht verfügen. Teilweise werden in den beschriebenen Beispielen auch konkretere Angaben dazu gemacht, wie die Technik zu gestalten ist, um dem Menschen die ihm zugeschriebene Rolle zu ermöglichen.

Was hingegen fehlt, sind Angaben dazu, was seitens des Menschen gegeben sein muss, damit er die ihm zugeschriebene Rolle auch tatsächlich wahrnehmen kann (Wäfler & Schmid, 2020). Die Rolle, die in den oben beschriebenen Beispielen für den Menschen vorgesehen ist, ist anspruchsvoll. Es stellt sich die Frage, woher die entsprechenden Kompetenzen und auch die notwendige Motivation kommen, die für eine effektive Wahrnehmung dieser Rolle Voraussetzung sind. Es ist nicht ausreichend, die technischen Voraussetzungen zu schaffen. Vielmehr ist es eine Frage der Arbeitsgestaltung und damit der Gestaltung des soziotechnischen Systems, ob der Mensch seine Rolle wahrnehmen kann und will (Ulich, 2011). Darauf soll im nächsten Abschnitt eingegangen werden.

## 5 Vier Thesen zur soziotechnischen Systemgestaltung der Zukunft

Der soziotechnische Systemansatz versteht Arbeitsorganisationen als Systeme, die aus einem technischen und einem sozialen Teilsystem bestehen (Ulich, 2011). Entscheidend ist, dass die Systemleistung davon abhängig ist, wie gut die beiden Teilsysteme aufeinander abgestimmt sind. Studien im englischen Kohlebergbau (Trist & Bamforth, 1951) haben gezeigt, dass die Einführung neuer Technologien dazu führen kann, dass sich die Systemleistung insgesamt verschlechtert. Zwar war die neue Technologie besser als die vorherige. Im Zuge ihrer Einführung hat sich jedoch als Nebeneffekt die Arbeitsorganisation derart verschlechtert, dass der Nutzen des technischen Fortschritts mehr als zunichte gemacht wurde. Diese Verschlechterung bestand im Wesentlichen darin, dass viele neue Schnittstellen zwischen voneinander abhängigen Tätigkeiten entstanden, sodass erhebliche Reibungsverluste und Abstimmungsprobleme auftraten. Erst nachdem im sozialen Teilsystem auch die Arbeitsabläufe und -aufgaben sorgfältig konzipiert wurden, konnten sich die Potenziale der neuen Technologien entfalten. In der Folge trat dann auch tatsächlich Leistungssteigerung ein. Haupterkenntnis daraus war, dass die Einführung neuer Technologien nie ein technisches sondern immer ein soziotechnisches Projekt ist. Es wird hier die Annahme vertreten, dass dies auch bezüglich der Künstlichen Intelligenz und der Autonomen Systeme zutrifft.

Grote (2015) diskutiert Ansätze der Gestaltung des komplementären Zusammenwirkens von Mensch und Maschine in Bezug auf die neuen Technologien. Stellvertretend für entsprechende Methoden beschreibt sie die Methode KOMPASS (Grote et al. 1999, 2000; Wäfler et al. 1999, 2003). Diese bietet normative Gestaltungskriterien auf drei Ebenen. Ebene 1: Mensch-Maschine Funktionsteilung. Das Gestaltungsziel auf dieser Ebene ist sicherzustellen, dass der Mensch Kontrolle über die automatisierten Prozesse behält. Ebene 2: Individuum. Hier soll durch eine entsprechende Aufgabengestaltung beim Menschen Aufgabenorientierung gefördert werden, ein Zustand des Interesses und Engagements für die Aufgabe (Ulich, 2011). Ebene 3: Organisation. Auf dieser Ebene soll entsprechend dem soziotechnischen Systemansatz Selbstregulation gefördert werden, mit dem Ziel, die Organisation zu befähigen, Schwankungen und Störungen mittels kleiner Regelkreise lokal regulieren zu können.

