Barton, Marie-Christin; Pöppelbuß, Jens

**Article — Published Version**

# Prinzipien für die ethische Nutzung künstlicher Intelligenz

HMD Praxis der Wirtschaftsinformatik

**Provided in Cooperation with:**
Springer Nature

*Suggested Citation:* Barton, Marie-Christin; Pöppelbuß, Jens (2022) : Prinzipien für die ethische Nutzung künstlicher Intelligenz, HMD Praxis der Wirtschaftsinformatik, ISSN 2198-2775, Springer Fachmedien Wiesbaden GmbH, Wiesbaden, Vol. 59, Iss. 2, pp. 468-481, https://doi.org/10.1365/s40702-022-00850-3

HMD (2022) 59:468–481
https://doi.org/10.1365/s40702-022-00850-3

SCHWERPUNKT

# Prinzipien für die ethische Nutzung künstlicher Intelligenz

**Marie-Christin Barton · Jens Pöppelbuß**

Eingegangen: 15. Oktober 2021 / Angenommen: 13. Februar 2022 / Online publiziert: 10. März 2022


**Zusammenfassung** Die voranschreitende Digitalisierung im Allgemeinen und die zunehmende Anwendung von künstlicher Intelligenz (KI) im gesellschaftlichen Alltag bietet einige Vorteile. Allerdings werden von verschiedenen Interessensgruppen ethische Bedenken geäußert. Damit KI langfristig angewendet und KI-Lösungen erfolgreich vertrieben werden können, ist es daher unerlässlich, dass Organisationen eine ethisch-angemessene Richtung einschlagen. So erhält das relativ neue Feld der KI-Ethik wachsendes Interesse sowohl in der Wissenschaft als auch in der Praxis. Dieser Beitrag skizziert ethische Herausforderungen, denen Organisationen begegnen und somit auch Bedenken, die in der Gesamtgesellschaft durch die Digitalisierung im KI-Zeitalter aufkommen. Diese werden auf Basis verschiedener ethischer Modelle untersucht. Zudem wird ein modifiziertes Modell mit sechs Prinzipien für die ethische Nutzung künstlicher Intelligenz vorgeschlagen. Dieses umfasst die sechs KI-Ethik-Prinzipien: Wohltätigkeit, Transparenz, Nicht-Boshaftigkeit, Autonomie, Gerechtigkeit und Datenschutz. Basierend auf diesen sechs Prinzipien werden Handlungsanweisungen im Umgang mit KI-Anwendungen skizziert.

**Schlüsselwörter** Künstliche Intelligenz · KI-Ethik · PAPA · Digitale Ethik · Ethische Richtlinien · Datenschutz

Marie-Christin Barton (✉) · Jens Pöppelbuß
Lehrstuhl für Industrial Sales and Service Engineering, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Deutschland
E-Mail: marie-christin.barton@isse.rub.de

Jens Pöppelbuß
E-Mail: jens.poeppelbuss@isse.rub.de

## Principles for the Ethical Use of Artificial Intelligence

**Abstract** Digitization in general and the increasing application of artificial intelligence (AI) in our everyday life offers some advantages. However, ethical concerns are raised by various stakeholders. For AI to be successfully applied in the long term, it is essential that organizations adopt an ethically appropriate direction. Thus, the relatively new field of AI ethics is receiving growing interest from both academia and practitioners. This paper outlines some of the ethical challenges faced by organizations and points to concerns raised in society related to AI. These are examined based on various ethical models. In addition, a modified model with six AI ethics principles is proposed. These include: beneficence, transparency, non-maleficence, autonomy, justice, and data privacy. We further derive directions for action in dealing with AI applications based on the six AI ethics principles.

**Keywords** Artificial Intelligence · AI Ethics · PAPA · Digital Ethic · Ethical Guidelines · Data Privacy

## 1 Einleitung

Die Berücksichtigung ethischer Aspekte bei der Nutzung künstlicher Intelligenz (KI) ist ein Thema von wachsender praktischer Relevanz. Kunden reagieren sehr sensibel auf unethisches Verhalten – beabsichtigt oder unbeabsichtigt – von Unternehmen, weshalb eine ethische Ausrichtung für Organisationen strategisch relevant wird (Sena und Nocker 2021). Die Forderung nach ethisch angemessenem Verhalten hat das relativ neue Feld der KI-Ethik entstehen lassen, über das sowohl in der Wissenschaft wie auch in der Praxis noch wenig bekannt ist (Mayer et al. 2021).

Erste Arbeiten haben sich bereits mit ethischen Aspekten von Informationssystemen im Allgemeinen beschäftigt (z. B. McBride 2014; Rothenberger et al. 2019; Vermanen et al. 2019). Die Untersuchung von Vermanen et al. (2019) konzentriert sich beispielsweise auf die ethischen Fragen im Zusammenhang mit dem Einsatz des Internets der Dinge in kleinen und mittleren Unternehmen unter Verwendung des PAPA-Models von Mason (1986). Der Name des PAPA-Modells ist ein Akronym für die vier ethischen Fragen, die mit dem Eintritt in das Informationszeitalter aufkamen und sich teilweise überschneiden: Privatsphäre (*privacy*), Genauigkeit (*accuracy*), Eigentum (*property*) und Zugänglichkeit (*accessibility*). Organisationen müssen eine Balance aus maximalem Umsatz- und Gewinnstreben sowie ethisch korrektem KI-Design finden (Rothenberger et al. 2019).

In den letzten Jahren steigt die Anzahl an Veröffentlichungen, die ethische Fragen explizit für den KI-Kontext thematisieren, wobei nach wie vor Uneinigkeit über einen konzeptionellen Rahmen zum Verständnis dieser Fragestellungen besteht. Die Arbeiten orientieren sich meist an ethischen Richtlinien aus angrenzenden Disziplinen, wie z. B. Informationssysteme im Allgemeinen, Big Data und Cybersicherheit (Formosa et al. 2021). Zu beachten ist, dass sich Ethik aufgrund von technischen, ökonomischen und anderen gesellschaftlichen Wandlungen beständig verändert (vgl. Manzeschke 2021).

