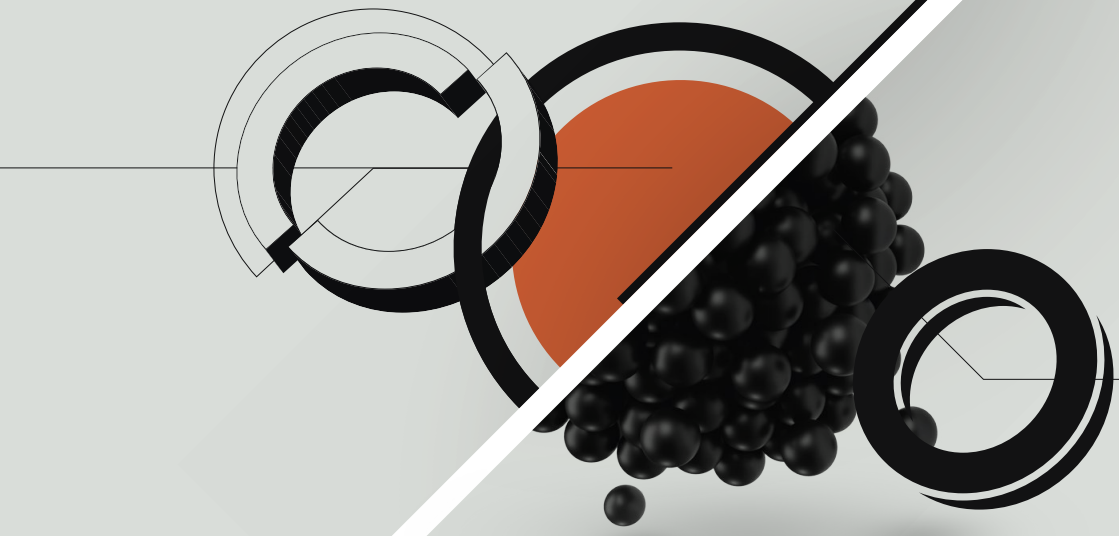
zv
ki

Wie gestalten wir vertrauenswürdige Künstliche Intelligenz?

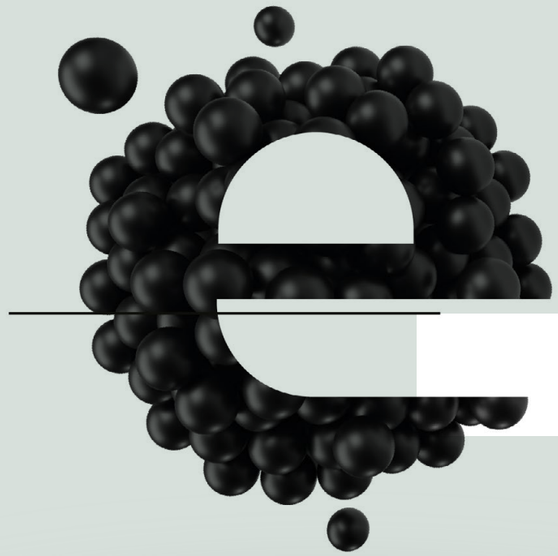
Error 404



Heft #1 / Mai 2022

Missing Link

Magazin für vertrauenswürdige Künstliche Intelligenz



EDITORIAL

WIE GESTALTEN WIR VERTRAUENSWÜRDIGE KÜNSTLICHE INTELLIGENZ?

Vertrauenswürdige Künstliche Intelligenz (KI) ist etwas, das zutiefst mit uns Menschen verbunden ist.

Die Frage, wie wir Künstliche Intelligenz so gestalten können, dass wir sie als vertrauenswürdig beurteilen, provoziert zunächst einmal Rückfragen: Was ist Vertrauen? Wozu ist es da? Vertrauen entsteht in einem Prozess zwischen Menschen. Dadurch reduzieren wir Komplexität, um handlungsfähig zu bleiben. Wir bauen Vertrauen auf Grundlage von Wissen auf, aber nicht auf Basis vollumfänglicher Informiertheit. (Ob es Letztere überhaupt geben kann und ob wir bei Entscheidungen stets Zugriff auf alle Informationen haben, ist eine Frage, die wir an anderer Stelle diskutieren werden.) Um zu erfassen, wodurch sich vertrauenswürdige KI-Software auszeichnet, müssen wir uns ebenfalls fragen: Was macht KI-Systeme aus? KI ist eine Technologie, die wir einsetzen, um bestimmte Vorgänge zu automatisieren. Entwickler:innen gestalten sie auf Basis von Annahmen über die Realität und ziehen dafür bestimmte Daten aus der Vergangenheit heran. Dies geschieht im Rahmen existierender gesellschaftlicher Strukturen und innerhalb sozialer Kontexte. (Entwickler:innen sind übrigens nicht nur die, die den Code schreiben. Vielmehr sind das alle, die an der Entwicklung von KI-Systemen beteiligt sind.)

Wenn wir diese ersten Hinweise auf Vertrauen und KI-Systeme auf uns wirken lassen, wird eines deutlich: Vertrauenswürdige KI ist etwas, das zutiefst mit uns Menschen verbunden ist – etwas, das nicht losgelöst von uns und den Strukturen existiert, die wir schaffen. Wenn wir ein KI-System hinsichtlich seiner Vertrauenswürdigkeit betrachten möchten, müssen wir uns demnach auch die Strukturen ansehen, in denen dieses System entsteht und wirkt.

Es gibt noch eine weitere Gemeinsamkeit von KI-Systemen und ihrer Entwicklung einerseits sowie Vertrauen und seiner Zuschreibung andererseits. In beiden Fällen reduzieren wir Komplexität. Wenn wir ein KI-System hinsichtlich seiner Vertrauenswürdigkeit erfassen wollen, müssen wir diese Komplexität anerkennen und zu einem gewissen Grad rekonstruieren, um sie im Anschluss wieder dekonstruieren zu können. Das klingt zunächst abstrakt, nach komplizierter Theorie, nach Philosophieren und überhaupt nicht nach Praxis. Letztere wünschen sich einige, um ins Gestalten vertrauenswürdiger KI-Systeme zu kommen.

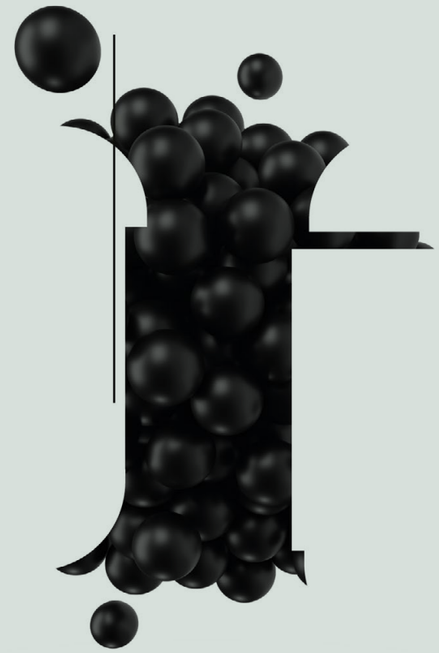
Vertrauen Sie uns: Es wird auf den nächsten Seiten auch konkret. Wir wandern zwischen den Polen der Abstraktion und Konkretion hin und her, um uns Antworten auf die Frage „Wie gestalten wir vertrauenswürdige KI?“ zu nähern. Das geschieht mithilfe von Definitionen, Zahlen und Darstellungen dessen, wie sich Akteur:innen aus Wirtschaft, Politik, Zivilgesellschaft und Wissenschaft dieser Frage nähern.

Warum machen wir das? Eine Antwort hierauf ist schnell gefunden: Die Verwobenheit von KI-Systemen mit uns Menschen bedeutet, dass sie Auswirkungen auf unseren Alltag, auf unser Leben als Individuen und in Gesellschaften, als Verbraucher:innen und Bürger:innen haben. Diese Effekte sind ein Resultat dessen, wie KI-Systeme gestaltet sind. Es ist demnach notwendig, eine Verbindung zwischen dem Gestalten von KI-Systemen und ihren Wirkweisen herzustellen. Dabei betrachten wir nicht einfach Technik, sondern Menschen und ihr Zusammenleben.

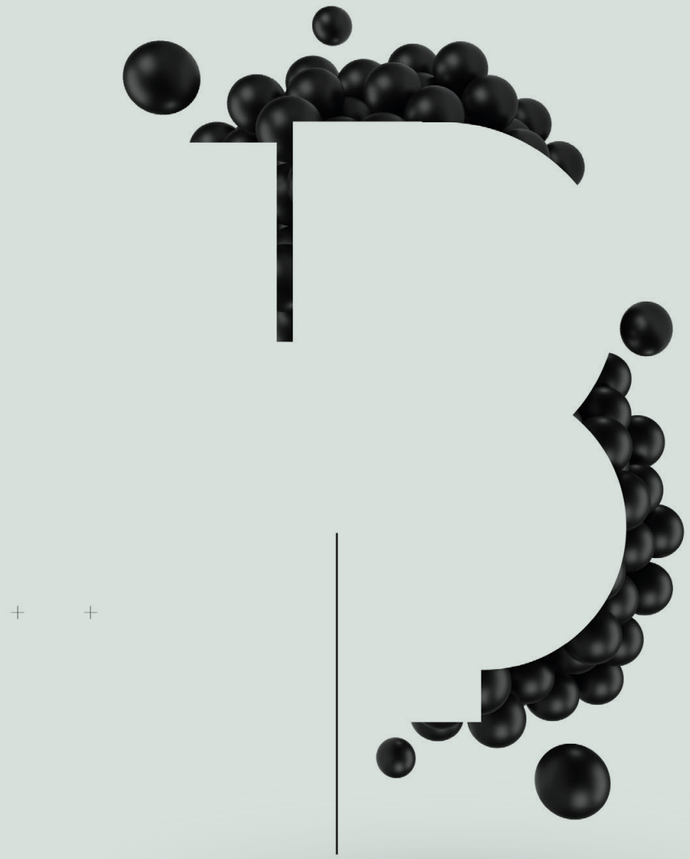
Spoiler-Alert: Trotz der formulierten Notwendigkeit, Komplexität rekonstruieren zu müssen, reduzieren wir sie in dieser Ausgabe ebenfalls. Die Rekonstruktion ist Teil der Arbeit des *Zentrums für vertrauenswürdige Künstliche Intelligenz (ZVKI)* und findet im Rahmen verschiedener Schwerpunkte statt. In diesem Magazin stellen wir Erkenntnisse dieser Arbeit und zentrale Fragestellungen vor. Die kommenden Seiten fügen sich demnach in ein Gesamtwerk ein, das Komplexität anerkennen möchte, um mit ihr umzugehen.

Mitmachen
Das ZVKI versteht sich als neutrale Schnittstelle zwischen Disziplinen und Akteur:innen, zwischen Nutzer:innen und Expert:innen. Treten Sie mit uns in Kontakt und in den Austausch. Sie erreichen uns per E-Mail unter zvki@irights-lab.de.

Mehr über unsere Aktivitäten und Themen erfahren Sie auf unserer Webseite www.zvki.de.



INHALTE



BENENNEN

WAS IST VERTRAUEN? WAS IST VERTRAUENSWÜRDIGE KI?

Was wir meinen, wenn wir von Vertrauen und Vertrauenswürdigkeit sprechen.

Vertrauen kann als Zustand zwischen Wissen und Nichtwissen definiert werden und stellt einen Mechanismus dar, um Komplexität zu reduzieren. Mithilfe von Vertrauen sind wir in der Lage, Situationen einzuschätzen, um Entscheidungen zu treffen.[1]

Vertrauen ≠ blindes Vertrauen

Vertrauen ist nicht gleichzusetzen mit „blindem Vertrauen“, das sich jeder Wissensgrundlage entzieht. Im skizzierten Verständnis basiert Vertrauen auf einem bestimmten Grad des Wissens. Alles andere wäre Hoffnung.[2]

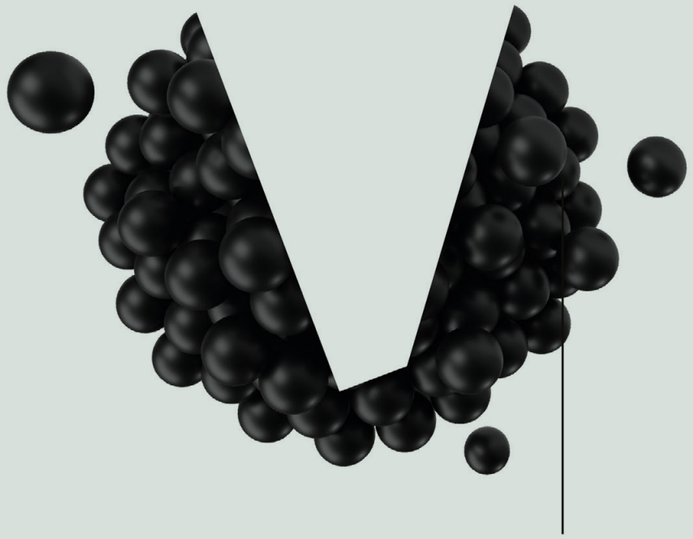
Vertrauen ist in der Regel eine Eigenschaft, die wir Beziehungen zwischen Menschen zuschreiben. Mithilfe von Vertrauen können wir zusammenarbeiten.[3] Es wird schrittweise aufgebaut und erfordert ständige wechselseitige Interaktionen.[4] Demnach entwickelt sich Vertrauen in einem Prozess zwischen Menschen auf Basis eines Zustands zwischen Wissen und Nichtwissen. Dieses Verständnis macht Vertrauen in Bezug auf KI-Technologien nicht zu einem mangelhaften Konzept, sondern rückt vielmehr die Akteur:innen und Beteiligten, die KI ersinnen, entwickeln, verkaufen, kaufen, nutzen und regulieren, in den Fokus. Vertrauen, das zwischen diesen Akteur:innen entstehen und aufrechterhalten werden kann, ist die Grundlage für vertrauenswürdige KI.