Es ist Kern entsprechender Methoden sicherzustellen, dass Menschen die Kontrolle über automatisierte Prozesse und damit über die Technik behalten. Gemäß Grote (2015) besteht in den Arbeitswissenschaften Einigkeit darüber, dass diese Kontrolle für die Verantwortungsübernahme wichtig ist. Ohne Kontrolle kann der Mensch sich nicht mehr verantwortlich fühlen, unabhängig davon, ob ihm die Verantwortung zugewiesen wird oder nicht. In Bezug auf Künstliche Intelligenz und Autonome Systeme verliert der Mensch jedoch per Definition zumindest einen Teil der Kontrolle. Dies weil Entscheidungsmodelle der selbstlernenden Künstlichen Intelligenz für Menschen nicht nachvollziehbar sind und weil Autonome Systeme eben autonom entscheiden. Grote (2015) schließt, dass dort, wo keine menschliche Kontrolle mehr besteht, auch keine menschliche Verantwortung mehr verlangt werden kann. Daher ist aus ihrer Sicht wichtig, dass die Grenzen der Kontrolle möglichst klar identifiziert werden. Man muss also verstehen und beschreiben, wo menschliche Kontrolle noch gegeben ist und wo nicht mehr. Gemäß Grote (2015) könnte ein erfreulicher Nebeneffekt davon auch sein, dass gesellschaftlicher Druck entsteht, Situationen ohne Kontrolle so weit wie möglich zu vermeiden. Dies würde die Berücksichtigung arbeitswissenschaftlicher Hinweise bei der Gestaltung und Implementierung neuer Technologien möglicherweise fördern.

Vor diesem Hintergrund und im Sinne eines Fazits werden im Folgenden vier Thesen zur soziotechnischen Systemgestaltung der Zukunft formuliert.

***These 1:*** *Die neuen Technologien führen nicht nur zu teilweisem Kontrollverlust sondern sie erhöhen das Potenzial für Systemversagen zusätzlich.*
Per Definition verlieren Menschen beim Einsatz Künstlicher Intelligenz und Autonomer Systeme teilweise Kontrolle über automatisierte Prozesse. Darüber hinaus erhöhen diese neuen Technologien ebenfalls per Definition das Potenzial für Systemversagen im Sinne von Perrow (1992) zusätzlich. Er geht davon aus, dass sich die Wahrscheinlichkeiten von Unfällen bei zunehmender Systemkomplexität erhöht, weil es in komplexeren Systemen zu unerwarteten und ungünstigen Interaktionen zwischen Systemteilen kommen kann (vgl. auch Dekker, 2011). Versteht man unter Komplexität u. a. Eigendynamik und Vernetzung, dann ist beides bei den neuen Technologien gegeben. Autonome Systeme vernetzen sich eigendynamisch. Künstliche Intelligenz besteht aus eigendynamischen Verknüpfungen in der Software. Entsprechend stellt Mitchell (2019) fest, dass Künstliche Intelligenz immer wieder auf unerwartete Weise versagt. Kuhn und Liggesmeyer (2019) beschreiben, dass Autonome Systeme auf unerwartete Weise miteinander interagieren können, weshalb z. B. prospektive Tests hinsichtlich Systemsicherheit gar nicht möglich sind.

Betont sei hier, dass diese These nicht technikfeindlich gemeint ist. Die Annahme beispielsweise, dass auch selbstfahrende Autos Unfälle verursachen, spricht nicht zwingend gegen selbstfahrende Autos. Vielleicht sind sie ja trotzdem sicherer unterwegs als menschliche Autofahrende. Vielleicht verursachen sie ja weniger Unfälle als die Menschen. Trotzdem besteht die Gefahr systemischer Unfälle, die möglicherweise größer sind, als menschenverursachte Unfälle sein können. In diesem Sinne steckt darin ein erhöhtes Potenzial für Systemversagen, weil vernetzte Autonome Systeme auch vernetzte Unfälle verursachen können.

***These 2:*** *Der Kontrollverlust soll möglichst gering gehalten werden, v. a. aber soll die soziotechnische Resilienz erhöht werden.*
Im Sinne von Grote (2015) soll der Kontrollverlust möglichst gering gehalten werden. Dem verbleibenden Kontrollverlust wie auch dem erhöhten Potenzial für Systemversagen muss jedoch trotzdem etwas entgegengesetzt werden. Naheliegend ist, die Resilienz des soziotechnischen Systems zu erhöhen. Resilienz wird dabei im Sinne von Hollnagel (2011) verstanden: „Resilience is defined as the intrinsic ability of a system to adjust its functioning prior to, during, or following changes and disturbances, so that it can sustain required operations under both expected and unexpected conditions.“ (Hollnagel, 2011, S. xxxvi).