Organisationen im Allgemeinen und Unternehmen im Speziellen sind wichtige Vermittler in modernen, pluralen Gesellschaften, da in ihnen eine Vielzahl von Entscheidungen getroffen werden, die das Leben der Individuen wie auch das der Gemeinschaft nicht unwesentlich formen (Manzeschke 2021). Vor allem der private Sektor spielt eine zentrale Rolle in der Entwicklung und Verbreitung von KI-Technologien, sodass sich kritische Anmerkungen und Handlungsbedarfe insbesondere an die Unternehmenswelt richten (Mayer et al. 2021).

Dieser Beitrag skizziert ethische Herausforderungen, denen Organisationen begegnen und somit auch Bedenken, die in der Gesamtgesellschaft durch die Digitalisierung im KI-Zeitalter aufkommen. Diese werden auf Basis ethischer Modelle, wie dem PAPA-Modell (Mason 1986), dem AI4People-Modell (Floridi et al. 2018) sowie veröffentlichter KI-Ethik-Prinzipien (Jobin et al. 2019) untersucht und Handlungsempfehlungen im Umgang mit KI-Technologien daraus abgeleitet. In diesem Zuge ergeben sich auch Vorschläge für eine Weiterentwicklung bislang bestehender Modelle im Hinblick auf aktuelle KI-Entwicklungen.

## 2 Merkmale künstlicher Intelligenz

KI wird zunehmend als Schlagwort in den Medien, der Wissenschaft und der Wirtschaft verwendet. Oftmals findet dann auch der Begriff Machine Learning (ML, deut. maschinelles Lernen) Verwendung oder wird sogar synonym verwendet. Laut Kreutzer und Sirrenberg (2019) beschreibt KI im Allgemeinen die Fähigkeit einer Maschine zur Ausführung kognitiver Aufgaben, die dem menschlichen Verstand ähnelt. Eine mögliche Definition bieten Kaplan und Haenlein (2019, S. 3): „[AI is] defined as a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation". Aus Sicht der Datenwissenschaft (Data Science) wird ML lediglich als ein Teilbereich von KI angesehen, der sich auf die maschinellen Lernverfahren konzentriert, die Vorhersagen oder Entscheidungen aus erkannten Mustern und Gesetzmäßigkeiten treffen (Akerkar 2019).

KI lässt sich durch drei Schlüsselkomponenten charakterisieren, nämlich *Inputdaten, ML-Algorithmus* und *Outputentscheidung*. Inputdaten, wie beispielsweise Bilder und Sprache (unstrukturierter Input) oder Transaktionsdaten (strukturierter Input) sind unerlässlich für das Funktionieren einer KI (Canhoto und Clear 2020). In der heutigen digital fortgeschrittenen, schnelllebigen Welt werden riesige Mengen an Daten (Big Data) produziert, die für KI-Systeme potenziell nutzbar sind. Die Daten können historisch Daten (z. B. vergangene Kundentransaktionsdaten), Echtzeitdaten (z. B. Tracking während des Onlineshoppings) oder in Form von Wissen (z. B. abgelehnte oder akzeptierte Produktempfehlungen aus der Vergangenheit) vorhanden sein (Canhoto und Clear 2020). In der Praxis haben ein Großteil der Unternehmen und Entwickler:innen keinen Zugriff auf ausreichende Trainingsdatenmengen. Deshalb setzen sie im Sinne einer Demokratisierung von KI-Technologien zunehmend auf Small Data (kleinere Datenmengen und angepasste ML-Methoden) sowie Wide Data (Synergienutzung aus verschiedenen Datenquellen und -typen). Beide Ansätze erleichtern robustere Analysen, verringern die Abhängigkeit eines Unternehmens von

Big Data und ermöglichen ein umfassenderes, vollständigeres Situationsbewusstsein (Goasduff 2021).

Der *ML-Algorithmus* beschreibt das Rechenverfahren, das den Dateninput verarbeitet. Dabei werden auf Basis der Trainingsdaten Muster und Gesetzmäßigkeiten erkannt, die es der KI ermöglichen, ihre Leistung zu verbessern, ohne dass sie explizit darauf programmiert wird. Es gibt drei wesentliche Arten von Verarbeitungsalgorithmen: *Supervised Learning, Unsupervised Learning* und *Reinforcement Learning*. Beim Erstgenannten nutzt die KI einen Trainingsdatensatz, der gegebene Paare von Input und Output enthält, sodass der Algorithmus Muster lernt und die entwickelten Regeln auf zukünftige Fälle desselben Problems anwendet. Dieses Verfahren findet beispielweise bei der Krebserkennung Anwendung. Beim Unsupervised Learning verwendet die KI Trainingsdaten, die nicht mit dem richtigen Output gekennzeichnet sind und identifiziert selbst Muster oder Beziehungen zwischen den Datenpunkten. Dies ist beispielswese der Fall, wenn verwandte Bilder in einer Fotodatenbank aufzuspüren sind. Beim dritten Algorithmus, dem Reinforcement Learning, erhält die KI einen Trainingsdatensatz sowie ein Ziel, sodass die beste Kombination von Aktionen zur Erreichung dieses Ziels gefunden werden muss (z. B. bei einem Brettspiel).

Die dritte Schlüsselkomponente von KI ist die aus dem ML-Prozess resultierende *Outputentscheidung*, bei der zwischen einem Ergebnis (z. B. Bonitätswert), einer Ergebnisauswahl (z. B. Auswahl von Videos, die möglicherweise gegen die Nutzungsbedingungen der Videoplattform verstoßen und die von einem Mitarbeitenden weiter analysiert werden müssen) sowie einer Aktion (z. B. selbstfahrendes Auto) unterschieden werden kann (Canhoto und Clear 2020). Die drei Schlüsselkomponenten von KI sind in Abb. 1 dargestellt. An dieser Stelle ist anzumerken, dass es beispielsweise über die Genannten hinaus weitere ML-Methoden gibt und KI-Anwendungen sehr komplex sein können, jedoch hier nur in Grundzügen darge-