Das Wissen, auf dem Vertrauen zu einem gewissen Grad basiert, bezieht sich auf die beteiligten Akteur:innen, aber auch auf die Anwendungen und ihre Eigenschaften. Nutzer:innen machen die Vertrauenswürdigkeit einer Maschine beziehungsweise eines automatisierten Systems unter anderem an ihrer Beschaffenheit und ihrer technologischen Verwundbarkeit fest.[5] Dazu zählen beispielsweise die Sicherheit eines Systems und der Schutz der verwendeten Daten. Neben der Beurteilung der Eigenschaften der Software sind auch der Einsatzzweck und -kontext sowie das soziokulturelle Umfeld und persönliche Eigenschaften der Beteiligten bedeutsam dafür, dass Vertrauen entstehen kann.[6]

FAQ: Was hat Vertrauen mit KI-Software zu tun?
Da KI-Anwendungen sehr komplex sind, können wir als Nutzer:innen sie nicht bis ins Detail verstehen und bewerten. Das bedeutet aber nicht, dass wir ihnen blind vertrauen müssen. Um Vertrauen in KI-Software zu entwickeln, brauchen wir ein gewisses Maß an Wissen. Wir müssen die Funktionsweise, Einsatzgebiete, Chancen und Herausforderungen sowie die gesetzlichen Regulierungen von KI-Systemen zu einem bestimmten Grad verstehen können.

Außerdem benötigen wir die Sicherheit, dass unsere Rechte gewahrt bleiben, wenn KI zum Einsatz kommt. Maßnahmen und gesetzliche Regulierungen können dazu beitragen, sicherzustellen, dass KI-Systeme unsere Rechte wahren. Zu diesen Maßnahmen gehört beispielsweise, dass KI-Anwendungen nach entsprechenden Kriterien und Anforderungen gebaut werden müssen und dass der rechtmäßige Einsatz von KI-Systemen kontrolliert wird.

Weitere Antworten auf häufig gestellte Fragen gibt es hier: www.zvki.de > KI-Navigator > Unsere Inhalte > FAQs.



VERMESSEN

HALTEN NUTZER:INNEN KI-SOFTWARE FÜR VERTRAUENSWÜRDIG?

Nutzer:innen haben kein Vertrauen in KI-Systeme. Mit dieser Ausgangsthese wird häufig die Notwendigkeit begründet, vertrauenswürdige KI-Systeme zu fördern und entsprechende Maßnahmen dafür zu finden. Lässt sich diese Annahme belegen?

Fachliteratur und empirische Untersuchungen führen in diesen Zusammenhang vor allem die Begriffe ‚algorithmic aversion', ‚algorithmic appreciation' sowie ‚automation bias' an. Das Konzept der ‚algorithmic aversion' geht davon aus, dass Menschen algorithmischen Empfehlungen vor allem dann skeptisch begegnen, wenn damit eine gewisse Unsicherheit verbunden ist. Diese entsteht beispielsweise hinsichtlich der Datengrundlage, die das System nutzt, um persönliche Präferenzen zu ermitteln. Literatur zu ‚algorithmic appreciation' schlüsselt hingegen auf, dass vor allem Lai:innen Ratschläge, die Ergebnisse algorithmischer Entscheidungsprozesse sind, mehr wertschätzen als Ratschläge von Menschen. Der Begriff ‚automation bias' erfasst, dass Menschen dazu neigen können, Entscheidungen weniger zu hinterfragen, wenn sie das Ergebnis von Software sind.[7]

Wie erwünscht sind KI-Systeme in einzelnen Anwendungsfeldern?

Auszug aus den Ergebnissen einer Bevölkerungsbefragung in Deutschland des *Center for Advanced Internet Studies – CAIS* von April 2021[8]

Anwendungsfeld	voll und ganz dafür (Angaben in Prozent)	voll und ganz dagegen (Angaben in Prozent)
in der industriellen Produktion	32,5	2
im Gesundheitswesen	12,7	12,1
im persönlichen Alltag	6,9	10,5
bei politischen Entscheidungen	2,3	42,2
bei Gericht	2,2	40

Frage: Wie erwünscht ist KI in einzelnen Anwendungsfeldern? (n=1.009)

→ KI-Systeme waren bei den Befragten in manchen Anwendungsfeldern erwünschter als in anderen. Die Befragten lehnten den Einsatz von KI-Software vor allem bei politischen Entscheidungen und bei Gericht ab. Im Rahmen der industriellen Produktion wiederum befürworteten deutlich mehr Menschen den Einsatz, als ihn ablehnten. Demnach schienen der Einsatzkontext und -zweck eine wichtige Rolle dabei zu spielen, ob KI-Systeme als wünschenswert gelten.

Was denkt die deutsche Bevölkerung über algorithmische Empfehlungssysteme?

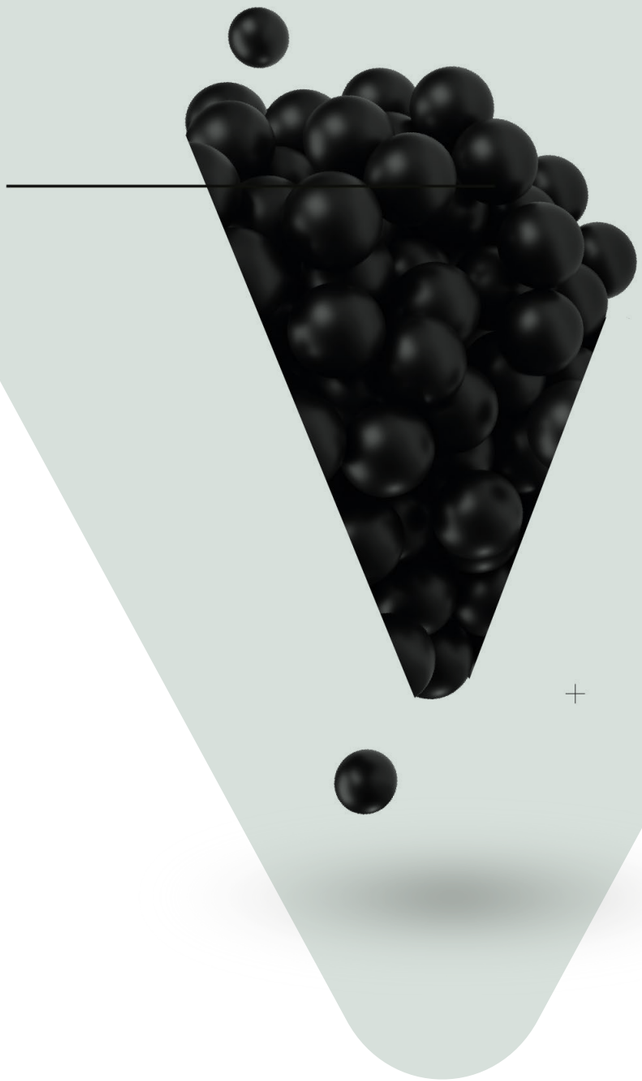
Ergebnisse des „Meinungsmonitors Künstliche Intelligenz" des *CAIS* von Mai 2021[9]

- 26 Prozent der Befragten attestierten algorithmischen Empfehlungssystemen eine nachvollziehbare Funktionsweise.
- 11 Prozent glaubten, dass die Systeme alle Nutzer:innen gleichbehandeln.
- 6 Prozent der Befragten hielten algorithmische Empfehlungssysteme für vertrauenswürdig.
- 67 Prozent sagten hingegen, dass diese Systeme nicht vertrauenswürdig sind.
- Darüber hinaus dachten 26 Prozent, dass algorithmische Empfehlungssysteme die Präferenzen der Nutzer:innen beeinflussen

→ KI-Technologien erschienen der deutschen Bevölkerung laut dieser Umfrage als eher nicht vertrauenswürdig, da sie bestimmte Werte und Prinzipien wie Nachvollziehbarkeit oder Gleichbehandlung als nicht erfüllt wahrnahmen. Chancen für Verbesserungen sahen sie in den Bereichen Regulierung und Kontrolle. Zugleich empfanden die Befragten KI-Technologien als nützlich und ihre Ergebnisse wurden nicht immer hinterfragt.[10]

Bevölkerungsbefragung: Wann beurteilen Nutzer:innen KI-Anwendungen als vertrauenswürdig?
In einer repräsentativen Umfrage möchten wir mehr darüber herausfinden, wie sich Nutzer:innen über KI-Systeme informieren und wann ihnen eine Anwendung vertrauenswürdig erscheint.

Die Ergebnisse der Befragung finden Sie voraussichtlich im Juni auf unserer Webseite www.zvki.de.



VERSTEHEN

WORAN MACHEN WIR VERTRAUENSWÜRDIGKEIT FEST?

Seit 2018 begegnen sich KI-Technologien und ethische Überlegungen zunehmend in Kodizes von Unternehmen, politischen Strategien, zivilgesellschaftlichen Statements und wissenschaftlichen Betrachtungen. Der Begriff der Vertrauenswürdigkeit wird vor allem auf EU-Ebene, aber auch in Deutschland bemüht. Bei Betrachtung der gegenwärtigen Auseinandersetzungen in den Bereichen Wirtschaft, Politik, Zivilgesellschaft und Wissenschaft zeichnet sich eine Herausforderung deutlich ab: Vertrauenswürdigkeit im Sinne bestimmter ethischer Prinzipien umzusetzen und zugleich die komplexen Kontexte zu berücksichtigen, in denen KI-Systeme unser Leben als Individuen und Gesellschaften, als Verbraucher:innen und Bürger:innen maßgeblich beeinflussen. Es lohnt sich ein genauerer Blick.

Wirtschaftliche Praxis: Vertrauenswürdigkeit - mehr als ein Versprechen

Wenn große IT-Konzerne uns versichern, die Welt mit ihren Produkten besser zu machen, reicht das nicht aus, um vertrauenswürdige KI-Anwendungen zu garantieren. Zahlreiche Unternehmen haben das bereits erkannt und beispielsweise Selbstverpflichtungen entwickelt. Sie sollen sicherstellen, dass ihre KI-Systeme bestimmte Rechte und Werte wahren. Vor allem große IT-Unternehmen bringen ihre Perspektiven in aktuelle politische Debatten ein.

Wenn Unternehmen KI-Systeme einsetzen oder anbieten, spielen Vertrauensbildungsprozesse auf zwei Ebenen eine Rolle: gegenüber den Kund:innen und Verbraucher:innen sowie gegenüber den eigenen Mitarbeiter:innen. Fehlendes Vertrauen auf beiden Ebenen kann verhindern, dass ein Unternehmen von der Entwicklung, dem Einsatz beziehungsweise dem Verkauf von KI-Anwendungen profitiert.[11] Je nach Einsatzbereich ist es unterschiedlich, ob das Vertrauen der Kund:innen oder der Mitarbeiter:innen für die Unternehmensziele bedeutsamer ist. Investitionen in vertrauensbildende Maßnahmen erscheinen in beiden Fällen lohnenswert.

Sich selbst Orientierung geben

Um das Vertrauen in die eigenen KI-Anwendungen zu stärken, arbeiten große Unternehmen wie die *Deutsche Telekom* oder *SAP* mit eigenen Ethikkodizes oder Handbüchern. Die *Deutsche Telekom* entwickelte beispielsweise interne Leitlinien für KI-Anwendungen. Sie basieren auf bestimmten Kriterien wie Verantwortlichkeit, Transparenz, Verlässlichkeit und auch auf Vertrauenswürdigkeit.[12] Bereits 2018 verabschiedete das Unternehmen eine Strategie unter den Schlagwörtern „Digital Ethics“ und „Digital Responsibility“.[12] *SAP* veröffentlichte ebenfalls 2018 „Grundsätze für Künstliche Intelligenz“, die beispielsweise festhalten, dass bei der Entwicklung von KI-Software stets die Selbstverpflichtung gilt, die Menschenrechte der *Vereinten Nationen* zu achten. Das Unternehmen hat darüber hinaus einen KI-Ethikrat ins Leben gerufen.[13]

Auch andere Unternehmen wie *Google*[14], *Microsoft*[15], *Bosch*[16] oder *BMW*[17] haben KI-Leitlinien, die Ethikstandards setzen und vor allem das Vertrauen ihrer Kund:innen in ihre Produkte stärken sollen. Im Zusammenhang mit den unternehmerischen Selbstverpflichtungen oder Kodizes bleibt oft unklar, wie die Unternehmen die einzelnen Kriterien genau umsetzen und ihre Produkte diesbezüglich überprüfen. Zudem ist bislang offen, inwiefern Verbraucher:innen die Maßnahmen überhaupt wahrnehmen und als vertrauensfördernd beurteilen.

Für viele Unternehmen spielen vor allem Transparenz, Diskriminierungsfreiheit und Vielfalt sowie Verantwortung und Sicherheit eine Rolle. Um die Bedeutung ethischer Erwägungen zu ermitteln, befragte die *Plattform Lernende Systeme* der *Deutschen Akademie der Technik-*

wissenschaften (acatech) im Jahr 2020 ausgewählte Unternehmen in Deutschland, die Mitglieder der Plattform sind und selbst KI-Systeme entwickeln und einsetzen. Die Befragten sahen in der Umsetzung ethischer Werte kein Wettbewerbshindernis, sondern eher einen Vorteil, der als Verkaufsargument dienen kann.[18]

Wenn der Standard in der Masse untergeht

Solche Maßnahmen der „Corporate Digital Responsibility" – der verantwortungsvollen Digitalisierung in einem ökologischen und sozialen Verständnis[19] – können als eine Säule angesehen werden, um unternehmensseitig das Vertrauen in KI-Anwendungen zu stärken. Ergänzend müssen weitere Maßnahmen und Hilfsmittel greifen.[20] Dazu zählen Ansätze aus dem zivilgesellschaftlichen, wissenschaftlichen und politischen Bereich wie Checklisten, Arbeitsvorlagen oder Leitlinien. Sie sollen den Verantwortlichen in der wirtschaftlichen Praxis Anhaltspunkte geben, wie sie bei der Entwicklung und Anwendung von KI-Software bestimmte ethische Werte berücksichtigen sowie Entwicklungsprozesse und Einsatzkontexte entsprechend gestalten können.[21]

Des Weiteren sollen beispielsweise Standards, Normen und Zertifikate dabei helfen, dass Unternehmen ihre KI-Systeme hinsichtlich ethischer Werte einheitlich überprüfen können. Die „Deutsche Normungsroadmap Künstliche Intelligenz" hält fest, dass Normen und Standards Transparenz und Vertrauen herstellen und die Kommunikation zwischen den Beteiligten durch einheitliche Begriffe und Konzepte erleichtern.[22] Der Ansatz der „Normungsroadmap" basiert auf einer Risikobewertung von KI-Systemen, die eine Einstufung erlauben soll. Je nach Bewertung ergeben sich verschiedene Transparenz- und Nachvollziehbarkeitspflichten.[23]

Neben diesem Ansatz des *Deutschen Instituts für Normung e. V. (DIN)*, der *Deutschen Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE)* und des *Bundesministeriums für Wirtschaft und Energie (BMWi)* gibt es zahlreiche weitere Standardisierungsansätze. Der Bericht „AI Watch" des *Joint Research Centre* der *Europäischen Kommission* stellt allein für das Jahr 2021 die Neuerscheinung von 21 Standards fest. Seit 2019 ist die Zahl an KI-Standards rasant gestiegen. Ein deutlicher Rückgang der Neuerscheinungen ist laut Bericht erst im Jahr 2024 zu erwarten.[24]

Ob diese vielfältige Landschaft an Standards, Leitlinien und anderen Ideen, um vertrauenswürdige KI zu fördern, für Unternehmen gegenwärtig bereits eine klare Orientierung darstellt, erscheint fraglich. So hält Bernd Beckert vom *Fraunhofer-Institut für System- und Innovationsforschung ISI* in einem Forschungsbericht fest, dass es eine große Lücke zwischen dem konzeptionellen Angebot zu vertrauenswürdigen KI-Systemen und der praktischen Umsetzung gibt. Dieses Defizit ist nicht auf Deutschland beschränkt.