***These 3:*** *Es müssen neue Formen resilienzförderlicher, soziotechnischer Systemgestaltung gesucht werden.*
Soziotechnische Systemgestaltung hat schon immer zum Ziel, organisatorische Resilienz zu fördern, indem es lokale Regulation von Schwankungen und Störungen fördert und damit die kompetente Bewältigung von Unerwartetem. Dazu werden beispielsweise funk-

tional integrierte Organisationseinheiten gebildet, die möglichst unabhängig sein sollen (Ulich, 2011), und es wird eine angemessene Balance von zentraler Steuerung und lokaler Autonomie angestrebt (Grote, 1997). Es stellt sich die Frage, ob diese Formen der Resilienzförderung in einer digitalisierten Arbeitswelt mit Künstlicher Intelligenz und Autonomen Systemen noch ausreichend sind. Darüber, wie die Arbeitswelt der Zukunft aussehen wird, kann man natürlich nur spekulieren. Auffallend ist, dass Organisationsformen auftauchen, die agil sind (Laloux, 2014). Dabei bezieht sich Agilität auf die Fähigkeit, die eigenen Organisationsstrukturen kontinuierlich zu verändern. Dies sehen die klassischen soziotechnischen Konzepte nicht vor. Sie fordern zwar auch Selbstregulation aber innerhalb stabiler organisatorischer Strukturen. Demgegenüber werden organisatorische Grenzen in agilen Organisationen nicht nur dynamisiert. Sie werden auch zumindest teilweise aufgelöst, beispielsweise indem Individuen Rollen in mehreren Organisationseinheiten (z. B. Kreisen) einnehmen können. In klassischen Organigrammen ist dies nicht gegeben. Dort gehört ein Individuum in aller Regel auch einem Kästchen im Organigramm an (was zum vielerwähnten Gärtchendenken mit beiträgt).

Neben dieser teilweisen Auflösung innerbetrieblicher Grenzen könnte die Digitalisierung noch weitere, größere Veränderungen in der Arbeitswelt bewirken (Scheer, 2016; Wäfler, 2017; Kuhn, & Liggesmeyer 2019). Dazu könnte beispielsweise (a) eine Zunahme der Sharing Economy gehören, die auch überbetriebliche Organisationsgrenzen teilweise auflöst, sowie (b) die Entstehung von Geschäftsmodellen um digitale Ökosysteme, die einen effizienten Zugang zu Dienstleistungen Autonomer Systeme ermöglichen, sodass nicht mehr Technik, sondern Services verkauft werden, und auch (c) eine Verflachung von Hierarchien, weil hierarchische Dienstwege mit der Geschwindigkeit der Informationsverarbeitung in den vernetzten technischen Systemen nicht mehr mithalten und ihrerseits zum Flaschenhals werden. In der Folge solcher Veränderungen nimmt nicht nur das Arbeiten in inner- und überbetrieblichen Netzwerken zu, sondern beispielsweise auch die Heterogenität organisationaler Einheiten, die Verbreitung von Heterarchien, die Notwendigkeit interdisziplinärer Zusammenarbeit oder das Spannungsfeld von Kooperation und Konkurrenz, das sich jedem einzelnen Arbeitenden stellt. Natürlich sind dies Spekulationen. Da die neuen Technologien wie Künstliche Intelligenz und Autonome Systeme in der betrieblichen Realität noch kaum verbreitet sind (Arvanitis, Grote, Spescha, Wäfler & Wörter, 2018; Bienefeld, Grote, Stoller, Wäfler, Wörter & Arvanitis, S. (2018); Franken & Wattenberg, 2019) haben sich auch ihre Auswirkungen noch nicht wirklich manifestiert. Entscheidend ist jedoch, dass diese Auswirkungen gestaltbar sind, da Technik immer Optionen der Organisationsgestaltung ermöglicht (Ulich, 2011). Ziel muss es daher sein, (neue) Formen resilienzförderlicher soziotechnischer Systemgestaltung zu finden, welche geeignet sind, den durch Künstliche Intelligenz und Autonome Systeme verursachten Kontrollverlust zu kompensieren. Dazu müssen die neuen Technologien und ihre Auswirkungen aber unbedingt besser erforscht und verstanden werden.

