Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

White Paper:
Mensch und KI – gemeinsam besser

Hinweise für eine erfolgreiche Nutzung der künstlichen Intelligenz in wissensintensiven Bereichen

Prof. Dr. Toni Wäfler, BSc Andrina Eisenegger, MSc Samira Hamouche, BSc Nicholas Magee

Inhaltsverzeichnis

1. Geltungsbereich des White Papers

Dieses White Paper fokussiert auf das Zusammenwirken von Mensch und künstlicher Intelligenz (KI) bei wissensintensiven Arbeitsaufgaben. Darunter werden Aufgaben verstanden, die auf menschlicher Seite eine hohe Fachexpertise sowie ein ausgeprägtes Erfahrungswissen voraussetzen. Das Paper beschreibt, wie eine Zusammenarbeit von Mensch und KI gestaltet werden sollte, sodass Menschen ihr Wissen kontinuierlich weiterentwickeln können. Damit fokussiert das Paper explizit nicht auf den Einsatz von KI für simple Alltagsaufgaben.

Das Paper beginnt mit einer Diskussion von zwei grundsätzlich unterschiedlichen Ansätzen der Kombination von Mensch und Technik (Kap. 2). Darauf folgen eine Definition des Begriffs «Intelligenz» sowie eine Beschreibung der Unterschiede zwischen menschlicher und künstlicher Intelligenz (Kap. 3). Aus dieser Unterschiedlichkeit werden zum einen typische Probleme der Zusammenarbeit von Mensch und KI hergeleitet (Kap. 4) und zum anderen wird dargestellt, was Menschen von KI hinsichtlich Entscheidungsunterstützung, kontinuierlichem Lernen, Vertrauen in KI sowie Motivation benötigen (Kap. 5). Abschliessend werden Tipps für eine menschenzentrierte Gestaltung des Zusammenwirkens von Mensch und KI gegeben. Der Anhang weist auf eine vertiefende Literatur hin (Anhang 1) und beschreibt bestehende Instrumente (Anhang 2).

Dieses White Paper wendet sich an ein breites Publikum und verzichtet daher so weit wie möglich auf eine ausgeprägte technische oder psychologische Fachsprache.

Das White Paper basiert auf Erkenntnissen aus dem angewandten Forschungsprojekt «Mensch-KI Teaming im industriellen Teile-Management» der Hochschule für Angewandte Psychologie FHNW (www.fhnw.ch/mensch-ki-teaming).

2. Gestaltungsansätze: KI mit Mensch vs. KI gegen Mensch

Grundsätzlich können mit der Einführung von KI zwei fundamental unterschiedliche Strategien verfolgt werden: Automatisieren oder Informatisieren.

Hinsichtlich Entscheidungsfindung bedeutet dies:
- Strategie der Automatisierung durch KI: Die KI entscheidet. Ziel ist es, menschliche Fähigkeiten durch KI zu ersetzen.
- Strategie der Zusammenarbeit von Mensch und KI: Ziel ist es, menschliche Fähigkeiten durch Fähigkeiten der KI so zu ergänzen, dass die Kombination von Mensch und KI bessere Entscheide trifft, als es Menschen oder die KI je alleine können.

«Leftover»: die Rolle, für die Menschen nicht geeignet sind

Falls die KI immer die richtige Entscheidung trifft, spricht aus Sicht der Entscheidungsqualität nichts gegen Automatisierung. Falls es jedoch noch immer Menschen braucht, welche die KI überwachen, die KI-generierten Entscheidungen prüfen oder die Verantwortung dafür übernehmen müssen, ergeben sich fundamentale Probleme:
- Menschen müssen eine KI überwachen, die schneller entscheidet und dabei mehr Faktoren berücksichtigt, als sie es können. Dies übersteigt ihre Fähigkeiten grundsätzlich, sodass sie z. B. Prozesszustände falsch einschätzen.
- Menschen können als Folge von Automatisierung auch Fähigkeiten verlieren. Dies weil Menschen ihre Fähigkeiten durch Training aufbauen und aufrechterhalten. Müssen sie Entscheide nicht mehr selbst treffen, dann verlernen sie die entsprechenden Fähigkeiten oder bauen sie gar nicht erst auf.
- Selbst wenn Menschen noch über ausreichend Fähigkeiten verfügen, eine KI zu überwachen, so ist diese Aufgabe sehr monoton, weshalb Menschen dabei schnell ermüden. Es ist keine menschliche Stärke, einen Prozess hoch aufmerksam zu überwachen und dabei die Aufmerksamkeit fokussiert zu behalten.
- Zudem entstehen auch Über- und Untervertrauen in die KI. Beides führt zu Fehleinschätzungen.

Solche unerwünschten Auswirkungen sind Folge davon, dass den Menschen eine Rolle übertragen wird, für die sie nicht geeignet sind. Sie müssen dabei jene Funktionen übernehmen, deren Automatisierung nicht gelungen ist. Dies wird auch als «Leftover»-Prinzip bezeichnet: Menschen müssen übernehmen, was übrig bleibt.

«Komplementarität»: die clevere Kombination von Mensch und KI

Wenn es nicht gelingt, Entscheidungsprozesse vollständig zu automatisieren, und dass Menschen darin also noch eine wichtige Rolle zu erfüllen haben, dann muss diese Rolle menschengerecht sein. Dabei bedeutet menschengerecht, dass diese Rolle den menschlichen Eigenschaften und Fähigkeiten entsprechen muss.

Diesem Anspruch wird die Strategie der Zusammenarbeit gerecht[1]. Sie zielt auf eine optimale Gestaltung der Zusammenarbeit von Mensch und KI ab, sodass sie gemeinsam bessere Entscheide fällen, als dies Mensch oder KI je alleine können.

Die Zusammenarbeit von Mensch und KI kann unterschiedlich ausgeprägt sein:

- Beratend: Die KI schlägt Entscheide vor. Menschen prüfen diese Entscheidungsvorschläge inhaltlich und auf Einhaltung von Vorschriften. Voraussetzung dafür ist, dass die von der KI vorgeschlagenen Entscheidungen für Menschen nachvollziehbar sind. Man spricht in diesem Zusammenhang auch von «Explainable AI (XAI)», also von einer KI, die sich erklären kann.
- Unterstützend: Die KI unterstützt den menschlichen Entscheidungsprozess. Sie schlägt also nicht einfach etwas vor, sondern unterstützt Menschen dabei, selbst zu entscheiden, indem sie beispielsweise unterstützt, Informationen zu sammeln, Annahmen zu prüfen oder Argumente zu bewerten. Man spricht in diesem Zusammenhang auch von «Joint Decision-Making», also von gemeinsamem Entscheiden.
- Gemeinsam lernend: Mensch und KI werden gemeinsam immer besser, indem sie ihre Fähigkeit, gemeinsam zu entscheiden, kontinuierlich weiterentwickeln. Einerseits unterstützt die KI dabei explizit menschliches Lernen. Andererseits unterstützt auch der Mensch explizit maschinelles Lernen. Man spricht in diesem Zusammenhang auch von «Co-Learning», also von gemeinsamem Lernen.

Gemeinsam ist diesen drei Ansätzen der Zusammenarbeit, dass Mensch und KI aufeinander abgestimmt sind. Dabei wird explizit berücksichtigt, dass Mensch und KI ganz unterschiedlich zu Entscheidungen gelangen. Während selbst erfahrene Menschen beispielsweise Mühe damit bekunden, viele Details systematisch in ihre Entscheidungsfindung miteinzubeziehen und dabei nicht hin und wieder das eine oder andere zu übersehen, fällt es ihnen vergleichsweise einfach, in neuartigen Situationen zu improvisieren. Vorausgesetzt natürlich, sie verfügen über die entsprechende Expertise. Bei KI verhält es sich genau umgekehrt. KI entscheidet aufgrund sehr grosser Datenmengen, aus der sie systematisch gelernt hat. Dabei ist sie sehr auf bestimmte Zielstellungen spezialisiert, etwa auf die Erkennung von Gesichtern. Innerhalb dieses Bereichs kann sie die Fähigkeiten von Menschen bei Weitem übertreffen. Ausserhalb dieses Bereichs ist sie aber unfähig, auch nur die einfachsten Aufgaben zu bewältigen, etwa eine Tomate zu erkennen. So kann sie in diesem Beispiel ihre Fähigkeit der Bilderkennung nicht auf neuartige Situationen anwenden.

Mensch und Technik verfügen also über ganz unterschiedliche Stärken und Schwächen. Sie können sich jedoch in ihren Stärken gegenseitig fördern und in ihren Schwächen gegenseitig unterstützen. Damit ergänzen sie sich gegenseitig, sind also komplementär.

Komplementarität bedeutet, sich aufgrund der Unterschiedlichkeit gegenseitig zu ergänzen. Welches genau die Unterschiede zwischen Mensch und KI sind, wird im folgenden Abschnitt beschrieben.

1 Wäfler, T., Windischer, A., Ryser, C., Weik, S. & Grote, G. (1999). Wie sich Mensch und Technik sinnvoll ergänzen. Die GESTALTUNG automatisierter Produktionssysteme mit KOMPASS. Zürich: vdf Hochschulverlag.

3. Intelligenz: KI rechnet, Menschen denken

Der Begriff «KI» weist darauf hin, dass Intelligenz künstlich erreicht werden kann. Vor diesem Hintergrund sollen hier vorerst die Begriffe «Intelligenz» und «KI» definiert werden. Dies im Wissen, dass es zu beiden Begriffen keine einheitliche, allgemeingültige Definition gibt. Umso wichtiger ist es, explizit zu beschreiben, was in diesem White Paper darunter verstanden wird.

In der Psychologie ist Intelligenz definiert als die «Fähigkeit, aus Erfahrung zu lernen, Probleme zu lösen und Wissen einzusetzen, um sich an neue Situationen anzupassen[2]». Intelligenz befähigt Menschen also dazu, sich in ihrer Umwelt kompetent zu verhalten und ihre Ziele auch dann zu erreichen, wenn sie mit neuartigen Situationen konfrontiert sind. Damit ist Intelligenz auch immer umweltbezogen. Ein*e Ureinwohner*in beispielsweise verfügt über die Intelligenz, im Urwald des Amazonas zu überleben. Dazu wäre ein*e Europäer*in kaum fähig, trotz Bestnoten in der Schule. Je nach Umwelt werden unterschiedliche Erfahrungen und Problemlösefähigkeiten benötigt.

KI ist definiert als die «Nachbildung menschlicher Intelligenz innerhalb der Informatik[3]». KI ist also die künstlich hergestellte Fähigkeit, aus Erfahrung zu lernen und damit auch in neuartigen Situationen Ziele zu erreichen.

Teilweise werden auch technische Regler bereits als intelligent bezeichnet, etwa Thermostatventile, die die Temperatur messen und davon abhängig die Heizintensität regeln, um die Raumtemperatur stabil zu halten. Nach dieser Auffassung besteht Intelligenz in der Fähigkeit, sich anzupassen bzw. dynamische Temperaturschwankungen ausgleichen zu können. Für das vorliegende White Paper reicht dies noch nicht aus, um als intelligent bezeichnet zu werden. Zumindest müsste auch noch eine Lernfähigkeit gegeben sein.