FAQ: Welche Rolle spielen ethische Prinzipien bei der KI-Entwicklung und dem Einsatz von KI?
Es gibt seit einigen Jahren eine öffentliche Debatte über die Bedeutung unserer Werte für die Gestaltung von KI-Systemen. Wie sich KI-Systeme verhalten, hängt von den Menschen ab, die sie entwickelt und trainiert haben. Alle Personen, die über den Einsatz von KI-Anwendungen entscheiden und sie gestalten, tragen deshalb eine große Verantwortung.

Aktuell fehlen jedoch noch verbindliche ethische Vorgaben. Die Unternehmen entscheiden selbst, ob und welche ethischen Prinzipien sie bei der Entwicklung von KI-Systemen berücksichtigen. Mit der KI-Verordnung der EU sollen Teile der ethischen Anforderungen verbindlich werden.

Eine weitere Herausforderung besteht darin, dass nicht alle Gestalter:innen von KI-Systemen dazu ausgebildet wurden, ethische Prinzipien zu beachten. Ethik ist bisher noch kein wesentlicher Bestandteil der entsprechenden Studiengänge und Berufsausbildungen, etwa in der Informatik.

Weitere Antworten auf häufig gestellte Fragen finden Sie hier: www.zvki.de > KI-Navigator > Unsere Inhalte > FAQs.

Politische Maßnahmen: Vertrauenswürdigkeit – eine Verpflichtung

Der Entwurf der KI-Verordnung der *Europäischen Union* gilt als Meilenstein in den politischen Debatten, um ethische Prinzipien in die Praxis zu bringen. Das Ziel ist, KI-Systeme verantwortungsvoll zu gestalten, um Vertrauen zu ermöglichen. Die Verordnung könnte zusätzlich Orientierung für einheitliche Standards und darauf basierende Zertifizierungsmaßnahmen bieten. Die Herausforderung dabei ist, dass ethische Prinzipien auf dem Weg in die Umsetzung nicht in den Hintergrund geraten.

Vertrauenswürdige KI ist in den politischen Diskursen um KI-Systeme zu einer besonders häufig bemühten Prämisse herangewachsen.[26] Sie taucht etwa in den „OECD AI Principles"[27] der *Organisation für wirtschaftliche Entwicklung und Zusammenarbeit (OECD)* von 2019 oder im Entwurf der KI-Verordnung der *Europäischen Union (EU)* („Artificial Intelligence Act – AIA"[28]) von 2021 auf. Letztere sieht vor, sichere und vertrauenswürdige KI-Systeme zu gewährleisten, um möglichen Risiken zu begegnen, die mit der Entwicklung, dem Einsatz und der Nutzung solcher Technologien verbunden sein können.[29]

Was als riskant gilt

Der Entwurf der KI-Verordnung der *EU* bezieht Risikobewertungen von KI-Technologien ein. Das Ziel ist, auf dieser Grundlage feststellen zu können, welche KI-Systeme und Anwendungseinsätze mit einem besonderen „Risiko der Schädigung, Beeinträchtigung oder negativer Auswirkungen"[30] einhergehen. Dabei sind vier Risikostufen vorgesehen. Ein unannehmbares Risiko liegt vor, wenn die Grundrechte der *EU*-Bürger:innen verletzt werden. Diese Einstufung führt zu einem Verbot der Systeme. Hochrisiko-KI-Anwendungen stellen „ein hohes Risiko für die Gesundheit und Sicherheit oder für die Grundrechte natürlicher Personen"[31] dar. Der Entwurf schlägt bestimmte Anforderungen vor, die diese Systeme erfüllen müssen. Dazu zählen etwa Transparenz-, Dokumentations- und Informationspflichten. Ein geringes Risiko ist gegeben, wenn beispielsweise eine eindeutige Manipulationsgefahr besteht. Dann sollen spezifische Transparenzpflichten gelten. Alle anderen KI-Systeme stellen laut Entwurf ein minimales Risiko dar. Sie müssen das geltende Recht beachten. Software mit unannehmbarem Risiko verstoßen gegen Grundrechte oder die Werte der EU und sind verboten.[32] Ein Kriterienkatalog soll es ermöglichen, die Risiken einzelner Systeme genauer zu beurteilen.[33]

Der Entwurf des „AIA" der *EU* bezieht sich auf die „Ethik-Leitlinien für eine vertrauenswürdige KI"[34], die die *Hochrangige Expertengruppe für Künstliche Intelligenz (AI HLEG)* bereits 2019 entwickelte. Als zentrale Kriterien für vertrauenswürdige KI definiert die *AI*

KI-SYSTEME POLITISCH ERFASSEN UND STEUERN

Vor allem seit 2018 entstehen auf internationaler Ebene sowie auf EU- und Bundesebene Strategien, Gutachten, Leitlinien und erste Regulierungsentwürfe, um KI-Systeme so zu entwickeln, dass sie ethische Prinzipien berücksichtigen.

2018

November 2018
KI-Strategie der Bunderegierung

Die Strategie definiert drei zentrale Ziele: Deutschland als Standort für KI-Technologien etablieren, verantwortungsvolle und gemeinwohlorientierte KI-Systeme sicherstellen sowie einen gesellschaftlichen Dialog und eine politische Gestaltung von KI-Software verwirklichen, die auch ethische Aspekte berücksichtigen.

April 2019
Ethik-Leitlinien der HLEG AI

Die Europäische Kommission beauftragte im Juni 2018 eine unabhängige Expert:innengruppe damit, ethische Leitlinien für KI-Systeme zu erarbeiten. Die Expert:innen definieren sieben Prinzipien, um vertrauenswürdige KI-Systeme zu entwickeln und einzusetzen.

2019

Mai 2019
OECD-Empfehlung zu KI

Expert:innen erarbeiteten die Empfehlungen als internationales Policy-Instrument in Arbeitsgruppen und bezogen dabei Regierungen, Industrie, Zivilgesellschaft und Gewerkschaften ein. So entstanden Grundsätze für eine verantwortungsvolle Steuerung vertrauenswürdiger KI-Systeme.

Oktober 2019
Gutachten Datenethikkommission

Die Bundesregierung setzte im Juli 2018 die Datenethikkommission ein, die ein Gutachten erarbeitete. Es befasst sich mit der Gestaltung datenbasierter Technologien einschließlich KI-Technologien und nennt ethische Maßstäbe und Leitlinien für den Schutz der Einzelnen und des gesellschaftlichen Zusammenlebens.

2020

Februar 2020
KI-Strategie der EU

Im „Weißbuch zur Künstlichen Intelligenz“ formuliert die Europäische Kommission den Anspruch, die EU-Bürger:innen vor unbeabsichtigten Auswirkungen oder Missbrauch zu schützen sowie dafür zu sorgen, dass die Grundrechte gewahrt bleiben. Außerdem nennt das Weißbuch das Ziel, die EU zu einem zentralen wirtschaftlichen Standort der KI-Entwicklung zu machen.

Dezember 2020
CAHAI-Studie des Europarats

Im September 2019 mandatierte der Europarat ein Ad-hoc-Komitee für Künstliche Intelligenz (CAHAI). Im Dezember 2020 veröffentlichte es die Studie zur Machbarkeit eines rechtlichen Rahmenwerks und seinen möglichen Elementen.

Dezember 2020
Fortschreibung KI-Strategie der Bundesregierung

Die Fortschreibung will verantwortungsvolle und gemeinwohlorientierte KI-Entwicklungen und -Anwendungen zu einem Markenzeichen europäischer KI-Technologien machen. Damit soll ein Ordnungsrahmen für sichere und vertrauenswürdige KI möglich werden.

2021

Februar 2021
Kriterienkatalog für cloudbasierte KI-Systeme des BSI

Das Bundesamt für Sicherheit in der Informationstechnik (BSI) veröffentlichte den Kriterienkatalog „AIC4 (Artificial Intelligence Cloud Service Compliance Criteria Catalogue)“, um die Informationssicherheit von KI-Cloud-Diensten transparent darzustellen und Mindeststandards zu formulieren.

April 2021
Entwurf KI-Verordnung der EU

Der Entwurf für eine KI-Regulierung in der Europäischen Union hat das Ziel, sichere und vertrauenswürdige KI-Systeme zu gewährleisten, möglichen Risiken zu begegnen, die mit der Entwicklung, dem Einsatz und der Nutzung von KI-Technologien verbunden sein können sowie die Grundrechte der EU-Bürger:innen zu wahren. Dabei setzt der Entwurf auf Risikobewertungen.

November 2021
UNESCO-Empfehlungen zur Ethik von KI

Die UNESCO bezeichnet ihre Empfehlungen als ersten globalen Völkerrechtstext zur Ethik von KI, der ein normatives Instrument darstellen soll. Er wurde von 193 Staaten unterschrieben und verknüpft ethische Regeln mit menschenrechtlichen Verpflichtungen.

2022

April 2022
CAI nimmt Arbeit auf

Das Committee on Artificial Intelligence (CAI) nahm als Nachfolger des CAHAI seine Arbeit auf. Es soll bis November 2023 ein Rechtsinstrument für KI-Systeme entwickeln, das auf den Standards für Menschenrechte, Demokratie und Rechtsstaatlichkeit des Europarats basiert.

HLEG den Vorrang menschlichen Handelns und menschlicher Aufsicht, technische Robustheit und Sicherheit, den Schutz der Privatsphäre und Datenqualitätsmanagement, Transparenz, Vielfalt, Nichtdiskriminierung und Fairness, gesellschaftliches und ökologisches Wohlergehen sowie eine Rechenschaftspflicht.[35]

Im Jahr 2020 veröffentlichte die *Europäische Kommission* darauf aufbauend das „Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen“. Es gilt als KI-Strategie der *EU* und ebnete den Weg von den Empfehlungen der *AI HLEG* hin zum Entwurf der KI-Verordnung der *EU*. Das Strategiepapier hält die Notwendigkeit fest, EU-Bürger:innen vor unbeabsichtigten Auswirkungen oder Missbrauch zu schützen und dafür zu sorgen, dass sie ihre Grundrechte wahrnehmen können. Das Weißbuch formuliert zudem das Ziel, die *EU* als Wirtschaftsstandort für KI-Entwicklung zu stärken. Neben fehlenden Investitionen und Kenntnissen ist vor allem das mangelnde Vertrauen ein Grund dafür, dass Bürger:innen KI-Anwendungen nicht ausreichend akzeptieren – so das Weißbuch.[36] Deutschland verabschiedete bereits 2018 eine KI-Strategie, die sie 2020 mit dem Ziel fortschrieb, einen Ordnungsrahmen für sichere und vertrauenswürdige KI zu ermöglichen.[37]

Menschenrechte im Blick und der Blick vieler Menschen

Im September 2019 mandatierte der *Europarat* das *Ad Hoc Committee on Artificial Intelligence (CAHAI)*, das im Dezember 2020 eine Studie veröffentlichte, die die Machbarkeit eines rechtlichen Rahmenwerks und dessen mögliche Bestandteile untersucht. Sie stellt fest, dass bislang kein verbindliches Rechtsinstrument existiert, das speziell an die Herausforderungen angepasst ist, die mit KI-Systemen einhergehen. Die bestehenden und sehr unterschiedlichen Instrumente können keinen umfassenden Schutz der Menschenrechte gewährleisten – so die Studie.[38] Sie stellt darüber hinaus fest, dass dreißig Mitgliedsstaaten Strategien und Maßnahmen für KI-Systeme eingeführt haben. Diese würden das Ziel verfolgen, das Vertrauen in KI-Technologien und ihre Akzeptanz zu stärken, die Kompetenzen für ihre Entwicklung zu fördern sowie die Forschung und die Unternehmen zu unterstützen.

