UNABHÄNGIGE

HOCHRANGIGE EXPERTENGRUPPE FÜR KÜNSTLICHE INTELLIGENZ

EINGESETZT VON DER EUROPÄISCHEN KOMMISSION IM JUNI 2018

ETHIK-LEITLINIEN FÜR EINE VERTRAUENSWÜRDIGE KI

ETHIK-LEITLINIEN FÜR EINE VERTRAUENSWÜRDIGE KI

Hochrangige Expertengruppe für künstliche Intelligenz

Das vorliegende Dokument wurde von der Hochrangigen Expertengruppe für KI (HEG-KI) verfasst. Die im vorliegenden Dokument genannten Mitglieder der HEG-KI unterstützen insgesamt den Rahmen für eine vertrauenswürdige KI, der mit diesen Leitlinien vorgelegt wird, auch wenn sie nicht notwendigerweise mit jeder einzelnen im Dokument enthaltenen Aussage einverstanden sind.

Die in Kapitel III des vorliegenden Dokuments enthaltene Bewertungsliste für vertrauenswürdige KI wird eine Pilotphase unter Beteiligung der betroffenen Kreise durchlaufen, deren Zweck das Sammeln von Rückmeldungen aus der Praxis ist. Eine überarbeitete Fassung der Bewertungsliste unter Berücksichtigung der während der Pilotphase eingegangenen Rückmeldungen wird der Europäischen Kommission Anfang 2020 vorgelegt werden.

Die HEG-KI ist eine unabhängige Expertengruppe, die im Juni 2018 von der Europäischen Kommission eingesetzt wurde.

Kontakt Nathalie Smuha – Koordinatorin für die HEG-KI
E-Mail CNECT-HLG-AI@ec.europa.eu

Europäische Kommission
B-1049 Bruxelles/Brüssel

Veröffentlichung des Dokuments am 8. April 2019.

Ein erster Entwurf dieses Dokuments wurde am 18. Dezember 2018 veröffentlicht und war Gegenstand einer offenen Konsultation, in deren Rahmen Rückmeldungen von mehr als 500 Beteiligten eingingen. Wir möchten uns ausdrücklich und herzlich bei allen Beteiligten für die Rückmeldungen zum ersten Entwurf dieses Dokuments bedanken, die bei der Erstellung dieser überarbeiteten Version berücksichtigt worden sind.

Weder die Europäische Kommission noch Personen, die in ihrem Namen handeln, sind für mögliche Verwendungen der folgenden Informationen verantwortlich. Für den Inhalt dieser Arbeitsunterlage ist allein die Hochrangige Expertengruppe für künstliche Intelligenz (HEG-KI) verantwortlich. Auch wenn die Ausarbeitung dieser Leitlinien unter Beteiligung von Mitarbeitern der Kommissionsdienststellen erfolgte, entsprechen die in diesem Dokument zum Ausdruck gebrachten Ansichten dem gemeinsamen Standpunkt der HEG-KI und stellen keinesfalls einen offiziellen Standpunkt der Europäischen Kommission dar.

Weitere Informationen über die Hochrangige Expertengruppe für künstliche Intelligenz sind online abrufbar (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).
Die Weiterverwendung von Dokumenten der Europäischen Kommission ist im Beschluss 2011/833/EU (ABl. L 330 vom 14.12.2011, S. 39) geregelt. Für die Verwendung oder den Nachdruck von Fotos oder anderem Material, an dem die EU kein Urheberrecht hält, ist eine Genehmigung direkt bei den Urheberrechtsinhabern einzuholen.

PDF	ISBN 978-92-76-11987-6	doi:10.2759/22710	KK-02-19-841-DE-N
Print	ISBN 978-92-76-12799-4	doi:10.2759/856513	KK-02-19-841-DE-C

INHALTSVERZEICHNIS

ZUSAMMENFASSUNG **2**

A. EINLEITUNG **5**

B. EIN RAHMEN FÜR EINE VERTRAUENSWÜRDIGE KI **7**

I. Kapitel I: Fundamente einer vertrauenswürdigen KI **11**

1. Grundrechte als moralischer und rechtlicher Anspruch 12
2. Von Grundrechten zu ethischen Grundsätzen 12

II. Kapitel II: Verwirklichung einer vertrauenswürdigen KI **17**

1. Anforderungen an eine vertrauenswürdige KI 17
2. Technische und nicht-technische Methoden zur Schaffung einer vertrauenswürdigen KI 25

