Ulla Coester, Norbert Pohlmann

Relevanz von Werten für den Einsatz von KI

Der zunehmende Einsatz von Künstlicher Intelligenz (KI) im Rahmen der Digitalisierung ist evident. Aufgrund der stetigen Durchdringung unserer Gesellschaft mit moderner Technologie steht die Frage im Raum, was ist von größter Relevanz, wenn es um den Einsatz von KI geht. Die Antwort lautet: Eine vertrauensvolle Kooperation. Warum diese auf Werten basieren muss, wird im Artikel dargelegt.

Die Frage, wie eine angemessene KI-Strategie im Sinne der Allgemeinheit beziehungsweise die – unter allen Aspekten sowohl wirtschaftlicher als auch gesellschaftlicher respektive ethischer Natur – adäquate Nutzung von KI-Systemen ausgelegt sein soll, wird bereits seit geraumer Zeit auf verschiedenen Ebenen diskutiert. Dieser Dialog ist unerlässlich, auch wenn mittlerweile – unter anderem aufgrund der globalen technologischen Entwicklung – keine prinzipiellen Zweifel mehr darüber bestehen, dass deren Einsatz erforderlich ist und die Digitalisierung zudem als Ziel für Deutschland im Koalitionsvertrag 2025 (vgl. hierzu [13]) definiert wurde. Die Verflechtung von Digitalisierung und KI ist dabei naheliegend: Digitalisierung „bedeutet die Verwendung von Daten und algorithmischen Systemen als Produktionsfaktor oder als Bestandteil neuer oder verbesserter Prozesse und Produkte[.] Kennzeichen sind die Virtualisierung [...] und das Teilen von Daten[.]“ [7]. Der KI werden insgesamt viele Eigenschaften und Fähigkeiten zugesprochen, um die Intention, die mit der Digitalisierung verfolgt wird, effizienter und unmittelbarer zu erreichen.

Dr. Ulla Coester

ist Wirtschaftsethikerin und Leiterin des Forschungsschwerpunktes Vertrauenswürdigkeit im Institut für Internet-Sicherheit an der Westfälischen Hochschule Gelsenkirchen

E-Mail: coester@internet-sicherheit.de

Prof. Dr. Norbert Pohlmann

Professor für Cyber-Sicherheit und Leiter des Instituts für Internet-Sicherheit – if(is) an der Westfälischen Hochschule in Gelsenkirchen sowie Vorstandsvorsitzender des Bundesverbands IT-Sicherheit – TeleTrusT und im Vorstand des Internetverbandes – eco.

E-Mail: pohlmann@internet-sicherheit.de

Trotz der festgelegten Ziele verläuft die Diskussion in der breiten Öffentlichkeit jedoch zumeist nicht unvoreingenommen, da es hierbei oftmals einer konsistenten Definition des Begriffs KI und somit entsprechend einer einheitlichen Grundlage dafür entbehrt. Denn KI umfasst eine Vielzahl von Konzepten: Hierin inkludiert sind statistische Modelle zur Erkennung von Malware-Mustern [Maschinelles Lernen], generative Systeme zur Simulation von Sprache und logischem Denken [Large Language Models (LLMs)], agentenbasierte KI-Systeme [Agentic AI] – eine autonome Softwareeinheit, die eigenständig Entscheidungen trifft sowie Handlungen ausführt, um ein Ziel zu erreichen – sowie die (künftige) ‚Allgemeine Künstliche Intelligenz‘ [Singularität], die eine noch weitere Agilität der KI impliziert. Insbesondere letztere führt zu weiteren kontroversen Debatten (vgl. hierzu u. a. [15], [2]), da dieser Zeitpunkt „das Erreichen der starken Künstlichen Intelligenz beschreibt“, was bedeutet, „dass sich KI-Systeme rasant selbstständig verbessern und signifikant intelligenter als Menschen sind“ [20].