***These 4:*** *Die Rolle des Menschen in der digitalisierten Arbeitswelt – und damit das Menschenbild der Arbeitsgestaltung – ist mit Bezug auf Auswirkungen der neuen Technologien auf den Menschen und unter Berücksichtigung der Grenzen dieser neuen Technologien zu hinterfragen.*

Lehrbücher der Arbeitswissenschaften nennen die Reihenfolge von „economic man", „social man", „self-actualizing man" bis zum „complex man" als Entwicklung der Menschenbilder, die die Arbeitsgestaltung maßgeblich beeinflusst haben und noch immer beeinflussen (z. B. Ulich, 2011). Kauffeld (2014) ergänzt den „virtual man", dessen Arbeit wie auch dessen Leben insgesamt von Informations- und Kommunikationstechnologien geprägt ist. Der „virtual man" agiert flexibel, passt sich neuen Technologien an und hat eine Neigung zu Kooperation und Aktivität in Netzwerken.

Zehnder (2019) zitiert Heinrich Pestalozzi, der gesagt habe, der Mensch sei Kopf, Hand und Herz. Inzwischen zeichne sich ab, dass die Künstliche Intelligenz dem Kopf und der Roboter der Hand überlegen sein werden. Bleibt für den Menschen also das Herz. Das Emotionale und vielleicht auch das Nicht-Rationale differenziert den Menschen. Kravcik, Ullrich und Igel (2019) halten die Fähigkeit, die richtigen Fragen zu stellen, als eine der wichtigsten Kompetenzen des Menschen in der künftigen Arbeitswelt. Aber auch andere Kompetenzen der Wissensarbeit wie beispielsweise kritisches Denken, Problemlösung, Kollaboration und Kommunikation werden noch wichtiger. Insgesamt nimmt die Wichtigkeit von Bildung zu (Floridi, 2015). Dies alles passt zur Annahme von Mason (2015), der meint, die Digitalisierung habe einen neuen Menschen geschaffen: „… the educated and connected human being." (Mason, 2015, S. 21). Laloux (2014) wiederum geht davon aus, dass für Menschen Purpose und Self-Actualization zunehmend wichtiger werden.

Wie auch immer – die Frage, welche Rolle der Mensch in der Arbeitswelt mit Künstlicher Intelligenz und Autonomen Systemen übernehmen soll, ist zu diskutieren. Dabei ist zu berücksichtigen, wie sich diese Technologien und die Digitalisierung generell auf Individuen und soziale Systeme auswirken, und welche Grenzen diese Technologien haben. Dies nicht, um dem Menschen eine Lückenbüßer-Rolle zuzuweisen, sondern im Sinne eines komplementären Mensch-

Maschine Teamings, in dem sich Mensch und Maschine mit ihren qualitativ unterschiedlichen Stärken und Schwächen gegenseitig so weit wie möglich ergänzen. „Educated“ und „connected“ gefällt da sehr gut.

**Literatur**


Abbass, H. A. (2019). Social Integration of Artificial Intelligence: Funktions, Automation Allocation logic an Human-Autonomy Trust. *Cognitive Computation*, *11*, 159-171.

Arvanitis, S., Grote, G., Spescha, A., Wäfler, T. & Wörter, M. (2018). Digitalisierung in der Schweizer Wirtschaft: Technologiestand und Auswirkungen auf Beschäftigung und Qualifikation der Beschäftigten. *KOF Analysen*, *2*, 49-59. https://doi.org/10.3929/ethz-b-000270799

Bainbridge, L. (1987). Ironies of Automation. In J. Rasmussen, K. Duncan & J. Leplat (Eds.), *New Technology and Human Error* (pp. 271-283). Chichester: Wiley.

Banker, S. & Khetani, S. (2019). Algorithm Overdependence: How the Use of Algorithmic Recommendation Systems Can Increase Risks to Consumer Well-Being. *Journal of Public Policy & Marketing*, *38*, 500-515.