III. Kapitel III: Bewertung einer vertrauenswürdigen KI **30**

C. BEISPIELE FÜR DIE MÖGLICHKEITEN, DIE DIE KI BIETET, UND KRITISCHE ERWÄGUNGEN **42**

D. FAZIT **45**

GLOSSAR **47**

ZUSAMMENFASSUNG

(1) Ziel der vorliegenden Leitlinien ist die Förderung einer vertrauenswürdigen KI. Eine vertrauenswürdige KI zeichnet sich durch **drei Komponenten** aus, die während des gesamten Lebenszyklus des Systems erfüllt sein sollten: a) Sie sollte **rechtmäßig** sein und somit alle anwendbaren Gesetze und Bestimmungen einhalten, b) sie sollte **ethisch** sein und somit die Einhaltung ethischer Grundsätze und Werte garantieren und c) sie sollte **robust** sein, und zwar sowohl in technischer als auch sozialer Hinsicht, da KI-Systeme selbst bei guten Absichten unbeabsichtigten Schaden anrichten können. Jede Komponente an sich ist notwendig, jedoch nicht ausreichend, um das Ziel einer vertrauenswürdigen KI zu erreichen. Idealerweise wirken alle drei Komponenten harmonisch zusammen und überlappen sich in ihrer Funktionsweise. Sollte es in der Praxis zu Spannungen zwischen diesen Komponenten kommen, sollte die Gesellschaft daran arbeiten, sie in Einklang zu bringen.

(2) In diesen Leitlinien wird ein **Rahmen für die Verwirklichung einer vertrauenswürdigen KI** festgelegt. Der Rahmen beschäftigt sich nicht ausführlich mit der ersten Komponente der vertrauenswürdigen KI (rechtmäßige KI)[1]. Vielmehr will er Orientierungen für die Förderung und Sicherung einer ethischen und robusten KI (Komponenten zwei und drei) bieten. Die vorliegenden Leitlinien richten sich an alle betroffenen Kreise. Es soll sich dabei nicht nur um eine Liste mit ethischen Grundsätzen handeln, sondern um eine Hilfestellung für die mögliche Umsetzung dieser Prinzipien in soziotechnischen Systemen. Die Hilfestellung gliedert sich in drei Abstraktionsebenen, und zwar von den am stärksten abstrahierten Überlegungen in Kapitel I bis hin zu konkreten Hinweisen in Kapitel III, wobei Beispiele für Chancen und kritische Fragen, die KI-Systeme aufwerfen, den Abschluss bilden.

I. Anhand eines auf Grundrechten beruhenden Ansatzes werden in Kapitel I die **ethischen Grundsätze** und die damit in Verbindung stehenden Werte benannt, die bei der Entwicklung, Einführung und Verwendung von KI-Systemen gewahrt werden müssen.

Wichtige Leitlinien aus Kapitel I:

- ✓ Die Entwicklung, Einführung und Nutzung von KI-Systemen muss so erfolgen, dass die folgenden ethischen Grundsätze eingehalten werden: *Achtung der menschlichen Autonomie, Schadensverhütung, Fairness und Erklärbarkeit.* Die möglichen Spannungen zwischen diesen Grundsätzen müssen zur Kenntnis genommen und gelöst werden.
- ✓ Besondere Berücksichtigung von Situationen, in denen besonders schutzbedürftige Gruppen wie Kinder, Menschen mit Behinderungen und andere betroffen sind, die schon in der Vergangenheit Benachteiligung erfahren haben oder die einem besonders hohen Exklusionsrisiko ausgesetzt sind. Gleiches gilt für Situationen, die sich durch ungleiche Macht- oder Informationsverteilung auszeichnen, etwa zwischen Arbeitgebern und Arbeitnehmern oder Unternehmen und Verbrauchern.[2]
- ✓ Es gilt anzuerkennen und zu berücksichtigen, dass KI-Systeme dem Einzelnen und der Gesellschaft zwar einen erheblichen Nutzen bringen, gleichzeitig jedoch bestimmte Risiken bergen und möglicherweise negative, mitunter schwer absehbare, erkennbare oder messbare Auswirkungen (z. B. im Hinblick auf Demokratie, Rechtsstaatlichkeit, Verteilungsgerechtigkeit oder den menschlichen Geist als solchen) haben können. Zur Abwendung dieser Gefahren müssen gegebenenfalls angemessene und in Anbetracht der Höhe des Risikos verhältnismäßige Maßnahmen getroffen werden.