Hinzu kommt, dass sich die Bewertung des verhältnismäßigen Einsatzes nicht nur anhand der Handlungen der KI herleiten lassen kann, sondern auch die Annäherung an Autonomie und den angemessenen Entscheidungsraum der Anwender berücksichtigen muss: Ein Sicherheitsmodell, das Schwachstellen erkennt, operiert in einem definierten präzisen Rahmen, in dem messbare Ergebnisse erzielt und eine signifikante menschliche Aufsicht gewährleistet wird. Im Gegensatz dazu lässt sich konstatieren, dass smarte Systeme wie große Sprachmodelle (LLMs) respektive agentenbasierte KI-Systeme diese Grenze verwischen, da hier ein Prozess der Generierung von Inhalten, der Interpretation von Absichten und der Handlung initiiert wird [16].

Letztlich stellt die Komplexität, die mit der Einführung von KI-Systemen einhergeht, insgesamt eine Herausforderung dar. Die technologischen Entwicklungen sind bereits heute vielschichtig und opak, was die Identifizierung der potenziellen Verwendungszwecke beziehungsweise der intendierten Funktionen seitens der Anwender-(Unternehmen) erschwert. Insbesondere liegen diesen weder einschlägige noch jeweils exakt auf ihren Bedarf sowie ihre Kompetenz zugeschnittene Informationen darüber vor, welche Faktoren im jeweiligen Anwendungsfall maßgeblich sind und inwiefern diese Technologie den angestrebten Zielen entspricht oder wo mögliche (Rest-)Risiken entstehen und wie ein adäqua-

ter Umgang damit sein müsste. Entsprechendes gilt auch für die Abwägung dahingehend, ob die jeweilige KI-Anwendung mit den Werten der Anwender(-Unternehmen) konform geht.

1 Perspektive der Gesellschaft

Die Entwicklung zeigt, dass insgesamt eine grundlegende Einigung in der breiten Öffentlichkeit bezüglich Art und Umfang des Einsatzes von KI notwendig erscheint, allein aufgrund der Annahme, dass KI hypothetisch dazu geeignet ist, nahezu jeden Aspekt der konventionellen Rahmenbedingungen einer Gesellschaft beziehungsweise des Miteinanders zu verändern, denn sie wird bereits jetzt als eine der disruptivsten Technologien eingestuft ([8], [24], [1], [19]).

Die Ambivalenz in der Meinungsbildung hinsichtlich der Realisierung im Detail sowie der Implikationen der technologischen Entwicklung ergibt sich – neben dem uneinheitlichen Ausgangspunkt im Hinblick auf die Interpretation von KI – zudem aus der Tatsache, dass die Zuordnung, ob im Einzelfall daraus eine Chance oder ein Risiko resultiert, aktuell nicht eindeutig definiert werden kann. Das liegt einerseits daran, dass aus der bestehenden Informationsasymmetrie zwischen Anbietern von KI-Lösungen und Anwendern, die sich aufgrund der zunehmenden Komplexität der KI-Systeme noch vergrößern wird, Unsicherheitsfaktoren resultieren: Zum einen lässt sich für Letztere nicht ohne Weiteres sicherstellen, ob die eigenen sowie auch die gesellschaftlichen Werte bei dem Einsatz der jeweiligen KI noch erfüllt werden. Diese (empfundene) Unwägbarkeit kann unter anderem darauf gründen, dass bereits heute Möglichkeiten genutzt werden, entgegen den Interessen der Anwender unethisch zu agieren, zum Beispiel im Bereich E-Commerce – etwa, dass durch Manipulation, falsche Versprechungen oder dem Einsatz von Dark Pattern eine Kaufentscheidung herbeigeführt wird, die der Anwender ohne jegliche Beeinflussung keinesfalls treffen würde. Andererseits ist es zunehmend diffizil, die langfristigen Konsequenzen der innovativen Technologien zu prognostizieren und entsprechend einzuordnen – unter anderem bedingt durch die hohe Entwicklungsgeschwindigkeit sowie aufgrund des Fakts, dass es nur wenige Referenzpunkte gibt, um hier eine genaue Einschätzung machen zu können.