Die folgenden Abschnitte beschreiben, auf welche unterschiedlichen Weisen Mensch und KI intelligentes Verhalten hervorbringen und welchen Bezug dies zu Entscheidungsprozessen hat.

2 Myers, D. G. (2005). Psychologie. Heidelberg: Springer. (S. 460)

3 Wikipedia, 2024

3.1. Wie rechnet KI?

Gemäss obiger Definition umfasst Intelligenz die Fähigkeit zu lernen und sich neuartigen Situationen anzupassen. Im Folgenden ist beschrieben, wie KI diese Fähigkeit hervorbringt. Dabei geht es nicht um die technischen Details der entsprechenden Algorithmen (z. B. Deep Learning oder Reinforcement Learning). Vielmehr geht es darum, die Stärken und Schwächen von KI herauszuarbeiten, um später daraus ableiten zu können, inwiefern KI und menschliche Intelligenz komplementär sind und wie die beiden bezüglich Entscheidungsprozessen clever kombiniert werden können.

3.1.1. Fähigkeiten der KI

Hinsichtlich der zentralen Merkmale von Intelligenz, zu lernen und das Gelernte auf neuartige Situationen anzupassen, zeigt KI im Wesentlichen drei Fähigkeiten:

- Mustererkennung: KI kann lernen, in Daten Muster zu erkennen. Dies befähigt sie beispielsweise zu Bilderkennung, wo sie in der Verteilung von Pixeln Muster findet und so Gesichter erkennen oder Hunde von Katzen unterscheiden kann. Diese Lernfähigkeit kann sie auf verschiedene Arten von Daten anwenden. Beispielsweise kann sie in Daten von Mitarbeitenden erkennen, welche Mitarbeitenden in ihrem Job am meisten Erfolg haben.
- Kategorisierung: Hat die KI aus Daten, mit denen sie trainiert wurde, gelernt, Muster zu erkennen, kann sie diese Fähigkeit auf neue Daten anwenden. Hat sie beispielsweise gelernt, welche Mitarbeitenden im Job am meisten Erfolg haben, kann sie die identifizierten Muster auf Bewerbungsdossiers anwenden, um diese zu kategorisieren. Damit ist sie in der Lage, erfolgsversprechende Kandidierende von anderen zu unterscheiden. Ist sie entsprechend trainiert, kann KI auch für andere Problemstellungen erfolgversprechende Entscheidungen erkennen.
- Generative KI: KI kann Muster, die sie in Daten erkannt hat, auch reproduzieren. Dies wenden Sprachmodelle wie ChatGPT an. Hier hat KI beispielsweise das Muster gelernt, dass auf das Wort «Hund» wahrscheinlicher die Worte «Leine» oder «bellen» folgt als das Wort «Eisberg». Aufgrund solcher Wahrscheinlichkeiten ist sie fähig, neue Texte zusammenzusetzen, die erstaunlich plausibel klingen. Entsprechend kann sie auch Muster in Musik oder Bildern von Künstler*innen reproduzieren, sodass das Resultat zwar neu ist, den Originalen jedoch entspricht.

KI revolutioniert die Automatisierbarkeit

Die oben beschriebenen intelligenten Fähigkeiten der KI sind disruptiv. Sie durchbrechen die Grenzen dessen, was automatisierbar ist. Ohne die Fähigkeit der Maschine, selbst zu lernen und das Gelernte auf neuartige Situationen anzuwenden, ist nur automatisierbar, was Menschen auf klassische Weise programmieren können. Solange dies gilt, stösst die Automatisierbarkeit an die Grenzen des Polanyi-Paradoxons. Dieses besagt, dass Menschen mehr wissen, als sie sagen können. Insbesondere Menschen mit hoher Fachexpertise treffen oft Entscheide aufgrund ihres Erfahrungswissens, ohne diese im Detail begründen zu können. Solche Entscheide sind meist von hoher Qualität, die Expert*innen können jedoch nicht im Detail beschreiben, wie sie darauf kommen (vgl. unten im Text, wie dieses Phänomen psychologisch erklärt wird).

Solches Erfahrungswissen wird auch als implizites Wissen bezeichnet. Dies weil es nicht gesagt, also nicht explizit gemacht werden kann. Damit kann es auch nicht als klassischer Algorithmus beschrieben und in eine Maschine programmiert werden. 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 KI in Daten selbstständig Muster erkennen, so werden auch Aufgaben automatisierbar, für welche Menschen ihr Wissen nicht explizieren können. Dazu gehört beispielsweise das Übersetzen von Texten in andere Sprachen, wo es sich als nur sehr begrenzt möglich erwiesen hat, manuell zu programmieren, wie Texte zu übersetzen sind. Dies unter anderem weil zwischen Wörtern unterschiedlicher Sprachen keine Eins-zu-eins-Beziehung besteht. Demgegenüber ist KI fähig, in vielen Texten unterschiedlicher Sprachen selbstlernend Muster zu erkennen, was zu einer sehr viel besseren Übersetzungsqualität geführt hat. Die Fähigkeit, zu lernen und das Gelernte auf neue Situationen anzuwenden, hat die Automatisierbarkeit also revolutioniert.

3.1.2. Schwächen der KI

KI bildet also menschliche Intelligenz nach und eröffnet damit ganz neue Möglichkeiten der Automatisierbarkeit, darunter auch die Automatisierung von Entscheidungsprozessen. Dennoch zeigt KI im Vergleich zum Menschen auch einige Schwächen, die im Folgenden beschrieben sind.

- KI erkennt Muster, nicht Fakten: KI kann zwar Muster erkennen, die Bedeutung der Muster erkennt sie jedoch nicht. In der Bilderkennung beispielsweise lernt KI, Muster in Pixeln zu erkennen und damit Bilder zu klassifizieren, also etwa Schäferhunde von Huskys zu unterscheiden. Allerdings hat sich gezeigt, dass KI dabei auch völlig falsche Unterscheidungsmerkmale identifiziert. So hat sie auch schon Schäferhunde, die im Schnee standen, für Huskys gehalten. Erst da ist aufgefallen, dass KI gar nicht wirklich gelernt hat, Schäferhunde von Huskys zu unterscheiden. Vielmehr unterscheidet sie Bilder mit viel Weiss von Bildern ohne Weiss. Schäferhund oder Husky ist für eine KI nicht ein Tier, das bellen kann, Fleisch frisst und stinkt, wenn es nass ist. Für KI ist «Hund» ein Pixelmuster, zu dem sie im Training gelernt hat, dass dieses als Hund bezeichnet wird. Ähnlich verhält es sich beim Erkennen anderer Verzerrungen in den erlernten Mustern. So kann KI beispielsweise Bewerbungsdossiers danach klassifizieren, ob die betreffende Person im ausgeschriebenen Job Erfolg haben wird. Derart kann sie Selektionsentscheidungen unterstützen. Ihre Entscheidungsempfehlung basiert die KI auf Mustern, die sie aus Daten gelernt hat. Diese Muster zeigen ihr, über welche Merkmale erfolgreiche Bewerbende verfügen. Wenn nun in den Trainingsdaten Verzerrungen stecken, wenn also beispielsweise in der Vergangenheit Männer gegenüber Frauen bevorzugt wurden, dann lernt die KI aus diesen Daten, dass Männer im besagten Job erfolgreicher sind als Frauen. Sie wird also einen Mann empfehlen und tradiert damit die Diskriminierung.
- KI kann Fakten nicht interpretieren: Unterstützt eine KI beispielsweise den ärztlichen Entscheid, ob ein*e Patient*in im Spital bleiben soll oder nach Hause entlassen werden kann, so beruht dies ebenfalls auf Mustererkennung. Auch hier lernt die KI, aus Daten der Vergangenheit zu erkennen, welche Krankheitssymptome problematisch und welche unproblematisch sind. Entsprechend wird sie empfehlen, Patient*innen mit Symptomen, die zu keinen Komplikationen geführt haben, nach Hause gehen zu lassen. In den Daten sind aber möglicherweise keine Gründe dafür enthalten, weshalb bestimmte Symptome nicht zu Komplikationen geführt haben. So kann es beispielsweise sein, dass bestimmte Symptome nur deshalb zu keinen Komplikationen geführt haben, weil die betroffenen Personen im Krankenhaus behalten und besonders umsorgt wurden. Dennoch wird die KI die falsche, potenziell gefährliche Empfehlung geben, Personen mit diesen Symptomen nach Hause zu entlassen. In den Mustern erkennt die KI nur Zusammenhänge zwischen Symptomen und Komplikationen. Sie kann diese Zusammenhänge jedoch nicht interpretieren.
- KI erfindet vermeintliche Fakten: Generative KI kann in Daten erkannte Muster auch reproduzieren. So können beispielsweise Stimmen oder Gesichter nachgemacht werden. Sprachmodelle wie ChatGPT können Text generieren und damit Fragen vermeintlich beantworten. Bei der Beantwortung von Fragen generiert das Sprachmodell einen Text aufgrund sprachlicher Muster, welche es aus vielen Texten gelernt hat. Die derart generierte Antwort kann zutreffend sein, aber auch teilweise oder vollständig erfunden.

Rechnen ist nicht dasselbe wie verstehen

Die beschriebenen Schwächen zeigen deutlich, was KI nicht kann: KI versteht die gelernten Muster nicht. Sie kann zwar hervorragend rechnen und damit Muster in Daten identifizieren. Die Bedeutung dieser Muster kann KI hingegen nicht erkennen. Dies weil sie keine Fähigkeit besitzt, Inhalt von Daten bzw. dessen Sinnhaftigkeit wahrzunehmen. Daher bleibt der Begriff «Hund» für KI eine Folge von vier Buchstaben bzw. von vier Bytes. Eine Vorstellung der Kreatur Hund hat die KI nicht. Eine KI kann überhaupt nicht beurteilen, ob Entscheidungsvorschläge, die sie macht, sinnvoll sind. Dies obliegt dem Menschen. Allerdings setzt dies nicht nur hohe Fachkompetenz voraus, sondern auch Nachvollziehbarkeit des KI-generierten Vorschlages. Beides ist anspruchsvoll, insbesondere wenn KI den Eindruck macht zu verstehen, was sie vorschlägt. Doch dazu weiter unten mehr.

3.2. Wie denkt Mensch?

Intelligenz wurde oben definiert als die Fähigkeit zu lernen und sich neuartigen Situationen anzupassen. Im Folgenden ist beschrieben, wie Menschen diese Fähigkeit hervorbringen. Dabei geht es auch hier nicht um die psychologischen Details der entsprechenden kognitiven Prozesse. Vielmehr sollen hier menschliche Stärken und Schwächen herausgearbeitet werden, um danach zu beschreiben, wie KI und menschliche Intelligenz komplementär miteinander kombiniert werden können, um Entscheidungsprozesse zu verbessern.

3.2.1 Fähigkeiten des Menschen

Über menschliche Denk- und Entscheidungsfähigkeit könnte (oder müsste) man ganze Bücher schreiben. Hier wird insbesondere auf zwei Aspekte eingegangen, welche menschliches Denken besonders vom Rechnen der KI unterscheidet: Expertise und Wertehaltung.