[1] Alle normativen Aussagen im vorliegenden Dokument sollen eine Hilfestellung zur Verwirklichung der Komponenten zwei und drei einer vertrauenswürdigen KI (Ethische und robuste KI) darstellen. Diese Aussagen sind daher nicht als Rechtsberatung oder als Hilfestellung zur Einhaltung geltenden Rechts zu betrachten, obwohl anerkannt wird, dass sich viele dieser Aussagen bis zu einem gewissen Grad bereits in bestehenden Gesetzen widerspiegeln. Diesbezüglich sei auf Absatz 21 ff verwiesen.

[2] Siehe Artikel 24 bis 27 der Charta der Grundrechte der Europäischen Union (EU-Grundrechtecharta) über die Rechte von Kindern und älteren Menschen, die Integration von Menschen mit Behinderungen sowie die Arbeitnehmerrechte. Siehe auch Artikel 38 über Verbraucherschutz.

II. Aufbauend auf Kapitel I wird im Kapitel II anhand von **sieben Anforderungen** an KI-Systeme erläutert, wie sich eine vertrauenswürdige KI realisieren lässt. Bei der Umsetzung können sowohl technische als auch nicht-technische Methoden angewendet werden.

Wichtige Leitlinien aus Kapitel II:

- ✓ Es muss gewährleistet sein, dass die Entwicklung, Einführung und Nutzung von KI-Systemen die Anforderungen an vertrauenswürdige KI erfüllen: 1) Vorrang menschlichen Handelns und menschliche Aufsicht, 2) technische Robustheit und Sicherheit, 3) Schutz der Privatsphäre und Datenqualitätsmanagement, 4) Transparenz, 5) Vielfalt, Nichtdiskriminierung und Fairness, 6) gesellschaftliches und ökologisches Wohlergehen sowie 7) Rechenschaftspflicht.
- ✓ Berücksichtigung technischer und nichttechnischer Methoden, um die Umsetzung dieser Anforderungen sicherzustellen.
- ✓ Förderung von Forschung und Innovation als Beitrag zur Bewertung von KI-Systemen und zur weiteren Erfüllung der Anforderungen; Bekanntmachung von Ergebnissen und offenen Fragen in der breiten Öffentlichkeit sowie systematische Ausbildung einer neuen Generation von Experten auf dem Gebiet der KI-Ethik.
- ✓ Klare und proaktive Informationsübermittlung an betroffene Kreise über die Fähigkeiten und Grenzen der KI-Systeme, die realistische Erwartungen ermöglichen, sowie über die Art und Weise der Implementierung der Anforderungen. Für die Anwender muss klar erkennbar sein, dass sie es mit einem KI-System zu tun haben.
- ✓ Möglichkeit der Rückverfolgbarkeit und Nachprüfbarkeit von KI-Systemen, insbesondere in kritischen Zusammenhängen oder Situationen.
- ✓ Beteiligung der Interessenträger während des gesamten Lebenszyklus des KI-Systems. Schulungs- und Ausbildungsförderung mit dem Ziel, allen Interessenträgern Kompetenzen auf dem Gebiet der vertrauenswürdigen KI zu vermitteln.
- ✓ Zwischen den verschiedenen Grundsätzen und Anforderungen können möglicherweise wesentliche Spannungen auftreten. Diesbezügliche Kompromisse und Lösungen müssen kontinuierlich ermittelt, bewertet, dokumentiert und mitgeteilt werden.

III. Kapitel III enthält eine konkrete und nicht erschöpfende Bewertungsliste für vertrauenswürdige KI, die dazu dienen soll, die in Kapitel II genannten Anforderungen operativ umzusetzen. Diese **Bewertungsliste** muss zukünftig auf den spezifischen Anwendungsfall des KI-Systems zugeschnitten werden[3].

Wichtige Leitlinien aus Kapitel III:

- ✓ Aufstellung einer Bewertungsliste für vertrauenswürdige KI in Bezug auf die Entwicklung, Einführung oder Nutzung der KI-Systeme und deren Anpassung an den konkreten Anwendungsfall, in dem das System eingesetzt wird.
- ✓ Eine solche Bewertungsliste kann niemals erschöpfend sein. Bei der Gewährleistung von vertrauenswürdiger KI geht es nicht um das Abhaken von Punkten einer Liste, sondern um einen kontinuierlichen Prozess der Ermittlung und Umsetzung von Anforderungen, der Bewertung von Lösungen und der Erzielung besserer Ergebnisse über den gesamten Lebenszyklus des KI-Systems unter Beteiligung der relevanten Interessenträger.

