Wegweiser Künstliche Intelligenz: Verstehen, anwenden und zuversichtlich Zukunft gestalten

Thilo Stadelmann

Zusammenfassung Wäre Künstliche Intelligenz eine Strasse, wie sähe die Beschilderung aus? "Kein Tempolimit" / "Achtung, Gefahr" / "Mautstation voraus" / "Beginn des mehrspurigen Ausbau"? Entsprechende Narrative werden in Bezug auf KI tatsächlich erzählt, teilweise schliessen sie sich gegenseitig aus. Welchen man traut, ist entscheidend für die Zukunft, geschäftlich, gesellschaftlich und privat. Wir werfen daher einen analytischen Blick auf den Zustand dieser Strasse und lichten den Schilderwald. Dazu ergründen wir die wissenschaftliche, philosophische, geschäftliche und gesellschaftliche Ebene der Metapher "Künstliche Intelligenz", um uns deren Kern zu nähern. Wir betrachten die zugrunde liegende Technologie und werfen einen Blick auf primäre Anwendungsfelder und häufig genannte Risiken. Hieraus leiten wir konkrete Handlungsoptionen für den Umgang mit und den Einsatz von KI ab und fassen einen Ausblick auf unsere Zukunft.

1 Eine Strassenkarte für die High-Tech Landschaft

Viele Narrative sind aktuell im Umlauf über Künstliche Intelligenz (KI): Von einer Nützlichkeits-Revolution für die Wirtschaft [1] mit Marktpotential in den Billionen [2] ist die Rede; bis zum potentieller Killer der menschlichen Art durch ein plötzliches Verselbständigen hypothetischer "Artificial General Intelligence" (AGI) [3] und anderen dystopischen Szenarien [4] ist alles dabei [5]. Wie zu viele Schilder in einer Autobahn-Baustelle kann das verwirren. Welchen schenkt man Beachtung?

Dieser Beitrag begreift KI als Zweierlei: *Mächtiges Werkzeug* einerseits, das man verstehen, wissenschaftlich greifen und vorteilhaft nutzen kann; und *plakative Worthülse* andererseits, mit der sich Hypes anfeuern und je nach Weltanschauung Ängste

Prof. Dr. Thilo Stadelmann
ZHAW Centre for Artificial Intelligence, Technikumstrasse 71, CH-8400 Winterthur, Switzerland
AlpineAI AG, Obere Strasse 22b, CH-7270, Davos, Switzerland
E-mail: stdm@zhaw.ch

schüren oder Utopien kreieren lassen. Beide Bedeutungs-Varianten müssen betrachtet werden, da sie Einfluss auf unsere wirtschaftliche, private und gesellschaftliche Lebensrealität haben. Wie kam es dazu?

Die Wurzeln des wissenschaftlichen Fachgebiets reichen zurück in die 1950er Jahre. Die Wahl des Namens für die neue Disziplin fiel dabei aus monetären Erwägungen auf "Artificial Intelligence": Man wollte viele Fördergelder einwerben mit den neuen "komplexen Computeranwendungen" [6]. Dies sorgt bis heute für extremere Emotionen—die sich in extremerem Hype und Enttäuschung niederschlagen—als in der Wissenschaft sonst üblich, da Intelligenz und alles damit Einhergehende dem Menschen sehr nahe geht. So kam die KI erst Mitte der 2010er Jahre wieder aus einem Winterschlaf heraus, in den sie durch öffentliche und fachliche Strafung durch Nichtbeachtung geschickt worden war: Noch 2014 etwa bekannten sich laut einer Studie in der Schweiz nur wenige Forschungsgruppen zu diesem Thema [7] (heute sind es hunderte); die Einführung eines KI-Curriculums [8] in einem Informatik-Studiengang rief im gleichen Jahr noch grössere Bedenken hervor ("Alter Zopf, braucht man das?"). Gleichzeitig war das maschinelle Lernen bereits dabei, durch Nützlichkeit Einzug in die Unternehmen zu finden [9]. Dieser Nutzen spiegelt die kontinuierlich fortschreitende Entwicklung in einem Fachgebiet wieder, das inhärent anwendungsorientiert agiert und so nützliche Werkzeuge produziert [10]. Die Vorbehalte, Hypes und Winterschläfe andererseits sind Ausdruck der extremen Erwartungen, welche durch die Antropomorphisierung der Technologie von Mensch und Gesellschaft in sie hineinprojiziert werden.

Im Folgenden gehen wir den Grundlagen beider Bedeutungen von "KI" nach (Kapitel 2) und erklären die Funktionsweise. Dann spannen wir den Bogen heutiger Anwendungsmöglichkeiten exemplarisch auf und beleuchten Risiken (Kapitel 3). So gewappnet mit solidem Grundverständnis zu Möglichkeiten und Begrenzungen der Technologie werfen wir schliesslich einen Blick in die Zukunft (Kapitel 4)—die wir als Menschheit selber gestalten.

2 Hintergrund: Was ist Künstliche Intelligenz?

2.1 Wissenschaftliche Grundlage

KI ist das wissenschaftliche Fachgebiet, welches sich mit der *Erzeugung intelligent wirkenden Verhaltens mittels des Computers* beschäftigt. Als solches gehört es zur Disziplin der Informatik—der "Computerwissenschaft". Während Vorbild für die Qualität der Ergebnisse oft der Mensch ist, geht es bei KI nicht um eine Simulation der dahinterstehenden biologischen Prozesse, noch greift KI methodisch auf eine einheitliche Theorie von Intelligenz zurück. Vielmehr handelt es sich um eine methodische *Werkzeugkiste* unterschiedlichster Verfahren, welche die Vorteile moderner Computer ausnutzen (grosser Speicher für Daten, auf denen schnelle, einfache Rechenoperationen ausgeführt werden) [11]. Mit diesen Methoden lassen sich

spezifische, mehr oder weniger isolierte Aspekte intelligenten Verhaltens simulieren. Das definitive Lehrbuch der KI stammt von Russell & Norvig [12].

Historisch bietet die "Werkzeugkiste KI" zwei grosse Fächer. Im ersten finden sich oft als *wissensbasiert* bezeichnete Methoden, welche letztendlich darauf abzielen, eine Faktenbasis mittels unbestechlicher Logik zu bearbeiten und damit zu neuen Aussagen über die Welt zu kommen. Der Versuch, alles intelligente Verhalten auf Logik zurückzuführen und so ein Weltmodell aufzubauen, kann als gescheitert betrachtet werden [13]. Trotzdem finden die zugrunde liegenden Methoden täglich millionenfach Anwendung: Schnelle Suche über alle Kombinationen von ersten Schritten über nächste Schritte usw. hinweg nach einer Schrittfolge, die zum gewünschten Ziel führt, verhalf 1997 nicht nur IBMs KI-System "Deep Blue" zum Sieg gegen den amtierenden Schachweltmeister Garri Kasparov. Es ermöglicht heute auch die Wegfindung in unseren Navigationssystemen.