Menschen entscheiden mit Erfahrung und Wissen
Der Nobelpreisträger Daniel Kahnemann[4] unterscheidet zwischen schnellem Denken und langsamem Denken. Schnelles Denken erfolgt intuitiv. Dabei ist Intuition nicht etwas Magisches, sondern basiert auf konkreter Erfahrung mit bestimmten Situationen. Diese Erfahrung befähigt Menschen gewissermassen, Situationen zu lesen und daraus Schlussfolgerungen abzuleiten. Erfahrene Ärzt*innen beispielsweise können Symptombilder lesen und erkennen darin sofort die entsprechenden Krankheitsursachen. Ebenso können erfahrene Verkäufer*innen das Verhalten ihrer Kund*innen lesen und die passende Verkaufsstrategie wählen. Erfahrung lässt Menschen also Situationen interpretieren und die passende Bewältigungsstrategie auswählen. Dies kann sehr schnell erfolgen und ist nicht zwingend ein bewusster Prozess. Expert*innen gehen oft so vor. Es befähigt sie insbesondere in kritischen Situationen, in denen ein schneller Entscheid erforderlich ist, kompetent zu agieren. Allerdings können sie dann nicht immer erklären, weshalb sie so vorgegangen sind (und können ihre Vorgehensweise daher auch nicht in eine Maschine programmieren, s. o.).

Demgegenüber ist langsames Denken eher ein wissensbasiertes Vorgehen, wobei Situationen bewusst analysiert und daraus Schlussfolgerungen abgeleitet werden. Dieses Vorgehen ist viel systematischer als schnelles Denken, braucht jedoch Zeit und Wissen.

Menschen unterscheiden sich beim Entscheiden also von KI insofern, als sie Situationen aufgrund von Erfahrung und Wissen interpretieren. Welche kognitiven Funktionen und Prozesse dazu notwendig sind, ist weiter unten beschrieben.

Menschen unterscheiden Wichtiges von Unwichtigem und übernehmen Verantwortung
Für die meisten Entscheide gibt es nicht die eine richtige Lösung, die man berechnen kann. Vielmehr gibt es immer mehrere Lösungen, die alle Vor- und Nachteile haben.
Daher muss zwischen verschiedenen möglichen Lösungen priorisiert werden. Dies bedingt, dass die jeweiligen Vor- und Nachteile gewichtet werden. Dabei können solche Gewichtungen auch dynamisch sein. Beispielsweise kann es bei der Auftragsplanung für bestimmte Aufträge wichtig sein, diese schnell zu bearbeiten, koste es, was es wolle. Demgegenüber sind bei anderen Aufträgen die Kosten zu minimieren. Möglicherweise ist dies nicht produktspezifisch, sondern abhängig vom jeweiligen Kunden. Priorisierungsmerkmale können also vielfältig und variabel sein, in Abhängigkeit davon, was in einer konkreten Situation als wichtig betrachtet wird. Menschen nehmen in der Entscheidungsfindung aufgrund von Werthaltungen Abwägungen vor. KI verfügt über kein entsprechendes Wertesystem.

Mit Werthaltungen und Abwägen verbunden ist auch die Fähigkeit, Verantwortung zu übernehmen, also für die Konsequenzen eines Entscheides einzustehen. Damit ist nicht gemeint, dass eine KI für ihr Entscheiden weder belohnt noch bestraft werden kann. Vielmehr ist ein Verantwortungsbewusstsein gemeint, welches es sinnvoll macht, einen im Detail suboptimalen Entscheid zu priorisieren, beispielsweise eine pünktliche Zugabfahrt zu verpassen, um damit einen im Weg stehenden Menschen nicht zu gefährden. Um verantwortungsvolle Entscheide treffen zu können, braucht es ein übergeordnetes Wertesystem, welches in der Regel kulturell verankert ist. Solche übergeordneten Werte relativieren die Ziele, auf die eine KI trainiert ist. Auch über ein solches übergeordnetes Wertesystem verfügt eine KI nicht.

Im Gegensatz zu KI sind Menschen also nicht darauf beschränkt, aufgrund von Mustern, die aus Daten berechnet wurden, Lösungen zu generieren. Vielmehr bewerten sie Situationen aufgrund von Erfahrungen und Wissen, suchen mit gesundem Menschenverstand nach Lösungen.

3.2.2 Schwächen des Menschen

Auch Menschen sind nicht perfekt in der Entscheidungsfindung. Zum einen können sie weniger Daten berücksichtigen als KI. Denn wer kann schon alle wissenschaftlichen Publikationen lesen, um einen Entscheid vorzubereiten? Vielmehr unterliegen sie auch einer Vielzahl von Fehlern und Verzerrungen.

Anfälligkeit für Fehler und Verzerrungen
Wenn Menschen Entscheide treffen, kann ihnen eine Vielzahl unterschiedlicher Fehler unterlaufen. So können sie Wichtiges übersehen, weil sie unaufmerksam sind oder weil ihre Aufmerksamkeit haarscharf am Wesentlichen vorbeigeht. Auch sind sie anfällig für Vergesslichkeit und können daher Wichtiges auslassen. Auch Denkfehler sind möglich, sodass sich Menschen irren, weil sie von falschen Annahmen ausgehen. All dies passiert

4 Kahnemann, D. (2012). Schnelles Denken, langsames Denken (7. Auflage). Random House.

nicht nur, wenn Menschen sich nicht bemühen. Solche Fehler treten auch auf, obwohl sich Menschen anstrengen und fokussiert arbeiten. Manchmal vielleicht gerade deswegen. So kann konzentriertes Arbeiten zu einem Tunnelblick führen, wodurch anderes aus dem Blick fällt.

Menschliches Denken und Entscheiden ist auch durch verschiedene Verzerrungseffekte geprägt. Die wichtigsten sind im Folgenden beschrieben:

- Ankereffekte: Informationen, die man als Erstes sieht, «verankern» die Aufmerksamkeit und das Denken. Man denkt dann in die Richtung des Ankers und vernachlässigt alternative Denkrichtungen.
- Bestätigungseffekte: Hat man einen Eindruck gewonnen, dann sucht man einäugig nach Bestätigungen für diesen Eindruck. Dabei werden Argumente, die dafürsprechen, überbewertet und Gegenargumente vernachlässigt.
- Fixierungseffekte: Nachdem man einen Entscheid getroffen hat, ist man darauf «fixiert» und verliert die Fähigkeit, den Entscheid zu hinterfragen.

Solche Verzerrungen beeinflussen die Qualität von Entscheidungen sehr. Teilweise unterlaufen sie Expert*innen sogar mehr als Anfänger*innen. Dies weil Anfänger*innen weniger überzeugt von ihren Fähigkeiten sind und sich daher eher hinterfragen.

Insgesamt sind Menschen mit hoher Expertise erstaunlich fähig, aufgrund ihres Wissens und ihrer Erfahrung gute Entscheide zu treffen. Sehr oft liegen sie damit richtig. Wenn sie sich aber irren, dann haben sie aufgrund der beschriebenen Verzerrungseffekte eher Mühe, ihre Entscheide zu hinterfragen.

3.3 Mensch und KI im Vergleich

Die oben beschriebenen unterschiedlichen Eigenschaften und Fähigkeiten von Mensch und KI können folgendermassen auf den Punkt gebracht werden: KI kann hervorragend rechnen und damit Muster in grossen Datenmengen erkennen. Sie kann diese jedoch nicht verstehen. Demgegenüber fällt es Menschen schwer, Daten systematisch auszuwerten. Wenn sie erfahren sind, können sie Situationen jedoch lesen und interpretieren. Für Menschen haben Situationen damit Bedeutung. Für KI sind sie Datenmuster.

4. Warum die Zusammenarbeit von Mensch und KI schwierig ist

Die oben beschriebenen Fähigkeiten von KI bringen zwar erstaunliche Leistungen hervor und bergen wohl auch noch grössere Potenziale. Für den Menschen ist jedoch eine Zusammenarbeit mit KI gerade wegen dieser Fähigkeiten schwierig. Dies aus folgenden Gründen[5]:

- KI spielt Intelligenz vor: KI ist nicht so intelligent, wie sie scheint. Vorschläge, die KI generiert, können ungenau, voreingenommen oder frei erfunden sein. Dies zu erkennen ist sehr schwierig.
- KI ist undurchschaubar: KI ist für Menschen weitgehend undurchschaubar. Es ist für Menschen schwer nachzuvollziehen, wie KI auf Vorschläge kommt. Ohne eine gute Vorstellung davon, wie KI funktioniert, können Menschen das Verhalten von KI weder vorhersehen noch verstehen. Dies erschwert es für sie, mit KI zusammenzuarbeiten und ihre Irrtümer zu erkennen. Entsprechend steigt die Gefahr, dass Menschen vom Verhalten der KI überrascht werden und darauf nicht reagieren können.
- KI ist zunehmend undurchschaubarer: Da KI lernfähig ist, kann sie sich kontinuierlich weiterentwickeln. Dies macht es für Menschen noch schwieriger, sie zu verstehen. Die Möglichkeit, dass Menschen von der KI überrascht werden, nimmt damit ebenfalls kontinuierlich zu.
- KI verstärkt menschliche Fehler: Wenn KI einen Vorschlag macht, verstärkt sie bei den Menschen Fehler und Verzerrungen (s. o.). Dies weil der Vorschlag die Aufmerksamkeit der Menschen in eine bestimmte Richtung lenkt (Ankereffekt), sie selektiv nach Argumenten suchen lässt (Bestätigungseffekt) und auf eine Lösung fixiert (Fixierungseffekt). KI sollte eigentlich genau das Gegenteil bewirken, nämlich Menschen dabei unterstützen, Verzerrungen in ihrem Denken zu überwinden.
- KI gaukelt vor, menschlich zu sein: Je natürlicher KI mit Menschen kommuniziert, je mehr sie also beispielsweise mit einer menschlichen Stimme grammatikalisch korrekte, plausibel klingende Aussagen formuliert, desto grösser ist die Gefahr, dass ihr menschliche Eigenschaften zugeschrieben werden, die sie nicht hat. Dies verstärkt die Annahme, die KI würde nicht nur rechnen, sondern tatsächlich denken und verstehen, was sie sagt. So erschwert es sich für den Menschen, Aussagen von KI korrekt einzuordnen.

5 Bradshaw, J. M., Hoffman, R.R., Johnson, M., & Woods, D.D. (2013). The seven deadly myths of "autonomous systems". IEEE Intelligent Systems, 28 (3), pp. 54–61. Endsley, M. R. (2023). Ironies of artificial intelligence. Ergonomics, 66(11), 1656–1668.Mitchell, M. (2019). Artificial Intelligence. New York: Farrar, Strauss and Giroux.

– KI ist nicht kooperationsfähig: KI ist darauf trainiert, bestimmte Probleme zu lösen. Sie ist nicht darauf trainiert, mit Menschen zusammenzuarbeiten. Damit wird sie für den Menschen zu einer Kooperationspartnerin, die sich nicht ins Gegenüber hineinversetzen und damit auch nicht mit ihm abstimmen kann. KI ist wie eine Kollegin, die ihre Aufgabe vielleicht sehr gut bewältigt, jedoch unfähig ist, sich mit anderen Kolleg*innen zu koordinieren.