(3) Ein letzter Abschnitt des Dokuments soll einige der Punkte, die innerhalb des Rahmens angesprochen werden, durch Beispiele zu ergreifender vorteilhafter Möglichkeiten konkretisieren und kritische Fragen behandeln, die KI-Systeme aufwerfen und die sorgfältig geprüft werden sollten.

(4) Zwar bieten diese Leitlinien eine Hilfestellung für KI-Anwendungen im Allgemeinen, bei der Schaffung eines

[3] Im Einklang mit dem Geltungsbereich des in Absatz 2 festgelegten Rahmens dient die Bewertungsliste nicht als Informationsquelle für den Zweck, die Einhaltung geltender Gesetze zu gewährleisten (rechtmäßige KI), sondern beschränkt sich darauf, Hilfestellungen bei der Einhaltung der Komponenten zwei und drei der vertrauenswürdigen KI (ethische und robuste KI) anzubieten.

Fundaments für eine vertrauenswürdige KI führen jedoch unterschiedliche Situationen zu unterschiedlichen Problemstellungen. Es ist daher zu prüfen, inwieweit im Kontext der Spezifität von KI-Systemen ein sektorbezogenes Konzept in Ergänzung zu diesem horizontalen Rahmen benötigt wird.

(5) Diese Leitlinien sollen weder dazu dienen, aktuelle oder zukünftige politische Entscheidungen oder Regulierungsvorhaben zu ersetzen, noch sollen sie deren Einführung verhindern. Sie sollten als ein dynamisches Arbeitspapier betrachtet werden, das im Laufe der Zeit regelmäßig zu überarbeiten sein wird, damit es mit der Entwicklung der Technik, unseres sozialen Umfelds und unseres diesbezüglichen Wissens Schritt halten kann. Diese Leitlinien sollen daher als Ausgangspunkt für die Diskussion über „vertrauenswürdige KI für Europa“[4] verstanden werden. Über Europa hinaus sollen diese Leitlinien auch die Forschung, Reflexion und Diskussion über einen ethischen Rahmen für KI-Systeme auf weltweiter Ebene fördern.

[4] Dieses Ideal soll für in den EU-Mitgliedstaaten entwickelte, eingeführte und verwendete KI-Systeme sowie für anderswo entwickelte oder produzierte Systeme gelten, die in die EU eingeführt und hier verwendet werden. Wenn im vorliegenden Dokument der Begriff „Europa“ genannt wird, sind damit die EU-Mitgliedstaaten gemeint. Dennoch erheben diese Leitlinien den Anspruch, auch außerhalb der EU von Bedeutung zu sein. Diesbezüglich sei darauf hingewiesen, dass sowohl Norwegen als auch die Schweiz Teil des koordinierten Plans für die KI sind, der im Dezember 2018 von der Kommission und den Mitgliedstaaten vereinbart und veröffentlicht wurde.

A. EINLEITUNG

(6) In ihren Mitteilungen vom 25. April 2018 und 7. Dezember 2018 hat die Europäische Kommission (die Kommission) ihre Vision für die künstliche Intelligenz (KI) dargelegt, die eine „ethische, sichere und hochmoderne KI made in Europe“ unterstützt[5]. Die Vision der Kommission ruht auf drei Säulen: i) Erhöhung der öffentlichen und privaten Investitionen in KI, um ihre Verbreitung zu beschleunigen, ii) Vorbereitung auf sozio-ökonomische Veränderungen und iii) Gewährleistung eines angemessenen ethischen und rechtlichen Rahmens zur Stärkung der europäischen Werte.

(7) Zur Unterstützung der Umsetzung dieser Vision hat die Kommission eine hochrangige Expertengruppe für künstliche Intelligenz (HEG-KI) eingesetzt, eine unabhängige Gruppe, die damit beauftragt wurde, 1) KI-Ethik-Leitlinien und 2) KI-Politik- und -Investitionsempfehlungen zu erarbeiten.

(8) Das vorliegende Dokument enthält die KI-Ethik-Leitlinien, die nach weiteren Beratungen unserer Gruppe unter Berücksichtigung der aus den öffentlichen Konsultationen zu dem am 18. Dezember 2018 veröffentlichten Entwurf eingegangenen Rückmeldungen überarbeitet wurden. Es baut auf der Arbeit der Europäischen Gruppe für Ethik der Naturwissenschaften und der Neuen Technologien[6] auf und bezieht seine Inspiration aus vergleichbaren Bemühungen[7].