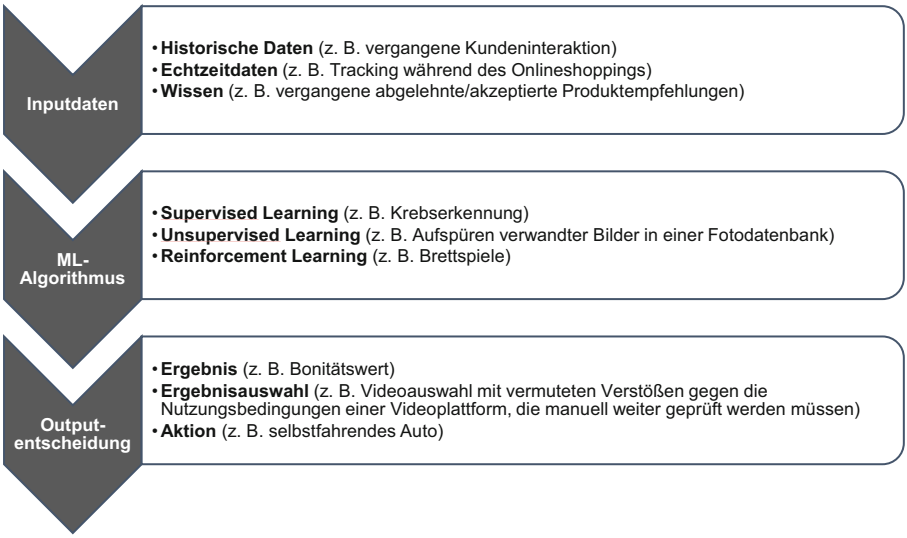


**Abb. 1** Schlüsselkomponenten von KI. (Modifiziert nach Canhoto und Clear 2020, S. 184)

stellt sind. Weitere technische Ausführungen würden über den Schwerpunkt dieses Beitrags auf KI-Ethik hinausgehen.

Die Simulation menschlicher Intelligenz durch KI ist in ihrer simpelsten Form nahezu allgegenwärtig. Die oben beschriebenen Merkmale von KI decken eine große Bandbreite von Technologien ab, angefangen von Maschinen, die Objekte in Bildern wiedererkennen können (Natural-Image-Processing), über Software-Applikationen, die den nächstbesten Verkaufsakt empfehlen (Recommender Systems), bis hin zu Systemen, die eine Art von Bewusstsein haben und ihren aktuellen Zustand verarbeiten können (Sena und Nocker 2021). Die Nutzung von KI-Technologien kann zur menschlichen Selbstverwirklichung beitragen, die menschliche Handlungsfähigkeit stärken, gesellschaftliche Fähigkeiten steigern und den Zusammenhalt fördern (Floridi et al. 2018). In der Praxis profitieren z. B. Vertriebsmitarbeitende von Zeitersparnissen, indem KI-Assistenten die Planung und Organisation des Arbeitsalltags übernehmen oder sogar Verkaufsprozesse teils automatisiert werden. Auch können durch KI-Unterstützung potenzielle Kunden identifiziert (Lead Generation) und nach ihrem Kundenwert priorisiert werden (Customer Scoring), wodurch Unternehmen erhöhte Umsätze verzeichnen können. Individualisierte Kaufempfehlungen (Next-Product-to-Buy) können ebenfalls die Verkaufsrate eines Anbieters erhöhen, aber gleichzeitig einer Privatperson langes Herumirren auf Internetseiten ersparen (Kreutzer und Sirrenberg 2019). Im Gesundheitssektor wird KI beispielsweise bereits in der Kardiologie genutzt, wo Herzraten automatisch interpretiert werden oder mittels Bildmaterial Herzerkrankungen diagnostiziert werden. In Produktionsstätten werden intelligente Lösungen sowie Robotertechnik ebenfalls in die Arbeitsprozesse integriert. KI findet auch in vielen weiteren Branchen, wie z. B. der Agrarwirtschaft, der Finanzwirtschaft, der Automobilindustrie, der Maschinenindustrie und dem öffentlichen Sektor vorteilhafte Anwendung, die schließlich in Effizienz- und Effektivitätsgewinn resultieren (Radhakrishnan und Gupta 2020; Sidorenko et al. 2021).

## 3 Ethische Bedenken im KI-Zeitalter

Trotz der zahlreichen, vielversprechenden Potenziale und Vorteile der Anwendung von KI birgt dieser technologische Fortschritt auch einen destruktiven Charakter, der Bedenken und Debatten sowohl auf Seiten des öffentlichen als auch privaten Sektors schürt. Eine besondere Stellung nimmt das Thema u. a. im Zusammenhang mit der Privatsphäre der Anwender:innen ein. Mit Privatsphärebedenken und weiteren ethischen Fragestellungen sowie Vorurteilen bei der Erstellung von Modellen für ML werden Data Scientists und KI-Entwickler:innen regelmäßig konfrontiert (Blackman 2020). Um diesen Bedenken zu entgegnen, reagieren nationale und internationale Organisationen mit einer Ad-hoc-Entwicklung von KI-Experten-Komitees, die häufig mit der Ausarbeitung von Strategiedokumenten beauftragt sind. Zu diesen Ausschüssen gehört beispielsweise die „High-Level Expert Group on Artificial Intelligence“, eingesetzt durch die Europäische Kommission (Jobin et al. 2019). Noch scheinen diese Entwicklungen die Bedenken in der Gesellschaft nicht aufzulösen. Bedenken gegenüber bzw. Risiken von KI-Technologien hängen meist mit einer versehentli-

chen Überbeanspruchung oder einem vorsätzlichen Missbrauch von Daten, getrieben z. B. von Gier oder geopolitischer Feindschaft, zusammen. Durch den böswilligen Einsatz von KI könnten beispielsweise der E-Mail-Betrug zunehmen oder sogar Cyberkriegsführung intensiviert werden. Die Entwertung menschlicher Fähigkeiten, die Beseitigung der menschlichen Verantwortung, die Verringerung menschlicher Kontrolle und die Untergrabung menschlicher Selbstbestimmung sind Risiken, die die Würde des Menschen adressieren (Floridi et al. 2018).

Skeptiker hinterfragen weiterhin die Verlässlichkeit der KI-Technologien sowie die Sicherheit für den Menschen. Beispielsweise stellt sich die Frage, wer im Falle von Fehlentscheidungen oder fälschlichen Maßnahmen auf Basis der KI, z. B. in Produktionsstätten, haften muss; sei es in rechtlicher wie auch moralischer Hinsicht. Ist die Programmiererin, der zuständige Informatiker im Support, die Betriebsleiterin oder gar der ausführende Mitarbeitende für den KI-Output verantwortlich? Warum eine KI zu einem bestimmten Output kommt, kann nicht immer erklärt werden, denn ein Merkmal, das die KI ausmacht, ist schließlich, dass sie mehr oder weniger eigenständig dazulernt (Kreutzer und Sirrenberg 2019). Derartige Unsicherheit führt zu Widerständen bei der Etablierung von KI in Unternehmen und der Gesamtgesellschaft (vgl. Sidorenko et al. 2021).