KI muss erst lernen, mit Menschen zusammenzuarbeiten

Für Menschen ist die Zusammenarbeit mit KI aus den beschriebenen Gründen anspruchsvoll. Zur Verbesserung dieser Zusammenarbeit könnte eine entsprechende Gestaltung der KI hilfreich sein. Beispielsweise sollte eine KI ihre Vorschläge erklären können, sie sollte transparent machen, wie sie auf ihre Vorschläge kommt, dabei sollte sie Fähigkeiten, aber auch ihre Grenzen aufzeigen, und sie sollte auch nicht gescheiter tun, als sie ist. Letzteres bedeutet vor allem, dass sie explizit als KI identifizierbar sein sollte und keine Menschlichkeit vortäuscht.

KI soll also lernen, mit Menschen zusammenzuarbeiten. Der folgende Abschnitt beschreibt, was Menschen in diese Zusammenarbeit einbringen können und was sie von der KI dazu benötigen.

5. Was der Mensch einbringt und was der Mensch benötigt

Menschen verfügen über Eigenschaften und Fähigkeiten, welche die KI nicht hat (s. o.). Vor diesem Hintergrund können sie insbesondere Folgendes in die Zusammenarbeit mit KI einbringen:

- Bedeutung verstehen: Menschen können im Gegensatz zu KI Situationen interpretieren und deren Bedeutung verstehen. Damit können sie Vorschläge der KI überprüfen und verifizieren. Dazu müssen die Vorschläge für sie nachvollziehbar sein.
- Sich engagieren: Menschen können sich im Gegensatz zu KI engagieren und dabei Prioritäten setzen. Dazu muss die Zusammenarbeit mit KI motivierend sein.
- Verantwortung übernehmen: Menschen können im Gegensatz zu KI Verantwortung übernehmen. Dazu müssen Vorschläge der KI beeinflussbar sein. Dies weil sich Menschen nicht für Entscheidungen verantwortlich fühlen, die von anderen – hier von der KI – behauptet werden und bei welchen sie nicht zumindest mitgeredet haben.

Damit Menschen ihre Stärken in die Zusammenarbeit mit KI einbringen können, müssen also einige Voraussetzungen erfüllt sein. Diese sind im Folgenden beschrieben.

5.1 Menschen beim Entscheiden unterstützen

Menschen beim Entscheiden zu unterstützen, geht darüber hinaus, Entscheide vorzuschlagen und Menschen darüber befinden zu lassen. Dies nur schon deshalb, weil die von KI generierten Vorschläge für Menschen meist nicht nachvollziehbar sind. Dies wird vielleicht etwas besser, wenn KI zu den Vorschlägen Argumente liefert. Eine viel effektivere Entscheidungsunterstützung besteht aber darin, den eigentlichen Prozess des menschlichen Entscheidens[6] zu unterstützen. Dies bedeutet zum Beispiel, Menschen zu helfen, eigene Argumente zu finden und zu prüfen. Die KI soll also nicht Vorschläge liefern, über die Menschen nachdenken sollen. KI soll vielmehr beim Nachdenken helfen.

KI soll beim Nachdenken helfen

Expert*innen, also erfahrene Menschen, können Situationen lesen und interpretieren und gelangen so in aller Regel zu guten Entscheiden (s. o.). Im Folgenden ist etwas detaillierter ausgeführt, welche kognitiven Prozesse daran unter anderem beteiligt sind.

6 Klein, G., & Wright, C. (2016). Macrocognition: From Theory to Toolbox. Frontiers in Psychology, 7.

Um kompetent zu entscheiden, müssen Menschen zuallererst Situationen korrekt einschätzen. Sie müssen sich also ein möglichst präzises Bild der aktuellen Situation machen. Beim Autofahren beispielsweise bedeutet dies, die Verkehrssituation inklusive der spielenden Kinder am Strassenrand, das Verhalten der anderen Verkehrsteilnehmenden, aber auch die Witterungsverhältnisse richtig einzuschätzen, um sich für eine angepasste Geschwindigkeit zu entscheiden. Eine analoge Situationseinschätzung benötigt auch ein*e Betriebsdisponent*in, um eine gute Auftragsplanung vorzunehmen, oder ein*e Ärzt*in, um eine angemessene Diagnose zu stellen. Dazu ist es wichtig, die Situation zu beobachten und dabei die eigene Aufmerksamkeit auf das Wesentliche zu richten. Man muss also zwischen wichtigen und unwichtigen Informationen unterscheiden können. Hat man einmal das Wesentliche im Blick, muss man immer auch ein wenig vorausblicken können, sodass man von der Situationsentwicklung nicht überrascht wird. Derart kann ein*e Betriebsdisponent*in vorwegnehmen, welcher Arbeitsplatz für den Auftragsdurchlauf zu einem Flaschenhals werden könnte, um sich bereits präventiv für eine geeignete Auftragsplanung zu entscheiden. Auf der Basis der Situationseinschätzung erfolgt eine angemessene Situationsbeeinflussung. Hören beispielsweise Maschinenbedienende, dass ihre Werkzeugmaschinen eigenartige Geräusche von sich geben, müssen sie wissen, welche Hebel sie haben, um die Maschineneinstellung zu optimieren. Erfahrene Personen kennen die entsprechenden Einflussmöglichkeiten.

Die beschriebenen Prozesse der Aufmerksamkeitssteuerung, der Situationseinschätzung und der Erkennung geeigneter Einflussmöglichkeiten sind Teil guten Entscheidens. Sie sind verbunden mit kontinuierlicher Sinnzuschreibung, um aufgenommene Informationen richtig einzuordnen und entsprechend angemessene Entscheide zu fällen. KI kann solche Denkprozesse unterstützen, beispielsweise indem sie auf relevante Merkmale hinweist, indem sie Menschen dabei unterstützt, ihre Annahmen zu überprüfen, oder indem sie Menschen die Konsequenzen ihrer Entscheide transparent macht. Solche Funktionen von KI wären eine echte Unterstützung menschlicher Entscheidungsprozesse und nicht ein blosses Vorschlagen von Entscheiden, die für den Menschen schwer zu beurteilen sind.

5.2 Menschen beim Lernen unterstützen

Um gute Entscheide zu fällen bzw. um in ihrem Entscheidungsverhalten immer besser zu werden, müssen auch Expert*innen kontinuierlich dazulernen. KI kann sie dabei massgeblich unterstützen.

<u>KI soll es ermöglichen, Erfahrungen zu machen und daraus zu lernen</u>

Lernen bedeutet, Erfahrungen zu machen, Lehren daraus zu ziehen und diese in das eigene Wissen zu integrieren. Dabei gehen Menschen durch vier zyklische Phasen[7]:

- Konkrete Erfahrung machen: In dieser ersten Phase geht es um die praktische Auseinandersetzung mit Aufgaben, um bestehende Erfahrungen zu hinterfragen oder neue zu machen. Es muss also Gelegenheiten geben, in Situationen zu pröbeln und Neues auszuprobieren.
- Effekte beobachten: In der zweiten Phase geht es darum, aus den Erfahrungen Schlüsse zu ziehen. Man beobachtet, was erfolgreich war oder was verbessert werden könnte. Solche Beobachtungen sind eine entscheidende Voraussetzung für die Verinnerlichung des Gelernten. Es muss also Gelegenheiten geben, die Erfahrungen aus mehreren Perspektiven zu betrachten und Vor- sowie Nachteile abzuwägen.
- Beobachtungen verallgemeinern: In der dritten Phase werden die Beobachtungen aus den konkreten Erfahrungen verallgemeinert. Das bestehende Wissen wird dadurch angepasst und um die neuen Erkenntnisse weiterentwickelt. Es muss also Gelegenheit bestehen, verallgemeinerte Schlussfolgerungen aus den konkreten Beobachtungen zu ziehen.
- Aktives Ausprobieren: In der abschliessenden Phase werden die verallgemeinerten Schlussfolgerungen geprüft. Man wendet sie in der realen Welt an, um Probleme zu lösen und Entscheide zu treffen. Dabei erkennt man, was funktioniert, und verfeinert so das Wissen. Es muss also Gelegenheit bestehen, aktiv zu experimentieren und Feedback dazu zu bekommen.

Die beschriebenen Phasen werden zyklisch durchlaufen. So folgen auf Erfahrungen Beobachtungen, Verallgemeinerung, Experimentieren und wieder neue Erfahrungen.

7 Kolb, D. A. (1984). Experiential learning: Experience as the source of learning and development. Prentice-Hall.

Um mit KI das eigene Entscheidungsverhalten kontinuierlich verbessern zu können, müssen Menschen zu drei verschiedenen Themenfeldern lernen:
– Lernen über die Aufgabe: Dazu gehört insbesondere eine zunehmende Kompetenz des eigenen Fachwissens beziehungsweise der eigenen Expertise.
– Lernen über die KI: Je mehr man über die KI lernt, desto besser weiss man, was sie kann und was sie nicht kann und desto gezielter kann man sie nutzbringend einsetzen.
– Lernen über sich selbst: Das eigene Entscheidungsverhalten kann sehr unterschiedlich sein. Einige Menschen entscheiden beispielsweise risikofreudiger, andere vorsichtiger. Vielleicht entscheidet man auch zu Wochenbeginn anders als vor dem Wochenende. Je mehr man dazu über sich selbst weiss, desto bewusster kann man entscheiden.

Zu allen drei Themenfeldern kann KI die unterschiedlichen Phasen der Lernzyklen gezielt unterstützen. Beispielsweise können Menschen neue Fachkompetenzen zur Aufgabe gewinnen, wenn KI ihnen Feedback darüber gibt, welche Effekte ihre Vorgehensweise hat. Über die KI können sie lernen, wenn die KI transparent macht, was sie kann und was sie nicht kann. Und über sich selbst können Menschen lernen, wenn die KI ihnen einen Spiegel vorhält, indem sie Variabilität im persönlichen Entscheidungsverhalten aufzeigt.

5.3 Das Vertrauen von Menschen gewinnen

Die KI gut zu kennen, ist auch Voraussetzung dafür, Vertrauen in sie zu gewinnen. Sobald Prozesse automatisiert werden, ist angemessenes Vertrauen zentral für einen richtigen Umgang mit der entsprechenden Technik. Wenn Menschen Über- oder Untervertrauen in die Technik entwickeln, können daraus verschiedene Fehler entstehen. Bei Übervertrauen beispielsweise erkennt man nicht mehr, wenn die Technik nicht angemessen funktioniert, und greift dann entsprechend auch nicht ein. Bei Untervertrauen nutzt man die Technik gar nicht erst oder zumindest nicht so, wie das vorgesehen wäre. Über- und Untervertrauen führen entsprechend zu einem unangemessenen Umgang mit KI.

Menschen sollen Möglichkeiten und Grenzen einer KI erkunden

Es ist nicht Ziel, dass Menschen möglichst viel Vertrauen in die KI haben. Ziel ist vielmehr, dass sie ein angemessenes Vertrauen in KI haben. Dies bedeutet, dass sie ein realistisches Verständnis des Anwendungsbereichs und der Grenzen der jeweiligen KI aufbauen. Angemessenes Vertrauen besteht also darin, beurteilen zu können, welche Möglichkeiten und Grenzen eine bestimmte KI hat, sodass ihr für bestimmte Aufgaben oder Ziele in bestimmten Kontexten oder Problemsituationen angemessen vertraut und für andere Aufgaben oder Ziele in bestimmten Kontexten oder Problemsituationen angemessen misstraut wird.