(9) In den vergangenen Monaten hat sich unsere 52-köpfige Gruppe zu Beratungen getroffen, diskutiert und interagiert und dabei getreu dem europäischen Motto „In Vielfalt geeint“ gearbeitet. Wir sind der Überzeugung, dass die KI das Potenzial hat, die Gesellschaft signifikant zu transformieren. Die KI ist kein Selbstzweck, sondern ein vielversprechendes Mittel, um das menschliche Gedeihen und somit das Wohlbefinden von Individuum und Gesellschaft und das Gemeinwohl zu steigern sowie zur Förderung von Fortschritt und Innovation beizutragen. Insbesondere können KI-Systeme dabei helfen, die Ziele für nachhaltige Entwicklung der Vereinten Nationen zu erreichen, z. B. beim Thema Geschlechtergerechtigkeit, bei der Bekämpfung des Klimawandels, beim rationalen Umgang mit natürlichen Ressourcen, bei der Gesundheitsförderung sowie bei Mobilität und Produktionsverfahren; des Weiteren können uns solche Systeme auch bei der Überwachung von Indikatoren helfen, die Fortschritte in den Bereichen Nachhaltigkeit und sozialer Zusammenhalt messen.

(10) Um dies zu erreichen, müssen KI-Systeme[8] **auf den Menschen ausgerichtet** sein und auf der verpflichtenden Grundlage stehen, dass ihre Nutzung im Dienste der Menschheit und des Gemeinwohls steht, mit dem Ziel, menschliches Wohl und menschliche Freiheit zu mehren. Obwohl KI-Systeme großartige Chancen bieten, entstehen durch sie auch bestimmte Risiken, die angemessen und verhältnismäßig behandelt werden müssen. Wir haben jetzt die wichtige und günstige Gelegenheit, auf die Entwicklung dieser Systeme gestalterischen Einfluss zu nehmen. Wir wollen gewährleisten, dass wir den sozio-technischen Umgebungen, in die sie eingebettet sind, vertrauen können, und wir wollen erreichen, dass die Hersteller von KI-Systemen dadurch einen Wettbewerbsvorteil erlangen, dass sie die vertrauenswürdige KI in ihre Produkte und Dienstleistungen integrieren. Hierzu gehört, dass die **Vorteile von KI-Systemen maximiert** und gleichzeitig **ihre Risiken ausgeschlossen bzw. minimiert** werden sollen.

(11) Wir sind der Ansicht, dass es im Kontext des schnellen technologischen Wandels unabdingbar ist, dass Vertrauen auch in Zukunft das Element bleibt, das Gesellschaften, Gemeinschaften, Wirtschaftsräume und nachhaltige Entwicklung zusammenhält. Deshalb bestimmen wir **vertrauenswürdige KI als unsere**

---

[5] COM(2018) 237 und COM(2018) 795. Bitte beachten Sie, dass der Begriff „made in Europe“ in der gesamten Mitteilung der Kommission verwendet wird. Trotzdem sollen die vorliegenden Leitlinien nicht nur für die in Europa hergestellten KI-Systeme gelten, sondern auch für anderswo entwickelte Systeme, die nach Europa eingeführt und hier verwendet werden. Im gesamten vorliegenden Dokument soll deshalb unser Ziel zum Ausdruck kommen, eine vertrauenswürdige KI „für“ Europa zu fördern.

[6] Die Europäische Gruppe für Ethik der Naturwissenschaften und der Neuen Technologien (EGE) ist eine Beratungsgruppe der Kommission.

[7] Siehe Abschnitt 3.3 in COM(2018) 237.

[8] Das Glossar am Ende dieses Dokuments enthält eine Definition von KI-Systemen, die für die Zwecke dieses Dokuments verwendet wird. Auf diese Definition wird in einem gesonderten, diesen Leitlinien beiliegenden Dokument näher eingegangen. Es wurde von der HEG-KI ausgearbeitet und trägt den Titel „Eine Definition der KI: Wichtigste Fähigkeiten und Wissenschaftsgebiete“.

grundlegende Ambition, denn Menschen und Gemeinschaften können der Entwicklung und Anwendung von Technologien nur dann vertrauen, wenn ein klarer und umfassender Rahmen existiert, der Vertrauenswürdigkeit gewährleistet.

(12) Wir glauben, dass Europa diesen Weg beschreiten sollte, um sich selbst als Heimat und Vorreiter für innovative und ethische Technologien in der Welt zu positionieren. Als europäische Bürgerinnen und Bürger wollen wir die Vorteile der vertrauenswürdigen KI so nutzen, dass unsere Grundwerte und die Achtung von Menschenrechten, Demokratie und Rechtsstaatlichkeit jederzeit gewahrt bleiben.