Des Weiteren fallen durch die Digitalisierung massenhaft personenbezogenen Daten an, die in eine KI eingespeist werden können. In diesem Zuge kommen Bedenken zur Privatsphäre, Dateneigentum, Zugänglichkeit und Sicherheit auf. Die Verletzung der Privatsphäre ist nicht die einzige Gefahr, die es bei der Einführung von der KI zu vermeiden gilt (Floridi et al. 2018; McBride 2014). Fehlende Transparenz zur Datenverarbeitung sowie Datenspeicherung verschrecken potenzielle Anwender:innen. Gleichsam kommen Forderungen zur Erklärbarkeit und Nachvollziehbarkeit der Datenverarbeitung und des KI-Outputs (Explainable AI) auf. KI-Technologie bedeutet auch, dass Sensoren, Datenbanken usw. miteinander vernetzt sind, wodurch tendenziell mehr Schwachstellen geboten werden, die für Hacker anfällig sind. Es ist notwendig, dass eine geregelte, sichere Zugänglichkeit gewährleistet wird (vgl. Mason 1986). An dieser Stelle ist anzumerken, dass Menschen in der Regel einen sensiblen Umgang mit ihren Daten erwarten, um die Technologien zu akzeptieren, gleichzeitig aber selbst teils leichtfertig und unreflektiert ihre Daten, z. B. auf Internetseiten, in Umlauf bringen (Albayrak et al. 2018). Weitere Bedenken betreffen die mögliche Substituierbarkeit menschlicher Arbeitskräfte und die Mensch-Technik-Interaktion allgemein (Manzeschke 2021).

Sidorenko et al. (2021) sehen insgesamt zwei Blöcke von ethischen Risiken, die aus der Nutzung von KI erwachsen. Zum einen gibt es Risiken, die im Zusammenhang mit der Erhebung und Verarbeitung personenbezogener Daten stehen, wie z. B. Verzerrungen in den Datensätzen, die zum Training der KI genutzt werden und letztlich zu Fehlentscheidungen führen können. Zum anderen gibt es Risiken, die die Ethik der Outputentscheidungen und deren Übereinstimmung mit allgemeinen, gesellschaftlichen Normen und Werten umfassen (Sidorenko et al. 2021). Es ist letztlich stets zu bedenken, wie viel Entscheidungsgewalt an die KI delegiert wird. Um den ethischen Bedenken in der Gesellschaft zu begegnen, reicht es schließlich nicht aus, lediglich gesetzeskonform zu handeln und rechtliche Vorgaben bei der Entwicklung sowie Anwendung von KI zu berücksichtigen. Öffentliche Akzeptanz

und Annahme von KI-Technologien wird es nur geben, wenn die Vorteile als sinnvoll und die Risiken als möglich und denkbar, aber gleichzeitig als vermeidbar oder minimierbar, z. B. durch Risikomanagement (z. B. Versicherungen) oder Wiedergutmachung, erachtet werden (Floridi et al. 2018).

## 4 Ethische Ansätze im Bereich der KI

In der Literatur werden zumeist drei Theorien genannt, auf denen ethische Modelle im Kontext von Informationssystemen und Digitalisierung beruhen (Spiekermann 2021). Zum einen ist die *Deontologie* zu nennen, die universal-moralische Gesetze zur Begrenzung der Aktionen aller rationalen Individuen identifiziert (Schmidt 2011). Die zweite Theorie ist die *Tugendethik*, deren Ausgangspunkt die Tugend bzw. das tugendhafte Verhalten des Menschen ist. Sie gleicht den persönlichen Charakteristiken eines Menschen, die ihn zu einer guten Person machen – z. B. Eigenschaften wie Geduld und Aufrichtigkeit (Schramme 2011). Die dritte und wohl am häufigsten genannte Theorie in diesem fachlichen Kontext ist der *Utilitarismus*, eine Form der konsequentialistischen Ethik. Dieser ethische Entscheidungsansatz berücksichtigt die Folgen, d. h. Schaden und Nutzen, des Handelns. Eine ethische Handlung bringt die größtmöglichen guten Folgen und zugleich kleinstmöglichen schlechten Folgen hervor (McBride 2014; Schroth 2011). Da die ethischen Aspekte in diesem Kontext mehrere Interessensgruppen betreffen, stellt die Diskursethik darüber hinaus einen möglichen Ansatz zur Lösung ethischer Problemstellungen dar. Sie hebt hervor, dass ethische Ableitungen aus einem intersubjektiven Diskurs resultieren (Mingers und Walsham 2010; Spiekermann 2021).

Ein häufig auftretendes Modell zur Betrachtung ethischer Aspekte im Fachgebiet der Informationssysteme, das auf der Theorie des Utilitarismus beruht, ist das sogenannte *PAPA-Modell* von Mason (1986). Der Name des Modells ist ein Akronym für die vier Bereiche ethisch relevanter Fragestellungen, die mit dem Eintritt in das Informationszeitalter aufkamen und sich teilweise überschneiden: *Privatsphäre (privacy), Genauigkeit (accuracy), Eigentum (property)* und *Zugänglichkeit (accessibility)*. Folgende Fragestellungen lassen sich diesen vier Bereichen zuordnen (Mason 1986):

- *Privatsphäre:* Welche Informationen über die eigene Person oder über sich selbst oder seine Verbindungen muss eine Person anderen preisgeben, unter welchen Bedingungen und mit welchen Schutzmaßnahmen? Welche Dinge kann man für sich behalten und nicht gezwungen werden, sie anderen zu offenbaren?
- *Genauigkeit:* Wer ist verantwortlich für die Authentizität, Treue und Genauigkeit von Informationen? Wer ist verantwortlich für Fehler in Informationen und wie kann die geschädigte Partei entschädigt werden?
- *Eigentum:* Wem gehören die Informationen? Was sind die gerechten und fairen Preise für ihren Austausch? Wem gehören die Kanäle, durch die Informationen übertragen werden? Wie sollte der Zugang zu dieser knappen Ressource zugewiesen werden?