Die KI-Forschung investiert viel Aufwand in die Vertrauenswürdigkeit von KI. In der Fachsprache spricht man dabei von «Trustworthiness». Das Ziel ist, KI möglichst zuverlässig zu machen, sodass sie an Vertrauenswürdigkeit gewinnt. Auch wenn es selbstverständlich wichtig ist, KI möglichst zuverlässig zu machen, löst dieser Ansatz das Vertrauensproblem nicht. Dies weil auch die beste KI – wie jede hervorragende Technik – Grenzen haben wird. Auch wenn diese Grenzen deklariert werden, erwartet man von den Menschen, dass sie dieser Deklaration blind glauben. Echtes Vertrauen ist nicht blind, sondern basiert auf eigener Erfahrung.

Damit Menschen eigene Erfahrung machen können, damit sie also erkennen können, wann und wie sie sich auf eine KI verlassen können, müssen sie diese aktiv erkunden. Sie müssen sie also anwenden und ausprobieren, sodass sie ihre Grenzen kennenlernen. Nur so ist sichergestellt, dass Vertrauen nicht blind gegeben wird, sondern auf direkter Erfahrung und einem tiefen Verständnis der Fähigkeiten der KI beruht. Dies führt zu einer differenzierteren und effektiveren Zusammenarbeit zwischen Mensch und KI.

5.4 Motivation des Menschen fördern

Die Tendenz, IT-Werkzeuge nicht zu nutzen, ist weit verbreitet. Diese Erfahrung wurde schon in vielen IT-Projekten gemacht. Die User*innen nutzen die ihnen zur Verfügung gestellten Tools nicht oder jedenfalls nicht so, wie von ihnen erwartet wird. Gründe dafür wurden oben bereits angesprochen. Wenn eine KI den Menschen Entscheidungen vorschlägt, die sie nicht verstehen, für die sie aber Verantwortung übernehmen müssen, dann nutzen sie diese nicht und entscheiden lieber selbst. Daher ist es wichtig, die Zusammenarbeit von Mensch und KI motivationsförderlich zu gestalten.

KI soll Möglichkeiten eröffnen, nicht einschränken

Drei Aspekte der Zusammenarbeit von Mensch und KI müssen gezielt gestaltet werden, um die Motivation, mit der KI zusammenzuarbeiten, gezielt zu fördern[8]:

- Erlebte Sinnhaftigkeit: Menschen müssen ihre Arbeitsaufgaben als sinnhaft erleben. Sie müssen also nicht nur wissen, was und wie sie etwas tun (know-how), sondern auch warum sie dies tun (know-why). Für die Zusammenarbeit mit KI bedeutet dies beispielsweise, dass die Vorgehensweise der KI nicht nur nachvollziehbar, sondern auch begründet sein soll.
- Erlebte Verantwortung: Ein gewisser Grad an Autonomie ist Voraussetzung dafür, dass sich Menschen für ihr Tun verantwortlich fühlen. Können sie nicht beeinflussen, was und wie sie etwas tun, dann erleben sich Menschen nicht als dafür verantwortlich. Für die Zusammenarbeit mit KI bedeutet dies beispielsweise, dass sie die Art und Weise, wie sie KI nutzen, um ihre Aufgaben zu bearbeiten, beeinflussen können müssen.
- Feedback: Engagiert arbeiten Menschen nur, wenn sie in Bezug auf ihre Leistung Feedback bekommen. Dieses Feedback muss unmittelbar in der täglichen Arbeit ersichtlich sein. Man sollte also jederzeit sehen können, ob man auf dem richtigen Weg ist und ob man die Arbeitsziele erreicht hat. Für die Zusammenarbeit mit KI bedeutet dies beispielsweise, dass die KI den Menschen transparent macht, wie gut die KI-gestützte Entscheidung dazu beigetragen hat, die Arbeitsziele zu erreichen.

Diese drei Aspekte motivationsförderlicher Arbeitsgestaltung stellen sicher, dass KI Menschen nicht einschränkt, sondern ihnen neue Möglichkeiten der Aufgabenbewältigung schafft. Dies ist Voraussetzung für interessiertes und engagiertes Arbeiten. Interessieren und engagieren können sich Menschen nur, wenn sie wissen, warum sie etwas tun (erlebte Sinnhaftigkeit), wenn sie beeinflussen können, was sie tun (erlebte Verantwortung), und wenn sie beurteilen können, ob sie erfolgreich waren (Feedback). Voraussetzung für Motivation ist also, dass KI nicht einschränkt, sondern Möglichkeiten schafft.

8 Ulich, E. (2011). Arbeitspsychologie. Stuttgart: Schäffer Poeschel.
Hackman, J. R., &, Oldham, G.R. (1976). Motivation through the design of work: Test of a theory. Organizational Behavior and Human Performance, 16(2), 250–279.

6. Tipps zur Gestaltung von KI

Die bisherigen Ausführungen zeigen auf, dass Mensch und KI grundsätzlich unterschiedlich sind und dass eine gelingende Zusammenarbeit von Mensch und KI diese Unterschiedlichkeit berücksichtigen muss. Insbesondere muss die Gestaltung sowohl der KI als auch der Zusammenarbeit von Mensch und KI auf menschliche Eigenschaften eingehen. Sie sollte also menschengerecht sein. Mica Endsley[9] hat eine Reihe grundsätzlicher Tipps für eine menschengerechte Gestaltung von KI erarbeitet, die im Folgenden zusammenfassend dargestellt sind:

- KI muss Menschen unterstützen: Wenn eine KI Vorschläge generiert, die Menschen nicht verstehen, aber beurteilen müssen, dann ist dies keine Unterstützung. Eine wirkliche Unterstützung hilft den Menschen beim Denken, indem sie einzelne Denkprozesse unterstützt und dabei menschliche Eigenheiten berücksichtigt.
- Zuschreibung von Eigenschaften: KI-Systeme wie beispielsweise Bots sollen explizit als KI-Systeme identifizierbar sein. Sie sollen nicht so tun, als ob sie Menschen wären. Dies, um zu verhindern, dass Menschen ihnen menschliche Eigenschaften zuschreiben, die sie gar nicht haben.
- Erklärbarkeit von KI: Die KI muss den Menschen in jedem Einzelfall vermitteln, warum sie bestimmte Empfehlungen ausspricht oder Massnahmen ergreift. Diese Erklärungen müssen sowohl auf die spezifischen Umstände als auch auf die Bedürfnisse der jeweiligen Nutzenden zugeschnitten sein, damit sie verständlich und hilfreich sind.
- Transparenz der KI: Menschen müssen sich jederzeit ein angemessenes Bild (a) über den aktuellen Zustand der gemeinsamen Aufgabenbearbeitung, (b) über die aktuelle Aufgabenverteilung zwischen Mensch und KI und (c) über die aktuelle Form der Zusammenarbeit mit der KI machen können.
- Verzerrungen offenlegen: KI muss dem Menschen transparent machen, aus welchen Daten sie gelernt hat und welche Einschränkungen und Verzerrungen sich daraus für ihr Funktionieren ergeben.
- Menschliche Kontrolle über KI: Die Zusammenarbeit von Mensch und KI muss derart gestaltet sein, dass Menschen jederzeit ausreichend informiert sind, um zu erkennen, wann eine Situation ausserhalb der Fähigkeiten der KI liegt. Dies als Voraussetzung dafür, rechtzeitig eingreifen zu können.
- Training und Aufrechterhaltung von Fähigkeiten: Menschen müssen jederzeit in der Lage sein, die Aufgaben mit oder ohne KI-Unterstützung ausführen zu können. Auch dies ist eine Voraussetzung dafür, rechtzeitig eingreifen zu können.
- KI und Mensch gemeinsam testen: Die Sicherheit und Zuverlässigkeit von KI-Systemen bei der Zusammenarbeit mit Menschen muss sorgfältig geprüft werden, bevor sie in sicherheitskritischen Umgebungen eingesetzt werden. Dabei muss auch getestet werden, ob Menschen die KI verstehen und Probleme unter realistischen Bedingungen erkennen und überwinden können. Bei diesen Tests sollten auch unvorhergesehene Verhaltensweisen der KI und die Fähigkeit der Menschen, diese zu erkennen und angemessen darauf zu reagieren, ermittelt werden.

9 Endsley, M. R. (2023). Ironies of artificial intelligence. Ergonomics, 66(11), 1656–1668.

Anhang

Anhang 1: Vertiefende Literatur

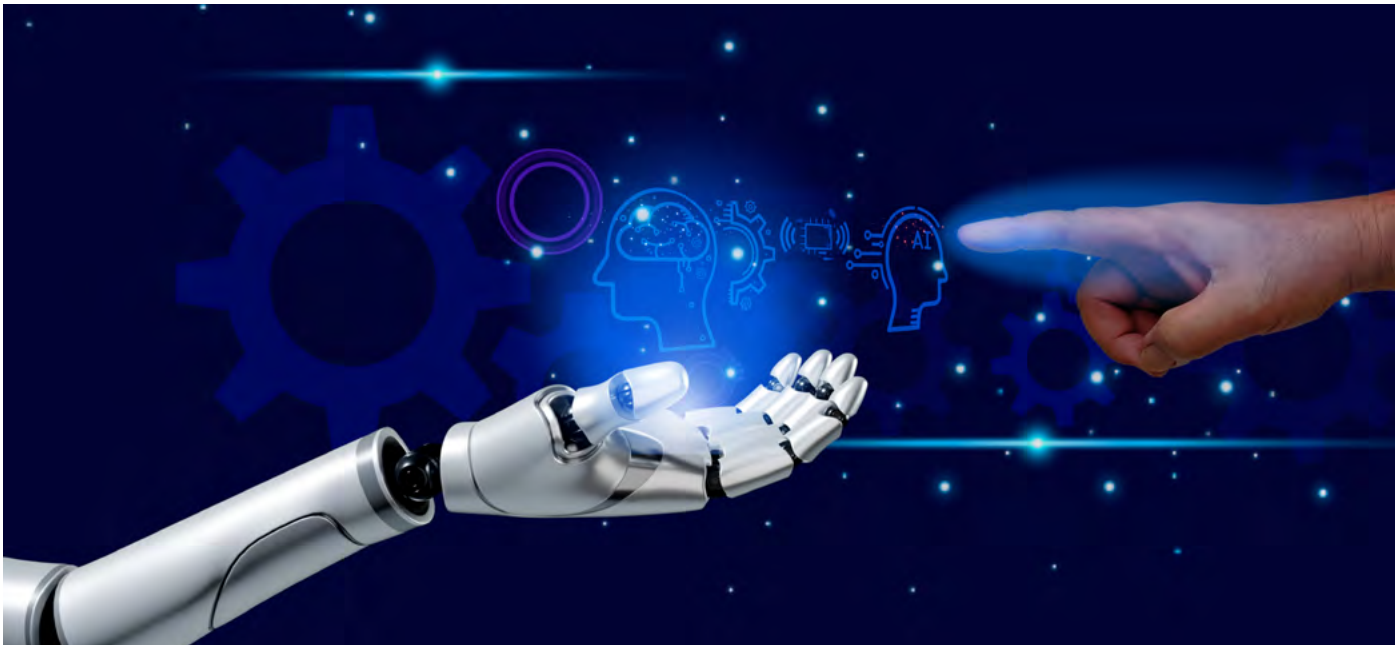
Mitchell, M. (2019). Artificial Intelligence. New York: Farrar, Strauss and Giroux.