Das *CAHAI* nennt neben der Notwendigkeit eines übergeordneten Rechtsrahmens Zertifizierungssysteme als mögliche Maßnahme, um der Entwicklung und dem Einsatz von KI Menschenrechte sowie ethische Prinzipien vorzuschreiben. Sie könnten einen Rechtsrahmen unterstützen und Unternehmen dazu dienen, Risiken im Zusammenhang mit KI-Software zu bewältigen. Für Expert:innen und Lai:innen wären Zertifikate hilfreich, um die Vertrauenswürdigkeit entsprechender Technologien beurteilen zu können. Die finalen Empfehlungen des *CAHAI* wurden im Dezember 2021 veröffentlicht. Darin spricht sich das Komitee unter anderem dafür aus, einen Multi-Stakeholder-Ansatz zu wählen, um einen Rechtsrahmen zu entwickeln. Darüber hinaus gelte es, die Gesellschaft für die Auswirkungen von KI-Systemen zu sensibilisieren. Dies könne durch öffentliche Beratungen auf Grundlage des „Übereinkommens über Menschenrechte und Biomedizin“ erfolgen.[41] Im April 2022 nahm das *Committee on Artificial Intelligence (CAI)* als Nachfolger des *CAHAI* seine Arbeit auf. Es soll bis November 2023 ein Rechtsinstrument für KI-Systeme entwickeln, das auf den Normen für Menschenrechte, Demokratie und Rechtsstaatlichkeit des *Europarats* beruht.[42]

Mangelnde Flexibilität und blinde Flecken

Gegen einen einheitlichen Ansatz, wie ihn der Entwurf zur KI-Verordnung der *EU* vorsieht, sprechen sich teilweise Unternehmen wie *Google* aus. Grund dafür seien die unterschiedlichen Anwendungsfälle, die dabei nicht ausreichend berücksichtigt würden. Zugleich betont das Unternehmen die Notwendigkeit von Regulierung.[43] Auch der IT-Konzern *Microsoft* hebt die Bedeutung von Regulierung hervor und fordert einen stärker ergebnis- und prinzipienorientierten Ansatz, als es der Entwurf bislang vorsieht. Es gelte, KI-Systeme als soziotechnische Systeme zu erfassen. Zugleich sollte der risikobasierte Ansatz zur Prüfung von

KI-Systemen überdacht werden, der hauptsächlich an den Anbieter:innen ansetze. Diese könnten den Einsatz der Technologien durch ihre Kund:innen nicht immer überprüfen.[44]

Die *Organisation der Vereinten Nationen für Bildung, Wissenschaft und Kultur (UNESCO)* stellt in derzeitigen politischen Papieren blinde Flecken fest, die sie mit ihren eigenen KI-Ethikempfehlungen adressieren möchte. Diese Empfehlungen seien ein grundlegendes normatives Instrument. Damit verfolgen sie keinen vergleichbaren regulativen Ansatz wie der Entwurf des „AIA“ der *EU*. Das Besondere an den *UNESCO*-Empfehlungen sei, dass es sich dabei um den ersten global verhandelten Völkerrechtstext zu KI handele, der von 193 Staaten angenommen wurde.[45] Außerdem wendeten sich die Empfehlungen bislang wenig beachteten Punkten zu, um ethische Regeln mit menschenrechtlichen Verpflichtungen zu verknüpfen. Dazu gehöre beispielsweise, Nichtdiskriminierung und Gleichbehandlung als strukturelle Probleme[46] zu adressieren, KI-Kapazitäten und Datenzugänge im Globalen Süden auszubauen[47] oder Bürger:innen durch moderne Wissensformate zu befähigen und zwar seitens der KI-Akteur:innen[48]. Zudem müsse die Zivilgesellschaft bei Standardisierungsprozessen stärker eingebunden werden.[49]

Zivilgesellschaftliche Perspektiven: sich dem Sozialen zuwenden

Stimmen aus der Zivilgesellschaft richten den Fokus in der Debatte um vertrauenswürdige KI auf strukturelle Machtungleichgewichte und den Charakter von KI-Software als soziotechnische Systeme. Diese Aspekte finden in den Diskursen gegenwärtig kaum Beachtung – ähnlich wie die Perspektiven zivilgesellschaftlicher Organisationen und die Einschätzungen von Betroffenen.

Risiko kennt keine Systemgrenzen

Im November 2021 veröffentlichen 114 zivilgesellschaftliche Organisationen in Europa ein gemeinsames Statement zum Entwurf der KI-Verordnung der *EU*. Darin stellen sie die Grundrechte der Menschen in den Vorder-

grund und bezeichnen den risikobasierten Ansatz des Verordnungsentwurfs als „dysfunktional“. Er berücksichtige nicht, dass bei einer Risikobewertung vor allem der Kontext bedeutsam sei, in dem ein KI-System zum Einsatz kommt. Darüber hinaus sei der Ansatz zu starr und erlaube zu wenig Ergänzungen. Ähnlich wie große IT-Konzerne bemängelt das Statement auch die Tatsache, dass der „AIA“ hauptsächlich Anbieter:innen von KI-Software Verpflichtungen auferlegen möchte und die Kund:innen unberücksichtigt bleiben. Gerade aus der Art und Weise, wie Systeme genutzt würden, könnten sich erhebliche Risiken ergeben. Weiterhin erfasse der Entwurf die Rechte betroffener Menschen nicht ausreichend und beziehe diese ebenfalls nicht in die Untersuchung von Hochrisiko-Systemen ein. Auch Nachhaltigkeitsaspekte und die Frage, wie die Entwicklung und der Einsatz von KI-Systemen ressourcenschonend erfolgen könnten, blieben unberücksichtigt. Ein erster Schritt in diese Richtung wäre, Transparenz in Bezug auf den Ressourcenverbrauch herzustellen, der mit der Entwicklung und dem Betrieb von KI-Software einhergeht.[50]

Expert:innen der *Plattform Lernende Systeme* der *Deutschen Akademie der Technikwissenschaften (acatech)* haben den Entwurf der KI-Verordnung ebenfalls geprüft und schlagen im November 2021 in einem Whitepaper Ergänzungen vor. Sie setzen an unberücksichtigten Punkten an, die für das Entstehen von Vertrauen in KI-Systeme bedeutsam sind. Die Autor:innen empfehlen, dass eine Risikobewertung immer auch daran gemessen werden muss, ob es Handlungsalternativen und -freiheiten gibt, ob Nutzer:innen also auf andere Systeme oder Angebote ausweichen können.[51] Darüber hinaus sprechen sich die Expert:innen dafür aus, genauer zu regulieren, wie die Verantwortung über Haftungsregeln aufgeteilt wird und hierbei auch nach Kritikalität des Anwendungsgebiets zu unterscheiden.[52]

Komplexität anerkennen, um sie zu bewältigen

Der Theologe und Ethiker Peter Dabrock, einer der Autor:innen des Whitepapers der *Plattform Lernende Systeme*, kritisiert das risikobasierte Kritikalitätskonzept

des „AIA“-Entwurfs ebenfalls im Hinblick auf daraus resultierende Zertifizierungsmaßnahmen. Damit werde eine Sicherheit vorgegeben, die ohne parallellaufende nicht technische Maßnahmen nicht existiere. Das Konzept verspreche mehr, als es halten könne, worunter vor allem die Nutzer:innen litten.[53]

Das Statement der 114 zivilgesellschaftlichen Organisationen kritisiert am Entwurf des „AIA“ zudem, dass er sich zu stark an die *EU*-Rechtsvorschriften zur Produktsicherheit anlehnt, um es den Anbieter:innen damit zu erleichtern, die Verordnung einzuhalten. Das würde jedoch dazu führen, dass wichtige politische und rechtliche Entscheidungen an die europäischen Normungsorganisationen delegiert würden, die stark von Akteur:innen aus der Industrie dominiert wären.[54]

Die Politikwissenschaftlerin Leonie Beining beschreibt in einem Impulspapier der *Stiftung Neue Verantwortung* die Herausforderung, dass KI-Technologien eine hohe Forschungs- und Entwicklungsdynamik aufweisen, die aufwendige und eher starre Standardisierungs- und Zertifizierungsprozesse nur schwer abbilden können. Zugleich betont auch sie den soziotechnischen Charakter von KI-Systemen, deren Auswirkungen davon abhängig sind, wer sie zu welchen Zwecken in welchen Kontexten einsetzt. Demnach spielt dieser Aspekt nicht nur bei Regulierungsmaßnahmen eine Rolle, sondern auch im Zusammenhang mit daran angelehnten Standards und Zertifikaten. Es entstehen hohe Ansprüche, denn es geht nicht allein um die Frage, welche ethischen Werte wichtig sind, um Vertrauen in KI-Anwendungen zu stärken und wie sie sich operationalisieren lassen. Vielmehr kommt die Frage hinzu, welche Werte in welcher Situation wichtig sind und wer sie in welcher Situation zuschreibt. Dafür gibt es in unserer Gesellschaft bislang kein geteiltes Wertekonzept – so Beining.[55]

Auch Nutzer:innen und von den Ergebnissen von KI-Software Betroffene finden in den bisherigen Debatten um ethische Aspekte und mögliche Maßnahmen vergleichsweise wenig Gehör. Um KI-Systeme als soziotechnische Systeme zu begreifen und eine gemeinsame Wertebasis auszubilden, ist es notwendig, die Debatte um vertrauenswürdige KI stärker in die Breite zu tragen.

FAQ: Welche ethischen Fragen diskutieren Politik, Öffentlichkeit, Zivilgesellschaft, Wissenschaft und Wirtschaft im Zusammenhang mit KI-Anwendungen?
Viele ethische Fragen beschäftigen sich damit, zu welchem Zweck wir KI einsetzen möchten: Dürfen KI-Systeme zur massenhaften Überwachung von Menschen verwendet werden? Ist es ethisch vertretbar, autonome Waffensysteme gegen Menschen einzusetzen? Bis zu welchem Grad möchten wir medizinische Entscheidungen einer KI-Anwendung überlassen?

Eine zweite Gruppe an Fragen thematisiert, was ethische Werte für die Gestaltung und Kontrolle von KI-Systemen bedeuten: Müssen wir stets darauf hingewiesen werden, dass KI-Anwendungen bestimmte Entscheidungen treffen oder mit uns kommunizieren? Wie können wir sicherstellen, dass der Einsatz eines KI-Systems fair ist?

Diese Fragen betreffen uns alle und müssen daher mit der gesamten Gesellschaft diskutiert werden. Das ZVKI soll unter anderem in diesem Zusammenhang als zentraler Ort der Debatte in Deutschland dienen.

Weitere Antworten auf häufig gestellte Fragen finden Sie hier: www.zvki.de > KI-Navigator > Unsere Inhalte > FAQs.

Wissenschaftliche Erkenntnisse: wie aus Defiziten Ansätze entwachsen

Die Fragen, was vertrauenswürdige KI-Systeme ausmacht und wie wir KI-Software vertrauenswürdig gestalten können, sind ebenso jung wie die Debatte über entsprechende Maßnahmen. Weitere wissenschaftliche Grundlagen erscheinen dringend notwendig, um einen Umgang mit KI-Systemen etablieren zu können, der den formulierten Anliegen gerecht wird.

Im Zusammenhang mit KI-Systemen spielen der Kontext ihrer Entwicklung und ihres Einsatzes eine Rolle sowie die sozialen Strukturen, in denen sie entstehen und wirken. Diese Auffassung ist nicht nur Bestandteil der bereits geschilderten zivilgesellschaftlichen Auseinandersetzungen, sondern auch Ergebnis wissenschaftlicher Betrachtungen. Die Bedeutung dessen, Maßnahmen der Regulierung, Standardisierung und Zertifizierung auf ein breites Fundament zu stellen, wird an dieser Stelle noch einmal deutlich.

Reichtum an Perspektiven fördern

Der IT-Rechtswissenschaftler Martin Ebers äußert in der Analyse „Standardizing AI“ von August 2021 Bedenken in Bezug auf die vorgesehene Art und Weise, wie beziehungsweise durch wen laut „AIA“ Hochrisiko-Systeme als solche bewertet werden sollen. Diese Bewertung erfolge durch die Unternehmen selbst in Kombination mit bestehenden privaten Standardisierungsorganisationen.[56] Kritik an diesem Vorgehen äußern auch einige zivilgesellschaftliche Organisationen. Da es sich bei KI-Systemen nicht um rein technisch zu verstehende Anwendungen handele, sondern um solche, bei denen eine Vielzahl rechtlicher und ethischer Entscheidungen anstünden, könnten die Beurteilungen nicht an Standardisierungsorganisationen ausgelagert werden. Dadurch käme es zu einer fehlenden demokratischen und rechtlichen Kontrolle sowie zu einer unzureichenden Beteiligung verschiedener Interessengruppen. Diese Form der Delegation stelle den „AIA“ auf eine wackelige rechtliche Grundlage.[57] Die Studie des *CAHAI* empfiehlt, partizipatorische Ansätze zu etablieren, wenn KI-Systeme im öffentlichen Sektor

eingesetzt werden sollen. Dabei gelte es vor allem, unterrepräsentierte und schutzbedürftige Personen einzubeziehen. Das könnte Vertrauen in die Technologien und ihre Akzeptanz fördern.[58] Ähnliches gilt laut einem Bericht der *Oxford Commission on AI & Governance* der *University of Oxford* von Dezember 2021 auch für Normungsvorhaben, die aus dem „AIA“ hervorgehen. Demnach müsse es eine substanzielle Beteiligung derjenigen geben, die am meisten auf den Schutz der grundlegenden Menschenrechte angewiesen seien. Frühzeitig veröffentlichte Normungsfahrpläne und nationale KI-Normungszentren könnten dabei helfen, verschiedene Interessengruppen einzubinden. Außerdem seien entsprechende Schulungs- und Weiterbildungsmaterialien für unterschiedliche Interessengruppen notwendig. Darüber hinaus müssten Normen eine entsprechende Flexibilität aufweisen, um die schnellen technologischen Entwicklungen anzuerkennen. Zu diesem Zweck müssten beschleunigte Normungsverfahren entwickelt werden.[59] In Bezug auf Normen und Standards stellt das *Fraunhofer IAIS* in einer vergleichenden Studie von Oktober 2021 fest, dass es lohnend sein kann, bei der Definition von Motivationen und Zielen auch übergeordnete Ansätze zu berücksichtigen, beispielsweise die Nachhaltigkeitsziele der *Vereinten Nationen*.[60]

Einen Schritt zurücktreten, um voranzukommen

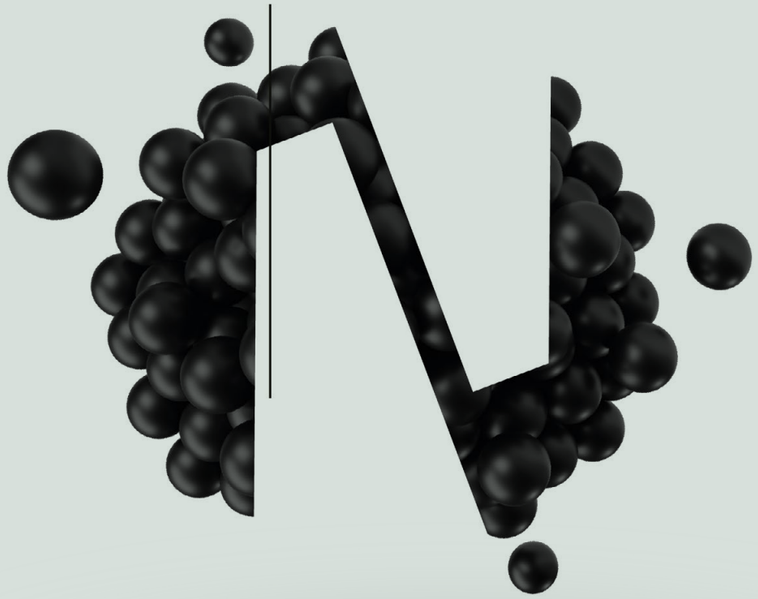
Es ist unerlässlich, die evidenzbasierte Auseinandersetzung damit zu stärken, inwiefern KI- Anwendungen ethische Werte, Rechte und unser Zusammenleben berühren. Die Forschung hierzu steht erst am Anfang.[61] Zielgerichtete Maßnahmen wie eine KI-Verordnung und Ansätze der Normierung und Zertifizierung müssen jedoch auf Forschungsergebnissen gründen, damit sie zu vertrauenswürdigen KI-Systemen beitragen. Dafür erscheinen die bisherigen wissenschaftlichen Erkenntnisse nicht ausreichend. Ähnliches gilt für die Bedingungen, unter denen Nutzer:innen KI-Anwendungen als vertrauenswürdig beurteilen. Die Evidenzbasis ist unter anderem deshalb eingeschränkt, weil bislang große Abhängigkeiten von kleinen Stichproben bestehen und Langzeituntersuchungen fehlen. Darüber hinaus haben Forscher:innen oft keinen Zugang zu relevanten Systemen, ihren Einsatzkontexten und den damit in Verbindung stehenden Daten. Viele Erkenntnisse beruhen auf kurzfristigen experimentellen Forschungsdesigns.[62]

In politischen Diskursen und Leitprinzipien von Unternehmen spielt der Begriff Vertrauen bezüglich KI-Systemen eine wichtige Rolle. Was ein vertrauenswürdiges KI-System ausmacht, wird entweder an teilweise unterschiedlichen Kriterien festgemacht oder bleibt offen. Vertrauen erscheint als Alltagsbegriff, dessen Verständnis oftmals vorausgesetzt wird. In der Auseinandersetzung mit Regulierung, Normung, Zertifizierung und Maßnahmen, die den Diskurs um ethische Prinzipien im Zusammenhang mit KI-Systemen in die Breite tragen sowie Wissen vermitteln sollen, ist es jedoch notwendig, Vertrauenswürdigkeit genauer zu erfassen, um darauf aufbauend Maßnahmen zu ergreifen.