- *Zugänglichkeit:* Auf welche Informationen haben eine Person oder eine Organisation das Recht oder das Privileg sie zu erhalten, unter welchen Bedingungen und mit welchen Garantien?

Zu dem Zeitpunkt als Mason (1986) sein Modell veröffentlicht hat, gab es noch deutlich andere Standards und Verbreitungsgrade von Informations- und Kommunikationstechnologien als heute. So war beispielsweise der Besitz von Computern in Privathaushalten selten und Computertechnologien wurden in Unternehmen häufig zentralisiert vorgehalten. Auch der von IBM entwickelte intelligente Schachcomputer erlangte erst 1997 weltweites Aufsehen (Kreutzer und Sirrenberg 2019). Dennoch bietet das PAPA-Modell eine erste Orientierung, indem es vor allem die ethische Handhabung einer der Schlüsselkomponenten von KI, nämlich des Dateninputs (vgl. Canhoto und Clear 2020), gewährleisten soll. So wird u. a. datenschutzrechtliche Bedenken und Risiken Rechnung getragen.

Das *AI4People-Modell* für ethische KI von Floridi et al. (2018) ist ein weiteres relevantes und deutlich jüngeres Modell zur Betrachtung ethischer Aspekte. Die Autoren übertragen und erweitern Prinzipien aus der traditionellen Bioethik – die in ihren Ansichten der KI-Ethik ähnelt – auf den KI-Kontext. Hierzu haben sie sechs Dokumente zu KI-Ethik-Prinzipien heranzogen. Folgende fünf Prinzipien bilden demnach einen ethischen Rahmen für KI (Floridi et al. 2018):

- *Wohltätigkeit (beneficence):* Förderung des Wohlbefindens, Bewahrung der Würde und Erhaltung des Planeten.
- *Nicht-Boshaftigkeit (non-maleficence):* Vermeidung von Verletzung der persönlichen Privatsphäre, Sicherheit und Limitierung der KI-Fähigkeiten.
- *Autonomie (autonomy):* Ausbalancierte Entscheidungsgewalt von Menschen und KI.
- *Gerechtigkeit (justice):* Wohlstand (fair) fördern, Solidarität bewahren, Diskriminierung und weitere Schäden verhindern.
- *Erklärbarkeit (explicability):* Ermöglichung der anderen Prinzipien durch Verständlichkeit, Transparenz und Verantwortlichkeit.

Die Wohltätigkeit bzw. das Erschaffen von KI-Lösungen, die einen Nutzen für die Menschheit bietet, steht an oberster Stelle dieser fünf Prinzipien (Floridi et al. 2018).

Bei einem Vergleich der beiden Modelle fällt auf, dass Nicht-Boshaftigkeit und Privatsphäre eine große Überdeckung zeigen. Des Weiteren kann der Aspekt Genauigkeit unter das Prinzip Gerechtigkeit gefasst werden, denn nur durch genaue, authentische Daten bzw. unvoreingenommen aufbereitete Datensätze besteht die Grundvoraussetzung, dass eine KI einen fairen und nicht-diskriminierenden Output liefert (vgl. Mayer et al. 2021). Zudem überschneidet sich Genauigkeit mit dem Prinzip Erklärbarkeit, da in beiden Fällen die notwendige Klärung von Verantwortlichkeit genannt wird (vgl. Floridi et al. 2018; Mason 1986).

In einer weiteren Studie haben Jobin et al. (2019) insgesamt 84 Dokumente, die von nationalen und internationalen Kommitees veröffentlicht wurden, untersucht. Darunter befinden sich auch die sechs Dokumente, die von Floridi et al. (2018) betrachtet wurden. Jobin et al. (2019) identifizieren insgesamt elf ethische Prinzipien,

die sich teils überschneiden und in der Häufigkeit ihres Vorkommens stark variieren (das meistgenannte Prinzip taucht 73 mal auf, während das elfte Prinzip lediglich sechsmal genannt wird). In ihren Ergebnissen legen sie weiter dar, dass bei fünf ethischen Prinzipien weitestgehend eine globale Übereinstimmung festzustellen ist. So erscheinen die Prinzipien *Transparenz, Gerechtigkeit und Fairness, Nicht-Boshaftigkeit, Verantwortlichkeit* sowie *Privatsphäre* als zentral, wobei es inhaltliche Unterschiede hinsichtlich ihrer Definition, Interpretation, Wichtigkeit und Umsetzungsweise gibt. Dies verdeutlicht, wie wichtig es ist, dass die Entwicklung von ethischen Richtlinien nicht getrennt von einer ethischen Analyse und angemessenen Umsetzungsstrategien betrachtet wird (Jobin et al. 2019). Das nach Jobin et al. (2019) wichtigste – weil am häufigsten vorkommende – Prinzip Transparenz gleicht dem Prinzip der Erklärbarkeit. Wohltätigkeit wird nicht als zentrales Prinzip erachtet, da es in nicht einmal der Hälfte der Dokumente genannt wird. Einigkeit besteht hingegen hinsichtlich der Prinzipien Gerechtigkeit (und Fairness) sowie Nicht-Boshaftigkeit. Während von Jobin et al. (2019) Verantwortlichkeit als eigenständiges Prinzip identifiziert wurde, wird diese bei den anderen beiden Modellen implizit bei Genauigkeit bzw. Erklärbarkeit genannt. Das Prinzip der Privatsphäre wird sowohl von Mason (1986) als auch Jobin et al. (2019) als zentral benannt. Obwohl es einige Überschneidungen gibt, zeigt die Vielfalt der Grundsätze, dass es bislang keinen Konsens über einen ordnenden Rahmen zur KI-Ethik gibt.