Vertrauenswürdige KI

(13) Vertrauenswürdigkeit ist eine Grundvoraussetzung dafür, dass Menschen und Gesellschaften KI-Systeme entwickeln, einführen und nutzen. Wenn KI-Systeme und die dahinterstehenden Menschen nicht bewiesenermaßen vertrauenswürdig sind, könnten daraus resultierende unerwünschte Konsequenzen zur Folge haben, dass ihre Akzeptanz möglicherweise untergraben und dadurch die Verwirklichung der potenziell gewaltigen sozialen und ökonomischen Vorteile von KI-Systemen verhindert wird. Mit unserer Vision, die Ethik als Grundpfeiler zur Gewährleistung und Skalierung der vertrauenswürdigen KI heranzuziehen, wollen wir Europa bei der Nutzbarmachung dieser Vorteile helfen.

(14) Das Vertrauen in die Entwicklung, Einführung und Nutzung von KI-Systemen ist nicht nur für die inhärenten Eigenschaften der Technologie, sondern auch für die Qualitätsmerkmale der sozio-technischen Systeme mit KI-Anwendungsmöglichkeiten von Bedeutung[9]. Analog zu Fragen des Vertrauens (bzw. des Vertrauensverlustes) in die Luftfahrt, Kernkraft oder Lebensmittelsicherheit sind es nicht bloß die einzelnen Bestandteile des KI-Systems, sondern das System in seinem gesamten Kontext, das Vertrauen entweder schafft oder zerstört. Die Arbeit an einer vertrauenswürdigen KI betrifft deshalb nicht nur die Vertrauenswürdigkeit des KI-Systems an sich, sondern sie bedarf eines holistischen und systemischen Ansatzes, der die Vertrauenswürdigkeit aller Beteiligten und der entsprechenden Prozesse als Teil des sozio-technischen Kontextes des Systems während seines gesamten Lebenszyklus umfasst.

(15) Eine vertrauenswürdige KI zeichnet sich durch **drei Komponenten** aus, die während des gesamten Lebenszyklus des Systems erfüllt sein sollten:

1. Sie sollte **rechtmäßig** sein und somit geltendes Recht und alle gesetzlichen Bestimmungen einhalten;
2. sie sollte **ethisch** sein und somit die Einhaltung ethischer Grundsätze und Werte garantieren; und
3. sie sollte **robust** sein, und zwar sowohl in technischer als auch in sozialer Hinsicht, da KI-Systeme möglicherweise unbeabsichtigten Schaden verursachen, selbst wenn ihnen gute Absichten zugrunde liegen.

(16) Jede dieser drei Komponenten ist notwendig, jedoch alleine nicht ausreichend, um das Ziel einer vertrauenswürdigen KI zu erreichen[10]. Idealerweise wirken alle drei Komponenten harmonisch zusammen und überlappen sich in ihrer Funktionsweise. In der Praxis kommt es jedoch möglicherweise zu Spannungen zwischen den Komponenten (z. B. stehen der Geltungsbereich und der Inhalt geltender Gesetze nicht notwendigerweise immer mit den ethischen Normen im Einklang). Es liegt in unserer individuellen und kollektiven Verantwortung als Gesellschaft, daran zu arbeiten, dass alle drei Komponenten zusammenwirken und zur Gewährleistung einer vertrauenswürdigen KI beitragen.[11]

(17) Ein vertrauenswürdiges Herangehen ist der Schlüssel zu einer „verantwortlichen Wettbewerbsfähigkeit“. Es schafft die Grundlage dafür, dass alle von KI-Systemen betroffenen Personen darauf vertrauen können, dass Konstruktion, Entwicklung und Nutzung dieser Systeme rechtmäßig, ethisch und robust sind. Diese Leitlinien sollen verantwortungsvolle und nachhaltige KI-Innovationen in Europa fördern. Durch sie soll die Ethik zu

[9] Bei diesen Systemen spielen u. a. Menschen, staatliche Stellen, Unternehmen, Infrastrukturen, Software, Protokolle, Normen, Governance, geltende Gesetze, Überwachungsmechanismen, Anreizsysteme, Auditverfahren und Berichterstattung über bewährte Methoden eine Rolle.

[10] Dies schließt nicht aus, dass zusätzliche Bedingungen möglicherweise notwendig sind bzw. werden.