Einblicke in die Arbeit der ZVKI-Partner: Wie können wir zertifizieren?
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS und Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC arbeiten gegenwärtig an einer umfassenden Studie zur Sicherheit von KI-Systemen und untersuchen in verschiedenen Projekten grundlegende Aspekte für eine Zertifizierung. Dabei wenden sie unter anderem einen interdisziplinären Ansatz zwischen Informatik, Rechtswissenschaft, Philosophie und Ethik an. Sie leiten auf dieser Basis sechs Handlungsfelder sowie ethische und rechtliche Anforderungen in Bezug auf KI-Systeme in den Bereichen Fairness, Transparenz, Autonomie & Kontrolle, Datenschutz, Verlässlichkeit und Sicherheit ab. Diese Handlungsfelder können als Grundlage für eine risikobasierte Prüfung von KI-Systemen genutzt werden.

Die Fraunhofer-Institute AISEC und IAIS sowie die Freie Universität Berlin erarbeiten aus verschiedenen Blickwinkeln wissenschaftliche Erkenntnisse zu den Fragen, was vertrauenswürdige KI ausmacht und wie sie gefördert werden kann.

Mehr über unsere Partner finden Sie hier: www.zvki.de > KI-Bibliothek > Partner.



NACHFRAGEN

Lorena Jaume-Palasí ist die Gründerin von *The Ethical Tech Society*, einer gemeinnützigen Nichtregierungsorganisation, die Prozesse der Automatisierung und Digitalisierung erforscht und in Bezug auf ihre gesellschaftliche Relevanz normativ einordnet. Das Europäische Parlament, die Europäische Kommission, Organisationen, Verbände und Regierungen weltweit konsultieren sie als Sachverständige zu ethischen Fragen im Zusammenhang mit KI-Systemen.

WAS VERLIEREN WIR AUS DEM BLICK?

Wir betrachten KI-Systeme generalistisch anhand bestimmter Kategorien, sagt die Philosophin Lorena Jaume-Palasí im Interview. Damit versuchen wir, ihren Einsatz mithilfe von Individualrechten ethisch zu rechtfertigen und handhabbar zu machen. Die gesellschaftliche Dimension des Ganzen gerät damit außer Acht, denn eine Gesellschaft ist mehr als die Summe ihrer Individuen.

Wie nehmen Sie die aktuellen Diskurse in Europa um ethische Fragen im Zusammenhang mit KI-Systemen wahr?
Die europäischen Debatten empfinde ich als kosmetische Diskurse. Sie sind sehr stark von einer europäischen Sichtweise geprägt. Wir denken, wir müssen wettbewerbsfähig bleiben und deshalb KI-Systeme erschaffen, die ethischen und vor allem europäischen Werten entsprechen, als wären sie somit „exportfähiger“. Allein das ist problematisch. Wenn wir Ethik als Strategie für die Wettbewerbsfähigkeit einsetzen, haben wir den Raum des Ethischen verlassen.

Wir wenden einen Jahrhunderte alten, kolonialistischen Blick an, wenn wir KI-Systeme bauen, aber auch normieren wollen: Unsere Wissenschaft, unsere Gesetze und Ethik sind von einer bestimmten Herangehensweise und Methodik geprägt. Wir denken, dass wir eine „objektive“ Rationalität in Maschinen integrieren können und dass diese Maschinen anhand bestimmter Kategorien „objektive“ Handlungen oder Empfehlungen vornehmen können. Diese Idee geht von einem Verständnis von Rationalität aus, das problematisch ist. Es ist so, als könnten wir unabhängig von der eigenen soziokulturellen Prägung, der Prägung der eigenen Hautfarbe etc. die Welt betrachten und daraus die Essenz der Menschen oder der Dinge objektiv identifizieren. Menschen, aber auch Sachen, sind jedoch im ständigen Wandel – in ihren Identitäten, Verständnissen, Kulturen, in ihren soziohistorischen und ökonomischen Gegebenheiten. Sie lassen sich nicht in Schubladen stecken. Der Kontext verändert die Nutzung und Bedeutung von Sachen. Diese Komplexität lässt sich nicht anhand einer Liste von Kategorien abbilden.

Was ist die Konsequenz dieses Blicks auf KI-Systeme?
Wenn wir KI-Technologie auf diese Weise erfassen, adressieren wir die tatsächlichen Probleme nicht. Ein Beispiel für ein solches Problem ist, dass KI-Anwendungen immer wieder zu diskriminierenden Ergebnissen kommen. Die Ursachen sind hier nicht einfach in den KI-Systemen zu suchen. Sie lassen sich auch nicht dadurch lösen, mehr Daten von marginalisierten Kollektiven zu erfassen und zu nutzen. Die Forscherin Zoé Samudzi merkt hierzu in einem Artikel in *The Daily Beast* an: „In einem Land, in dem die Verbrechensbekämpfung Schwarzsein bereits mit inhärenter Kriminalität in Verbindung bringt, ist es kein sozialer Fortschritt, Schwarze für eine Software sichtbar zu machen, die unweigerlich weiter gegen uns eingesetzt wird.“[63] Stattdessen sollten wir die gesamten Kontexte inklusive der Machtasymmetrien betrachten, in denen wir KI-Systeme entwickeln und einsetzen. Eine Ethik jenseits kolonialer Denkmethoden versteht die enorme Bedeutung von Kontext.

> „Wir nehmen die Bürger:innen in die Pflicht, informiert zu sein und auf Basis dessen reflektiert zu entscheiden. Dieses Verlagern von Verantwortung finde ich ethisch problematisch.“

Auch Transparenz würde hier wenig verändern. Mit der Forderung nach Transparenz in diesen machtasymmetrischen Verhältnissen schützen wir niemanden. Wir verlagern im Grunde nur Verantwortung und stellen Misstrauen in den Fokus, statt auf Aufsichtsverfahren zu setzen. Wir nehmen die Bürger:innen in die Pflicht, informiert zu sein und auf Basis dessen reflektiert zu entscheiden. Dieses Verlagern von Verantwortung finde ich ethisch problematisch.

In den europäischen Debatten ziehen wir Menschenrechte oder Grundrechte – also Individualrechte – heran, um KI-Systeme zu regulieren.

Es geht bei der Entwicklung, dem Einsatz und dem Umgang mit KI-Systemen jedoch um gesellschaftliche Infrastrukturen. Sie über Individualrechte zu regulieren, verkennt die gesellschaftliche Dimension. Denn das Gesellschaftliche, das Öffentliche, entsteht nicht aus der Summe aller Einzelinteressen. Es hat einen kollektiven Charakter. Damit lenken wir von den strukturellen Problemen unserer Gesellschaft ab.

Was können wir anders machen?

Wir müssen einen Schritt zurücktreten in unseren Gesamtüberlegungen und erstmal besser verstehen, welche Probleme im Zusammenhang mit KI-Technologie stehen. Wir müssen uns die Ursachen angucken, anstatt die Symptome behandeln zu wollen. Wie gesagt handelt es sich beim Umgang mit KI-Systemen um ein Infrastrukturthema. Das bedeutet, wir organisieren Gesellschaft. Das ist das politischste, was es gibt. Denn wir entscheiden, wer wie und unter welchen Umständen Zugang zu Dingen oder Leistungen erhält. Hier geht es um den Aufbau von Macht und Abhängigkeit. Solche Prozesse geschehen immer auf Kosten bestimmter Individuen und Gruppen. Wir müssen uns also stets fragen: Wen schließen wir gerade aus? Und welche Probleme ergeben sich daraus? Wir benötigen Verfahren, um das zu reflektieren, und Mechanismen, um es zu kompensieren. Dadurch kann eine Infrastruktur legitimer werden. Vor allem bedeutet das, dass wir die Betroffenen einbinden und danach fragen müssen, welche Probleme sie sehen und welcher Umgang damit sinnvoll wäre. Wir ignorieren Betroffene – Be_hinderte, Schwarze, Roma und Sinti, Regenbogenfamilien, Muslime, Sexarbeiter:innen, Migrant:innen etc. Unsere Gesellschaft ist bunt, doch wir denken sie weiß, blond und christlich. Betroffene kennen die Probleme, die strukturell sind und sich auch in KI-Anwendungen fortschreiben, auch wenn sie das Wort Algorithmus nicht definieren können. Denn darauf kommt es auch nicht an.

„Wir müssen uns stets fragen: Wen schließen wir gerade aus? Und welche Probleme ergeben sich daraus? Wir benötigen Verfahren, um das zu reflektieren, und Mechanismen, um es zu kompensieren.“

Diese Forderungen richten sich auch an uns als zivilgesellschaftliche Organisationen. Anstatt Betroffenengruppen zu involvieren, erheben wir Nichtregierungsorganisationen (NGOs) oft Anspruch auf die Deutungshoheit über den sozialen Effekt von KI. In vielen großen Nicht-Regierungsorganisationen sitzen weiße privilegierte Menschen, die beispielsweise über Diskriminierungen sprechen, die sie gar nicht kennen können. Und deren ebenfalls isolierter Blick auf die Überprüfung einzelner Systeme stellt einen verzerrten Blick dar. Das lenkt von den Betroffenengruppen ab, die einen ganz anderen systemischen Blick auf die Probleme haben. Auch die Zivilgesellschaft muss sich an das Prinzip „nichts über uns ohne uns“ halten. Hinzu kommt, dass es in diesem Bereich – anders als bei Regierungen und in der Wirtschaft – keinerlei Gewaltenteilung und Rechtfertigungsmechanismen gibt.

Wir treiben also die Probleme, die wir seit Jahrzehnten beziehungsweise Jahrhunderten geschaffen haben, weiter voran und übertragen sie auf Technologien. Stülpen wir denen ein scheinbar ethisches Konzept über, um uns dann darauf auszuruhen, dass wir gerecht handeln?

Ja. Es gibt zahlreiche Wissenschaftler:innen und Betroffene, die das thematisieren und dieses Wissen öffentlich gemacht haben. Seit mehr als hundert Jahren haben wir zum Beispiel unter den Stichworten Neokolonialismus und Kolonialität Zugang dazu. Doch unsere Organisationsformen – und KI ist eine Art der Organisation – sind geprägt von der eingangs geschilderten Denkweise. Wir schreiben diese Denkweise fort, indem wir sie in KI-Systeme integrieren wollen, um so wettbewerbsfähig zu bleiben. Wir denken dann weiterhin kolonialistisch und erheben diese Denkweise über die anderer Menschen und Kollektive sowie über die Gegebenheiten bei ihnen vor Ort. Wir beanspruchen eine moralische Erhabenheit, die wir gar nicht haben. Mit Blick auf die Zukunft müssen wir uns der Kritik öffnen, zuhören und daraus lernen. Wir haben als kolonialer Kontinent eine moralische Pflicht dazu.

„Wir schreiben diese Denkweise fort, indem wir sie in KI-Systeme integrieren wollen, um so wettbewerbsfähig zu bleiben. Wir denken dann weiterhin kolonialistisch … .“

Welchen Blick auf KI-Technologien wünschen Sie sich?

Ich wünsche mir, dass wir Einsicht zeigen und dass wir KI-Systeme nicht mehr isoliert betrachten, sondern eher als ein Transformationsprojekt ansehen, das multidimensional und multifaktoriell ist. Bislang tun wir das jedoch nicht. Es ist wichtig, dass wir von Anfang an verstehen, dass KI als Infrastruktur Machtverhältnisse regelt. Und wir müssen uns überlegen, ob diese Machtverhältnisse mehr gesellschaftliche Asymmetrien verursachen oder einen ausgleichenden Charakter haben. Das Bauen von Infrastruktur ist immer eine langfristige Maßnahme. Infrastrukturen lohnen sich nur, wenn wir sie pflegen können und wenn sie anpassungsfähig sind. Denn eine Infrastruktur ist auch der Bau einer gesellschaftlichen Abhängigkeit. Also müssen wir uns fragen, ob wir bestimmte Abhängigkeiten brauchen. Wir müssen uns fragen, was uns diese Transformation bringt. Dafür ist es notwendig, dass wir unsere gesellschaftlichen Machtasymmetrien, die daraus entstehenden Bedürfnisse und Friktionen kennen. Das bedeutet, wir müssen zunächst einmal denjenigen zuhören und die verstehen, die sonst nicht gesehen und gehört werden. Das ist nur möglich, wenn wir Betroffene einbinden und ihre Expertise dann auch berücksichtigen. Darüber können wir uns ein Bild davon machen, für welche Probleme KI-Technologien eine Lösung sein können und unter welchen Umständen.

VERARBEITEN

WIE PRÜFEN, ERKLÄREN UND SCHÜTZEN WIR?