Es ist jedoch das zweite grosse Fach in der Werkzeugkiste KI, dass für den aktuellen Boom rund um intelligente Systeme verantwortlich ist. Vermutlich jedes der aktuell öffentlich breit diskutierten, verblüffenden KI-Systeme, inklusive der generativen KI, basiert auf dessen Methoden des *maschinellen Lernens*. Hierbei handelt es sich um Verfahren, um Verhalten zu erzeugen, dass wir nicht zufriedenstellend in Regeln (oder Programmcode) ausdrücken könnten, aber mittels Beispielen exemplarisch beschreiben können. Nehmen wir die Herausforderung als Beispiel, auf Fotos Katzen von Hunden zu unterscheiden: Es ist unklar, welche Menge von Regeln diese Trennung eindeutig und korrekt beschreiben würde. Für jede Regel etwa bezüglich Fell oder Kopfform liessen sich Ausnahmen finden. Hingegen ist es einfach, einen *Datensatz* von Hunde- und Katzenbildern zusammenstellen, aus denen ein Mensch einfach den Zusammenhang (oder *Zielfunktion*, $f(x) = y$) erlernen würde. Hierbei symbolisierte x die Eingabe, also ein Bild, y entspräche der Ausgabe, etwa 1 für Hund und -1 für Katze, und $f()$ stände für die Funktion, welche Bilder x auf sogenannte *Labels* y abbildet.

Im Machine Learning (ML) übernimmt der Computer dieses *Training*: Er bekommt die Bilder des Datensatzes als $\mathbf{x}$ in einer geeigneten Kodierung (z.B. als Vektor, der alle Bildpunkte als Werte hintereinander aufreiht, Bildzeile für Bildzeile) und eine geeignete anpassbare Funktion $f()$ (beispielsweise eine, welche durch Wahl der *Parameter* viele verschiedene kurvige Flächen darstellen kann). Nun passt er ausgehend von einer initialen (z.B. zufälligen) Konfiguration die Parameter von $f()$ diese sukzessive so an, dass der *Fehler* zwischen *vorhergesagtem* y' und echtem y für ein gegebenes $\mathbf{x}$, und das für alle $(\mathbf{x}, y)$-Paare im Datensatz, *minimiert* wird.

Stellen wir uns das der Einfachheit halber mit Bildern repräsentiert nur mittels zweier Pixel $\mathbf{x} = (x_1, x_2)$ vor—damit sind sie lediglich Punkte in einem zweidimensionalen Koordinatensystem, das wir uns noch leicht visualisieren können (siehe Abbildung 1). Wir können annehmen, dass Hundebilder untereinander eine gewisse Ähnlichkeit aufweisen und daher einen Cluster im Koordinatensystem bilden, etwas entfernt von allen Katzenbildern. Deshalb könnten wir die beiden Cluster mit einer Gerade—einer geraden Linie—zu trennen versuchen. Die Zielfunktion $f()$ wäre dann, dass ein Hundebild oberhalb der durch $(\theta_0 + \theta_1 x_0 + \theta_2 x_1 = 0)$ definierten Gerade zu liegen kommt und ein Katzenbild darunter, also

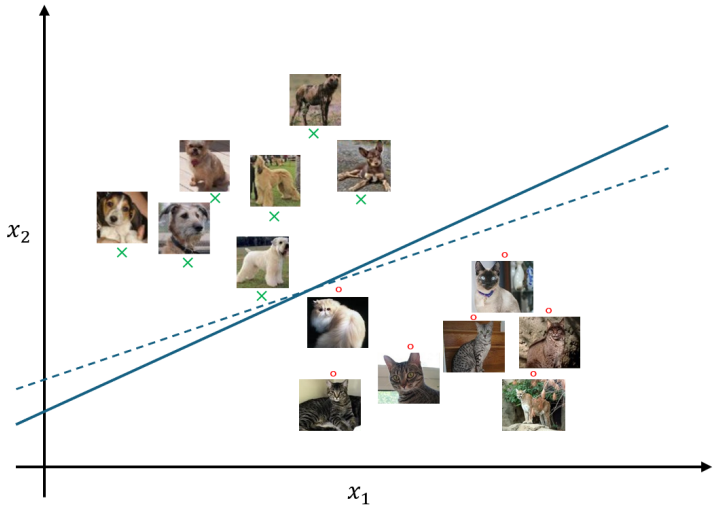

Abb. 1 Bilder von Hunden und Katzen (Quelle der Einzelbilder: ImageNet [14]), repräsentiert durch lediglich 2 Koordinaten (rote ∘, grüne ×). Diese werden anhand der Gerade $\theta_0 + \theta_1 x_0 + \theta_2 x_1 = 0$ perfekt in die beiden Kategorien getrennt (blaue, durchgezogene Linie): Alle Hundebilder finden sich oberhalb der Gerade ($\theta_0 + \theta_1 x_0 + \theta_2 x_1 > 0$), die Katzen kommen darunter zu liegen ($\theta_0 + \theta_1 x_0 + \theta_2 x_1 < 0$). Alternative Trenngeraden sind ebenfalls denkbar (etwa die gestrichelte, blaue Gerade). ML-Verfahren finden für gegebene Kodierung $\mathbf{x}$ der Datenpunkte und vorgegebene Funktionsklasse $f()$ (etwa Geradengleichungen) die optimalen Parameter $\boldsymbol{\theta}$.

$$f(\mathbf{x}) = \begin{cases} 1 & \text{wenn } \theta_0 + \theta_1 x_0 + \theta_2 x_1 \geq 0 \\ -1 & \text{wenn } \theta_0 + \theta_1 x_0 + \theta_2 x_1 < 0 \end{cases}$$

Hierbei sind die drei Parameter θ_1–θ_3 intuitiver bekannt als Steigung ($m = \frac{\theta_1}{\theta_2}$) und Kreuzungspunkt mit der vertikalen Achse ($b = \frac{\theta_0}{\theta_2}$), und die Geradengleichung selbst als $x_2 = mx_1 + b$. Durch Variieren der (zwei oder drei) Parameter allein lassen sich alle möglichen Geraden im Zweidimensionalen realisieren. Ziel des maschinellen Lernens ist es also, eine Konfiguration von m und b (bzw. θ_1–θ_3) zu finden (also die Gerade durch Drehen und Schieben so im Raum zu platzieren), dass möglichst alle $\mathbf{x}$, die Hunde repräsentieren ($y = 1$), über der Geraden zu liegen kommen, während alle Katzen-$\mathbf{x}$ ($y = -1$) unterhalb landen. Für ein neues Bild $\mathbf{x}'$ gibt das so trainiertes *Modell* $f()$ (mit anhand des Trainingsdatensatzes gefundenen idealen Parametern) nun direkt aus, ob es sich um Katze ($f(\mathbf{x}') < 0$) oder Hund ($f(\mathbf{x}') \geq 0$) handelt.