Auch die Beurteilung der Chancen im Sinne der Gesellschaft erweist sich als schwierig, wie der nachfolgende Dissens exemplarisch zeigt: Während seitens des Bundestags zur Stärkung der inneren Sicherheit eine Ausweitung technischer Befugnisse, unter anderem zur elektronischen Gesichtserkennung oder zum Videoschutz in Echtzeit an besonders kriminalitätsbelasteten Orten wie Bahnhöfen, auch der Einsatz moderner Software zur Analyse gefordert wird [6], stehen Datenaufsichtsbehörden diesem Ansinnen kritisch gegenüber; so vertritt beispielsweise der Landesbeauftragte für Datenschutz und Informationsfreiheit von Rheinland-Pfalz, Prof. Kugelmann, die Meinung, dass der Ausgangspunkt in der Diskussion um Freiheit und Sicherheit zunächst die Freiheit sein müsse, die Maßnahmen zur Stärkung der Sicherheit bedürften einer wissenschaftlichen und evidenzbasierten Untersuchung dahingehend, welche Auswirkungen das Bestehen und die praktische Anwendung der Überwachungsbefugnisse auf Freiheit und Demokratie habe [14].

2 Notwendigkeit von Rahmenbedingungen

Durch den Einsatz von KI – und hier insbesondere smarte Systeme wie große Sprachmodelle (LLMs) oder agentenbasierten KI-Systeme – können ethische Dilemmata insofern entstehen, dass zukünftig, insbesondere aufgrund des unbeaufsichtigten Handelns von KI-Agenten theoretisch eine Vielzahl unbeabsichtigter Konsequenzen resultieren ([18], [5]). Dies lässt sich anhand des folgenden Beispiels dokumentieren. „‚*Strike Back – Problem Unvollständigkeit der Daten*‘: Dem Konstrukt des ‚Strike Back' liegt die Hypothese zugrunde, dass ein Angriff durch einen Gegenangriff beendet werden kann. Unter Einsatz von KI wäre es theoretisch möglich, eine vollkommen neutrale Einschätzung zu ermitteln, was unternommen werden müsste, um den Angreifer zum Aufgeben zu motivieren. Die ethische Frage in diesem Kontext dreht sich um den Begriff der Neutralität und ob es überhaupt möglich ist, die Vollständigkeit der Daten zu erreichen, um eine ausgewogene und verantwortliche Entscheidung durch KI-Systeme errechnen zu lassen. Davon ausgehend, dass ein Strike Back automatisiert erfolgt, könnte möglicherweise ein Schaden auf einer höheren Ebene angerichtet werden, der moralisch nicht vertretbar wäre – wie etwa ein Gegenschlag, der gleichzeitig auch die Stromversorgung eines Krankenhauses lahmlegen würde." [3].

Aufgrund der Vielschichtigkeit, die mit dieser Technologie verbunden ist, sowie der Tatsache, dass der KI allgemein – wie vorweg festgestellt – ein hoher Einfluss auf die Gesellschaft insgesamt, auf Unternehmen und jedem Einzelnen zugeschrieben wird, erscheint eine Debatte in Bezug auf Prinzipien, Richtlinien, Anreizen sowie ethischen Rahmenbedingungen nicht nur sinnvoll, sondern erforderlich. Diesbezüglich gilt es einige Dimensionen näher zu betrachten.

3 Anerkennung von Werten

Die momentane Situation lässt erkennen, dass insgesamt bezüglich des Umgangs mit KI in der Gesellschaft ein gemeinsames (Spiel-)Verständnis zu schaffen ist[1]. Momentan kann dieses nicht als gegeben angesehen werden (vgl. hierzu [12]), weder in der breiten Öffentlichkeit noch zwischen Experten. Von daher wird es zunehmend notwendig, allgemeingültige Werte als gemeinsame Orientierungspunkte für die Abstimmung von Handlungen generell – und die Beurteilung des Einsatzes von KI im Speziellen – heranzuziehen. Grundsätzlich ermöglichen Werte Menschen ihre Präferenz für eine von mehreren Möglichkeiten zu bestimmen, das heißt jeder kann bezogen auf seine Handlung sowie die Abwägung der daraus resultierenden Konsequenzen jener Option den Vorzug geben, die als wertvoller anerkannt wird – zum Beispiel respektvoll agieren statt respektlos.

