1 Künstliche Intelligenz – mehr als eine Modeerscheinung?

Eine einheitliche Definition des Begriffs »künstliche Intelligenz« (KI) existiert bislang nicht. In der Regel versteht man darunter jedoch ein Teilgebiet der Informatik, das sich mit der Entwicklung von Systemen befasst, die Aufgaben übernehmen können, die typischerweise menschliche Intelligenz erfordern. Dazu gehören das Verstehen und Erzeugen natürlicher Sprache, das Erkennen von Mustern, das Lösen komplexer Probleme sowie das Treffen fundierter Entscheidungen.

Vereinfacht ausgedrückt geht es bei KI darum, Maschinen oder Softwarelösungen zu entwickeln, die in der Lage sind, Denk- und Lernprozesse zu simulieren. Moderne KI-Systeme können große Datenmengen verarbeiten, Informationen analysieren, eigenständig Schlüsse ziehen und durch Erfahrung – sei es durch ausgewertete Daten oder menschliches Feedback – kontinuierlich dazulernen. Der Fokus liegt dabei nicht nur auf der Automatisierung von Aufgaben, sondern zunehmend auf der Verbesserung und Optimierung von Prozessen, oft über das menschliche Leistungsvermögen hinaus.

Künstliche Intelligenz ist längst kein Zukunftsthema mehr, sondern fest in unserem Alltag verankert – oft unbemerkt. Sprachassistenten wie Siri, Alexa oder der Google Assistant reagieren auf gesprochene Befehle und passen sich mit der Zeit den Vorlieben ihrer Nutzer:innen an. Empfehlungsalgorithmen auf Plattformen wie Netflix, Spotify oder Amazon analysieren unser Verhalten und schlagen individuell zugeschnittene Inhalte oder Produkte vor. Navigationsdienste wie Google Maps oder in Fahrzeugen verbaute Navigationsgeräte nutzen KI, um auf Basis von Echtzeitdaten die schnellste Route zu berechnen oder Verkehrsstaus zu vermeiden. In sozialen Netzwerken filtern Algorithmen Inhalte vor und beeinflussen, was wir sehen. Übersetzungstools wie DeepL oder Google

Translate werden durch maschinelles Lernen immer präziser und ermöglichen eine nahezu barrierefreie Kommunikation über Sprachgrenzen hinweg.

Auch im Gesundheitswesen ist KI auf dem Vormarsch – etwa bei der Unterstützung von Diagnosen durch Bilderkennungssoftware oder der Entwicklung personalisierter Therapiepläne. In der Finanzwelt prüfen KI-Systeme Kreditwürdigkeit, entdecken betrügerische Transaktionen oder übernehmen automatisiert Aktienhandel. Im Smart Home optimiert KI das Raumklima, steuert Licht und Geräte und lernt aus dem Verhalten der Bewohner:innen. Selbst im Bereich der kreativen Produktion – etwa beim Komponieren von Musik, dem Schreiben von Texten oder dem Erstellen digitaler Kunstwerke – übernimmt KI zunehmend Aufgaben, die lange als nur von Menschen ausführbar galten.

All diese Beispiele zeigen: KI ist kein abstraktes Konzept mehr, sondern prägt – mal sichtbar, mal unsichtbar – zentrale Bereiche unseres Lebens. Diese technologische Entwicklung eröffnet enorme Potenziale, wirft zugleich aber auch grundlegende ethische, soziale und bildungspolitische Fragen auf.

1.1 Ein kurzer geschichtlicher Abriss

Die Vorstellung, dass Maschinen denken könnten, ist keine Erfindung des digitalen Zeitalters. Bereits im 18. Jahrhundert sorgte der Schachtürke, ein scheinbar selbstständig Schach spielender Automat, für Aufsehen. Obwohl sich später herausstellte, dass ein Mensch im Inneren die Züge ausführte, inspirierte diese Illusion Denker wie Charles Babbage, der im 19. Jahrhundert die Idee einer programmierbaren Rechenmaschine entwickelte. Diese frühen Überlegungen legten den Grundstein für das, was heute als künstliche Intelligenz bezeichnet wird.