Russell, S. (2019). Human Compatible. Artificial Intelligence and the Problem of Control. Penguin Books.

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

Wäfler, T. (2020). Gebildeter und vernetzter Mensch: Vier Thesen zur soziotechnischen Gestaltung der Zukunft. Journal Psychologie des Arbeitshandelns. 13(2), S. 5–21.

Anhang 2: Gestaltungshinweise mit Vor- und Nachteilen



Die Entwicklung und Verbreitung von unterschiedlichen Formen der künstlichen Intelligenz (KI) in zahlreichen Lebensbereichen bietet Chancen, aber auch Herausforderungen. Damit der Mensch seine Stärken einbringen und von den Potenzialen dieser fortschrittlichen Technologie profitieren kann, bedarf es fundierter Instrumente mit Gestaltungshinweisen, welche darauf abzielen, eine sinnvolle Interaktion zwischen Mensch und KI zu ermöglichen.

Nachfolgend werden fünf solcher Instrumente in Bezug auf ihre Stärken und Schwächen sowie ihre Potenziale, den Menschen zu unterstützen, beschrieben.

Übersicht über die beschriebenen Instrumente:

- Intervention User Interfaces (Schmidt & Herrmann, 2017)
- The Eight Golden Rules of Interface Design (Shneiderman et al., 2016)
- Guidelines for Human-AI Interaction (Amershi et al., 2019)
- Kriterien für die Mensch-Maschine-Interaktion bei KI (Huchler et al., 2020)
- The Explanation Goodness Checklist and Scale for Explainable AI (Hoffman et al., 2018)

Name des Gestaltungshinweises	**Intervention User Interfaces: A New Interaction Paradigm for Automated Systems (Schmidt & Herrmann, 2017)**
Literaturangabe	Schmidt, A. & Herrmann, T. (2017). Intervention User Interfaces: Interactions, 24(5), 40–45. https://doi.org/10.1145/3121357
Zweck, wozu einsetzbar	Autonome beziehungsweise automatisierte Systeme in komplexen Umwelten erfordern Schnittstellen («Intervention User Interfaces»), welche so gestaltet sind, dass Nutzende eingreifen können, um jederzeit Anpassungen vorzunehmen. «A challenge is that humans must be made aware and understand that there is an opportunity for interaction» (Schmidt & Herrmann, 2017, S. 42) Einsatzbereiche der Gestaltungshinweise: – Autonomes Fahren in vernetzten Verkehrssystemen – Smart Home und betreutes Wohnen – Automatisierte Fertigung und Smart Factories – Personalisierte Assistenzsysteme
Kriterien	Bei der Gestaltung der Schnittstellen für autonome Systeme sollten folgende sechs Gestaltungshinweise berücksichtigt werden. **Expectability and predictability** Nutzende müssen das Verhalten des automatisierten Systems jederzeit verstehen und voraussagen können, d. h., sie dürfen nicht überrascht werden. **Communicate options for interventions** Nutzende sollen kontextspezifische Optionen für eine Intervention in verständlicher und unaufdringlicher Form angezeigt erhalten. **Exploration of interventions** Nutzende sollen sichere Möglichkeiten (z. B. Simulation) für die Erkundung von Interventioånen und deren Auswirkungen erhalten. **Easy reversal of automated and intervention actions** Nutzende sollen Möglichkeit erhalten, automatisiertes Verhalten des Systems sowie Ergebnisse von Interventionen einfach rückgängig zu machen. **Minimize required attention** Minimierung der für die Bedienung des automatisierten Systems erforderlichen Aufmerksamkeit der Nutzenden. **Communicate how control is shared** Klare Kommunikation der Zuständigkeiten und tatsächlicher Kontrolle zwischen Nutzenden und automatisiertem System.

Name des Gestaltungshinweises	**Intervention User Interfaces: A New Interaction Paradigm for Automated Systems (Schmidt & Herrmann, 2017)**
Stärken	Die Stärke der Gestaltungshinweise von Schmidt und Herrmann (2017) liegt in der klaren Fokussierung auf der Schnittstellengestaltung von autonomen beziehungsweise zunehmend autonomen Systemen. Hierbei steht im Zentrum, wie die Nutzenden konkret Einfluss in Form von Interventionen ausüben können.
Schwächen	Als Schwäche kann die zu wenig ausführliche Operationalisierung der jeweiligen Gestaltungshinweise betrachtet werden.
Unterstützungsfunktionen	**Entscheidungsunterstützung:** Die menschliche Entscheidungsunterstützung wird durch die Gestaltungshinweise nicht explizit adressiert. **Unterstützung von menschlichem Lernen:** Die menschlichen Lernprozesse werden durch die Gestaltungshinweise nur teilweise durch «Exploration of interventions» unterstützt, jedoch wird die Thematik der Reflexion, welche für menschliches Lernen wichtig ist, nicht konkret adressiert. **Unterstützung von Vertrauen:** Insbesondere die Kriterien «Expectability and predictability» wie auch «Easy reversal of automated and intervention actions» unterstützen den Menschen, angemessenes Vertrauen in das System aufzubauen.

Name des Gestaltungshinweises	**The Eight Golden Rules of Interface Design (Shneiderman et al., 2016)**
Literaturangabe	Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., and Elmqvist, N., Designing the User Interface: Strategies for Effective Human-Computer Interaction: Sixth Edition, Pearson (May 2016) http://www.cs.umd.edu/hcil/DTUI6
Zweck, wozu einsetzbar	Die «Eight Golden Rules of Interface Design» von Shneiderman et al. (2016) sind grundlegende Hinweise für die Gestaltung von Schnittstellen, welche den Zweck verfolgen, die Effizienz, Benutzerfreundlichkeit und allgemeine Effektivität sicherzustellen. Einsatzbereiche der Gestaltungshinweise: – Softwareentwicklung – Webdesign
Kriterien	Gemäss Shneiderman et al. (2016) sollten Schnittstellen nach folgenden Kriterien gestaltet werden, um Benutzerfreundlichkeit sicherzustellen: **Strive for consistency** Terminologien, Layouts, Farben und Schriftfarben sollten innerhalb der Schnittstelle einheitlich verwendet werden. Dies ermöglicht den Nutzenden eine intuitive Nutzung. **Seek universal usability** Die Schnittstelle sollte für Nutzende mit unterschiedlichen Erfahrungen (z. B. Anfänger, Novizen) und Fähigkeiten nutzbar sein. Hierzu gehört beispielsweise die Berücksichtigung der Barrierefreiheit (WCAG 2.1). **Offer informative feedback** Die Nutzenden sollten für getätigte Aktionen angemessenes Feedback erhalten. Hierdurch erhalten Nutzende Sicherheit und Orientierung in der Nutzung der Schnittstelle. **Design dialogs to yield closure** Interaktionen sollten in Gruppen (Anfang, Mitte, Ende) organisiert werden. Nach einer Interaktionssequenz sollten Nutzende jeweils informatives Feedback erhalten. Dieses Feedback nach einer abgeschlossenen Interaktionssequenz sorgt für das Gefühl von Zufriedenheit und Leistung. **Prevent errors** Die Schnittstelle sollte so gestaltet sein, dass Nutzende möglichst keine gravierenden Fehler machen können. Das Risiko für Fehler sollte minimiert werden, zudem sollten Nutzende verständliche Möglichkeiten für die Fehlerkorrektur erhalten.

Name des Gestaltungshinweises	The Eight Golden Rules of Interface Design (Shneiderman et al., 2016)
Kriterien	**Permit easy reversal of actions** Nutzende sollten – sofern möglich – Aktionen (z. B. Dateneingaben, Gruppe von Aktionen) rückgängig machen können. Diese Funktion mindert Ängste der Nutzenden und sorgt für die Exploration unbekannter Optionen. **Keep users in control** Insbesondere erfahrene Nutzende möchten das Gefühl haben, die Schnittstelle zu beherrschen und dass die Schnittstelle auf ihre Aktionen reagiert. Überraschungen sollten in der Bedienung vermieden werden. Weiterhin sollten Nutzende keine langwierigen Dateneingaben oder auch aufwendige Informationsbeschaffung vornehmen müssen. **Reduce short-term memory load** Die Schnittstelle sollte so gestaltet sein, dass sie nicht das Kurzzeitgedächtnis der Nutzenden überfordert, indem Informationen klar dargestellt und leicht abrufbar gemacht werden. Hierzu gehört beispielsweise, dass Standorte von Websites sichtbar bleiben oder auch dass lange Formulare angepasst werden sollen, damit sie auf einen Bildschirm passen.
Stärken	Die grosse Stärke der Gestaltungshinweise von Shneiderman et al. (2016) liegt in ihrer universellen Anwendbarkeit für die Gestaltung von unterschiedlichsten Schnittstellen. Sie bieten eine gute Übersicht über zentrale Punkte, welche bei der Schnittstellengestaltung berücksichtigt werden sollen.
Schwächen	Die Gestaltungshinweise sind recht weit gefasst und allgemein gehalten, was es schwierig machen kann, sie ohne Beispiele oder weiteren Kontext für die konkrete Gestaltung einer Schnittstelle anwenden zu können.
Unterstützungsfunktionen	**Entscheidungsunterstützung:** Die menschliche Entscheidungsfindung wird teilweise durch die Reduzierung der kognitiven Belastung («Reduce short-term memory load») sowie durch die Möglichkeit, Fehler rückgängig zu machen («Permit easy reversal of actions»), adressiert. **Unterstützung von menschlichem Lernen:** Prozesslernen wird durch die Kriterien «Offer informative feedback» sowie «Permit easy reversal of actions» unterstützt. Das Feedback unterstützt den Nutzenden, den Bedienprozess nachzuvollziehen, das Rückgängigmachen von Fehlern ermöglicht den Nutzenden die Exploration und Reflexion. **Unterstützung bei Entwicklung von Vertrauen:** Vertrauen wird durch die Kriterien «Strive for consistency», «Prevent errors» sowie «Keep users in control» gefördert. Durch die konsistente Schnittstellengestaltung, welche Fehler vermeidet und den Nutzenden in Kontrolle hält, wird Vertrauen gefördert.