1 Um die Vielschichtigkeit, die sich im Rahmen gesellschaftlicher Interaktionen ergibt „heuristisch zu vereinfachen, kann es dienlich sein, die Metapher des Spiels zu nutzen[.]" (Suchanek, 2025, S.1). Hierbei werden drei Ebenen unterschieden: Spielzüge, Spielregeln und gemeinsames Spielverständnis. Die Ebene der ‚Spielzüge' meint die Handlungen der Akteure, sowohl der eigentlichen Spieler als auch aller weiteren Beteiligten. Die Ebene der ‚Spielregeln' konstituiert das Spiel, indem die Handlungen in einer Weise abgestimmt werden, dass das Spiel dauerhaft fortgesetzt werden kann. Die Ebene des ‚gemeinsamen Spielverständnisses' kann als Ordnung der je individuellen, subjektiven 'Spielverständnisse' gesehen werden – auch hier geht es um Abstimmung, unter anderem bezüglich der individuellen Vorstellungen, Überzeugungen, Erwartungen im Hinblick auf das gesellschaftliche Zusammenleben (vgl. hierzu [23]).

Im Kontext einer vertrauensvollen Kooperation – auch hinsichtlich des Einsatzes von KI – fällt den folgenden Werten eine besonders hohe Bedeutung zu (vgl. hierzu [22]):

- Solidarität: Mittels Solidarität lässt sich eine gesellschaftliche Zusammenarbeit zum gegenseitigen Vorteil gewähren, indem jedes Mitglied der Gesellschaft zum Gelingen beiträgt. Wichtig hierbei ist jedoch, dass jedem Einzelnen ungeachtet dessen ermöglicht werden muss, seine eigenen Interessen zu verfolgen. Besonders hervorzuheben ist, dass Solidarität den Bezugspunkt für weitere Werte bildet. Solidarität fällt zudem auf verschiedenen Ebenen eine zentrale Rolle zu, zum Beispiel im Gesundheitsbereich: Da sich die KI in der Medizin in einigen Bereichen, zum Beispiel im Rahmen der Diagnostik bestimmter Krankheiten, als wertvoll erwiesen hat, wäre es unter dem Aspekt der Solidarität hilfreich, wenn viele Menschen mit den entsprechenden Daten dazu beitragen würden, die Qualität der Trainingsdaten zu verbessern, um damit die Untersuchungsergebnisse für alle Patienten zu verbessern.
- Respekt: Respekt ist eine der elementarsten Formen, weil dieser in hohem Maße zu einem gelingenden Leben beiträgt und die richtige Einstellung einem anderen Menschen gegenüber darstellt. Somit ist das Bezeugen von Respekt eine der wesentlichen Bedingungen für die Schaffung von Vertrauen, denn dadurch lässt sich die Berücksichtigung legitimer Erwartungen seitens des Vertrauensgebers realisieren. Die Erwartungen sind, nicht geschädigt zu werden. Respekt bezeugt der Vertrauensnehmers – sprich der KI-Anbieter – dadurch, dass er bewusst auf Funktionalitäten verzichtet, die nicht im Sinne des Vertrauensgebers sind: zum Beispiel auf den Einsatz von Algorithmen, die bestimmte Verhaltensweisen beim Anwender forcieren, wie etwa eine längere Verweildauer auf einer Webseite.
- Fairness: In erster Linie wird durch Fairness die Anerkennung der Würde eines jeden Menschen dokumentiert. Denn es geht darum, die berechtigten (Vertrauens-)Erwartungen zu erfüllen und dies insbesondere dann, wenn es um die Vermeidung einer Benachteiligung oder der Schädigung anderer geht. Fairness bedeutet zudem, sich an bestehende Regeln und Vereinbarungen zu halten und – im Sinne der Selbstbindung – sich keine unfairen Vorteile zu Lasten des Vertrauensgebers zu verschaffen. Dies wäre leicht realisierbar, denn KI wird mittlerweile oftmals zur Unterstützung von Entscheidungen eingesetzt, zum Beispiel von Banken, Versicherungen oder im Personalwesen. Wird eine KI bei der Kreditvorgabe oder im Bewerbungsprozess (absichtlich) unsachgemäß eingesetzt, so kann dies zu einer Schädigung oder Benachteiligung führen – etwa, indem die legitimen Interessen bestimmter Personengruppen verletzt werden, weil im Rahmen einer Entscheidungsfin-

Datenschutz?
Zertifiziert!

Eine DSGVO-Zertifizierung bietet Ihnen entscheidenende Wettbewerbsvorteile und sichert das Vertrauen Ihrer Kunden.