Das Beispiel oben zeigt, wie ein bestimmter Anwendungsfall (Katzen von Hunden unterscheiden) zuerst für den Computer formalisiert werden muss (geeignete Kodierung von Eingabe x und Ausgabe y), ein Trainingsdatensatz aus (x, y)-Paaren zusammengestellt und eine geeignete Funktionsart $f()$ gewählt werden muss. Der Computer übernimmt nun die systematische Anpassung der Parameter von $f()$, so dass die Abweichung zwischen errechneter Ausgabe und echtem Label über alle Beispiele minimiert wird. War das Trainingsset gross genug und aussagekräftig auch für künftig erwartete Eingaben, besteht guter Grund zu der Annahme, dass das gelernte

Modell *generalisiert*. In realen Anwendungsfällen, in denen die Eingaben hochdimensional (etwa Bilder, Videos, Text, Sensoraufzeichnungen) und der Zusammenhang mit der gewünschten Ausgabe komplex ist (nichtlinear, wie etwa die Abbildung von visuellem Aussehen auf biologische Art), wird man eine anpassungsfähigere Zielfunktion als eine 2D-Gerade wählen—etwa ein sogenanntes neuronales Netz, womit man zu der Spielart *Deep Learning* des ML übergeht. Dessen bis zu Milliarden von Parametern sorgen letztlich für eine beliebige "Kurvigkeit" der Trennlinie von oben und damit für hohe Anpassungsfähigkeit an die Daten und den zu modellierenden Zusammenhang. Die Lernprinzipien aber bleiben die gleichen. Es empfiehlt sich, für eine kurze Einführung für Nichttechniker Kapitel 2 aus Segessenmann et al.'s Artikel [15] zu lesen; das definitive Lehrbuch legte kürzlich Prince vor [16].

Was für Eigenschaften hat ein solches ML-Modell? Es wurde ohne viel explizites Vorwissen direkt aus den Daten gelernt ("Daten-zentriert" [17, 18]). Das heisst, ‚was nicht in den Daten war, steckt nicht im Modell. Es besitzt weiterhin einen rein statistischen Blick auf die Welt, denn die oben beschriebene Art der Optimierung der Zielfunktion, "Maximum Likelihood", schätzt in der Tat implizit ab, wie sehr die Wahrscheinlichkeitsverteilung der Vorhersagen zur beobachteten Verteilung in den Trainingsdaten passt. Dieser statistische Blick funktioniert für sehr viele Anwendungsfälle im Schnitt über viele Vorhersagen (teilweise bedeutend) besser, als was ein Mensch manuell erreichen würde. Dies gilt für obigen Hunde-Katzen-Klassifikator genauso wie für Large Language Models (LLMs) wie ChatGPT, welche aus einem Kontext von Worten das wahrscheinlich nächste Wort vorhersagen. Allerdings kann das Ergebnis im Einzelfall völlig danebenliegen, da Aussagen über statistische Plausibilität—um beim Beispiel Sprache zu bleiben—keine Aussagen über Wahrheit sind. Dies ist bedenken, wenn einen die nächste beeindruckende KI-Demonstration in Staunen versetzt.

2.2 Philosophisch-narratives Umfeld

Seit es Technik und insbesondere Science Fiction Literatur gibt, fantasieren (im besten Sinne des Wortes) Menschen über "künstliche Intelligenz"—nicht als real existierende Technologie oder Wissenschaft, sondern als Container für futuristische Vorstellungen. Dass zwei Begriffe wie KI und "künstliche Intelligenz im Science-Fiction-Sinn Hollywoods" nichts gemein haben ausser die gleichen 22 Zeichen in identischer Konfiguration, ist normalerweise weder speziell noch beunruhigend. Der auch als "Black Box" bezeichneten Flugschreiber etwa wird gleich bezeichnet wie gänzlich unverwandte Phänomene, die man aufgrund ihrer Komplexität für gänzlich unverständlich hält ("Black Boxes"), doch dies führt nicht zu schwerwiegenden Missverständnissen. KI ist aber—wirtschaftlich und damit gesellschaftlich betrachtet—im Moment ein Spiel mit extrem hohen Einsätzen, so dass eine Vermischung der beiden Bedeutungsebenen in öffentlichkeitswirksame Statements problematisch wird: Die Projektion von Ereignissen aus der Science Fiction auf damit nicht verwandte heutige Technologien kann Angst und Unsicherheit hervorrufen.

Als Beispiel für ein solches Statement kann der offene Brief des Future of Life Institutes für ein KI-Moratorium dienen [19]: Allein der Hintergrund der zahlreichen Unterzeichnenden Forscher und Unternehmer verleiht dem Anliegen technisch-wissenschaftliche Glaubwürdigkeit. Die genannten Risiken und Lösungsvorschläge jedoch basieren auf Extrapolationen, die philosophisch-weltanschaulich begründet sind und von einflussreichen Teilen der Forschungsgemeinschaft als "hypothetisch" und "imaginär" bezeichnet werden [20].

Dass solche Narrative rund um sich selbst verbessernde, allgemeine oder superintelligente KI existieren, nahm seinen Ursprung spätestens mit Laplace' Gedankenexperiment, dass sich mittels vollständiger Daten und perfekter Modelle alle wahren Aussagen über die Welt ableiten liessen [21]. Interessanter Weise führte Laplace seinen "Dämon" lediglich ein, um die damit verknüpfte Vorstellung der vollständigen Vorhersagbarkeit ad absurdum zu führen. Ungeachtet dessen lebt diese Machbarkeitsfantasie bis heute fort und findet neue philosophische Formen, die Émile Torres und Timnit Gebru unter dem Akronym *TESTCREAL* zusammenfassen [22]: Materialismus führt so zum Glauben an die Singularität (dem "S" in TESCREAL) und damit der Möglichkeit, dass KI die ferne (Longtermismus, "L") aber glorreiche Zukunft der Menschheit im All (Extropianismus "E" und Cosmismus "C") als transhumane Wesen ("T") ruinieren könnte, da sich der Sprung zur Realisation quasi-menschlicher Eigenschaften in der Maschine jederzeit vollziehen könne und die Technologie unkontrollierbar machen könnten. Entsprechend legen Rationalismus und effektiver Altruismus ("R" und "EA") drastische Massnahmen im Hier und Jetzt nahe. Von diesen spricht Torres in letzter Konsequenz als der "Eugenik des 21. Jahrhunderts"—was auch nicht-Philosophen klar macht, dass die Science-Fiction-geschürte KI-Angst weltanschaulich eigentlich keine breite Basis haben sollte. Wie weit verbreitet sie in den für einen Grossteil heutiger KI-Fortschritte verantwortlichen Unternehmen trotzdem ist, zeigt exemplarisch (und drastisch) der Blog Leopold Aschenbrenners [23]: Eine ganze Community huldigt hier einer fragwürdigen Idee von Zukunft und beeinflusst auf dieser Glaubensbasis Politikgestaltung weltweit, obwohl sie von aussen betrachtet auch als "lächerlich" bezeichnet wird [24, 25].