Datensätze prüfen, ohne Spezialist:in zu sein; Rückschlüsse auf Personen verhindern, aber trotzdem Daten nutzen; die Funktionsweise von KI-Systemen nachvollziehen, ohne den Code zu verstehen:

Ein italienisches Start-up bietet einen KI-Kontrollraum, der das leisten will. Damit versucht das Unternehmen, Prinzipien wie Robustheit und Sicherheit, den Schutz der Privatsphäre und Datenqualitätsmanagement sowie Transparenz zu stärken. Diese Kriterien definierte die AI HLEG als zentrale Merkmale für vertrauenswürdige KI.

Wer?
CEO Shalini Kurapati gründete das Start-up *Clearbox AI Solutions* gemeinsam mit drei Mitgründern.
Das Team des Unternehmens arbeitet an Lösungen für erklärungsbedürftige KI-Systeme für Banken, Versicherungen sowie im Gesundheitswesen. Die Anwendung ist für Personen ohne datenwissenschaftliches und informationstechnisches Hintergrundwissen nutzbar.[64]

Wann?
- Juli 2019: Gründung
- Dezember 2019: das Unternehmen erhielt mehrere Preise, unter anderem den *Italian National Award for Innovation (PNI)*[65]
- März 2022: Auswahl von der *Europäischen Kommission* für das neue *Women TechEU* Pilotprogramm und damit verbunden Förderung durch das *Research Framework Program Horizon Europe*[66]

Wo?
Turin, Italien

Was?
Das Angebot des Start-ups unterstützt mithilfe eines KI-Kontrollraums dabei, Datensätze zu erweitern, zu bewerten und zu klonen. Die Anwendung erzeugt künstliche Daten auf Basis eines existierenden Datensatzes. Der so generierte synthetische Datensatz verfügt über die gleichen statistischen Eigenschaften und Verteilungen wie der Originaldatensatz. Die erzeugten künstlichen Daten können beispielsweise anstelle der Originaldaten verwendet werden, um Rückschlüsse auf Personen zu verhindern und damit den Datenschutz zu wahren. Zugleich sind synthetische Daten unter bestimmten Voraussetzungen hilfreich, wenn keine ausreichende Datenbasis existiert.

Wie?
Der KI-Kontrollraum ist eine Software, die Nutzer:innen sowohl cloudbasiert verwenden als auch auf eigenen Geräten oder Servern installieren können.

Beispielsweise liefert der KI-Kontrollraum verständliche Erklärungen zu den von komplexen Machine-Learning-Modellen ermittelten Ergebnissen, ohne deren Leistung zu beeinträchtigen. Nachdem Datensätze und Modelle geprüft wurden, ist es Kund:innen möglich, mithilfe einer Schnittstelle unter anderem die Sicherheit ihrer KI-Software zu verbessern. Über die Schnittstelle können sie ‚Adversarial Attacks' erkennen, also Manipulationsversuche durch Außenstehende.
Das Unternehmen *Clearbox AI Solutions* veröffentlicht Beiträge über die eigene Arbeit und widmet sich dabei beispielsweise den Fragen, wie KI-Software zum Wohl der Gesellschaft genutzt werden kann[67] oder wodurch sich erklärbare KI-Systeme auszeichnen[68]. Die Mitarbeiter:innen des Unternehmens orientieren sich bei ihrer Arbeit an der Forschung der *Stanford University* zu menschenzentrierter KI[69] und den „Ethikleitlinien für eine vertrauenswürdige KI" der *EU*.[70]

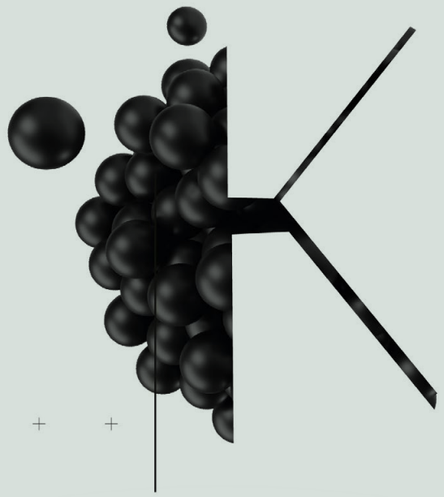
Definition: Machine Learning
Grundlegende Methode im Bereich Künstliche Intelligenz
Bei diesem Verfahren programmieren die Entwickler:innen den Algorithmus, der eine Aufgabe lösen soll, nicht selbst. Stattdessen legen die Entwickler:innen Zielvorgaben fest und bauen ein Computerprogramm, das den besten Lösungsalgorithmus selbstständig findet. Unterschieden werden drei grundlegende Lernmethoden: überwachtes Lernen, unüberwachtes Lernen und bestärkendes Lernen.

Weitere Begriffsdefinitionen finden Sie hier: www.zvki.de > KI-Navigator > Unsere Inhalte > Glossar.

Einblicke in die Arbeit der *ZVKI*-Partner: Wie können wir Privatsphäre schützen?
Beim Machine Learning werden Informationen von sogenannten Trainingsdaten genutzt. Diese Informationen oder Teile davon können wieder aus einem trainierten Modell extrahiert werden. Das kann – je nach Art der Daten – zu Verletzungen der Privatsphäre führen. Der Forschungszweig „Privatsphäre-bewahrendes Machine Learning" sucht nach Methoden, um die Informationsmengen zu begrenzen, die in ein solches Modell einfließen, um damit die Privatsphäre zu schützen. Auch das Erzeugen von synthetischen Daten wird erforscht, um auf echte – schützenswerte – Daten verzichten zu können. Darüber hinaus wird mit Hilfe von Angriffen versucht, sensible Daten aus Modellen und synthetischen Daten zu ziehen, um die Methoden zu überprüfen.

Die Fraunhofer-Institute AISEC und IAIS sowie die Freie Universität Berlin erarbeiten aus verschiedenen Blickwinkeln wissenschaftliche Erkenntnisse zu den Fragen, was vertrauenswürdige KI ausmacht und wie sie gefördert werden kann.

Mehr über unsere Partner finden Sie hier: www.zvki.de > KI-Bibliothek > Partner.



KOMBINIEREN

WAS NEHMEN WIR MIT?

Die erste Ausgabe von *Missing Link* erfasst den Ist-Zustand der Debatten und Auseinandersetzungen über vertrauenswürdige KI-Systeme in Deutschland und Europa und gleicht ihn mit dem Soll-Zustand ab, den Unternehmer:innen, Wissenschaftler:innen, zivilgesellschaftliche Akteur:innen und Politiker:innen fordern. Es bleibt die Erkenntnis, dass noch einiges zu tun ist: Wir müssen handeln und zugleich kritisch bleiben.

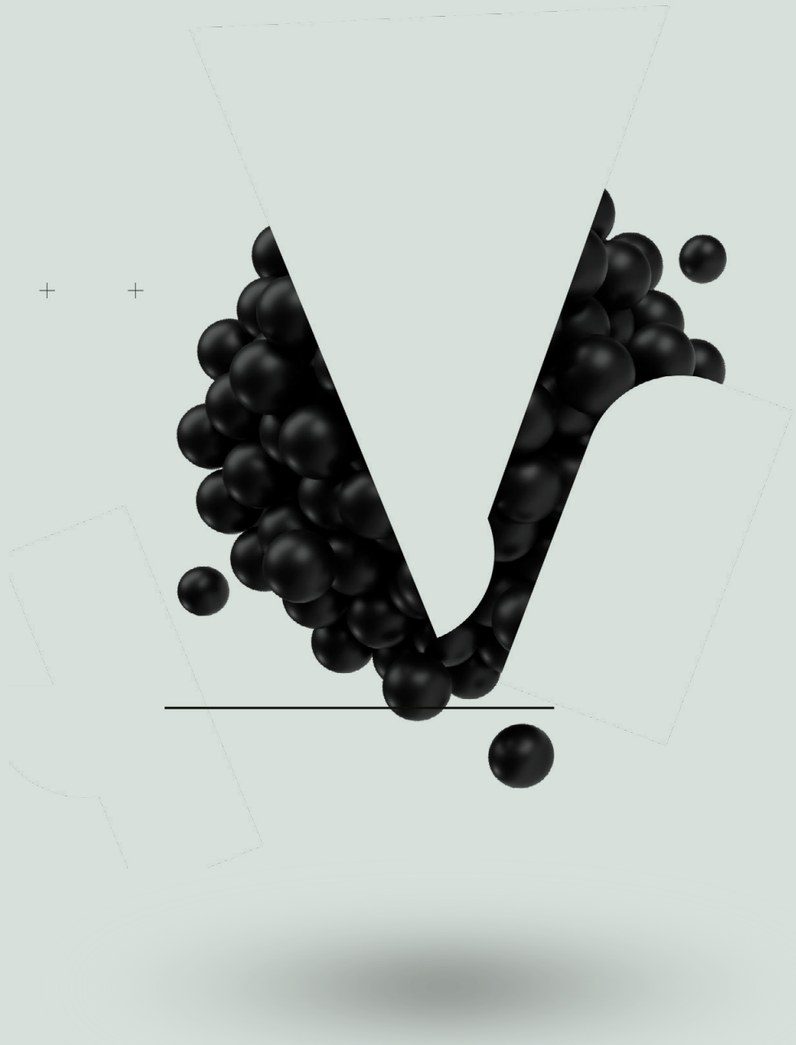
Wir halten fest, dass Vertrauenswürdigkeit immer auf einem gewissen Maß an Wissen basiert. Sie hat demnach nichts mit blindem Vertrauen zu tun, das auf reiner Hoffnung aufbaut. Aktuelle Umfragen weisen darauf hin, dass Nutzer:innen KI-Software eher weniger vertrauen, weil sie beispielsweise mangelnde Fairness vermuten. Demnach erscheint es sinnvoll, Maßnahmen zu ergreifen, die die Vertrauenswürdigkeit von KI-Software im dargelegten Verständnis stärken.

Gleichbehandlung, Nichtdiskriminierung, Sicherheit und Verantwortung sind Pflichten, die nahezu alle Akteur:innen im Diskurs um vertrauenswürdige KI-Systeme fordern. In der wirtschaftlichen Praxis arbeiten vor allem große IT-Unternehmen mit Selbstverpflichtungen, die dazu führen sollen, dass ihre KI-Software auf bestimmten Prinzipien beruht, etwa Fairness und das Einhalten von Menschenrechten. Wie sie ihre Verpflichtungen genau umsetzen, ist nicht überprüfbar. Zugleich finden wir in der Praxis bereits KI-Software, die bestimmte Kriterien wie Sicherheit oder Transparenz adressiert. Ein Praxisbeispiel haben wir in dieser Ausgabe vorgestellt. Weltweit, in Europa und Deutschland arbeiten etablierte Organisationen an Standards und Normen, an denen sich Unternehmen orientieren können, um vertrauenswürdige KI-Software zu entwickeln. Es existieren derzeit zahlreiche unterschiedliche Normungsinitiativen. Sie konzentrieren sich auf die Technik an sich – sowohl in Bezug auf die konkreten Standards als auch bezüglich der Annahmen, aus denen die Standards entstehen. Der Kontext gerät aus dem Blick.

Ein Missachten des Kontexts attestieren einige Wissenschaftler:innen und zivilgesellschaftliche Organisationen dem Entwurf der KI-Verordnung der *EU*. Demnach versucht er, Vertrauenswürdigkeit an einer Risikobewertung festzumachen, die bestimmte Pflichten nach sich zieht. Aus den dort formulierten Vorgaben könnten Standards und Zertifizierungsmechanismen abgeleitet werden. Jedoch hängt das Risiko einer Software stark davon ab, in welchen Zusammenhängen wir sie zu welchen Zwecken einsetzen. Der „AIA“ der *EU* erkennt KI-Systeme nicht als die soziotechnischen Systeme an, die sie sind, so die Kritik. Darüber hinaus sehen es einige Expert:innen kritisch, die Entscheidung, ob ein KI-System als vertrauenswürdig gilt, an Normungs- und Standardisierungsorganisationen auszulagern, da damit rechtliche und ethische Entscheidungen verbunden sind.

Insgesamt ist die Frage, inwiefern KI-Technologien unsere Werte, Rechte und unser Miteinander berühren, eine vergleichsweise neue Frage, die noch nicht ausreichend erforscht ist. Um zielgerichtete Maßnahmen ergreifen zu können, müssen wir deshalb an der einen oder anderen Stelle noch einmal einen Schritt zurücktreten, um genauer zu erfassen, was Vertrauenswürdigkeit in unseren Gesellschaften eigentlich ausmacht und wie wir sie greifbar machen und umsetzen können. Das bedeutet auch, dass wir uns klarmachen müssen, welche Abhängigkeiten und Machtverhältnisse wir mithilfe von KI-Systemen festigen und fortschreiben oder neu etablieren. In diesem Zusammenhang ist es vor allem wichtig, strukturell benachteiligte Menschen und von den Ergebnissen Betroffene einzubinden.

Kritische Perspektiven auf die aktuellen Schwerpunkte der Debatten um vertrauenswürdige KI und die Hinweise auf Leerstellen zeigen uns, dass noch viel zu tun ist. Wir müssen einerseits handeln, um den Einsatz von KI-Systemen in unseren Gesellschaften zu gestalten. Andererseits dürfen wir nicht in reinen Aktionismus verfallen und gegenwärtige Strukturen sowie gängige Formen der Organisation von Gesellschaft fortschreiben, ohne sie zu hinterfragen. An dieser Stelle zeigt sich noch einmal die Bedeutung dessen, Komplexität anzuerkennen, um sie zu dekonstruieren und handhabbar zu machen. Diese Erkenntnisse können wir als Quelle von vertrauenswürdiger KI betrachten.



VERBINDEN

WOZU ALL DAS?

Wozu gibt es dieses Magazin? Wozu dient dieses *Zentrum für vertrauenswürdige Künstliche Intelligenz (ZVKI)*? Wir möchten Wissen rund um den großen Themenkomplex Künstliche Intelligenz miteinander verbinden, um neue Erkenntnisse zu gewinnen. Im Fokus stehen dabei wir Menschen, unsere Grundrechte, unser Wohlergehen und unser Zusammenleben. Wir wollen herausfinden, wann KI-Systeme die Zuschreibung ‚vertrauenswürdig' verdienen.

Algorithmische Systeme und Verfahren der Künstlichen Intelligenz sind Begriffe aus Fachdiskussionen und zugleich Teil unseres Alltags. Mit ihrer Hilfe werden Entscheidungen getroffen, die Auswirkungen auf unser Leben haben. Um diese Auswirkungen zu erfassen, verständlich darzustellen und mit ihnen umzugehen, müssen wir aus Echokammern ausbrechen, Silodenken hinter uns lassen und Brücken zwischen Insellösungen bauen.