QR-Code scannen und mehr erfahren.

DSGVO
information
privacy
standard

datenschutz
cert

Wir prüfen und zertifizieren
Ihren Datenschutz und Ihre Informationssicherheit

dung oder eines Auswahlverfahrens per se eine Diskriminierung aufgrund unzureichender Trainingsdaten stattfindet.
- Nachhaltigkeit: Nachhaltigkeit bedeutet die moralische Würdigung der Zeitdimension, was die zeitliche Einbettung des eigenen Handelns ermöglicht. Dies beinhaltet, dass nicht nur die Betrachtung unmittelbarer Handlungsfolgen relevant sein kann, sondern ebenfalls die künftigen Handlungsbedingungen, die daraus resultieren respektive dadurch hervorgebracht werden. Somit muss eine Abwägung stattfinden zwischen dem, was kurzfristig attraktiv erscheint und den langfristigen Folgen, die sich daraus möglicherweise ergeben. Nachhaltigkeit im Kontext der KI ist von daher relevant, da sowohl die Entwicklung als auch die Nutzung von jeglicher Form von KI potenziell beträchtliche Auswirkungen auf die Umwelt, Wirtschaft und Gesellschaft hat. Aufgrund dessen ist es erforderlich, bei der Bewertung eines Einsatzes der KI nicht nur auf den kurzfristigen Nutzen zu fokussieren, sondern den daraus zu erwartenden Vorteil in Relation zu den möglichen langfristigen Folgen zu setzen.

Insbesondere die genannten Werte sind – im Sinne der bereits erwähnten Orientierungsfunktion – geeignet zur Schaffung gemeinsamer Maßstäbe im Kontext von KI und sollten zu diesem Zwecke herangezogen werden. Zum Beispiel bei der Entscheidung dahingehend, ob der Einsatz einer KI-Anwendungen vertretbar ist. Da sich dies jedoch nicht durchweg einzig seitens der Anwender(-Unternehmen) bestimmen lässt, ist es notwendig allgemeingültige Regeln – die auf den Werten der Gesellschaft gründen – zum Beispiel durch Gesetze wie die KI-Verordnung aufzustellen, um den Einsatz von Technologie bestmöglich im Sinne der Gesellschaft zu gestalten. Beispielhaft seien hier Deepfake-Technologien genannt, die einerseits eine hochpräzise Bild- und Stimmrekonstruktion bieten, wodurch diese andererseits jedoch Identitätsmissbrauch ermöglichen und zur Verbreitung von Falschinformationen genutzt werden können, oder die Technologie zur Gesichtserkennung, die mittlerweile eine extrem hohe Erkennungsgenauigkeit bietet und daher die Authentisierung etwa bei mobilen Endgeräten ermöglicht, aber dadurch massive Überwachungsmöglichkeiten schafft sowie potenziell sowohl die Privatsphäre von Personen verletzt als auch der Diskriminierung von bestimmten Personengruppen Vorschub leistet.

4 Vertrauenswürdigkeit als Wert

Doch warum fällt im Kontext der KI neben den bereits diskutierten Werten auch dem Wert der Vertrauenswürdigkeit eine hohe Bedeutung zu? Die entsprechende Notwendigkeit lässt sich folgendermaßen argumentieren: Generell besteht eine wichtige Aufgabe der Anwender(-Unternehmen) als Vertrauensgeber darin, ihr Vertrauen richtig zu platzieren, indem sie nach geeigneten Anhaltspunkten suchen mittels derer sich rechtfertigen lässt, vertrauen zu können. Momentan ist es gegebenenfalls möglich, sich auf KI-Systeme in einem bestimmten Maße – etwa bezüglich der Robustheit (der Ergebnisse), die sich validieren lässt – im Sinne von Zuverlässigkeit der Funktionalität zu verlassen. Im Kontext der zugeschriebenen Disruptivität kann es jedoch nicht als ausreichend angesehen werden, dass die KI-Systeme aus technischer Sicht einwandfrei funktionieren: Ihr Einsatz vermag trotzdem negative soziale oder ethische Auswirkungen verursachen, wenn eine kontextuelle Einordnung sowie die darauf basierende Bewertung unterbleibt.