Was ist also künstliche Intelligenz? Eine Sammlung von Methoden einerseits, mit denen sich heute (und in Zukunft besser) intelligentes Verhalten mit dem Computer nachahmen lässt; ein häufig mit weltanschaulich geprägten und an die Science Fiction angelehnten Inhalten gefüllter Begriff andererseits, mit dem sich viel Aufmerksamkeit und Angst erzeugen lässt. Beide Bedeutungsebenen sind im öffentlichen Diskurs wie auch in Produktversprechen teilweise verflochten.

3 Anwendung: Was kann Künstliche Intelligenz?

3.1 Geschäftlicher Nutzen

KI ist in aller Munde nicht wegen einer technologischen Revolution (wenn, dann wäre sie aktuell 8 Jahre alt: [26]), sondern einer Nützlichkeitsrevolution [1]: Für

generative KI, insbesondere Text- und Bilderzeugung, sind Einsatzmöglichkeiten im professionellen und privaten Kontext vielen Menschen schon nach kurzem Ausprobieren offensichtlich. Fünf Minuten reichen, um einen ersten Eindruck vom Potential zu erhalten, und die Einstiegshürden sind durch einfachen Onlinezugriff und Freemium-Services denkbar niedrig. Zu beachten ist freilich (und im professionellen Einsatz unbedingt erforderlich), auch im Ausprobieren nur datenschutzkonform zu operieren, um nicht etwa durch den Prompt-Verlauf Interna öffentlich zu machen. In erster Näherung sollten professionell nur die Bezahlversionen entsprechender Services genutzt werden, in denen die Geschäftsbedingungen meist ausschliessen, dass die Prompts vom Anbieter beliebig weiterverwendet werden dürfen. Mehr Kontrolle bieten, wenn es um unternehmensweiten Einsatz geht, spezialisierte Unternehmen wie ApineAI[1]. Alternativ lassen sich eigene lokale Instanzen betreiben [27].

Wenn die Idee fehlt, wo generative KI geschäftlich den grössten Nutzen bringen könnte, bieten die 100 Use Cases im Leitfaden des Harvard Business Review einen guten Einstieg [28]. Generell ist zu empfehlen, jeden Einsatz von KI (wie jede anderen Technologie auch) stur an echten Business Cases auszurichten: Ausgehend von den lukrativen, aber ungelösten Problemen in jedem Unternehmensbereich lässt sich am besten planen, wo ein Einsatz prinzipiell lohnenswert wäre. Durch Ausprobieren oder Austausch mit Spezialisten lässt sich dann im zweiten Schritt effizient die Machbarkeit klären[2] [29].

Nicht zu vernachlässigen ist bei allem Nutzen der generativen KI, dass die methodische Werkzeugkiste KI noch Weiteres zu bieten hat: Etwa als Teil einer Data Science Pipeline zum Auswerten von Datenpunkten und -Strömen [30] und somit im Verbund mit Datenbanken und Digitalisierungsprojekten [31]. Hier spielen klassische Verfahren des maschinellen Lernens genauso eine Rolle [29] wie Deep Learning, je nach Datenlage und Problemstellung [32]. Um die Möglichkeiten exemplarisch auszuloten, mag folgende Aufstellung von Anwendungen dienen, welche Teams des ZHAW Centre for AI in den letzten Jahren mit Unternehmen gemeinsam realisiert haben. Im Bereich der Dokumentenverarbeitung etwa die Segmentierung von Zeitungsseiten in einzelne Artikel [33], das Scannen von Musikalien in maschinenlesbare Form [34], oder die barrierefreie Erschliessung technischer Dokumentation [35, 36]. In der medizinischen Bildverarbeitung und dem Gesundheitssektor zum Beispiel die Verbesserung der bildgestützten Diagnose durch Datenpooling über Krankenhausgrenzen hinweg [37], die Beschleunigung von Krebsdiagnosen durch höhere Automatisierung [38], die Reduktion von Artefakten in CT-Bildern zur Strahlungsreduktion [39], oder die Überwachung von Intensivpatienten zur Vermeidung von Fehlalarmen in der Pflege [40]. Im industriellen Bereich Arbeiten zu Produktionsplanung für optimiertes Komplexitätsmanagement [41], Produktionsparameterschätzung für bessere Leistung von Photovoltaikmodulen [42], automatischer Qualitätskontrolle von Rotationsmaschinen [9], sowie Prozessüberwachung von Plastikspritzgussprozessen [43]. Weitere industrielle Anwendungen für Bild- und Zeitreihenanalyse werden besprochen in [44, 45].

---

[1] `https://alpineai.swiss/`

[2] `https://data-innovation.org/innovation/` verweist auf einen exemplarischen Ablauf.

Alle genannten KI-Systeme wurden in direkter Zusammenarbeit mit den Praxispartnern für den alltäglichen Einsatz innerhalb von ein bis zwei Jahren entwickelt, da es keine fertigen Methoden oder Produkte am Markt gab, um den zugrunde liegenden Businesscase zu ermöglichen. Insbesondere liesse sich keine dieser Anwendung durch ChatGPT und Konsorten ersetzen, obwohl einige der Systeme bereits seit 2017 im Einsatz sind. Die Eigenentwicklung in Partnerschaft zwischen KMUs und KI-erfahrenem Forschungspartner war wirtschaftlich nicht nur möglich, sondern auch sinnvoll. Falls hingegen für die angestrebte Lösung keine Forschungslücke zu füllen ist, sind Ergebnisse sogar innerhalb weniger Wochen bis Monate (teilweise sogar nur Tage) und ROIs im dreistelligen Bereich möglich [46]. Dies sind gute Nachrichten für jede Organisation, die erfolgsrelevante Fragestellungen hat, für welche *bessere Vorhersagen* im weitesten Sinn hilfreich sein könnten (KI ist "Vorhersagetechnologie": Das nächste Wort gegeben der Kontext; Produktgüte gegeben die Messung; Dokumenteninhalt gegeben den Scan; Objektart gegeben ein Foto; etc.). Ein Grundverständnis der Technologie hilft dabei, die erste Sichtung möglicher Anwendungsfälle nach Machbarkeit selst vorzunehmen, und für eine Auswahl mit guter Prognose in Detailabklärungen mit Partnern zu gehen.