Einen bedeutenden theoretischen Beitrag leistete der britische Mathematiker Alan Turing. In seinem 1950 veröffentlichten Aufsatz »Computing Machinery and Intelligence« (Bowen, 2017) stellte er die Frage, ob Maschinen denken können, und schlug vor, dies durch ein praktisches Experiment zu überprüfen. Anfänglich war dies von Turing nur als Gedankenexperiment angelegt, wurde aber im Laufe der Zeit (bis zur Gegenwart) zu einem Standardtest für künstliche Intelligenz. In diesem Test kommuniziert ein menschlicher Fragesteller schriftlich mit zwei Gesprächspartnern, wobei einer davon nicht menschlich ist. Die Versuchsperson weiß nicht, ob der Dialog mit Mensch oder Maschine geführt wird. Gelingt es der Maschine, mehrere Fragesteller:innen in einer signifikanten Anzahl von Fällen zu täuschen, gilt der Test als bestanden.

Der Philosoph John Searle beanstandete diese Art von Test in seinem Essay »The Chinese Room«, indem er ebenfalls ein Gedankenexperiment anführte (Moural, 2003): In einem isolierten Raum sitzt eine Person, die nur durch einen schmalen Schlitz Kontakt zur Außenwelt hat. Durch diese Öffnung erhält sie Zettel mit chinesischen Schriftzeichen, die Geschichten und Fragen enthalten. Die Person versteht kein Chinesisch und hat diese Zeichen nie zuvor gesehen. Im Raum befindet sich ein Handbuch in der Erstsprache der Person. Es erklärt nicht die Bedeutung der Zeichen, enthält aber detaillierte Anleitungen, wie auf die Geschichten zu antworten ist. Die Person folgt mechanisch diesen Regeln, kopiert unverständliche Zeichen und schiebt die Antworten nach draußen. Für einen chinesischsprachigen Beobachter außerhalb des Raums sehen die Antworten so aus, als kämen sie von jemandem, der Chinesisch versteht und den Inhalt der Geschichten begreift. Searle argumentiert: Obwohl perfekte Antworten produziert werden, versteht die Person im Raum weder die Sprache noch den Inhalt – sie befolgt nur Regeln. Obwohl Searle diesen Essay bereits 1980 veröffentlichte – also lange vor der Zeit der großen Sprachmodelle, wie wir sie heute kennen und nutzen –, kann dies sehr gut auf die Gegenwart umgelegt werden, denn Large Language Models (LLMs) wie beispielsweise

ChatGPT, Claude oder Gemini simulieren das Verstehen nur, sie verfügen über kein Weltwissen (▶ Kap. 2.2).

Die eigentliche Geburtsstunde der modernen KI-Forschung lässt sich auf 1956 datieren, als John McCarthy, Marvin Minsky, Nathaniel Rochester und Claude Shannon die Dartmouth Conference organisierten. Dort wurde der Begriff *artificial intelligence* geprägt und das Ziel formuliert, Maschinen zu entwickeln, die Aspekte menschlicher Intelligenz simulieren können. In den folgenden Jahren entstanden Programme wie der »Logic Theorist« von Allen Newell und Herbert A. Simon, der mathematische Theoreme beweisen konnte, und das Schachprogramm von Arthur Samuel, das durch selbstständiges Lernen seine Spielstärke verbesserte.

In den 1960er Jahren entwickelte Joseph Weizenbaum das Programm ELIZA[1], das einfache Gespräche simulieren konnte. Obwohl ELIZA keine wirkliche Sprachverarbeitung beherrschte, zeigte es, wie leicht Menschen Maschinen menschliche Eigenschaften zuschreiben. In den 1970er Jahren wurden Expertensysteme entwickelt, die spezialisiertes Wissen in bestimmten Bereichen, wie der medizinischen Diagnose, nutzten. Diese Systeme konnten jedoch nicht über ihre programmierten Regeln hinaus lernen, was ihre Anwendbarkeit einschränkte.