Theoretisch wäre es möglich, dass sich Anwender(-Unternehmen) gegen unerwünschte Nebenwirkungen, die bei dem Einsatz von KI-Anwendungen potenziell entstehen, vertraglich absichern. Jedoch kann ein Vertragswerk aufgrund unvorhersehbarer Kontingenzen, für die keine entsprechende Regeln vereinbart wurden beziehungsweise werden konnten, niemals vollständig sein. Aufgrund dessen ist es für Vertrauensgeber nicht möglich, sich darüber vor einer Ausbeutung respektive vor – durch den Vertrauensnehmer verursachte – Schädigungen zu schützen. Somit ist es im Sinne des Vertrauensgebers erstrebenswert, dass er und der Vertrauensnehmer ein gemeinsames Wertesystem haben. Dies ist faktisch im beiderseitigen Interesse: Dem Anwender wird so ermöglicht Vertrauen aufzubauen. Für den KI-Anbieter kann die daraus resultierende Verbesserung der Kooperation zur Senkung entsprechender Transaktionskosten – also alle Aufwendungen, die zur Anbahnung von Geschäftsbeziehungen und deren Aufrechterhaltung anfallen – führen.

Die konkrete Voraussetzung für eine gelingende Kooperation ist jedoch zum einen, dass – insbesondere beim Vertrauensnehmer – die Werte nicht nur verbal kommuniziert, sondern tatsächlich gelebt werden und zum anderen, dass beide Parteien die Werte auf die gleiche Art und Weise verstehen. Daraus lässt sich im Prinzip folgende Hierarchie ableiten: Werte sind die Grundlage für das ethische Handeln des Vertrauensnehmers, woraus entsprechend dessen Vertrauenswürdigkeit resultiert.

Als allgemein anerkannt gilt, dass eine Interdependenz zwischen Vertrauen und Vertrauenswürdigkeit besteht, da das Konzept der Vertrauenswürdigkeit[2] den Prozess des Vertrauens motiviert[3]. Denn im Prinzip entsteht eine praktische Verpflichtung, da der Vertrauensgeber im Regelfall erwartet, dass der Vertrauensnehmer sein – ihm gegenüber eingegangenes – Versprechen auch erfüllt ([9], [11]). Dieser Obligation kann der Vertrauensnehmer im Prinzip nur nachkommen, wenn er seine – für eine vertrauensvolle Kooperation relevanten – Werte auch tatsächlich dem Vertrauensgeber in seinen Handlungen zeigt. Unter anderem bezogen auf das Bezeugen von Respekt: hier sollte den Vertrauensgebern vermittelt werden, dass der Vertrauensnehmer auf den Einsatz bestimmter Funktionen – etwa dem umfangreichen Erheben von Kundendaten – verzichtet, die nur zu seinem eigenen Vorteil wären und nicht im Interesse des Vertrauensgebers.

Doch Vertrauen setzt auch immer einen gewissen Grad an Information über den Vertrauensnehmer voraus, denn gemäß Luhmann „[kann man] ohne jede vorherige Information [..] kaum vertrauen“ [17]. Daraus lässt sich ableiten, dass den Anwender(-Unternehmen) Informationen über KI-Anbieter zur Verfügung gestellt werden müssen, damit erstere eine informierte Entscheidung treffen können. Dafür bedarf es einer verlässlichen Quelle, denn mit jeder Entscheidung für den Einsatz einer KI-Anwendung übernehmen Anwender auch eine gewisse Verantwortung im Hinblick auf mögliche Auswirkungen.

2 Vertrauenswürdigkeit basiert auf den allgemein in der Wissenschaft anerkannten Bedingungen (conditions) Fähigkeit (ability), Wohlwollen (benevolence) und Integrität (integritiy). Somit ist Vertrauenswürdigkeit eine Eigenschaft, die jeweils zugeschrieben wird – Vertrauen hingegen ist eine Haltung. (vgl. hierzu [21]).

3 Die Beziehung beschreibt Hardin folgendermaßen: “Hence whatever can secure their trustworthiness enough for us to trust them will help us manage complexity[.]” [10].