3.2 Gesellschaftliche Herausforderung

Den oben skizzierten Chancen durch die Technologie stehen Risiken gegenüber. Viele hiervon sind ernstzunehmen und gleichzeitig gut handhabbar, insbesondere solche, welche auf Fehlleistungen mit *individuellen Auswirkungen* aktueller KI Systeme beruhen, basierend auf technischen Unzulänglichkeiten der eingesetzten Methodik. Hierzu zählt etwa das Problem des Bias von KI Systemen. Damit ist gemeint, dass Subgruppen der Gesellschaft durch den Einsatz eines KI-Systems benachteiligt werden, etwa weil ihre Gesichter schlechter von Gesichterkennungsmethoden erkannt werden (dies ist nachteilig, wenn es zum Beispiel zu einer schlechteren Behandlung durch die Polizei führt). Oder das Problem von Adversarial Attacks auf Bildklassifikatoren (das sich also durch die absichtliche unsichtbare Veränderung der Bilder das Ergebnis des Systems fast beliebig manipulieren lässt). Beide Probleme sind gut verstanden [47, 48] und es liegt absolut in der Macht jeder solche Systeme einsetzenden Organisation, deren negative Effekte zu verhindern. Der dazugehörige Aufwand muss jedoch verantwortungsvoll getätigt werden [49].

Ähnlich verhält es sich mit Risiken, die sich aus dem Werkzeug-Charakter von KI-Systemen generell ergeben und über den Einzelfall hinaus breiter ausstrahlen: Als Werkzeug kann KI für unterschiedliche Zwecke eingesetzt werden, gute (wie den Kampf gegen Krebs, s.o.) und schlechte. Zu letzteren gehört das Erzeugen von Fake Inhalten (z.B. Fake News, Deep Fakes). Bei der Betrachtung dieses Risikos hilft es, sich zu vergegenwärtigen—und das hat Methodencharakter auch für ähnlich gelagerte Fragen—, was daran eigentlich neu ist und was bestehen bleibt: Menschen fälschen Inhalte, seit es Originale gibt. Was sich ändert, ist die Grössenordnung. Das Erstellen beliebig professionell wirkender Inhalte ist nun einfach und kostengüns-

tig machbar für jedermann. Es wird befürchtet, das eine Flut von Fake Inhalten ein schweres Problem für demokratische Prozesse und gesellschaftlichen Zusammenhalt darstellen können. Aber Menschen sind schnell lernfähig und sehr intolerant gegenüber dem Gefühl, betrogen zu werden. Es ist denkbar, dass sie sehr schnell lernen, beliebigen anonymen Internetquellen nicht mehr in gewohnter Weise zu vertrauen, und stattdessen realen, vertrauten Beziehungen wieder zu mehr Stellenwert verhelfen. Flankierend wird KI natürlich auch eingesetzt, um Fake Inhalte zu erkennen und zu filtern, was einen Teil der Flut entschärft. Dies bedeutet nicht, dass die genannten Herausforderungen kein Problem darstellen würden. Aber es setzt sie ins Verhältnis zu anderen gesellschaftlichen Herausforderungen, welche zwar ernstzunehmen, aber eben auch handhabbar sind.

Sorge macht Vielen potentiell negative Auswirkungen von KI auf den Arbeitsmarkt. Stellenweise prognostizierte massive Arbeitslosigkeit wird von Experten jedoch vielfach angezweifelt: KI als Werkzeug automatisiert keine Jobs, sondern einzelne Tätigkeiten. Vor Jahren sagte man beispielsweise dem Beruf des Radiologen den Niedergang voraus, nachdem Computer Vision Systeme hervorragende Ergebnisse in der Röntgendiagnostik erzielt hatten. Heute gibt es mancherorts sogar Radiologenknappheit, da die Bildanalyse nur einen (kleinen) Teil des Berufs eines Radiologen ausmacht [50]. Die Art, wie KI erfolgreich eingesetzt wird, ist im Team mit dem Menschen, als kognitive Orthese. Sie erweitert menschliche Fähigkeiten, aber wie eine Brille in der Tasche ist sie auf sich allein gestellt recht wertlos [51].

Die Metapher von KI als Werkzeug hilft auch beim Einordnen der spezifischen Herausforderungen in der Bildung. Zunächst sind bessere Werkzeuge eine positive Entwicklung. Mit ihnen kommen wir weiter im Leben. Grundlegende Fertigkeiten jedoch, die uns erst zum erfolgreichen Werkzeugeinsatz befähigen, erlernen wir *ohne* sie. Beispielsweise müssen wir rechnen lernen, um ein solides Gespühr für Quantitäten aufzubauen, Taschenrechner hin oder her; und schreiben, da es das Denken schult. Der jetzige Kanon an Lehrstoff bleibt also weiterhin relevant. Ganz besonders Sprachfähigkeit wird immer wichtiger: Sie bildet nun sogar das Benutzerinterface zum Computer. Hinzukommt die Notwendigkeit eines grundsätzlichen Technikverständnisses für Alle, um die immer neu vermeldeten Durchbrüche nicht als magisch anzusehen, sondern Möglichkeiten und Begrenzungen realistisch einzuschätzen und für sich nutzbar zu machen. Da Fake von Handarbeit zu unterscheiden zunehmend schwierig wird, werden Leistungskontrollen besser mündlich vollzogen.

Die Frage stellt sich allerdings: Wie motivieren sich Lernende zukünftig zum Investieren der notwendigen Zeit und Kraft, wenn die KI-Lösung (machbar, aber dem Lernziel nicht dienlich) nur wenige Tastenanschläge entfernt ist? Lernen ist mit Schmerz verbunden, und einen guten Teil des *Guten Lebens* erreichen wir nur durch freiwilliges Annehmen bestimmten Schmerzes. Hier scheint die grösste Herausforderung im Zusammenhang mit KI für die Gesellschaft zu liegen: Wenn KI im Idealfall enorme neue Bequemlichkeiten und Erleichterungen bietet, wir das Leben, dass wir uns wünschen, aber nur erreichen, wenn wir selber aktiv bleiben (beim Lernen, in Beziehungen, durch sinnstiftende Tätigkeit)—wie raffen wir uns dazu auf? Wie bleiben wir menschlich (Neil Lawrence sieht als Kern des Menschen auch seine Begrenztheit, die als Stärke zu umarmen ist [52]), wenn Superkräfte locken

[53], die gezielt eingesetzt nützlich und überstrapaziert entmenschlichend sind? Den Metaphern zu KI liesse sich zugespitzt diejenige von ihrer Dualität als Medikament und Droge zur Seite stellen.