Name des Gestaltungshinweises	**Guidelines for Human-AI Interaction (Amershi et al., 2019)**
Literaturangabe	Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019, May). Guidelines for Human-AI Interaction. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1–13).
Zweck, wozu einsetzbar	Die «Guidelines for Human-AI Interaction» von Amershi et al. (2019) bieten Designern, Entwickelnden, Forschenden einen strukturierten Ansatz bei der Schnittstellengestaltung von Systemen, welche mit KI angereichert (infused) werden – sogenannten «AI-infused systems». Einsatzbereiche der Gestaltungshinweise: – Gestaltung von AI-infused systems
Kriterien	**Interaktionszeitpunkt: Initially** (Anfänglich): **Make clear what the system can do** Nutzende sollen dabei unterstützt werden zu verstehen, was die Fähigkeiten und Limitationen des KI-Systems sind. **Make clear how well the system can do what it can do** Nutzende sollen dabei unterstützt werden zu verstehen, wie häufig das KI-System Fehler macht. **Interaktionszeitpunkt: During Interaction** (Während der Interaktion): **Time services based on context** Der Zeitpunkt, an welchem die Nutzenden interagieren oder die Interaktion unterbrechen können, soll an die Aufgabe und Umgebung angepasst werden. **Show contextually relevant information** Nutzende sollen Informationen, welche für ihren Kontext und ihre Aufgabe relevant sind, angezeigt erhalten. **Match relevant social norms** Das KI-System soll im Verhalten und in der Präsentation von Informationen den sozialen und kulturellen Normen der Nutzenden angepasst sein. **Mitigate social biases** Es muss sichergestellt werden, dass die Sprache und das Verhalten des KI-Systems keine ungewünschten und unfairen Stereotypen und Biases reproduziert. **Interaktionszeitpunkt: When wrong** (Wenn falsch): **Support efficient innovation** Nutzende sollen bei Bedarf eine einfache Möglichkeit erhalten, die Dienste des KI-Systems aufzurufen oder anzufordern. **Support efficient dismissal** Für Nutzende sollte es einfach sein, unerwünschte Dienste des KI-Systems abzulehnen oder zu ignorieren. **Support efficient correction** Nutzende sollen einfache Möglichkeiten für die Bearbeitung, Überarbeitung oder Wiederherstellung haben, falls das KI-System falsch liegt. **Scope services when in doubt** Bei Unsicherheiten, welche durch mehrdeutige Situationen entstehen können, wenn das Ziel des Nutzenden unklar ist, soll das System versuchen, diese Mehrdeutigkeiten aufzulösen (z. B. Nachfrage an Nutzenden). Falls dies nicht möglich ist, sollten die Leistungen des Systems reduziert werden (z. B. Anzahl Empfehlungen). **Make clear why the system did what it did** Nutzende sollen Zugriff auf Erklärungen erhalten, warum sich das KI-System verhalten hat, wie es sich verhalten hat.

Name des Gestaltungshinweises	**Guidelines for Human-AI Interaction (Amershi et al., 2019)**
Kriterien	**Interaktionszeitpunkt: Over time** (Im Laufe der Zeit): **Remember recent interactions** Das System soll sich an kürzlich getätigte Interaktionen erinnern, um damit den Nutzenden zu ermöglichen, effizient darauf Bezug zu nehmen. **Learn from user behavior** Anhand von getätigten Handlungen des Nutzenden soll das System lernen und somit die Erfahrung für den Nutzenden personalisieren. **Update and adapt cautiously** Änderungen und Weiterentwicklungen des KI-Systems sollen begrenzt und nicht disruptiv sein. **Encourage granular feedback** Nutzende sollen die Möglichkeit erhalten, während der regelmässigen Interaktion mit dem KI-System Feedback zu ihren Präferenzen zu geben. **Convey the consequences of user actions** Nutzenden sollte mitgeteilt werden, wie ihre Aktionen zukünftige Verhaltensweisen des KI-Systems beeinflussen werden. **Provide global controls** Nutzende sollen die Möglichkeit erhalten, global zu bestimmen, was das KI-System überwacht und wie es sich verhält. **Notify users about changes** Nutzende sollen informiert werden, sobald das KI-System seine Fähigkeiten erweitert oder aktualisiert.
Stärken	Die Stärken der Gestaltungshinweise von Amershi et al. (2019) liegen in der Fokussierung auf Interaktionen in Mensch-KI-Systemen beziehungsweise «AI-infused systems». Eine weitere Stärke liegt in der Bandbreite von relevanten Aspekten sowie der Berücksichtigung unterschiedlicher Interaktionszeitpunkte, welche die Hinweise für die Gestaltung von Mensch-KI-Systemen abdecken. Zudem bieten die Gestaltungshinweise eine hohe Praxistauglichkeit, welche die umfassende Validierung durch Praktiker gezeigt hat.
Schwächen	Als Schwäche der Gestaltungshinweise von Amershi et al. (2019) kann die zu wenig ausgeführte Operationalisierung verstanden werden. Die konkrete Anwendung insbesondere für komplexe Anwendungsszenarien könnte eine Herausforderung darstellen.
Unterstützungsfunktionen	**Entscheidungsunterstützung:** Die menschliche Entscheidungsunterstützung wird teilweise durch das Kriterium «show contextually relevant information» adressiert. Adäquat gestaltete Erklärungen (XAI) unterstützen den Menschen beim Prozess von «Sensemaking», d. h. dabei, Erlebnisse in der Umwelt wahrzunehmen, zu verstehen und in sinnvolle Einheiten zu ordnen. **Unterstützung von menschlichem Lernen:** Durch die Gestaltungshinweise wird insbesondere das Prozesslernen des Menschen unterstützt. Durch die Kriterien «Make clear what the system can do», «Make clear how well the system can do what it can do», «Make clear why the system did what it did» sowie «Notify users about changes» werden zum einen das menschliche Prozesslernen, zum anderen die realistische Einschätzung der Fähigkeiten und Limitationen des Systems unterstützt. Dies wiederum unterstützt den Menschen bei der Entwicklung einer Vorstellung betreffend Prozess und System. **Unterstützung bei Entwicklung von Vertrauen:** Die Kriterien, welche in obigem Abschnitt bei der Unterstützung von menschlichem Lernen erwähnt wurden, unterstützen durch das Aufzeigen der Fähigkeiten und Limitationen des Systems auch die Entwicklung von angemessenem Vertrauen.

Name des Gestaltungshinweises	**Kriterien für die Mensch-Maschine-Interaktion bei KI (Huchler et al., 2020)**
Literaturangabe	Huchler, N., Adolph, L., André, E., Bauer, W., Bender, N., Müller, N., … & Suchy, O. (2020). Kriterien für die Mensch-Maschine-Interaktion bei KI. Ansätze für die menschengerechte Gestaltung in der Arbeitswelt. Plattform Lernende Systeme, München.
Zweck, wozu einsetzbar	Die «Kriterien für die Mensch-Maschine-Interaktion bei KI» von Huchler et al. (2020) eignen sich für die menschzentrierte Gestaltung von Mensch-KI-Systemen. Hierbei liegt der Fokus der Kriterien auf der Komplementarität der Interaktionspartner Mensch und KI. Einsatzbereiche der Gestaltungshinweise: – Gestaltung von AI-infused systems
Kriterien	**Cluster 1: Schutz des Einzelnen** **Sicherheit und Gesundheitsschutz** KI-Systeme sollen u. a. durch eine menschengerechte Gestaltung negative physische oder psychische Beanspruchungsfolgen (z. B. Monotonie oder psychische Sättigung) reduzieren. **Datenschutz und verantwortungsbewusste Leistungserfassung** Bei der Gestaltung von Mensch-Maschine-Interaktion im Kontext künstlicher Intelligenz sollen bereits beim Design dieser Systeme vertrauenswürdige Verfahren und absichernde Regelungen (z. B. DSGVO) berücksichtigt werden. **Vielfaltssensibilität und Diskriminierungsfreiheit** KI-Systeme sollten so gestaltet werden, dass der Schutz vor Diskriminierung von Individuen oder Gruppen sichergestellt werden kann. Weiterhin sollte darauf geachtet werden, mögliche Verzerrungen zu vermeiden. **Cluster 2: Vertrauenswürdigkeit** **Qualität der verfügbaren Daten** Die Mensch-Maschine-Interaktion innerhalb von KI-Systemen erfordert Datenmaterial, welches von hoher Qualität (Konsistenz, Vergleichbarkeit, Reliabilität, inhaltlicher Validität) ist, damit Fehlinterpretationen und Verzerrungen vorgebeugt werden kann sowie die Qualität von statistischen Vorhersagen sichergestellt wird. **Transparenz, Erklärbarkeit und Widerspruchsfreiheit** KI-Systeme stellen Nutzende vor eine oftmals nicht mehr nachvollziehbare Komplexität. Hier bedarf es Ansätze von einer erklärbaren KI (XAI), welche den Nutzenden Basisinformationen über u. a. die prinzipielle Funktionsweise, eingeschriebenen Zwecke und Zielsetzungen, den Datenfokus und die Datengrundlage der KI-Systeme liefert. **Verantwortung, Haftung und Systemvertrauen** Nutzende sollen Informationen, welche relevant für ihren Kontext und ihre Aufgabe sind, angezeigt erhalten. Wichtig für die menschliche Kontrolle des technischen Systems ist es, dass die Nutzenden ein zukünftiges Systemverhalten abschätzen können. Zu beachten beim Systemdesign ist, dass Prozesse (Input, Verarbeitung und Output) so angelegt sind, dass diese auch im Nachhinein nachvollzogen werden können, um eventuelle Problematiken zu identifizieren. **Cluster 3: Sinnvolle Arbeitsteilung** **Angemessenheit, Entlastung und Unterstützung** KI-Systeme sollen so gestaltet werden, dass die unterschiedlichen Fähigkeiten und Eigenschaften von Mensch und Technik optimal – d. h. komplementär – aufeinander abgestimmt werden. Hierdurch wird eine sinnvolle Arbeitsteilung ermöglicht, welche auf die Qualifikationen und Kompetenzen der jeweiligen Akteure zugeschnitten ist. Der Nutzende solcher KI-Systeme soll u. a. in schwierigen Entscheidungssituationen unterstützt werden. **Handlungsträgerschaft und Situationskontrolle** Bei der Mensch-Maschine-Interaktion innerhalb von KI-Systemen ist eine gezielte und transparente Gestaltung der Handlungsträgerschaft und Situationskontrolle wichtig. Dies