Das *ZVKI* ist ein zentraler Ort der Debatte in Deutschland. Es macht die Entwicklungen rund um gesellschaftliche Fragen zu Künstlicher Intelligenz und algorithmischen Systemen greifbar. Zugleich ist es eine neutrale Schnittstelle zwischen Wissenschaft, Wirtschaft, Politik und Zivilgesellschaft, die gemeinsam mit ihren Partnern Instrumente entwickelt, um vertrauenswürdige KI zu bewerten.

Die Ziele des *ZVKI* sind unter anderem:

- Informieren, Wissen vermitteln und aufklären: Verständnis ist die Voraussetzung, um Vertrauen aufzubauen. Deshalb bündeln wir Informationen und bereiten sie für verschiedene Zielgruppen auf.
- Forschen und wissenschaftliche Erkenntnisse verständlich darstellen: Wir untersuchen unter anderem, welche Schritte unternommen werden müssen, um negative Auswirkungen von KI-Systemen zu erkennen und ihnen zu begegnen.
- Evaluieren und prüfen: Wir schaffen Konzepte, um zu überprüfen, ob KI-Systeme Kriterien der Vertrauenswürdigkeit entsprechen.
- Zertifizieren: Wir entwickeln Instrumente zur Bewertung von KI und erarbeitet Anforderungen für deren Zertifizierung.
- Netzwerken und unterstützen: Um möglichst viele Stakeholder sowie deren Ansätze und Ideen zusammenzubringen, bieten wir verschiedene Formate des Austauschs an.

Um diese Ziele zu erreichen, arbeiten wir als interdisziplinäres Team und mit verschiedenen Partnern zusammen: Mit Unterstützung des *Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV)* baut der unabhängige Think Tank *iRights.Lab*, in Zusammenarbeit mit den *Fraunhofer-Instituten AISEC* und *IAIS* sowie der *Freien Universität Berlin*, das *ZVKI* auf.

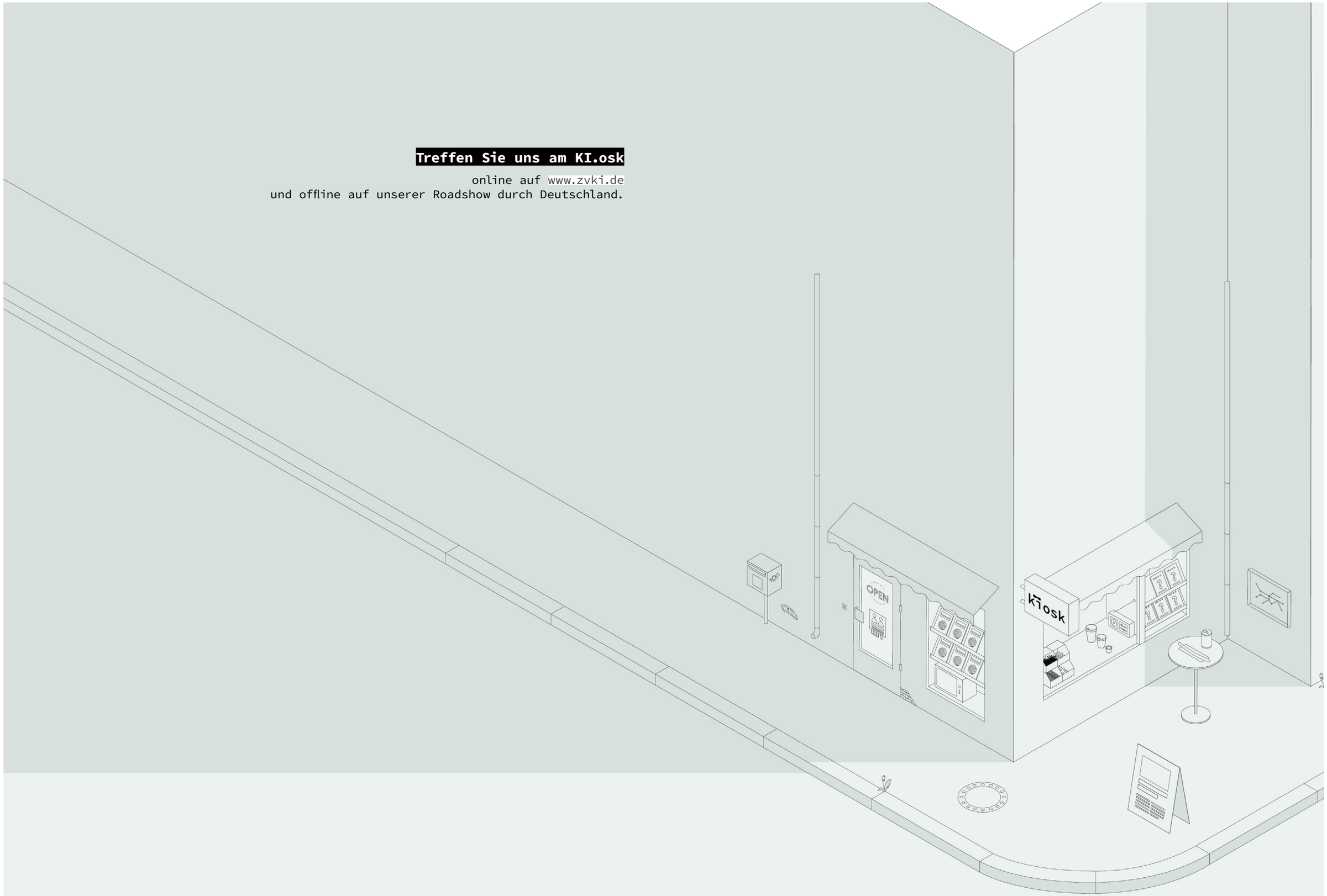
Mitmachen
Das Zentrum für vertrauenswürdige KI – ZVKI versteht sich als neutrale Schnittstelle zwischen Disziplinen und Akteur:innen, zwischen Nutzer:innen und Expert:innen. Treten Sie mit uns in Kontakt und in den Austausch. Sie erreichen uns über zvki@irights-lab.de.

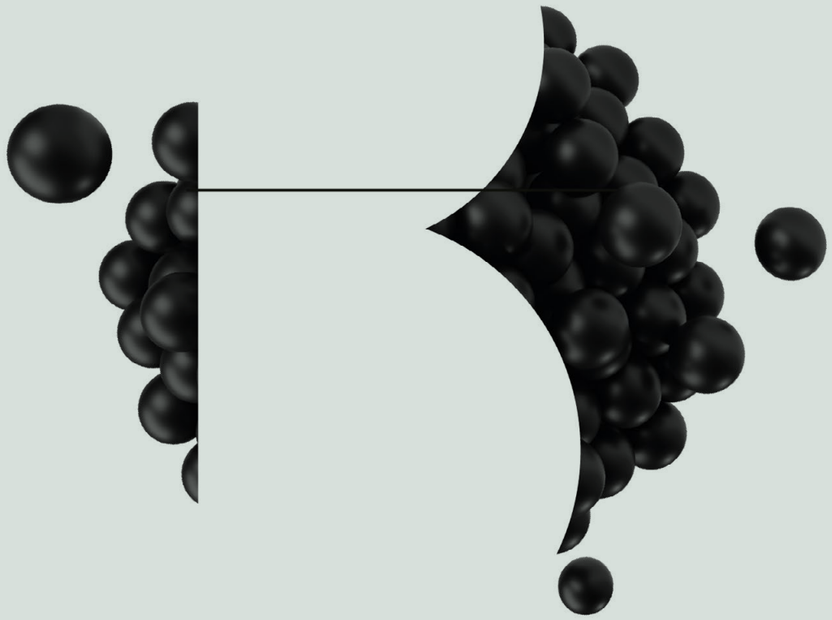
Mehr über unsere Aktivitäten und Themen erfahren Sie auf unserer Webseite www.zvki.de.

Anzeige

Treffen Sie uns am KI.osk

online auf www.zvki.de
und offline auf unserer Roadshow durch Deutschland.





BELEGEN

WOHER STAMMEN DIE INFORMATIONEN?

Quellen

[1] Luhmann, Niklas: Vertrauen. Ein Mechanismus der Reduktion sozialer Komplexität. 2014. Hg. v. UVK Verlagsgesellschaft mbH. Konstanz und München, S.10 ff., S. 32.

[2] Ebd. S. 30.

[3] Tschopp, Marisa; Ruef, Marc; Monett, Dagmar: Vertrauen Sie KI? Einblicke in das Thema Künstliche Intelligenz und warum Vertrauen eine Schlüsselrolle im Umgang mit neuen Technologien spielt. 2022. In: Miriam Landes, Eberhard Steiner und Tatjana Utz (Hg.): Kreativität und Innovation in Organisationen, Bd. 81. Berlin, Heidelberg: Springer, S. 346-319, hier S. 331.

[4] Siau, Keng; Wang, Weiyu: Building Trust in Artificial Intelligence, Machine Learning, and Robotics. Trust is hard to come by. 2018. Hg. v. Cutter Business Technology Journal (Vol. 31, No. 2), S. 47-53, hier S. 47. Online unter: https://www.cutter.com/article/building-trust-artificial-intelligence-machine-learning-and-robotics498981 (letzter Aufruf: 02.05.2022).

[5] Tschopp; Ruef; Monett, S. 333.

[6] Siau; Wang, S. 50 f.

[7] Kieslich, Kimon; Došenović, Pero; Marcinkowski, Frank: Algorithmische Empfehlungssysteme. Was denkt die deutsche Bevölkerung über den Einsatz und die Gestaltung algorithmischer Empfehlungssysteme? 2021. Center for Advanced Internet Studies (Factsheet Nr. 5), o. S. Online unter: https://www.cais.nrw/wp94-fa-4content/uploads/05/2021/Factsheet-5-KI-Recommender.pdf (letzter Aufruf: 02.05.2022).

[8] Center for Advanced Internet Studies GmbH (CAIS): KI Meinungsmonitor Künstliche Intelligenz. Was denkt die Bevölkerung über Künstliche Intelligenz? Wie berichten die Medien darüber? 2022, o. S. Online unter https://www.cais.nrw/memoki/3-16214363435 90#fd1101d5133 (letzter Aufruf: 02.05.2022).

[9] Kieslich; Došenović; Marcinkowski, o. S.

[10] Ebd., o. S.

[11] Dowling, Michael; Klinkenberg, Ralf; Köpcke, Hanna; Liebl, Andreas; Löser, Alexander; Mordvinova, Olga; Morik, Katharina; Rabe, Martin; Schlunder, Philipp; Schmidt, Fabian: KI im Mittelstand. Potenziale erkennen, Voraussetzungen schaffen, Transformation meistern. 2021. Hg. v. Lernende Systeme – Die Plattform für Künstliche Intelligenz, S. 5. Online unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/PLS_Booklet_KMU.pdf (letzter Aufruf: 02.05.2022).

[12] Fulde, Verena: Guidelines for Artificial Intelligence. Deutsche Telekom AG. O. D., o. S. Online unter https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline524366- (letzter Aufruf: 02.05.2022).

[13] Machmeier, Corinna; Madigan, Mary Carol: Verantwortungsvoll mit künstlicher Intelligenz umgehen – ein Jahr des Lernens. 2020, o. S. Online unter: https://news.sap.com/germany/01/2020/ki-kuenstliche-intelligenz-ethik/ (letzter Aufruf: 02.05.2022).

[14] Pichai, Sunder: AI at Google: our principles. Google Ireland Limited. 2018, o. S. Online unter: https://www.blog.google/technology/ai/ai-principles/ (letzter Aufruf: 02.05.2022).

[15] Microsoft Corporation: Responsible AI. O. D., o. S. Online unter: https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6 (letzter Aufruf: 02.05.2022).

[16] Robert Bosch GmbH: KI-Kodex von Bosch im Überblick. 2020, o. S. Online unter: https://assets.bosch.com/media/de/global/stories/ai_codex/bosch-code-of-ethics-for-ai.pdf (letzter Aufruf: 02.05.2022).

[17] Bayerische Motoren Werke AG: Ethik-Kodex der BMW Group für den Einsatz von Künstlicher Intelligenz. 2020, o. S. Online unter: https://www.bmwgroup.com/content/dam/grpw/websites/bmwgroup_com/downloads/DEU_PM_CodeOfEthicsForAI_Kurz.pdf (letzter Aufruf: 02.05.2022).

[18] Heesen, Jessica et. al.: Ethik-Briefing. Leitfaden für eine verantwortungsvolle Entwicklung. 2020. Hg. v. Lernende Systeme – Die Plattform für Künstliche Intelligenz, S. 24. Online unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_EB_200831.pdf (letzter Aufruf: 02.05.2022).

[19] Bundesministerium für Arbeit und Soziales (BMAS), Referat „CSR“ – Gesellschaftliche Verantwortung von Unternehmen: Corporate Digital Responsibility. O. D., o. S. Online unter: https://www.csr-in-deutschland.de/DE/CSR-Allgemein/Corporate-Digital-Responsibility/corporate-digital-responsibility.html (letzter Aufruf: 02.05.2022).

[20] Horn, Nikolai: Vertrauen, das Handlungsoptionen schafft, ohne naiv zu sein. 2021. Hg. v. Corporate Digital Responsibility (Online-Magazin), Reihe Philosophischer Zwischenruf, o. S. Online unter: https://corporate-digital-responsibility.de/article/philosophischer-zwischenruf-vertrauen/#_ftn1 (letzter Aufruf: 02.05.2022).

[21] Beckert, Bernd: Vertrauenswürdige künstliche Intelligenz. Ausgewählte Praxisprojekte und Gründe für das Umsetzungsdefizit. In: Zeitschrift für Technikfolgenabschätzung Theorie und Praxis. Ausgabe 30: KI-Systeme gestalten und erfahren. 2021, S. 22-17, hier S. 18. Online unter: https://www.tatup.de/index.php/tatup/issue/view/20%176/171S.2018% (letzter Aufruf: 02.05.2022).

[22] Wahlster, Wolfgang; Winterhalter, Christoph (Hg.): Deutsche Normungsroadmap Künstliche Intelligenz. 2020, S. 5. Online unter: https://www.din.de/resource/blob/6/772438b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf(letzter Aufruf: 02.05.2022).

[23] Ebd., S. 102.