4 Zukunft: Wie wollen wir leben?

KI wird sich in den kommenden Jahren weiter entwickeln, vermutlich nicht langsamer als bisher. Gleichzeitig wird sich allein durch die Durchdringung der Wirtschaft durch entsprechende Methoden und die damit einhergehenden Verschiebungen von Geld unsere Gesellschaft weiter wandeln. Das ist sogar unabhängig von jedem grösseren Fortschritt der Technologie, denn diese ist in ihrer aktuellen Nützlichkeit noch nicht ausgereizt. Hätten wir eine Wahl, was sollte das Ziel dieses Wandels sein?

4.1 Individuelle Handlungsempfehlungen

Auf der Ebene des Individuums und der Organisation macht es Sinn, sich mit KI auseinanderzusetzen und entsprechende Systeme, wo sinnvoll und zweckmässig, in Alltag und Arbeit zu integrieren. *Probieren Sie es aus*! Da die Entwicklung schnell geht, sollte dies periodisch geschehen: Was heute noch nicht gut genug ist, kann in einem Quartal anders aussehen.

Auf Sicht, was in der KI höchstens fünf Jahren entspricht, kann man erwarten, dass KI-Systeme die Nützlichkeit fähiger persönlicher Assistenten erreichen werden. Dies bedingt keine AGI oder Superintelligenz, sondern ist eine plausible Extrapolation. Sie geht aus von den Möglichkeiten heutiger LLMs, erweitert um die Fähigkeit, dass sich die Systeme eine Weile lang selber prompten, bevor sie mit einem Feedback an den Mensch zurückgelangen (sogenannte “Agentic Workflows” [54]). Kombiniert mit Erkenntnissen aus den Neurowissenschaften ist zudem denkbar, dass sich der exorbitante Ressourcenhunger an Daten und Strom, den das aktuelle Entwicklungsparadigma der KI bedingt (nämlich, ML-Systeme zu “skalieren”), bessert [55, 56]. Entsprechende ML Verfahren könnten auch zu besseren “Weltmodellen” führen, die ihre Umgebung weniger rein statistisch begreifen, sondern mehr in Kategorien von Ursache und Wirkung—und so eine Art “gesunden Maschinenverstand” entwickeln, der weniger anfällig für offensichtlich unsinnige Ausgaben ist. Auch hier: ohne AGI.

Da stellt sich die Frage: Was macht dann eigentlich den Menschen aus [15, 52]? Die Gewinner der Zukunft sind sicherlich diejenigen Menschen, die sich nicht von den Fähigkeiten eines schnellen Computers einschüchtern und ihre Identität in Frage stellen lassen. Stattdessen wissen sie, was ihnen als sozialen Beziehungswesen gut tut, und um die Bedeutung sinnstiftenden Lebensinhalts. Hierzu trägt ein individuelles Stärken von und sich beschäftigen mit *humanistischen* Gedanken bei, die sich mit dem Wert und der Würde des Menschen an sich beschäftigen, anstatt ihn zu technomorphisieren [15].

4.2 Gesellschaftliche Weichenstellungen

Auf gesellschaftlicher Ebene stellt sich wie bei jeder grossen Veränderung die Frage nach Regulierung. Aktuelle Ansätze zielen häufig auf Compliance und Haftung. Der EU AI Act etwa gibt wenige konkrete Leitplanken vor, aber reguliert, was wie zu dokumentieren ist. Greifen wir die eingangs gewählte Metapher des Verkehrs nochmal auf, fällt auf, dass mit den Gesetzen rund um die aufkommende Mobilität damals auch eindeutige Signale gesetzt wurden bezüglich der Art von Welt, in der man leben wollte: Hochgeschwindigkeit nur auf Schnellstrassen, Verkehrsberuhigung in Siedlungszonen, Zulassungspflicht für Fahrzeuge, Führerscheinpflicht für Nutzer, etc. Umgekehrt wollte man offenbar nicht beliebig schnelle Vehikel, anonym und überall, oder untrainierte Fahrer. Es scheint, ähnlicher Gestaltungswille im Bereich KI würde bedingen, nicht Technologie oder Anwendungsdomäne zu regulieren, sondern spezifisch auch die damit verbundenen Geschäftsmodelle. So liessen sich vielleicht Kollateralschäden wie im Bereich der sozialen Medien vermeiden [57].

Oben wurde als grösstes KI Risiko identifiziert, dass sich ob der angebotenen "Superkräfte" viele Menschen aus dem notwendigen Schmerz des Menschseins *ausklinken* könnten. Dies würde Ihre Reife torpedieren, was sie zu Verlierern einer hochtechnisierten Welt machen und letztlich von dem Guten Leben abschneiden würde, dass sie sich wünschen. Viele Ausgeklinkte würden ausserdem zum Problem einer auf Kooperation angewiesenen Gesellschaft. Um dem entgegenzuwirken, braucht es gesellschaftliche Anstrengungen in *Charakterbildung*: Es braucht das Beste im Menschen und unser bestes Selbst, um mit den stärksten Werkzeugen verantwortungsvoll umzugehen. Dies hat Implikationen auf Lehrpläne (neben der direkt auf das berufliche ausgerichteten Kompetenzorientierung einen neuen Fokus auf Ethik, Philosophie und spirituelle Ressourcen, denen viel Wissen um den Wert des Menschen innewohnt). Vielleicht kann sogar KI, als Coach, den Menschen bei der notwendigen Selbstüberwindung unterstützen. Die Zukunft gestalten wir selbst.

Literatur

[1] Pascal Kaufmann, Thilo Stadelmann und Benjamin F Grewe. *ChatGPT läutet Tech-Revolution ein*. Finanzen und Wirtschaft, `https://www.fuw.ch/chatgpt-laeutet-tech-revolution-ein-303487897856`. 2023.

[2] Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, Kate Smaje, Alex Sukharevsky, Lareina Yee und Rodney Zemmel. *The economic potential of generative AI: The next productivity frontier*. McKinsey & Company, `https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier`. 2023.

[3] Ruth Fulterer. *Künstliche Intelligenz contra Menschheit: Diesen Kampf gibt es nicht. Er ist nur eine rhetorische Strategie*. NZZ, `https://www.nzz.ch/`

meinung/kuenstliche-intelligenz-vs-menschheit-diesen-kampf-gibt-es-nicht-er-ist-nur-eine-rhetorische-strategie-ld.1732360. 2023.

[4] Yuval Noah Harari. *Homo deus: Eine Geschichte von Morgen*. CH Beck, 2023.

[5] Kai-Fu Lee und Chen Qiufan. *AI 2041: Ten visions for our future*. Crown Currency, 2021.