## 5 Modell zu KI-Ethik-Prinzipien und Handlungsempfehlungen

Ausgehend von der Synthese der drei vorgestellten Arbeiten zu ethischen Prinzipien zur Nutzung von KI wird nachfolgend ein Vorschlag für ein integriertes Modell (siehe Abb. 2) präsentiert, das sechs KI-Ethik-Prinzipien beinhaltet. Diese integrieren die vier Prinzipien des PAPA-Modells, die fünf Prinzipien des AI4People-Modells sowie die fünf von Jobin et al. (2019) als zentral betrachteten Prinzipien zu den Folgenden: *Wohltätigkeit, Transparenz, Nicht-Boshaftigkeit, Autonomie, Gerechtigkeit* und *Datenschutz* (siehe Tab. 1). Diese sechs Prinzipien werden grundsätzlich als

**Tab. 1** Zusammenfassung von KI-Ethik-Prinzipien (eigene Darstellung)

| KI-Ethik-Prinzip | Herleitung |
|---|---|
| Wohltätigkeit | Wohltätigkeit (Floridi et al. 2018; Jobin et al. 2019) |
| Transparenz | Verantwortlichkeit (Jobin et al. 2019), Genauigkeit (Mason 1986) und Erklärbarkeit (Floridi et al. 2018) |
| Nicht-Boshaftigkeit | Nicht-Boshaftigkeit (Floridi et al. 2018; Jobin et al. 2019) und Privatsphäre (Mason 1986) |
| Autonomie | Autonomie (Floridi et al. 2018) |
| Gerechtigkeit | Gerechtigkeit (Floridi et al. 2018) und Genauigkeit (Mason 1986) |
| Datenschutz | Privatsphäre (Jobin et al. 2019; Mason 1986), Eigentum und Zugänglichkeit (Mason 1986) |

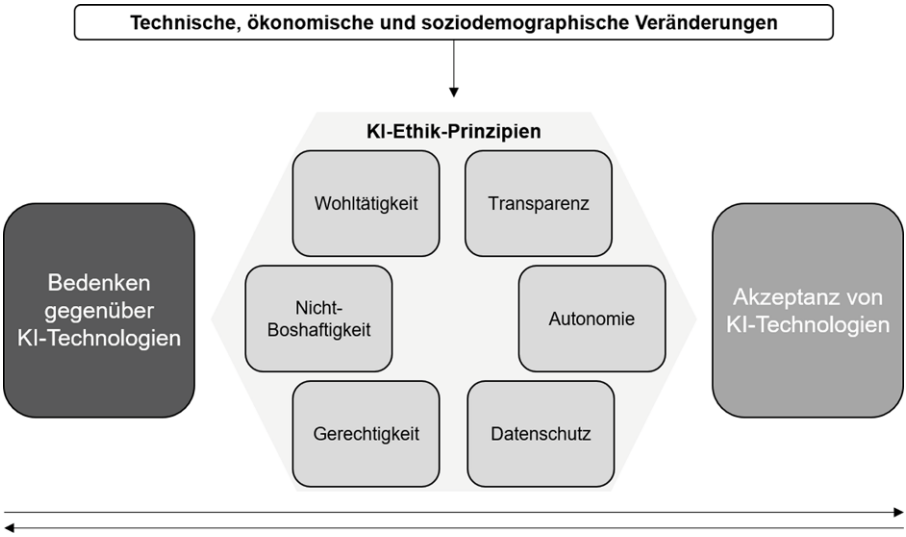


**Abb. 2** Ethik-Prinzipien im KI-Zeitalter. (Eigene Darstellung in Anlehnung an Floridi et al. 2018; Jobin et al. 2019; Manzeschke 2021; Mason 1986)

gleichberechtigt angesehen, können aber je nach Anwendungsszenario unterschiedlich gewichtet sein (vgl. Formosa et al. 2021).

Wohltätigkeit meint, dass KI zur Förderung des Guten, d. h. des menschlichen Wohlbefindens, des wirtschaftlichen Wohlstands, des Friedens usw. entwickelt und eingesetzt werden soll. Transparenz soll gewährleisten, dass die Interessensgruppen der KI-Technologie über Verantwortlichkeiten wie auch die Genauigkeit des Dateninputs, des ML-Algorithmus und des KI-Outputs aufgeklärt werden. Weiter wird mit der Nicht-Boshaftigkeit der Forderung nach Sicherheit und dem Vermeiden von Schäden nachgegangen. Die Einhaltung des Prinzips der Autonomie bedeutet, zwischen der menschlichen Entscheidungsbefugnis und Entscheidungsmacht, die der KI zugestanden wird, ein Gleichgewicht zu finden, sodass die Autonomie des Menschen nicht nachteilig eingeschränkt wird. Gerechtigkeit heißt, dass die KI gerecht, fair und gleichberechtigt funktioniert sowie unerwünschte Voreingenommenheit, Diskriminierung und Benachteiligung nicht vorkommt. Das sechste Prinzip Datenschutz vereint die Punkte datenbezogene Privatsphäre, Dateneigentum und die Zugänglichkeit zu personengebundenen Daten (vgl. Floridi et al. 2018; Jobin et al. 2019; Mason 1986). Des Weiteren werden KI-Ethik-Prinzipien durch technische, ökonomische und soziodemographische Veränderungen in der Gesellschaft beeinflusst, sodass sie nicht als starr und endgültig angenommen werden dürfen (vgl. Manzeschke 2021). Letztlich trägt die Einhaltung der Prinzipien dazu bei, dass ethische Bedenken und Risiken gegenüber der KI-Technologien zwar sichtbar, aber auch entkräftet werden, wodurch KI erst Akzeptanz in der Gesellschaft findet (vgl. Floridi et al. 2018).

Das dargestellte Modell (siehe Abb. 2) vermag nicht alle in der Literatur aufgeführten ethischen Prinzipien abbilden. Es stellt daher einen Kompromiss zwischen Übersichtlichkeit und Vollständigkeit dar. Des Weiteren wurde das Phänomen der Akzeptanzbildung vereinfacht aufgenommen. Neben der Erfüllung ethischer Prin-

zipien können weitere Aspekte, wie z. B. die wahrgenommene Nützlichkeit, einen Einfluss auf die Technologieakzeptanz ausüben (vgl. Technology Acceptance Model, kurz TAM, Davis et al. 1989).