Die hohen Erwartungen an die KI führten in den 1970er und -80er Jahren zu den KI-Wintern: Phasen, in denen das Interesse und die Finanzierung der KI-Forschung stark zurückgingen, weil die Technologie die gesteckten Ziele nicht erreichen konnte, was zu Enttäuschungen bei Investoren führte. In den späten 1980er Jahren durchlief die KI-Forschung eine Phase der Konsolidierung und Neuausrichtung. Ein zentrales Element dieser Phase war die Wiederentdeckung und Weiterentwicklung künstlicher neuronaler Netze. Bereits in den 1940er Jahren theoretisch beschrieben, erfuhren diese Modelle in den 1980er Jahren eine Renaissance. Forscher wie John Hopfield entwickelten das Hopfield-Netz, ein Modell

1 Es gibt immer noch Webseiten, auf denen ELIZA online ausprobiert werden kann, wie z. B. hier: https://www.med-ai.com/models/eliza.html.de

für assoziatives Gedächtnis, während Geoffrey Hinton die Boltzmann-Maschine vorstellte, die komplexe Muster erkennen und simulieren konnte. Diese Entwicklungen legten den Grundstein für das spätere Deep Learning.

Parallel dazu wurden Fortschritte im Bereich des maschinellen Lernens erzielt. Statt auf starre, regelbasierte Systeme zu setzen, begannen Wissenschaftler, probabilistische Modelle und statistische Lernverfahren zu erforschen. Dies ermöglichte es Maschinen, aus Daten zu lernen und sich an neue Informationen anzupassen, was die Flexibilität und Leistungsfähigkeit von KI-Systemen erheblich steigerte.

In den frühen 1990er Jahren begannen Unternehmen, KI-Technologien in kommerziellen Anwendungen zu integrieren. Spracherkennungssysteme, einfache Expertensysteme und erste Anwendungen des maschinellen Lernens fanden ihren Weg in Produkte und Dienstleistungen. Diese praktischen Erfolge trugen dazu bei, das Vertrauen in die KI-Forschung wiederherzustellen und neue Investitionen anzuziehen.

Schließlich markierte der Sieg von IBMs Deep Blue über den Schachweltmeister Garri Kasparow im Jahr 1997 einen symbolischen Wendepunkt. Dieses Ereignis demonstrierte die Fortschritte, die in den Jahren zuvor erzielt worden waren, und leitete eine neue Ära der KI-Forschung ein, die sich zunehmend auf lernfähige Systeme und datengetriebene Ansätze konzentrierte.

Ab dem Jahr 2000 erlebte die Künstliche Intelligenz (KI) eine Phase des kontinuierlichen Fortschritts, die durch bedeutende technologische Entwicklungen und eine zunehmende Integration in den Alltag gekennzeichnet war. Ein entscheidender Wendepunkt war der Durchbruch im Bereich des Deep Learning im Jahr 2012, als das neuronale Netzwerk AlexNet den ImageNet-Wettbewerb (dieser Wettbewerb war ein weltweiter Wettbewerb, bei dem Computer lernen sollten, Objekte auf Fotos zu erkennen – zum Beispiel Tiere, Fahrzeuge oder Alltagsgegenstände) gewann und die Genauigkeit in der Bildklassifikation erheblich steigerte. Diese Leistung demons-

trierte das Potenzial tiefer neuronaler Netze und leitete eine neue Ära in der KI-Forschung ein.

In den folgenden Jahren wurden KI-Technologien zunehmend in verschiedenen Bereichen eingesetzt. 2016 besiegte AlphaGo, entwickelt von DeepMind, den weltbesten Go-Spieler Lee Sedol, was die Fähigkeit von KI-Systemen, komplexe strategische Spiele zu meistern, unter Beweis stellte. Im selben Jahr stellte Google DeepMind WaveNet vor, ein neuronales Netzwerk, das realistisch klingende Sprachsynthese ermöglichte. Diese Entwicklungen führten zu einer breiteren Akzeptanz und Anwendung von KI in Bereichen wie Spracherkennung, Bilderkennung und personalisierten Empfehlungen.