Name des Gestaltungshinweises	Kriterien für die Mensch-Maschine-Interaktion bei KI (Huchler et al., 2020)
Kriterien	ermöglicht eine potenzielle Verantwortungszuschreibung sowie eine entsprechende Entlastung der Situation von unklaren Risiken. Zudem ermöglicht ein hohes Mass an Handlungsträgerschaft und Situationskontrolle die Vorbeugung von Unzufriedenheit der Nutzenden. **Adaptivität, Fehlertoleranz und Individualisierbarkeit** KI-Systeme sollen sich flexibel und situationsspezifisch an den Bedarfen und Bedürfnissen sowie der Arbeitspraxis von Nutzenden ausrichten können. Zum einen sollen solche KI-Systeme die Anforderungen aus der Umwelt in die eigene Systemlogik übersetzen («assimilierende Adaptivität»), zum anderen auch die eigene Bearbeitungslogik an die Bedarfe der Umwelt und insbesondere der Mitarbeitenden anpassen («komplementäre Adaptivität»). **<u>Cluster 4: Förderliche Arbeitsbedingungen</u>** **Handlungsräume und reichhaltige Arbeit** Bei der Gestaltung und dem Einsatz von KI-Systemen muss auf den Erhalt und falls möglich die Erweiterung von Handlungsspielraum geachtet werden. Dies betrifft einerseits die Ziele, die Arbeitsinhalte und die konkrete Ausführung, aber auch die strukturierenden, d. h. organisationalen und technischen, Rahmenbedingungen. Zudem soll darauf geachtet werden, dass die KI-Systeme nicht jene Arbeitsinhalte übernehmen, welche die Arbeit motivierend, qualifizierend und gesundheitsförderlich machen. **Lern- und Erfahrungsförderlichkeit** Die Mensch-Maschine-Interaktion innerhalb von KI-Systemen sollte für Nutzende lern- sowie erfahrungsförderlich gestaltet werden. Dazu gehört die nachvollziehbare («Explainable AI») und wechselseitige adaptive Gestaltung, um die Aneignung von Wissen und Erfahrung im Nutzungsprozess zu ermöglichen. Die wechselseitig lernförderliche Gestaltung erhöht die Wahrscheinlichkeit, dass Menschen bereit sind, ihr Wissen und ihre Erfahrungen in KI-Systeme einzubringen. **Kommunikation, Kooperation und soziale Einbindung** bei der Gestaltung der Mensch-Maschine-Interaktion KI-Systeme sollen im doppelten Sinne «sensibel» für soziale Kontexte und Strukturen sein. Einerseits sollen sie in sehr beschränktem Masse ganz als Kooperationspartner agieren; anderseits sollen sie notwendige und gewinnbringende Kommunikation, Kooperation und Verbundenheit nicht verhindern oder ersetzen, sondern idealerweise zielführend unterstützen. Kompetenzen des Menschen sollen vom technischen System erkannt und die Funktionalitäten innerhalb der Interaktion darauf adaptiert werden.
Stärken	Die Kriterien von Huchler et al. (2020) decken viele wichtige Aspekte ab, welche für den menschzentrierten Einsatz von künstlicher Intelligenz von Relevanz sind. Hierbei liegt der Fokus der unterschiedlichen Kriterien auf der Komplementarität von Mensch und KI.
Schwächen	Die Kriterien von Huchler et al. (2020) sind zwar gut und detailliert beschrieben, jedoch fehlt eine ausführliche Operationalisierung, was den Einsatz in der Praxis erschwert.
Unterstützungsfunktionen	**Entscheidungsunterstützung:** Die menschliche Entscheidungsunterstützung wird teilweise durch das Kriterium «Lern- und Erfahrungsmöglichkeiten» adressiert, welches mittels XAI die Nutzenden bei der Entscheidungsfindung unterstützen soll. **Unterstützung von menschlichem Lernen:** Die Unterstützung menschlicher Lernprozesse wird insbesondere mit den Kriterien «Transparenz, Erklärbarkeit und Widerspruchsfreiheit» sowie «Lern- und Erfahrungsmöglichkeit» adressiert. **Unterstützung bei Entwicklung von Vertrauen:** Die Unterstützung angemessener menschlicher Vertrauensentwicklung wird einerseits durch die Berücksichtigung der Kriterien im «Cluster 2: Vertrauenswürdigkeit» ermöglicht; andererseits durch das Kriterium «Handlungsträgerschaft und Situationskontrolle», welches mithilfe von Transparenz Klarheit und Vertrauen fördert.

Name des Gestaltungshinweises	**The Explanation Goodness Checklist and Scale for Explainable AI (Hoffman et al., 2018)**
Literaturangabe	Hoffman, R.R., Mueller, S.T., Klein, G., and Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects. Technical Report, DARPA Explainable AI Program.
Zweck, wozu einsetzbar	**Explanation Goodness Checklist:** Die Goodness-Checkliste richtet sich an Forschende oder Domänen-Experten, welche eine unabhängige Evaluation der Güte von Erklärungen durchführen wollen, die von Explainable-AI-Systemen (XAI) erzeugt werden. Damit kann herausgefunden werden, wie gut die Erklärungen des KI-Systems für Nutzende sind. **Explanation Satisfaction Scale:** Die Explanation Satisfaction ist eine Bewertung von Erklärungen durch Nutzende. Die Explanation Satisfaction Scale dient der Erfassung von Beurteilungen durch Forschungsteilnehmende, nachdem sie mit dem zu erklärenden XAI-System gearbeitet haben.
Kriterien	Die Explanation Goodness Checklist und die Explanation Satisfaction Scale sind sich sehr ähnlich, sie unterscheiden sich jedoch im jeweiligen Anwendungskontext (s. o.). Beide basieren auf nachfolgenden Kriterien: **Understandability** Bei diesem Kriterium geht es darum, dass die Erklärung den Nutzenden dabei hilft zu verstehen, wie das XAI-System funktioniert. **Sufficiency of detail** Bei diesem Kriterium geht es darum, ob der Detaillierungsgrad der Erklärung des XAI-Systems adäquat beziehungsweise ausreichend ist. **Completeness** Bei diesem Kriterium geht es darum, wie vollständig beziehungsweise unvollständig die Erklärung des XAI-Systems ist. **Usefulness** Bei diesem Kriterium geht es um die Umsetzbarkeit der Erklärung, d. h. darum, inwiefern diese Auskunft darüber gibt, wie das XAI-System genutzt werden kann. **Accuracy** Bei diesem Kriterium geht es darum, ob die Erklärung des XAI-System den Nutzenden hilft zu verstehen, wie verlässlich das XAI-System ist. Es adressiert also die Frage: Wie und wann ist das XAI-System verlässlich? **Trustworthiness** Bei diesem Kriterium geht es um die Vertrauenswürdigkeit der Erklärung, welche die Nutzenden des XAI-Systems erhalten. Es adressiert also die Frage: Wie und wann können Nutzende dem XAI-System vertrauen?

Name des Gestaltungshinweises	**The Explanation Goodness Checklist and Scale for Explainable AI (Hoffman et al., 2018)**
Stärken	**Explanation Goodness Checklist** Die Explanation Goodness Checklist bietet Forschenden und Domänen-Experten eine einfach nutzbare Checkliste, welche schnell ermöglicht, eine Evaluation der generierten Erklärungen des XAI-Systems a priori durchzuführen. **Explanation Satisfaction Scale** Die Explanation Satisfaction Scale ermöglicht es, mit 6 Items zu evaluieren, wie zufrieden die Nutzenden mit den generierten Erklärungen des XAI-Systems sind.
Schwächen	Die in der Checkliste und der Skala verwendeten Güte-Kriterien für Erklärungen sind bisher in der Praxis noch nicht systematisch validiert worden.
Unterstützungsfunktionen	**Entscheidungsunterstützung:** Die menschliche Entscheidungsfindung wird durch die Verwendung der Gestaltungshinweise teilweise unterstützt, da adäquat gestaltete Erklärungen den Menschen beim Prozess von «Sensemaking» unterstützen, d. h. dabei, Erlebnisse in der Umwelt wahrzunehmen, zu verstehen und in sinnvolle Einheiten zu ordnen. **Unterstützung von menschlichem Lernen:** Die Unterstützung menschlicher Lernprozesse wird durch die Verwendung der Gestaltungshinweise marginal gefördert, da adäquat gestaltete Erklärungen den Menschen bei der Entwicklung einer konkreten Vorstellung betreffend Aufgabe und Prozess unterstützen können. **Unterstützung bei Entwicklung von Vertrauen:** Durch die Verwendung des Kriteriums «Trustworthiness» wird die angemessene menschliche Vertrauensentwicklung adressiert, da transparent gemacht wird, wann vertraut werden kann.

Glossar

Begriff	Bedeutung
Intervention	Eine Intervention in der Mensch-Computer-Interaktion ist eine Aktion des Nutzenden, die während der Nutzung eines automatisierten Systems stattfindet und eine Abweichung vom vordefinierten Verhalten einleitet (Schmidt & Herrmann, 2017).
Autonome Systeme	Autonome Systeme sind technische Systeme, welche automatisch und selbstständig arbeiten, ohne dass der Mensch sie fortlaufend überwachen oder in sie eingreifen muss. Bei der Entwicklung dieser Systeme muss ein ausgewogenes Verhältnis zwischen Automatisierung und menschlicher Intervention gefunden werden (Schmidt & Herrmann, 2017).
Künstliche Intelligenz / Mensch-KI-Systeme	Gemäss Huchler et al. (2020) bezieht sich künstliche Intelligenz auf lernende Systeme, die immer komplexere Interaktionsfähigkeiten haben und in der Arbeitswelt eingesetzt werden. Die Zusammenarbeit zwischen Mensch und Technik erfordert eine Neujustierung der Aufgabenverteilung, wobei die Stärken und Potenziale beider Seiten berücksichtigt werden müssen. Eine koordinierte Balance kann dazu beitragen, dass KI sowohl den Beschäftigten als auch den technologischen und wirtschaftlichen Potenzialen gerecht wird. Die Gestaltung der Mensch-KI-Interaktion ist nur ein Teil der Anforderungen, die KI an den Wandel von Arbeit stellt. Diese Systeme werden Mensch-KI-Systeme genannt.
AI-infused systems	«AI-infused systems» sind benutzerseitige Anwendungen, welche verschiedene KI-Funktionen wie Spracherkennung, Übersetzung, Objekterkennung und Gesichtserkennung integrieren (Amershi et al., 2019).
Mensch-Maschine-Interaktion (MMI)	Die Mensch-Maschine-Interaktion, im Englischen «human-machine interaction» (HMI), bezieht sich auf die Interaktion zwischen Mensch und Maschine. Hierbei ist oftmals die Maschine ein Computer, daher gibt es enge Beziehungen und Überschneidungen mit der Disziplin Mensch-Computer-Interaktion (MCI), im Englischen «human-computer interaction» (HCI) (Huchler et al., 2020).
Explainable AI (XAI)	«Explainable AI» (XAI), auf Deutsch «Erklärbare KI» ist ein Konzept, das darauf abzielt, Nutzenden die Funktionsweise von KI-Systemen, die möglichen Fehler sowie die Gründe für Entscheidungen verständlich zu machen. Es geht darum, die Wirksamkeit der von KI-Systemen gelieferten Erklärungen zu messen und die Leistung von Mensch-Maschine zu evaluieren (Hoffman et al., 2018).

Impressum

Herausgeberin
Hochschule für Angewandte Psychologie FHNW

Redaktion
Doris Scholl, Prof. Dr. Toni Wäfler

Gestaltung und Satz
AnDiCoLab HGK Basel FHNW

Fotos
https://www.istockphoto.com

September 2024

Die Fachhochschule Nordwestschweiz FHNW
setzt sich aus folgenden Hochschulen zusammen:

- **Hochschule für Angewandte Psychologie FHNW**
- Hochschule für Architektur, Bau und Geomatik FHNW
- Hochschule für Gestaltung und Kunst Basel FHNW
- Hochschule für Life Sciences FHNW
- Hochschule für Musik Basel FHNW
- Pädagogische Hochschule FHNW
- Hochschule für Soziale Arbeit FHNW
- Hochschule für Technik FHNW
- Hochschule für Wirtschaft FHNW

Fachhochschule Nordwestschweiz FHNW
Hochschule für Angewandte Psychologie
Riggenbachstrasse 16
4600 Olten

www.fhnw.ch/psychologie