[11] Dies bedeutet auch, dass der Gesetzgeber oder politische Entscheidungsträger gegebenenfalls das geltende Recht daraufhin überprüfen müssen, ob es noch mit ethischen Grundsätzen im Einklang steht.

einem Grundpfeiler für die Entwicklung einer einzigartigen Herangehensweise an das Thema KI werden, die dem Vorteil, der Stärkung und dem Schutz des Wohls des einzelnen Menschen wie auch dem Gemeinwohl dient. Wir glauben, dass Europa sich auf diese Weise als führender Wirtschaftsraum auf dem Gebiet der hochmodernen KI, die unseres Vertrauens als Bürgerinnen und Bürger und als Gesellschaft würdig ist, in der Welt positionieren kann. Nur wenn die Vertrauenswürdigkeit gesichert ist, werden die europäischen Bürgerinnen und Bürger von den Vorteilen der KI-Systeme uneingeschränkt profitieren können, und zwar in der Gewissheit, dass Vorkehrungen zu ihrem Schutz vor möglichen Risiken getroffen worden sind.

(18) Genauso wie die Nutzung von KI-Systemen nicht an nationalen Grenzen halt macht, sind auch ihre Auswirkungen über Nationalstaaten hinweg spürbar. Globale Lösungen sind deshalb im Hinblick auf die durch KI entstehenden weltweiten Chancen und Herausforderungen geboten. Wir fordern deshalb alle Interessenträger dazu auf, an einem globalen Rahmen für eine vertrauenswürdige KI mitzuarbeiten und einen internationalen Konsens zu schaffen, der sich vor allem an der Förderung und Bewahrung unserer Grundrechte orientiert.

Zielgruppen und Umfang

(19) Diese Leitlinien richten sich an alle Akteure, die sich an der Gestaltung, Entwicklung, Einführung, Umsetzung oder Nutzung der KI beteiligen oder davon betroffen sind, also z. B. an Unternehmen, Organisationen, Forscherinnen und Forscher, öffentliche Dienste, Behörden, Institutionen, zivilgesellschaftliche Organisationen, Einzelpersonen, Arbeitnehmerinnen und Arbeitnehmer sowie Verbraucherinnen und Verbraucher. Die Interessenträger, die sich für die Verwirklichung einer vertrauenswürdigen KI engagieren, können sich freiwillig dafür entscheiden, diese Leitlinien als Methode zur Konkretisierung ihres Engagements im Rahmen der Entwicklung von KI-Systemen und Prozessen zu nutzen, wobei hier insbesondere auf die praktische Bewertungsliste in Kapitel III verwiesen wird. Die Bewertungsliste kann auch in bestehende Bewertungsverfahren integriert werden und diese ergänzen.

(20) Die Leitlinien bieten eine Hilfestellung für KI-Anwendungen im Allgemeinen und schaffen ein horizontales Fundament für eine vertrauenswürdige KI. **Unterschiedliche Situationen führen jedoch zu unterschiedlichen Herausforderungen.** KI-Systeme, die Musikempfehlungen erzeugen, werfen nicht dieselben ethischen Bedenken auf wie KI-Systeme, die kritische medizinische Behandlungen vorschlagen. Entsprechend ergeben sich aus KI-Systemen je nach Anwendungsbereich und Wirtschaftssektor unterschiedliche Chancen und Herausforderungen. So spielt es eine Rolle, ob Unternehmen und Kunden, Unternehmen untereinander, Arbeitgeber und Arbeitnehmer oder staatliche Stellen und Bürgerinnen und Bürger interagieren, wenn KI-Systeme zum Einsatz kommen. Angesichts der Spezifität von KI-Systemen ist es daher unstrittig, dass die Umsetzung der vorliegenden Leitlinien an die jeweilige KI-Anwendung angepasst werden muss. Es ist daher zu prüfen, ob ein zusätzlicher sektorenbezogener Ansatz als Ergänzung zum in diesem Dokument vorgeschlagenen, eher allgemein gehaltenen horizontalen Rahmen notwendig ist.

Um ein besseres Verständnis zu gewinnen, wie sich die Leitlinien auf horizontaler Ebene umsetzen lassen und welche Sachverhalte einen sektorbezogenen Ansatz erfordern, laden wir alle Beteiligten dazu ein, die Bewertungsliste für vertrauenswürdige KI (Kapitel III), welche den vorliegenden Rahmen konkretisiert, in einer Pilotphase zu testen und uns entsprechende Rückmeldungen zu geben. Anhand der während dieser Pilotphase gesammelten Rückmeldungen werden wir die Bewertungsliste der vorliegenden Leitlinien bis Anfang 2020 überarbeiten. Die Pilotphase beginnt spätestens im Sommer 2019 und dauert bis zum Jahresende. Alle Interessenträger können daran teilnehmen, wenn sie ihr Interesse über die Europäische KI-Allianz anmelden.