Die Einführung von Transformer-Architekturen im Jahr 2017 revolutionierte die Verarbeitung natürlicher Sprache. Modelle wie BERT (2018) verbesserten das Verständnis von Kontext in Texten erheblich. 2020 veröffentlichte OpenAI GPT-3, ein großes Sprachmodell, das in der Lage war, kohärente und kontextuell relevante Texte zu generieren. Dies markierte einen bedeutenden Fortschritt in der Fähigkeit von Maschinen, menschenähnliche Sprache zu erzeugen. Das Besondere an den Transformer-Modellen – der Begriff wurde in einem Forschungsbericht von Google geprägt (Vaswani et al., 2017) – liegt darin, dass sie Beziehungen zwischen den Daten (z. B. einzelnen Wörtern) herstellen können und damit in der Lage sind, Kontext und Bedeutung zu bestimmen. Somit ist es mit dieser Technologie möglich, längere Reihen und Wörter parallel zu verarbeiten und »sich zu merken«, was bereits produziert wurde, damit ein kohärenter Text entsteht. Der Aufmerksamkeitsmechanismus ermöglicht es dem Sprachmodell, dass bei jedem Datenelement die Wichtigkeit und Bedeutung verarbeitet, bestimmt und überwacht wird – somit können Beziehungen zwischen den Daten hergestellt werden und es erscheint, als würde das Sprachmodell beispielsweise den Unterschied zwischen den beiden Sätzen »Ich bringe Geld auf die Bank« und »Ich lese ein Buch auf der Bank« verstehen.

Die Jahre 2022 bis 2025 waren geprägt von einer rasanten Entwicklung und Verbreitung generativer KI. Die Veröffentlichung von

ChatGPT durch OpenAI am 30. November 2022 brachte KI-gestützte Textgenerierung einem breiten Publikum näher. Bereits innerhalb von fünf Tagen erreichte ChatGPT eine Million Nutzer, was es zur am schnellsten wachsenden Internetanwendung der Geschichte machte. Zwei Monate nach dem Start zählte ChatGPT über 100 Millionen monatlich aktive Nutzer und übertraf damit frühere Rekordhalter wie TikTok oder Instagram, die deutlich mehr Zeit benötigt hatten, um diese Marke zu erreichen. Bis April 2025 stieg die Zahl der wöchentlich aktiven Nutzenden des Sprachmodells von OpenAI auf ca. 800 Millionen. Ab 2023 folgten neben Sprachmodellen auch zahlreiche Bildgenerierungsmodelle wie DALL-E 2, Midjourney und Stable Diffusion, die es ermöglichten, aus Textbeschreibungen realistische Bilder zu erstellen. Diese Technologien finden Anwendung in Bereichen wie Design, Kunst und Werbung.

Im medizinischen Bereich erzielte DeepMind mit AlphaFold 2 einen Durchbruch in der Vorhersage von Proteinstrukturen, was erhebliche Auswirkungen auf die Medikamentenentwicklung hatte. Gleichzeitig wurden KI-Modelle wie Gato entwickelt, die in der Lage waren, eine Vielzahl von Aufgaben zu bewältigen, von der Steuerung von Robotern bis hin zur Textverarbeitung.

2024 kündigte OpenAI das Modell o3 an, das auf dem Abstraction and Reasoning Corpus (ARC) Benchmark eine Leistung erzielte, die über dem menschlichen Durchschnitt lag. ARC wird genutzt, um die Fähigkeit von KI-Systemen zu messen, neue Aufgaben zu lösen, die sie zuvor nicht trainiert haben. Es besteht aus Aufgaben, bei denen aus wenigen Beispielen Regeln abgeleitet werden müssen, ähnlich wie bei Intelligenztests für Menschen. Dies wurde als ein weiterer Schritt in Richtung einer allgemeinen künstlichen Intelligenz betrachtet. Parallel dazu veröffentlichte Google die Gemini-Modelle, die in verschiedenen multimodalen Aufgaben führend waren.