[6] Thilo Stadelmann, Martin Braschler und Kurt Stockinger. „Introduction to applied data science“. In: *Applied data science: lessons learned for the data-driven business*. Springer, 2019, S. 3–16.

[7] Jean-Daniel Dessimoz, Jana Koehler und Thilo Stadelmann. „Artificial intelligence research in Switzerland“. In: *AI Magazine* 36.2 (2015), S. 102–105.

[8] Thilo Stadelmann, Julian Keuzenkamp, Helmut Grabner und Christoph Würsch. „The AI-atlas: didactics for teaching AI and machine learning on-site, online, and hybrid“. In: *Education Sciences* 11.7 (2021), S. 318.

[9] Thilo Stadelmann, Vasily Tolkachev, Beate Sick, Jan Stampfli und Oliver Dürr. „Beyond ImageNet: deep learning in industrial practice“. In: *Applied data science: lessons learned for the data-driven business* (2019), S. 205–232.

[10] Thilo Stadelmann. „KI als Chance für die angewandten Wissenschaften im Wettbewerb der Hochschulen“. In: *Bürgenstock-Konferenz der Schweizer Fachhochschulen und Pädagogischen Hochschulen, Luzern, Schweiz, 20.-21. Januar 2023*. 2023.

[11] Rich Sutton. *The Bitter Lesson*. Incomplete Ideas, online verfügbar (01.10.2024): http://www.incompleteideas.net/IncIdeas/BitterLesson.html. 2019.

[12] Stuart J Russell und Peter Norvig. *Artificial intelligence: a modern approach, 4^{th} edition*. Pearson, 2022.

[13] Stephen Wolfram. *Remembering Doug Lenat (1950–2023) and His Quest to Capture the World with Logic*. Stephen Wolfram Writings, online verfügbar (01.10.2024): https://writings.stephenwolfram.com/2023/09/remembering-doug-lenat-1950-2023-and-his-quest-to-capture-the-world-with-logic. 2023.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li und Li Fei-Fei. „Imagenet: A large-scale hierarchical image database“. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, S. 248–255.

[15] Jan Segessenmann, Thilo Stadelmann, Andrew Davison und Oliver Dürr. „Assessing deep learning: a work program for the humanities in the age of artificial intelligence“. In: *AI and Ethics* (2023), S. 1–32.

[16] Simon JD Prince. *Understanding deep learning*. MIT press, 2023.

[17] Paul-Philipp Luley, Jan M Deriu, Peng Yan, Gerrit A Schatte und Thilo Stadelmann. „From concept to implementation: the data-centric development process for AI in industry“. In: *2023 10th IEEE Swiss Conference on Data Science (SDS)*. IEEE. 2023, S. 73–76.

[18] Thilo Stadelmann, Tino Klamt und Philipp H Merkt. „Data centrism and the core of Data Science as a scientific discipline“. In: *Archives of Data Science, Series A* 8.2 (2022).

[19] Future of Life Institute. *Pause Giant AI Experiments: An Open Letter*. FLI Open Letters, online verfügbar (02.10.2024): https://futureoflife.org/open-letter/pause-giant-ai-experiments/. 2023.

[20] Timnit Gebru, Emily M. Bender, Angelina McMillan-Major und Margaret Mitchell. *Statement from the listed authors of Stochastic Parrots on the "AI pause" letter*. DAIR Institute Blog, online verfügbar (02.10.2024): https://www.dair-institute.org/blog/letter-statement-March2023/. 2023.

[21] Pierre-Simon Laplace. *Essai philosophique sur les probabilités*. Courcier, 1814.

[22] Émile P Torres. *The Acronym Behind Our Wildest AI Dreams and Nightmares*. Truthdig, online verfügbar (01.10.2024): https://www.truthdig.com/articles/the-acronym-behind-our-wildest-ai-dreams-and-nightmares. 2023.

[23] Leopold Aschenbrenner. *The Decade Ahead*. Situational Awareness, online verfügbar (01.10.2024): https://situational-awareness.ai/. 2024.

[24] Melissa Heikkilä. *Meta's AI leaders want you to know fears over AI existential risk are "ridiculous"*. MIT Technology Review, online verfügbar (03.10.2024): https://www.technologyreview.com/2023/06/20/1075075/metas-ai-leaders-want-you-to-know-fears-over-ai-existential-risk-are-ridiculous/. 2023.

[25] Andrew Ng. *A Victory for Innovation and Open Source*. The Batch Letters, online verfügbar (03.10.2024): https://www.deeplearning.ai/the-batch/a-victory-for-innovation-and-open-source/. 2024.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser und Illia Polosukhin. „Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, S. 6000–6010. ISBN: 9781510860964.

[27] Lukas Tuggener, Pascal Sager, Yassine Taoudi-Benchekroun, Benjamin F Grewe und Thilo Stadelmann. „So you want your private LLM at home?: a survey and benchmark of methods for efficient GPTs". In: *11th IEEE Swiss Conference on Data Science (SDS), Zurich, Switzerland, 30-31 May 2024*. ZHAW Zürcher Hochschule für Angewandte Wissenschaften. 2024.

[28] Marc Zao-Sanders. *How People Are Really Using GenAI*. Harvard Business Review, online verfügbar (02.10.2024): https://hbr.org/2024/03/how-people-are-really-using-genai. 2024.

[29] Thilo Stadelmann. „Wie maschinelles Lernen den Markt verändert". In: *Digitalisierung: Datenhype mit Werteverlust? Ethische Perspektiven für eine Schlüsseltechnologie*. SCM Hänssler, 2019, S. 67–79.

[30] Martin Braschler, Thilo Stadelmann und Kurt Stockinger. „Data science". In: *Applied Data Science: Lessons Learned for the Data-Driven Business*. Springer, 2019, S. 17–29.

[31] Kurt Stockinger, Martin Braschler und Thilo Stadelmann. „Lessons learned from challenging data science case studies". In: *Applied Data Science: Lessons Learned for the Data-Driven Business*. Springer, 2019, S. 447–465.

[32] Thilo Stadelmann, Mohammadreza Amirian, Ismail Arabaci, Marek Arnold, Gilbert François Duivesteijn, Ismail Elezi, Melanie Geiger, Stefan Lörwald, Benjamin Bruno Meier, Katharina Rombach u. a. „Deep learning in the wild“. In: *Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19–21, 2018, Proceedings 8*. Springer. 2018, S. 17–38.

[33] Benjamin Meier, Thilo Stadelmann, Jan Stampfli, Marek Arnold und Mark Cieliebak. „Fully convolutional neural networks for newspaper article segmentation“. In: *2017 14th IAPR International conference on document analysis and recognition (ICDAR)*. Bd. 1. IEEE. 2017, S. 414–419.