In diesem Bezugsrahmen als vertrauenswürdige Quelle zu agieren und so einen Vertrauens-Mehrwert für Anwenderunternehmen zu bieten ist die grundlegende Intention, die maßgeblich war für die Etablierung der Vertrauenswürdigkeits-Plattform ‚Trust4Good'. Um diesem Anspruch gerecht werden zu können, ist es erforderlich, dass die Informationen der jeweiligen KI-Anbieter keine Falschaussagen beinhalten, weil dadurch die Vertrauenswürdigkeit einer Quelle sinkt. Das Bemühen um die Wahrhaftigkeit der Aussagen muss somit seitens der KI-Anbieter als grundlegende Prämisse anerkannt werden, wenn sie auf der Vertrauenswürdigkeits-Plattform „Trust4Good" im Rahmen eines Commitments den Anwenderunternehmen ihre Werte dergestalt dokumentieren, dass sie zu den einzelnen – für den Aufbau von Vertrauen maßgeblichen – Vertrauenswürdigkeits-Aspekten *Zutrauen*, *Zuverlässigkeit*, *Integrität* und *Sicherheit* (bezogen auf das Unternehmen) sowie *Transparenz*, *Leistungsfähigkeit* und *Zweckprägnanz* (bezogen auf die KI-Lösung) Auskunft geben (vgl. hierzu [4]). Denn unter der Annahme, dass das einzelne Anwenderunternehmen Vertrauen in den KI-Anbieter und dessen KI-Lösung aufbauen, beziehungsweise verstetigen kann, indem es prinzipiell in die Lage versetzt wird, einen Einblick in die jeweils notwendigen Details nehmen zu können, ist Ehrlichkeit eine zwingende Voraussetzung.

5 Fazit

Entscheidungen bezüglich KI basieren auf Vertrauen. Anwender(-Unternehmen) – und damit Vertrauensgeber – müssen sichergestellt wissen, dass die jeweiligen KI-Systeme weder Schädigungen beim Einsatz verursachen noch, dass sie bei deren Anwendung gegen die allgemeinen Werte der Gesellschaft verstoßen. Damit dieses Vertrauen gerechtfertigt ist, müssen die KI-Anbieter als Vertrauensnehmer vertrauenswürdig agieren – das heißt alles im Rahmen ihrer gegebenen Möglichkeiten tun, um Anwender(-Unternehmen) keinen Schaden zuzufügen. Um dieses Ziel zu erreichen ist es notwendig ein gemeinsames (Spiel-) Verständnis und darauf basierend übereinstimmende Erwartungen zu entwickeln. Hierfür bedarf es der Schaffung gemeinsamer Maßstäbe, die auf anerkannten Werten beruhen und von daher als Orientierungsfunktion dienen.

Literatur

[1] Almansour, M. (2023). Artificial intelligence and resource optimization: A study of Fintech start-ups, Resources Policy, Volume 80,103250, ISSN 0301-4207. https://doi.org/10.1016/j.resourpol.2022.103250. Zugriff am: 31.01.2026

[2] Bolstrom, N. (2014). Superintelligence Paths, Dangers, Strategies, Oxford University Press, ISBN 978-0199678112.

[3] Coester, U., Pohlmann, N. (2019). „Ethik und künstliche Intelligenz – Wer macht die Spielregeln für die KI?", IT & Production – Zeitschrift für erfolgreiche Produktion, TeDo Verlag

[4] Coester, U., Pohlmann, N. (2025). Warum Vertrauenswürdigkeit der Grundstein für die Digitalisierung ist. Datenschutz und Datensicherheit – DuD, Vol. 49, 1/2025, Seiten 11-15. 10.1007/s11623-024-2031-x.

[5] Cohen, M. K., Kolt, N. Bengio, Y., Hadfield, G. K., Stuart, R. (2024). Regulating advanced artificial agents, Science, Vol. 384, Issue 6691, pp. 36-38. DOI: science.org/doi/10.1126/science.adl0625. Zugriff am: 31.01.2026

[6] Deutscher Bundestag (2025). Drucksache 20/14699, Entschließungsantrag, https://dserver.bundestag.de/btd/20/146/2014699.pdf. Zugriff am: 31.01.2026

[7] Digitalisierung der Wirtschaft in Deutschland, Technologie- und Trendradar 2021, Seite 10. https://www.bundeswirtschaftsministerium.de/Redaktion/DE/Publikationen/Digitalisierungsindex/publikation-download-technologie-trendradar-2021.pdf?__blob=publicationFile&v=1. Zugriff am: 31.01.2026