Die Integration von KI in den Alltag setzte sich fort, mit Anwendungen in der Bildung, dem Gesundheitswesen und der Industrie. Gleichzeitig wuchs das Bewusstsein für ethische und regulatorische Fragen im Zusammenhang mit KI. Im Jahr 2023 unterzeichnete der US-Präsident eine Executive Order zur Regulierung von KI, und die

Europäische Union verabschiedete den AI Act[2], der Regeln für den Einsatz von KI-Systemen festlegte.

2025 erfolgte einerseits die Weiterentwicklung der verschiedenen generativen Modelle, andererseits rückten KI-Agenten in den Fokus, die in der Lage sind, komplexe Aufgaben autonom zu planen und auszuführen. Diese Systeme kombinieren Sprachverarbeitung, Entscheidungsfindung und Interaktion mit der Umgebung und finden Anwendung in Bereichen wie persönlicher Assistenz, Automatisierung und Forschung.

Wohin die Reise in Bezug auf (generative) KI geht, steht im Fokus vieler Zukunftsprognosen und -szenarien. Dabei gibt es Warnungen genauso wie übertriebenen Technik-Enthusiasmus, vor allem in Bezug auf allgemeine künstliche Intelligenz (engl. *artificial general intelligence*, kurz AGI). Dies bezeichnet eine Form von KI, die nicht nur einzelne Aufgaben lösen kann, sondern über ein breites, menschenähnliches Verständnis verfügt und flexibel in vielen Bereichen denken, lernen und handeln kann. Im Gegensatz zu heutigen KI-Systemen, die auf spezielle Anwendungen beschränkt sind, wäre eine AGI in der Lage, komplexe Probleme zu analysieren, sich neues Wissen anzueignen und eigenständig Entscheidungen in ganz unterschiedlichen Situationen zu treffen, ähnlich wie ein Mensch. Die Webseite https://ai-2027.com/ zeichnet unterschiedliche Szenarien, die aufgrund von Berechnungen der bisherigen Entwicklungen und Einholen von zahlreichen Expert:innenmeinungen durchaus realistisch eintreffen könnten.

2 Der AI Act (KI-Regulierung) verfolgt einen risikobasierten Ansatz, bei dem KI-Anwendungen je nach Gefahrenpotenzial in Risikostufen eingeteilt werden: von minimalem Risiko (z. B. Spamfilter) über hohes Risiko (z. B. KI in Medizin, Bildung oder kritischer Infrastruktur) bis hin zu verbotenen Anwendungen (z. B. Social Scoring nach chinesischem Vorbild). Für Hochrisiko-KI gelten strenge Anforderungen an Transparenz, Datenqualität, Aufsicht und Dokumentation. Der AI Act soll Innovation fördern, gleichzeitig aber Grundrechte, Sicherheit und den europäischen Binnenmarkt schützen.

1.2 Wie intelligent können Maschinen sein?

Der Begriff künstliche Intelligenz klingt beeindruckend – und ist aber gleichzeitig irreführend. Wer ihn hört, hat schnell ein Bild von denkenden Maschinen, von Robotern mit Bewusstsein oder von Computern, die eigenständig Entscheidungen treffen, vielleicht sogar Gefühle empfinden, vor Augen. Doch mit der Realität der heutigen KI-Systeme hat diese Vorstellung wenig zu tun. Der Ausdruck suggeriert eine Form von menschlicher Intelligenz, die künstlich nachgebaut wurde. Tatsächlich handelt es sich bei KI aber nicht um ein bewusstes, autonom handelndes System, sondern um Algorithmen und statistische Modelle, die auf der Grundlage großer Datenmengen Muster erkennen, Prognosen treffen und bestimmte Aufgaben automatisiert ausführen können, wenn auch in beeindruckender Geschwindigkeit.