[34] Lukas Tuggener, Raphael Emberger, Adhiraj Ghosh, Pascal Sager, Yvan Putra Satyawan, Javier Montoya, Simon Goldschagg, Florian Seibold, Urs Gut, Philipp Ackermann u. a. „Real world music object recognition“. In: *Transactions of the International Society for Music Information Retrieval* 7.1 (2024), S. 1–14.

[35] Felix M Schmitt-Koopmann, Elaine M Huang, Hans-Peter Hutter, Thilo Stadelmann und Alireza Darvishy. „FormulaNet: A benchmark dataset for mathematical formula detection“. In: *IEEE Access* 10 (2022), S. 91588–91596.

[36] Felix M Schmitt-Koopmann, Elaine M Huang, Hans-Peter Hutter, Thilo Stadelmann und Alireza Darvishy. „MathNet: A Data-Centric Approach for Printed Mathematical Expression Recognition“. In: *IEEE Access* (2024).

[37] Pascal Sager, Sebastian Salzmann, Felice Burn und Thilo Stadelmann. „Unsupervised domain adaptation for vertebrae detection and identification in 3D CT volumes using a domain sanity loss“. In: *Journal of Imaging* 8.8 (2022), S. 222.

[38] Peter R Jermain, Martin Oswald, Tenzin Langdun, Santana Wright, Ashraf Khan, Thilo Stadelmann, Ahmed Abdulkadir und Anna N Yaroslavsky. „Deep learning-based cell segmentation for rapid optical cytopathology of thyroid cancer“. In: *Scientific Reports* 14.1 (2024), S. 16389.

[39] Mohammadreza Amirian, Javier A Montoya-Zegarra, Ivo Herzig, Peter Eggenberger Hotz, Lukas Lichtensteiger, Marco Morf, Alexander Züst, Pascal Paysan, Igor Peterlik, Stefan Scheib u. a. „Mitigation of motion-induced artifacts in cone beam computed tomography using deep convolutional neural networks“. In: *Medical Physics* 50.10 (2023), S. 6228–6242.

[40] Raphael Emberger, Jens Michael Boss, Daniel Baumann, Marko Seric, Shufan Huo, Lukas Tuggener, Emanuela Keller und Thilo Stadelmann. „Video object detection for privacy-preserving patient monitoring in intensive care“. In: *2023 10th IEEE Swiss Conference on Data Science (SDS)*. IEEE. 2023, S. 85–88.

[41] Lukas Hollenstein, Lukas Lichtensteiger, Thilo Stadelmann, Mohammadreza Amirian, Lukas Budde, Jürg Meierhofer, Rudolf M Füchslin und Thomas Friedli. „Unsupervised learning and simulation for complexity management in business operations“. In: *Applied data science: lessons learned for the data-driven business* (2019), S. 313–331.

[42] Mattia Battaglia, Ennio Comi, Thilo Stadelmann, Roman Hiestand, Beat Ruhstaller und Evelyne Knapp. „Deep ensemble inverse model for image-based estimation of solar cell parameters“. In: *APL Machine Learning* 1.3 (2023).

[43] Peng Yan, Ahmed Abdulkadir, Giulia Aguzzi, Gerrit Schatte, Benjamin F Grewe und Thilo Stadelmann. „Automated process monitoring in injection molding via representation learning and setpoint regression“. In: *11th IEEE Swiss Conference on Data Science (SDS), Zurich, Switzerland, 30-31 May 2024*. ZHAW Zürcher Hochschule für Angewandte Wissenschaften. 2024.

[44] Niclas Simmler, Pascal Sager, Philipp Andermatt, Ricardo Chavarriaga, Frank-Peter Schilling, Matthias Rosenthal und Thilo Stadelmann. „A survey of un-, weakly-, and semi-supervised learning methods for noisy, missing and partial labels in industrial vision applications“. In: *2021 8th Swiss Conference on Data Science (SDS)*. IEEE. 2021, S. 26–31.

[45] Peng Yan, Ahmed Abdulkadir, Paul-Philipp Luley, Matthias Rosenthal, Gerrit A Schatte, Benjamin F Grewe und Thilo Stadelmann. „A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions“. In: *IEEE Access* (2024).

[46] Dorian Selz. *Where's the ROI for AI?* LinkedIn Post, online verfügbar (03.10.2024): `https://www.linkedin.com/posts/dselz\_wheres-the-roi-for-ai-lets-get-real-activity-7247176825972387840-rLWN`. 2024.

[47] Mohammadreza Amirian, Friedhelm Schwenker und Thilo Stadelmann. „Trace and detect adversarial attacks on CNNs using feature response maps“. In: *Artificial Neural Networks in Pattern Recognition: 8th IAPR TC3 Workshop, ANNPR 2018, Siena, Italy, September 19–21, 2018, Proceedings 8*. Springer. 2018, S. 346–358.

[48] Samuel Wehrli, Corinna Hertweck, Mohammadreza Amirian, Stefan Glüge und Thilo Stadelmann. „Bias, awareness, and ignorance in deep-learning-based face recognition“. In: *AI and Ethics* 2.3 (2022), S. 509–522.

[49] Eleonora Viganò, Corinna Hertweck, Christoph Heitz und Michele Loi. „People are not coins: Morally distinct types of predictions necessitate different fairness constraints“. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, S. 2293–2301.

[50] Saurabh Jha und Eric J Topol. „Upending the model of AI adoption“. In: *The Lancet* 401.10392 (2023), S. 1920.

[51] Kenneth M Ford, Patrick J Hayes, Clark Glymour und James Allen. „Cognitive Orthoses: Toward Human-Centered AI“. In: *AI Magazine* 36.4 (2015), S. 5–8.

[52] Neil D Lawrence. *The atomic human: Understanding ourselves in the age of AI*. Allen Lane, 2024.

[53] Andy Crouch. *The Life We're Looking for: Reclaiming Relationship in a Technological World*. Convergent Books, 2022.

[54] Andrew Ng. *Agentic Design Patterns Part 1*. The Batch Letters, online verfügbar (03.10.2024): `https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/`. 2024.

[55] Christoph von der Malsburg, Thilo Stadelmann und Benjamin F Grewe. „A theory of natural intelligence“. In: *arXiv preprint arXiv:2205.00002* (2022).

[56] Pascal J Sager, Jan M Deriu, Benjamin F Grewe, Thilo Stadelmann und Christoph von der Malsburg. „The Dynamic Net Architecture: Learning Robust and Holistic Visual Representations Through Self-Organizing Networks“. In: *arXiv preprint arXiv:2407.05650* (2024).

[57] Jeff Orlowski. *Das Dilemma mit den sozialen Medien*. Dokumentarfilm, verfügbar als Netflix Original, siehe auch (03.10.2024): `https://www.thesocialdilemma.com/`. 2020.