Behymer, K. J. & Flach, J. M. (2016). From Autonomous Systems to Sociotechnical Systems: Designing Effective Collaborations. She Ji: *The Journal of Design, Economics, and Innovation*, *2*, 105-114.

Bezemer, T., de Groot, M., Blasse, E., ten Berg, M., Kappen, T. H. & Bredenoord, A. L. (2019). A Human (e) in Clinical Decision Support Systems. *Journal of Medical Internet Research*, *21*, 1-9.

Bienefeld, N., Grote, G., Stoller, I., Wäfler, T., Wörter, M. & Arvanitis, S. (2018). Digitalisierung in der Schweizer Wirtschaft: Ergebnisse der Umfrage 2016, Teil 2. *KOF Studien*, *99*.

Both, G. & Weber, J. (2014). Hands-Free Driving? Automatisiertes Fahren und Mensch-Maschine Interaktion. In E. Hilgendorf (Hrsg.), *Robotik im Kontext von Moral und Recht* (S. 171-188). Baden-Baden: Nomos.

Brynjolfsson, E. & McAfee, A. (2014). *The Second Machine Age*. Kulmbach: Börsenmedien AG.

Chen, J. Q. (2019). Who should be the boss? Machines or humans? Proceedings of the European conference on the impact of artificial intelliegence and robotics (pp. 71-79). Oxford UK, 31.10 - 1.11 2019.

Crandall, J. W., Oudah, M., Chenlinangjia, T., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A. & Rahwan, I. (2018). Cooperation with machines. *Nature Communications*, *9*. Zugriff am 27.09.2018 von https://www.nature.com/articles/s41467-017-02597-8

Crowe, D., LaPierre, M. & Kebritchi, M. (2017). Knowledge Based Artificial Augmentation Intelligence Technology: Next Step in Academic Instructional Tools of Distance Learning. *TechTrends*, *61*, 494-506

Dekker, S. (2011). *Drift into Failure*. Farnham: Asgate.

EASA (2020). *Artificial Intelligence Roadmap*. A Human-Centric Approach to AI in Avition. easa.europa.eu/ai

Floridi, L. (2015). *Wie die Infosphäre unser Leben verändert*. Berlin: Suhrkamp.

Franken, S. & Wattenberg, M. (2019). The Impact of AI on Employment and Organisation in the Industrial Working Environment of the Future. In P. Griffiths & M. Nowshade Kabir (Eds.), Proceedings of the 1st European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR19) (pp. 141-148), doi: 10.34190/ECIAIR.19.096

Fraunhofer (2017). *Trends für künstliche Intelligenz*. München: Fraunhofer-Gesellschaft.

Gerst, D. (2019) Autonome Systeme und Künstliche Intelligenz. Herausforderungen für die Arbeitssystemgestaltung. In H. Hirsch-Kreinsen, & A. Karačić (Hrsg.), *Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt* (S. 101-137). Bielefeld: transcript Verlag.

Grote, G. (1997). Autonomie und Kontrolle. In E. Ulich (Hrsg.), *Mensch – Technik – Organisation*, Band 16. Zürich: vdf Hochschulverlag.

Grote, G. (2015). Gestaltungsansätze für das komplementäre Zusammenwirken von Mensch und Technik in Industrie 4.0. In H. Hirsch-Kreinsen, P. Ittermann & J. Niehaus (Hrsg.), *Digitalisierung industrieller Arbeit* (S. 131-146). Nomos: Baden-Baden.

Grote, G., Ryser, C., Wäfler, T., Windischer, A. & Weik, S. (2000). KOMPASS: A method for complementary function allocation in automated work systems. *International Journal of Human-Computer Studies*, *52*, 267-287.

Grote, G., Wäfler, T., Ryser, C., Weik, S., Zölch, M. & Windischer, A. (1999). Wie sich Mensch und Technik sinnvoll ergänzen. Die ANALYSE automatisierter Produktionssysteme mit KOMPASS. In E. Ulich (Hrsg.), *Schriftenreihe Mensch – Technik – Organisation*, Band 19. Zürich: vdf Hochschulverlag.