Besonders missverständlich ist der Begriff deshalb, weil er ein Verständnis von Intelligenz voraussetzt, das in den allermeisten KI-Anwendungen gar nicht gegeben ist. Die sogenannten intelligenten Systeme sind in der Regel hochspezialisiert. Eine KI, die medizinische Bilddaten auswertet, kann keine Sprache verstehen oder Musik komponieren. Eine Spracherkennungs-KI, die ein Diktat transkribiert, »weiß« nichts über die Bedeutung des Gesagten. Sie analysiert lediglich Schallwellen, gleicht diese mit bekannten Mustern ab und wählt das wahrscheinlichste Wort oder die wahrscheinlichste Satzstruktur aus. Es handelt sich dabei also nicht um Denken im menschlichen Sinne, sondern um die Simulation bestimmter Leistungen, die uns wie Intelligenz erscheinen.

Diese sprachliche Ungenauigkeit führt nicht nur zu falschen Erwartungen, sondern manchmal auch zu überzogenen Hoffnungen oder Ängsten. Während die einen glauben, dass KI bald viele menschliche Aufgaben vollständig übernehmen könne, fürchten andere einen Kontrollverlust durch vermeintlich übermächtige Maschinen. Dabei wird häufig übersehen, dass jede KI von Menschen entwickelt, trainiert und überwacht wird. Sie funktioniert nur so

gut, wie die Daten, auf denen sie basiert – und ist anfällig für Verzerrungen, Fehler und ethisch problematische Entscheidungen. Wenn etwa diskriminierende Daten verwendet werden, spiegelt die KI diese Verzerrung in ihren Entscheidungen wider. Dennoch wird die Verantwortung oft auf die KI abgeschoben, als handle es sich um ein unabhängiges Subjekt. Dabei sollte vielmehr gefragt werden: Wer hat die KI programmiert? Welche Ziele wurden verfolgt? Wer profitiert von ihrem Einsatz?

Insofern ist es wichtig, von Anfang an ein kritisches und differenziertes Verständnis des Begriffs künstliche Intelligenz zu fördern. Denn nur wer weiß, was sich tatsächlich hinter diesem Begriff verbirgt – und was nicht –, kann sinnvoll darüber diskutieren, wie KI in Bildung, Gesellschaft und Alltag eingesetzt werden soll. Häufig wird daher gefordert, statt von künstlicher Intelligenz von maschinellem Lernen zu sprechen. Doch unabhängig davon, welche Bezeichnung gewählt wird, ist ein reflektierter Umgang mit dem Begriff Voraussetzung für einen verantwortungsbewussten Umgang mit der Technologie selbst.

Lange Zeit war der Turing-Test ein Maßstab dafür, ob Maschinen Kommunikation so weit simulieren können, dass ein künstlicher Gesprächspartner nicht mehr von einem menschlichen unterscheidbar ist und somit als »intelligent« gilt. Heutige große Sprachmodelle bestehen den Turing-Test, wie von Forschenden im Frühjahr 2025 bewiesen wurde (Jones & Bergen, 2025). Doch dieser Test ist schon seit Langem umstritten, da er im Grunde nur eine spezielle Funktionalität der künstlichen Intelligenz abprüft und es eher um die »schauspielerischen Fähigkeiten« statt um Intelligenz geht (Thome, 2025).

Infolgedessen wurden immer wieder neue Tests entwickelt, um herauszufinden, wie es um die Fähigkeit von KI steht, Menschen überlegen zu sein. Im Frühjahr 2025 wurde Humanity's Last Exam (Phan et al., 2025) mit 2.500 fachübergreifenden Fragen entwickelt, die aus den unterschiedlichsten Fachbereichen von Quantenphysik über Philosophie bis zu Medizinethik reichen und daher nicht von einzelnen Menschen beantwortet werden könnten. Der Fragenka-