[8] DIN, DKE (2022). Deutsche Normungsroadmap, Künstliche Intelligenz (Ausgabe 2). www.din.de/go/normungsroadmapki. Zugriff am: 31.01.2026

[9] Durán, J.M., Pozzi, G. (2025). Trust and Trustworthiness in AI, Philosophy & Technology, Vol. 38, Article 16. https://doi.org/10.1007/s13347-025-00843-2. Zugriff am: 31.01.2026

[10] Hardin, R. (2002). Trust and Trustworthiness, New York: Russell Sage Foundation, Seite 30.

[11] Hawley, K. (2019). How to be trustworthy, Oxford University Press.

[12] Kettner, S. E., Thorun, C. und Cerulli-Harms, A. (2025), 2. Verbraucherperspektive, in: Brink, A., Fairness in Zeiten Künstlicher Intelligenz, Nomos-Verlag, Baden-Baden, ISBN print: 978-3-7560-3469-7, ISBN online: 978-3-7489-6503-9, Seite 89-102.

[13] Koalitionsvertrag, Verantwortung für Deutschland, 21. Legislaturperiode (2025). https://www.koalitionsvertrag2025.de/. Zugriff am: 31.01.2026

[14] Kugelmann, D. (2025). Interview Prof. Dr. Dieter Kugelmann, Landesbeauftragte für Datenschutz und Informationsfreiheit von Rheinland-Pfalz, 10.11.2025

[15] Kurzweil, R. (2005). The Singularity Is Near. When Humans Transcend Biology. Viking, New York, ISBN 0-670-03384-7.

[16] Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., Cestari, V., Rossi-Arnaud, C. & W. Quattrociocchi, W. (2025) The simulation of judgment in LLMs, Proc. Natl. Acad. Sci. (PNAS), USA, Vol. 122, No. 42, Seiten 1-11. https://www.pnas.org/doi/10.1073/pnas.2518443122. Zugriff am: 31.01.2026

[17] Luhmann, N. (2014). Vertrauen: Ein Mechanismus der Reduktion sozialer Kompetenz, 5. Auflage, UVK Verlagsgesellschaft mbH, Konstanz.

[18] Mitchell, M., Ghosh, A., Luccioni, A.S., Pistilli, G. (2025). Fully Autonomous AI Agents Should Not be Developed. https://doi.org/10.48550/arXiv.2502.02649. Zugriff am: 31.01.2026

[19] Păvăloaia, V.-D., & Necula, S.-C. (2023). Artificial Intelligence as a Disruptive Technology – A Systematic Literature Review. Electronics, Vol. 12, Issue 5. https://doi.org/10.3390/electronics12051102. Zugriff am: 31.01.2026

[20] Pohlmann, N (2024). Glossar 'Cyber-Sicherheit' https://norbert-pohlmann.com/glossar-cyber-sicherheit/singularitaet/. Zugriff am: 31.01.2026

[21] Stanford Encyclopedia of Philosophy (2020). Trust. https://plato.stanford.edu/entries/trust/. Zugriff am: 31.01.2026

[22] Suchanek, A. (2015). Unternehmensethik – In Vertrauen investieren, Mohr Siebeck, Seiten 153ff. https://www.mohrsiebeck.com/buch/unternehmensethik-9783161531651/. Zugriff am: 31.01.2026

[23] Suchanek, A. (2025). Digitalisierung und das Problem des gemeinsames Spielverständnisses, Studies in Philosophy of Economics | Studien zur Wirtschaftsphilosophie; Herausgegeben von Prof. Dr. Ludger Heidbrink, Verlag Karl Alber, Hinweis: die angegebenen Seitenzahlen entstammen dem Manuskript. https://doi.org/10.5771/9783495992128. Zugriff am: 31.01.2026

[24] UNESCO (2021). UNESCO-Empfehlung zur künstlichen Intelligenz, unesco. https://www.unesco.de/dokumente-und-hintergruende/publikationen/detail/unesco-empfehlung-zur-ethik-der-kuenstlichen-intelligenz/. Zugriff am: 31.01.2026