Grote, G., Weik, S. & Wäfler, T. (1996). KOMPASS: Complementary allocation of production tasks in sociotechnical systems. In S. A. Robertson (Ed.), *Contemporary Ergonomics* 1996 (pp. 306-311). London: Taylor & Francis.

Hager, G. D., Bryant, R., Horvitz, E., Matarić, M. & Honavar, V. (2017). Advances in Artificial Intelligence Require Progress Across all of Computer Science. Washington, D.C.: Computing Community Consortium Catalyst.

Hacker, W. (2018). Menschengerechtes Arbeiten in der digitalisierten Welt. Eine wissenschaftliche Handreichung. In E. Ulich (Hrsg.), *Schriftenreihe Mensch – Technik – Organisation*, Band 49. Zürich: vdf Hochschulverlag.

Heim, P. (2011). *Interaktive Angleichung als Modell für die Mensch-Computer-Interaktion im Semantic Web*. Unveröffentlichte Dissertation, Universität Stuttgart.

Hirsch-Kreinsen, H. & Karačić, A. (2019). Technologieversprechen Autonome Systeme. In H. Hirsch-Kreinsen & A. Karačić (Hrsg.), *Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt* (S. 9-24). Bielefeld: transcript Verlag.

Hollnagel, E. (2011). Prologue: The Scope of Resilience Engineering. In E. Hollnagel, J. Pariès, D. D. Woods & J. Wreathall, J. (Eds.), *Resileinece Engineering in Practice*. Farnham: Ashgate.

Hollnagel, E. & Woods, D. D. (2005). *Joint Cognitive Systems. Foundations of Cognitive Systems Engineering*. Boca Raton: Tylor & Francis Group.

Ittermann, P., Niehaus, J., Hirsch-Kreinsen, H., Dregger, J. & ten Hompel, M. (2016). S*ocial Manufacturing and Logistics. Gestaltung von Arbeit in der digitalen Produktion und Logistik*. Dortmund: Technische Universität Dortmund.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*, 577-586.

Kauffeld, S. (2014). *Arbeits-, Organisations- und Personalpsychologie für Bachelor* (2. Aufl.). Wiesbaden: Springer.

Kirste, M. (2019). Augmented Intelligence – Wie Menschen mit KI zusammen arbeiten. In V. Wittpahl (Hrsg.), *Künstliche Intelligenz*. Technologien, Anwendung, Gesellschaft. Berlin: Springer.

Kravcik, M., Ullrich, C. & Igel, C. (2019). Künstliche Intelligenz in Bildungs- und Arbeitsräumen. In H. Hirsch-Kreinsen & A. Karačić (Hrsg.), *Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt* (S. 27-45). Bielefeld: transcript Verlag.

Kuhn, T. & Liggesmeyer, P. (2019). Autonome Systeme, Potenziale und Herausforderungen. In H. Hirsch-Kreinsen & A. Karačić (Hrsg.), *Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt* (S. 27-45). Bielefeld: transcript Verlag.

Laloux, F. (2014). *Reinventing Organisations*. Brussels: Nelson Parker.

Ludwig, B. (2015). *Planbasierte Mensch-Maschine-Interaktion in multimodalen Assistenzsystemen*. Berlin-Heidelberg: Springer Verlag.

Lunze, J. (2016*). Künstliche Intelligenz für Ingenieure*. Berlin: Walter de Gruyter.

Manzey, D. (2012). Systemgestaltung und Automatisierung. In P. Badke-Schaub, G. Hofinger & K. Lauche (Hrsg.), *Human Factors. Psychologie sicheren Handelns in Risikobranchen* (S. 333-352). Berlin: Springer.

Mason, P. (2015). *Postcapitalism. A Guide for our Future*. Penguin Books.

Mitchell, M. (2019). *Artificial Intelligence*. New York: Farrar, Strauss and Giroux.

Norman, D. (2017). Design, Business Models, and Human-Technology Teamwork. *Research-Technology Management*, *60*, 26-29.

Parasuraman, R. & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, *52*, 381-410.

Parasuraman, R., Mouloua, M. & Molloy, R. (1996). Effects of Adaptive Task Allocation on Monitoring of Automated Systems. *Human Factors*, *38*, 665-679.