[24] AI Watch: AI Standardisation Landscape State of play and link to the EC proposal for an AI regulatory framework. Europäische Kommission. Joint Research Centre. 2021, S. 21. DOI: 10.2760/376602.

[25] Beckert, S. 19.

[26] Wilson, Christopher; van der Velden, Maja: Sustainable AI: An integrated model to guide public sector decision-making. 2022. In: Technology in Society 101926 ,68, S. 1–11, hier S. 6 f. DOI: 10.1016/j.techsoc.2022.101926.

[27] OECD: Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/2019 .0449. Online unter: https://legalinstruments.oecd.org/api/print?ids=648&lang=en (letzter Aufruf: 02.05.2022).

[28] Europäische Kommission: Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, COM/206/2021 final. 2021. Hg. v. Publications Office of the EU. Online unter: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX3%A52021PC0206 (letzter Aufruf: 02.05.2022).

[29] Ebd., S. 3.

[30] Ebd., S. 53.

[31] Ebd., S. 15.

[32] Europäische Kommission: Neue Vorschriften für künstliche Intelligenz – Fragen und Antworten. 2 /2021. O. S. Online unter: https://ec.europa.eu/commission/presscorner/detail/de/QANDA_1#1683_21 (letzter Aufruf: 02.05.2022).

[33] Europäische Kommission, 2021, S. 53.

[34] High-Level Expert Group on AI: Ethics Guidelines for Trustworthy Artificial Intelligence. 2019. Hg. v. Publications Office of the EU. Online unter: https://op.europa.eu/en/publication-detail/-/publication/d11-0434-3988569ea8-c1f01-aa75ed71a1 (letzter Aufruf: 02.05.2022).

[35] Ebd., S. 17 f.

[36] Europäische Kommission: Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen, COM(65 (2020 final. 2020. Hg. v. Publications Office of the EU, S. 10. Online unter: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf (letzter Aufruf: 02.05.2022).

[37] Nationale Strategie für Künstliche Intelligenz – AI Made in Germany. Strategie Künstliche Intelligenz der Bundesregierung. Fortschreibung 2020 .2020. Online unter: https://www.ki-strategie-deutschland.de/files/downloads/201201_Fortschreibung_KI-Strategie.pdf (letzter Aufruf: 02.05.2022).

[38] Europarat: Ad hoc Committee on Artificial Intelligence (CAHAI). Feasibility Study. 2020, S. 82 ff. Online unter: https://rm.coe.int/cahai-23-2020-final-eng-feasibility-study-/1680a0c6da (letzter Aufruf: 02.05.2022).

[39] Ebd., S. 21.

[40] Ebd., S. 52.

[41] Europarat: Ad hoc Committee on Artificial Intelligence (CAHAI). 1425th meeting, 16 February 2022: 10 Legal questions. 2022, o. S. Online unter: https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a4e8a5 (letzter Aufruf: 02.05.2022).

[42] Europarat: CAI – Committee on Artificial Intelligence. O. D., o. S. Online unter: https://www.coe.int/en/web/artificial-intelligence/cai(letzter Aufruf: 02.05.2022).

[43] Google Ireland Limited: Recommendations for Regulating AI. 2019, S. 2. Online unter: https://ai.google/static/documents/recommendations-for-regulating-ai.pdf (letzter Aufruf: 02.05.2022).

[44] Microsoft Corporation: Microsoft's Response to the European Commission's Consultation on the Artificial Intelligence Act. 2021, S. 1 f. Online unter: https://blogs.microsoft.com/wp-content/uploads/prod/sites/73/2021/09/microsoft-response-to-the-european-commission-consultation-on-the-artifical-intelligence-act.pdf (letzter Aufruf: 02.05.2022).

[45] Kettemann, Matthias C.: UNESCO-Empfehlung zur Ethik künstlicher Intelligenz. Bedingungen zur Implementierung in Deutschland. 2021. Hg. v. Deutsche UNESCO-Kommission, S. 6 und 26. Online unter: https://www.unesco.de/sites/default/files/03-2022/DUK_Broschuere_KI-Empfehlung_DS_web_final.pdf (letzter Aufruf: 02.05.2022).

[46] Ebd., S. 43.

[47] Ebd., S. 27.

[48] Ebd., S. 12.

[49] Ebd., S. 26.

[50] European Digital Rights (EDRi); Access Now; Panoptykon Foundation; epicenter.works; AlgorithmWatch; European Disability Forum (EDF); Bits of Freedom; Fair Trials; PICUM; ANEC: An EU Artificial Intelligence Act for Fundamental Rights. A Civil Society Statement. 2021, o. S. Online unter: https://edri.org/wp-content/uploads/12/2021/Political-statement-on-AI-Act.pdf (letzter Aufruf: 02.05.2022).

[51] Heesen, Jessica; Müller-Quade, Jörn; Wrobel, Stefan et al.: Kritikalität von KI-Systemen. Ein notwendiger, aber nicht hinreichender Baustein für Vertrauenswürdigkeit. 2021. Hg. v. Plattform Lernende Systeme, S. 18. Online unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_WP_Kritikalitaet_von_KI-Systemen.pdf (letzter Aufruf: 02.05.2022).

[52] Ebd., S. 16.

[53] Ebd., S. 29.

[54] European Digital Rights (EDRi) et. al., o. S.

[55] Beining, Leonie: Vertrauenswürdige KI durch Standards? Herausforderungen bei der Standardisierung und Zertifizierung von Künstlicher Intelligenz. 2020. Hg. v. Stiftung Neue Verantwortung, S. 14 f. Online unter: https://www.stiftung-nv.de/sites/default/files/herausforderungen-standardisierung-ki.pdf (letzter Aufruf: 02.05.2022).

[56] Ebers, Martin: Standardizing AI – The Case of the European Commission's Proposal for an Artificial Intelligence Act. 2021. In: Larry A. DiMatteo; Michel Cannarsa; Cristina Poncibò (Hg.): The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics, S. 1–23, hier S. 22. Online unter: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3900378 (letzter Aufruf: 02.05.2022).

[57] Ebd., S. 22 f.

[58] Europarat: 2020, S. 39.

[59] McFadden, Mark; Jones, Kate; Taylor, Emily; Osborn, Georgia: Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation. 2021. Hg. v. Oxford Commission on AI & Good Governance, S. 22. Online unter: https://oxcaigg.oii.ox.ac.uk/wp-content/uploads/sites/124/2021/12/Harmonising-AI-OXIL.pdf (letzter Aufruf: 02.05.2022).

[60] Mock, Michael; Becker, Daniel; Schmitz, Anna; Cremers, A.B.; Adilova, Linara; Poretschkin, Maximilian: Management System Support for Trustworthy Artificial Intelligence. A comparative study. 2021. Hg. v. Fraunhofer IAIS, S. 58 f. Online unter: https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/Fraunhofer_IAIS_Study_20%MSS.pdf (letzter Aufruf: 02.05.2022).

[61] Cave, Stephen; Dihal, Kanta: Standpunkt: Geschichte und Zukünfte Künstlicher Intelligenz. 2022, o. S. Online unter: https://background.tagesspiegel.de/digitalisierung/geschichte-und-zukuenfte-kuenstlicher-intelligenz (letzter Aufruf: 02.05.2022).

[62] Glikson, Ella; Woolley, Anita Williams: Human Trust in Artificial Intelligence: Review of Empirical Research. 2020. In: ANNALS 14 (2), S. 1–91, hier S. 2. DOI: 10.5465/annals.2018.0057.

[63] Samudzi, Zoé: Color photography was optimized for light skin. Facial recognition technologies are just carrying on color film's historical non-registry of black people. 2019, o. S. Online unter: https://www.thedailybeast.com/bots-are-terrible-at-recognizing-black-faces-lets-keep-it-that-way (letzter Aufruf: 02.05.2022).

[64] Fintech District. ClearBox AI: interview with Matteo Giovannetti. 2019, o. S. Online unter: https://www.fintechdistrict.com/clearbox-matteo-giovannetti/ (letzter Aufruf: 02.05.2022).

[65] i3P. PNI 2019: Clearbox AI Solutions, an I3P start-up, wins the ICT Award. 2019, o. S. Online unter: https://www.i3p.it/en/news/pni-2019-clearbox-ai-solutions-i3p-startup-wins-ict-award (letzter Aufruf: 02.05.2022).

[66] European Innovation Council: Women TechEU pilot – Selected companies and organisations. 2022, o. S. Online unter: https://eic.ec.europa.eu/system/files/02-2022/Women20%TechEU20%Ranking20%List.pdf (letzter Aufruf: 02.05.2022).

[67] Clearbox AI. How to use data and AI for good? 2021, o. S. Online unter: https://www.clearbox.ai/blog/-28-12-2021how-to-use-data-and-ai-for-good/ (letzter Aufruf: 02.05.2022).

[68] Clearbox AI: How does Explainable AI work? 2021, o. S. Online unter: https://www.clearbox.ai/blog/-28-12-2021how-to-use-data-and-ai-for-good/ (letzter Aufruf: 02.05.2022).

[69] Stanford University: Guiding Human-Centered AI. O. D., o. S. Online unter: https://hai.stanford.edu/research (letzter Aufruf: 02.05.2022).

[70] High-Level Expert Group on AI: 2019.

IMPRESSUM

Herausgeber und inhaltlich Verantwortlicher i. S. d. § 55 Abs. 2 RStV:
Philipp Otto
iRights.Lab GmbH
Oranienstr. 185
D-10999 Berlin

Telefon: +49 (0)30 40 36 77 230
Fax: +49 (0)30 40 36 77 260
E-Mail: zvki@irights-lab.de

Geschäftsführer:in: Philipp Otto, Dr. Wiebke Glässer

Registergericht: Amtsgericht Berlin-Charlottenburg
Registernummer: HRB 185640 B
Finanzamt für Körperschaften II
USt-IdNr.: DE311181302

Projektleitung: Philipp Otto

Inhaltliche Konzeption: Jaana Müller-Brehm, Philipp Otto, Verena Till

Gestalterische Konzeption und Layout: Christoph Löffler

Autorin: Jaana Müller-Brehm

Inhaltliche Mitarbeit: Franziska Busse, Michael Puntschuh, Paul Ritzka, Lisa Schmechel, Verena Till

Lektorat: text|struktur

Dieses Werk steht unter Creative Commons Lizenz CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0/deed.de. Ausgeschlossen davon sind die Fotos in dieser Ausgabe, die weiterhin urheberrechtlich geschützt sind.

Die Online-Version von *Missing Link* und weitere Informationen zum Projekt *ZVKI* finden Sie unter: www.zvki.de.

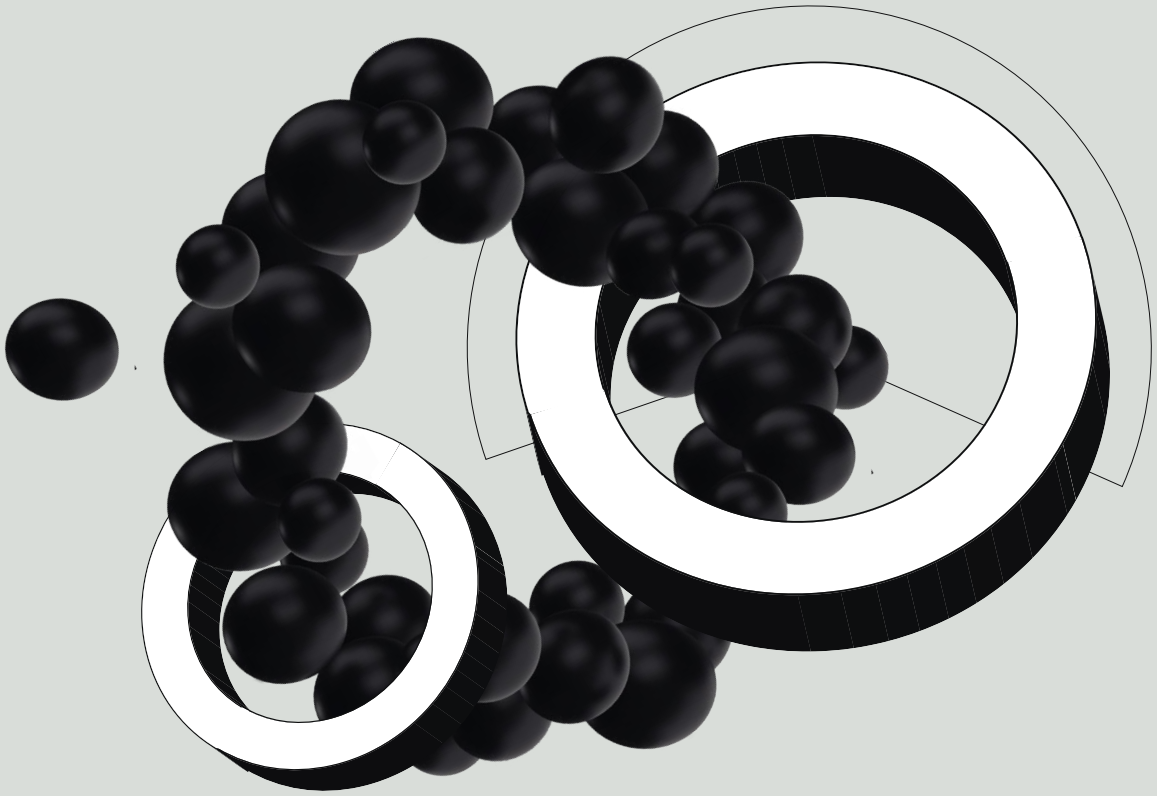
Projektpartner:innen ZVKI: *iRights.Lab*, *Fraunhofer AISEC*, *Fraunhofer IAIS*, *Freie Universität Berlin*

Gefördert durch: *Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV)*

Das Projekt *ZVKI* wird vom unabhängigen Think Tank *iRights.Lab* verantwortet und durchgeführt. Das *iRights.Lab* entwickelt Strategien und praktische Lösungen, um die Veränderungen in der digitalen Welt vorteilhaft zu gestalten. Wir unterstützen öffentliche Einrichtungen, Stiftungen, Unternehmen, Wissenschaft und Politik dabei, die Herausforderungen der Digitalisierung zu meistern und die vielschichtigen Potenziale effektiv und positiv zu nutzen. Weitere Informationen über das *iRights.Lab* finden Sie unter www.irights-lab.de.

Gefördert durch:
Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz
aufgrund eines Beschlusses des Deutschen Bundestages

iRights.Lab
Think Tank für die digitale Welt



Missing Link . Web Version
ISSN 2751-2207