Damit KI also eine weitverbreitete Anwendung und Akzeptanz in der Gesellschaft erfährt, muss neben der Verdeutlichung der positiven Potenziale und Vorteile von KI auch aktiv gehandelt werden, um die KI-Ethik-Prinzipien zu erfüllen. Handlungsprinzipien, die auf eine wertorientierte Entwicklung von Informationssystemen abzielen, bietet Spiekermann (2021). Entsprechende Handlungsempfehlungen zur Gewährleistung der KI-Ethik-Prinzipien können wie folgt lauten:

- *Wohltätigkeit*: Zur Realisierung einer „guten" KI ist die Ausrichtung an menschlichen Werten notwendig. Für eine gesellschaftlich nutzenstiftende KI können beispielsweise die 17 Ziele für nachhaltige Entwicklung (Sustainable Development Goals) der Vereinten Nationen, die Mehrwerte auf ökonomischer, sozialer und ökologischer Ebene verfolgen, herangezogen werden. Ebenfalls zum Wohlbefinden der Gesamtgesellschaft können eine minimierte Machtkonzentration, der positive Einsatz von Macht (z. B. von Großkonzernen und Regierungen) zur Einhaltung von Menschenrechten sowie die enge Zusammenarbeit mit betroffenen Menschen beitragen (vgl. Astobiza et al. 2021; Jobin et al. 2019).
- *Transparenz*: Relevante Kontextinformationen einer KI-Lösung müssen verständlich und zugänglich gehandhabt werden. Frühestmöglich sollten Verantwortlichkeiten, Datenquellen, -arten und -verarbeitung geklärt sowie dauerhaft festgehalten werden, damit die Interessensgruppen möglichst barrierefrei ein Verständnis von der KI erhalten können (vgl. Jobin et al. 2019; McBride 2014).
- *Nicht-Boshaftigkeit*: Maßnahmen zur Gewährleistung von Sicherheit und Vermeidung von Schaden fußen vornehmlich in technischen Maßnahmen, wie z. B. eingebaute Datenqualitätsauswertungen. Auch von der Regierung veranlasste Interventionen auf den Ebenen der KI-Forschung, -Entwicklung und -Einsatz, bis hin zu lateralen und kontinuierlichen Kontrollen tragen zur Sicherstellung bei. Unternehmen könnten aktiv um derartige Kontrollen bitten und somit der Gesellschaft das ordentliche Vorgehen signalisieren (vgl. Jobin et al. 2019).
- *Autonomie*: Die Entscheidungsbefugnis sowie der Handlungsfreiraum in der Mensch-KI-Interaktion sollte den Fähigkeiten des Menschen entsprechend zugeteilt werden. Gleichzeitig ist darauf zu achten, den KI-Anwender:innen nicht zu viel Kontrolle abzusprechen, damit sie nicht von der KI abhängig und in ihren Entscheidungs- und Handlungsspielräumen zu stark eingeengt werden. Ständige Fortbildungen erlauben es dem Menschen, weiter die Entscheidungshoheit zu behalten (vgl. McBride 2014).
- *Gerechtigkeit*: Die Erhaltung und Förderung von Fairness können beispielsweise durch technische Normen, transparente Vorgehen und eine Sensibilisierung der Öffentlichkeit erzielt werden. Die diverse Zivilgesellschaft sollte in ihrer Breite bei der KI-Entwicklung eingebunden und in den Daten abgebildet sein. Rechtsstaatliche Kontrollen tragen ebenso zur Wahrung dieses Prinzips bei (Floridi et al. 2018; Jobin et al. 2019).
- *Datenschutz*: Im Sinne des Datenschutzes müssen sich Anbieter von KI in Deutschland an die Datenschutz-Grundverordnung (DSGVO) und in anderen

Ländern an ähnliche Verordnungen halten. In Datenschutzrichtlinien und den allgemeinen Geschäftsbedingungen (AGB) wird i. d. R. über die datenbezogene Privatsphäre, Dateneigentum und die Zugänglichkeit zu personengebundenen Daten informiert (vgl. Spiekermann 2021).

Letzten Endes ist es wichtig, ein Gleichgewicht zwischen den Interessen von Einzelpersonen, Unternehmen und der Gesamtgesellschaft zu wahren. Durch eine international einvernehmliche und gleichzeitig landesspezifische Integration ethischer Prinzipien in Rechtsvorschriften, ähnlich wie es durch die DSGVO auf europäischer Ebene bereits geschehen ist, vermag ein derartiges Gleichgewicht verbindlich zu regeln oder zumindest Wege dorthin aufzuzeigen. Zu große, exzessive Restriktionen können allerdings die Entwicklung von KI-Technologien verlangsamen oder gar verhindern (vgl. Sidorenko et al. 2021).

## 6 Fazit

KI, verstanden als die technische Simulation menschlicher Intelligenz, bietet so wie andere revolutionäre Technologien Vorteile wie Nachteile. Die Nutzung von KI-Technologien kann z. B. zur menschlichen Selbstverwirklichung beitragen, die menschliche Handlungsfähigkeit stärken und in Effektivitäts- wie auch Effizienzsteigerungen resultieren. Deshalb findet KI auch in zahlreichen Branchen bereits wachsende Anwendung. Dem gegenüber stehen Risiken und ethische Bedenken, die die Erhebung und Verarbeitung personenbezogener Daten (Datenmissbrauch, Diskriminierung usw.) oder den KI-Output und seine Übereinstimmung mit allgemeinen gesellschaftlichen Normen und Werten betreffen. Damit KI nachhaltig Anwendung und Akzeptanz in der Gesellschaft findet, müssen die Vorteile als sinnvoll und die Risiken als möglich, aber vermeidbar oder minimierbar erachtet werden. Es muss ein angemessenes Maß der KI-Nutzung gefunden werden. Um den Bedenken zu entgegnen und Akzeptanz zu schüren müssen ethische Prinzipien bei der KI-Entwicklung und Anwendung eingehalten werden. Ausgehend von dem PAPA-Modell, dem AI4People-Modell sowie der Studie von Jobin et al. (2019) wurden sechs zentrale ethische Prinzipien im KI-Zeitalter ausgewählt: *Wohltätigkeit, Transparenz, Nicht-Boshaftigkeit, Autonomie, Gerechtigkeit* und *Datenschutz*. Zur Beachtung dieser Prinzipien gilt es praktische Maßnahmen umzusetzen, wie z. B. die transparente Darlegung von KI-Kontextinformationen oder die Sensibilisierung der Betroffenen hinsichtlich der KI-Ethik-Thematik. Es ist anzumerken, dass die in diesem Beitrag dargelegten Bedenken, Prinzipien und Handlungsempfehlungen keinen Anspruch auf Vollständigkeit erheben. Der Beitrag ist u. a. dadurch limitiert, dass lediglich drei Arbeiten hauptsächlich für die konzeptionelle Herleitung eines integrierten Vorschlags mit sechs ausgewählten Prinzipien betrachtet wurden. Nachfolgende Forschungsarbeiten könnten eine deutlich umfangreichere und systematische Literaturrecherche und -analyse mit Blick auf bereits veröffentliche Ethik-Modelle speziell für KI durchführen. Außerdem ist eine empirische Arbeit, in Form einer Befragung verschiedener Interessensgruppen zu den Erwartungen an KI-Ethik-Prin-