Parasuraman, R. & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39* (2), 230-253.

Perrow, C. (1992). *Normale Katastrophen. Die unvermeidbaren Risiken der Grosstechnik*. Frankfurt: Campus Verlag.

Rodriguez, S. S., Schaffer, J. A., O'Donovan, J. & Höllerer, T. (2019). *Knowledge Complacency and Decision Support Systems*. IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA).

Russell, S. (2019*). Himan Comaptible. Artificial Intelligence and the Problem of Control*. Penguin Books.

Samek, W., Wiegand, T. & Müller K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal*: ICT Discoveries, Special Issue The Impact of AI on Communication Networks and Services, 1, 1-10.

Scheer, A.-W. (2016). Industrie 4.0: Von der Vision zur Implementierung. In R. Obermaier (Hrsg.), *Industrie 4.0 als unternehmerische Gestaltungsaufgabe*. Wiesbaden: Springer.

Scherk, J., Pöchhacker-Tröscher, G. & Wagner, K. (2017). *Künstliche Intelligenz – Artificial Intelligence*. Linz: Pöschhacker Innovation Consulting.

Schmid, U. & Finzel, B. (2020). *Mutual Explanations for Cooperative Decision Making in Medicine*. KI - Künstliche Intelligenz. https://doi.org/10.1007/s13218-020-00633-2

Schmidt, A. & Herrmann, T. (2017). User Interfaces: A New Interaction Paradigm for Automated Systems. *Interactions*, *25/5*, 41-46.

Schulte, A. & Donath, D. (2018) A Design and Description Method for Human-Autonomy Teamind Systems. In W. Karwowski & T. Ahram (Eds.), *Intelligent Human Systems Integration*. Proceedings oft he 1st Intenational Conference on Intelligent Humans System Integration (IHSI 2018): Integrating People and Intelligent Systems, January 7-9, Dubai, United Arab Emirates.

Sheridan, T. B. & Parasuraman, R. (2005). Human-Automation Interaction. *Reviews of Human Factors and Ergonomics*, *1*, 89-129.

Spath, D., Ganschar, O., Gerlach, S., Hämmerle, M., Krause, T. & Schlund, S. (2013). *Studie Produktionsarbeit der Zukunft-Industrie 4.0*. Stuttgart: Fraunhofer Verlag.

Trist, E. L. & Bamforth, K. (1951). Some Social and Psychological Conse- quences of the Longwall Method of Coalgetting. *Human Relations*, *4*, 3-38.

Ulich, E. (2011). *Arbeitspsychologie*. Stuttgart: Schäffer Poeschel.

Wäfler, T. (2017). Industrie 4.0: Mehr als Technik. In M. K. Peter (Hrsg.), *KMU-Transformation*. Olten: FHNW.

Wäfler, T. & Schmid, U. (2020). Explainability is not Enough - Requirements for Human-AI-Partnership in Complex Sociotechnical Settings. Proceedings of European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR20), University of Lisbon, Portugal, 22 - 23 Oct. 2020.

Wäfler, T., Grote, G., Windischer, A. & Ryser, C. (2003). KOMPASS: A Method for Complementary System Design. In E. Hollnagel (Ed.), *Handbook of Cognitive Task Design* (pp. 477-502). Mahwah, NJ: Lawrence Erlbaum.

Wäfler, T., Windischer, A., Ryser, C., Weik, S. & Grote, G. (1999). Wie sich Mensch und Technik sinnvoll ergänzen. Die GESTALTUNG automatisierter Produktionssysteme mit KOMPASS. In E. Ulich (Hrsg.), *Schriftenreihe Mensch – Technik – Organisation*, Band 18. Zürich: vdf Hochschulverlag.

Zehnder, M. (2019) *Die digitale Kränkung*. Basel: NZZ Libro, Schwabe Verlagsgruppe.

Zuboff, S. (1988). *In the age of the smart machine*. New York: Basic Books.

Korrespondenz-Adresse:
Prof. Dr. Toni Wäfler
Fachhochschule Nordwestschweiz FHNW
Hochschule für Angewandte Psychologie
Riggenbachstrasse 16
D-4600 Olten
toni.waefler@fhnw.ch