zipien sowie zur Priorisierung dieser denkbar, z. B. unter Zuhilfenahme des Kano-Modells zur Unterscheidung von unterschiedlich gewichteten Anforderungen.

**Förderung** Die Arbeit entstand im Kompetenzzentrum HUMAINE, gefördert durch das Bundesministerium für Bildung und Forschung (BMBF; Förderkennzeichen: 02L19C200) im Programm „Zukunft der Wertschöpfung – Forschung zu Produktion, Dienstleistung und Arbeit".

**Funding** Open Access funding enabled and organized by Projekt DEAL.



## Literatur

Akerkar R (2019) Artificial intelligence for business. Springer, Cham

Albayrak CA, Ren O, Teille K (2018) Leitlinien für das menschliche Handeln in einer digitalisierten Welt. HMD Prax Wirtschaftinform 55:1048–1064. https://doi.org/10.1365/s40702-018-00450-0

Astobiza AM, Aparicio M, Toboso M, López D (2021) AI ethics for sustainable development goals. IEEE Technol Soc Mag 40(2):66–71

Blackman R (2020) A practical guide to building ethical AI. Harvard Business Review. https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai. Zugegriffen: 9. Sept. 2021

Canhoto AI, Clear F (2020) Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. Bus Horiz 63(2):183–193. https://doi.org/10.1016/j.bushor.2019.11.003

Davis F, Bagozzi P, Warshaw P (1989) User acceptance of computer technology—A comparison of two theoretical models. Manage Sci 35(8):982–1003. https://doi.org/10.1287/mnsc.35.8.982

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Yayena E (2018) AI4People—An ethical framework for a good AI society: Opporunities, risks, principles, and recommendations. Mind Mach 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

Formosa P, Wilson M, Richards D (2021) A principlist framework for cybersecurity ethics. Comput Secur 109:1–15. https://doi.org/10.1016/j.cose.2021.102382

Goasduff L (2021) Gartner says 70 % of organizations will shift their focus from big to small and wide data by 2025. https://www.gartner.com/en/newsroom/press-releases/2021-05-19-gartner-says-70-percent-of-organizations-will-shift-their-focus-from-big-to-small-and-wide-data-by-2025. Zugegriffen: 25. Jan. 2022

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2

Kaplan A, Haenlein M (2019) Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz 62(1):15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kreutzer RT, Sirrenberg M (2019) Künstliche Intelligenz verstehen. Grundlagen – Use-Cases – unternehmenseigene KI-Journey. Springer Gabler, Wiebaden

Manzeschke A (2021) Digitalisierung und Organisationsethik. Ethische und technikphilosophische Skizzen. Ethik Med 33:219–232. https://doi.org/10.1007/s00481-021-00630-5

Mason R (1986) Four ethical issues of the information age. MIS Q 10(1):5–12. https://doi.org/10.2307/248873
Mayer AS, Haimerl A, Strich F, Fiedler M (2021) How corporations encourage the implementation of AI ethics. ECIS 2021 research papers, Bd. 27
McBride NK (2014) ACTIVE ethics: An information systems ethics for the internet age. J Inf Commun Ethics Soc 12(1):21–43. https://doi.org/10.1108/JICES-06-2013-0017
Mingers J, Walsham G (2010) Towards ethical information systems: The contribution of discourse ethics. MIS Q 34(4):833–854. https://doi.org/10.2307/25750707
Radhakrishnan J, Gupta S (2020) Artificial intelligence in practice—Real-world examples and emerging business models. In: Sharma SK, Dwivedi YK, Metri B, Rana NP (Hrsg) Re-imagining diffusion and adoption of information technology and systems: A continuing conversation. Springer, Cham
Rothenberger L, Fabian B, Arunov E (2019) Relevance of ethical guidelines for artificial intelligence—A survey and evaluation. In: Twenty-seventh European Conference on Information Systems (ECIS 2019), Stockholm-Uppsala, Schweden
Schmidt T (2011) Deontologische Ethik. In: Stoecker R, Neuhäuser C, Raters ML (Hrsg) Handbuch Angewandte Ethik. J. B. Metzler, Stuttgart, S 43–49
Schramme T (2011) Tugendethik. In: Stoecker R, Neuhäuser C, Raters ML (Hrsg) Handbuch Angewandte Ethik. J. B. Metzler, Stuttgart, S 49–53
Schroth J (2011) Konsequentialistische Ethik. In: Stoecker R, Neuhäuser C, Raters ML (Hrsg) Handbuch Angewandte Ethik. J. B. Metzler, Stuttgart, S 49–53
Sena V, Nocker M (2021) AI and business models: The good, the bad and the ugly. Foundations and trends in technology. Inf Oper Manag 14(4):324–397. https://doi.org/10.1561/0200000100
Sidorenko EL, Khisamova ZI, Monastyrsky UE (2021) The main ethical risks of using artificial intelligence in business. In: Ashmarina SI, Mantulenko VV (Hrsg) Current achievements, challenges and digital chances of knowledge based economy. Springer Nature, Cham, S 423–429
Spiekermann S (2021) Value-based Engineering: Prinzipien und Motivation für bessere IT-Systeme. Inform Spektrum 4:247–256. https://doi.org/10.1007/s00287-021-01378-4
Vermanen M, Rantanen MM, Harkke V (2019) Ethical challenges of IoT utilization in SMEs from an individual employees perspective. In: Twenty-seventh European Conference on Information Systems (ECIS 2019), Stockholm-Uppsala, Schweden