B. EIN RAHMEN FÜR EINE VERTRAUENSWÜRDIGE KI

(21) Diese Leitlinien formulieren den Rahmen zur Verwirklichung einer vertrauenswürdigen KI auf der Basis der in der Charta der Grundrechte der Europäischen Union (EU-Grundrechtecharta) und in einschlägigen internationalen Menschenrechtsbestimmungen verankerten Grundrechte. Im Folgenden sprechen wir die drei Komponenten der vertrauenswürdigen KI kurz an.

Rechtmäßige KI

(22) KI-Systeme verrichten ihre Funktion nicht in einer rechtslosen Welt. Eine Reihe rechtverbindlicher Vorschriften gilt bereits heute auf europäischer, nationaler und internationaler Ebene oder sind für die Entwicklung, Einführung und Nutzung von KI-Systemen von Bedeutung. Relevante Rechtsquellen sind unter anderem das EU-Primärrecht (die EU-Verträge und die EU-Grundrechtecharta), EU-Sekundärrecht (z. B. die Datenschutz-Grundverordnung, Antidiskriminierungsrichtlinien, die Produktsicherheitsrichtlinie, die Produkthaftungsrichtlinie, die Verordnung über den freien Verkehr nicht-personenbezogener Daten, Richtlinien über Verbraucherschutz sowie Gesundheitsschutz und Sicherheit am Arbeitsplatz), aber auch die Menschenrechtsverträge der Vereinten Nationen und die Übereinkommen des Europarates (z. B. die Europäische Menschenrechtskonvention) sowie zahlreiche Gesetze der EU-Mitgliedstaaten. Neben den horizontal geltenden Vorschriften existieren auch verschiedene fachspezifische Vorschriften, die für bestimmte KI-Anwendungen gelten (z. B. die Verordnung über Medizinprodukte im Gesundheitssektor).

(23) Das geltende Recht schreibt sowohl positive wie auch negative Verpflichtungen vor. Das heißt, dass Gesetze nicht nur im Sinne von Verboten (Was *darf* man nicht?), sondern auch im Sinne von Geboten (Was *soll* man tun?) betrachtet werden sollten. Das geltende Recht verbietet nicht nur bestimmtes Handeln, sondern es bietet auch Handlungsoptionen. In diesem Zusammenhang sei erwähnt, dass die EU-Grundrechtecharta neben Artikeln, die sich mit uns im Hinblick auf die Gewährleistung einer vertrauenswürdigen KI eher vertrauten Bereichen wie Datenschutz und Nichtdiskriminierung beschäftigen, auch Artikel über die „unternehmerische Freiheit“ und die „Freiheit der Kunst und der Wissenschaften“ enthält.

(24) Die Leitlinien beschäftigen sich nicht explizit mit der ersten Komponente der vertrauenswürdigen KI (rechtmäßige KI), sondern sie sollen vielmehr eine Hilfestellung für die Förderung und Sicherung der Komponenten zwei und drei (ethische und robuste KI) bieten. Während sich die beiden letzteren bis zu einem gewissen Grad bereits häufig in bestehenden Gesetzen niederschlagen, geht ihre vollständige Verwirklichung möglicherweise über die bestehenden gesetzlichen Verpflichtungen hinaus.

(25) Die im vorliegenden Dokument getroffenen Aussagen dürfen keinesfalls dahin gehend ausgelegt werden, dass sie eine Form der Rechtsberatung oder rechtliche Leitlinien darstellen, wenn es um die Frage geht, wie geltende rechtliche Normen und Anforderungen eingehalten werden können. Keine der im vorliegenden Dokument getroffenen Aussagen begründet irgendwelche Rechte bzw. gesetzliche oder rechtliche Verpflichtungen gegenüber Dritten. Wir weisen jedoch darauf hin, dass natürliche oder juristische Personen zur Einhaltung von Gesetzen verpflichtet sind. Dies gilt für bereits heute geltende Gesetze und für Gesetze, die in Zukunft hinsichtlich der KI-Entwicklung erlassen werden. Diese Leitlinien basieren auf der Annahme, dass **alle gesetzlichen Rechte und Pflichten, die für die Prozesse und Aktivitäten zur Entwicklung, Einführung und Nutzung von KI-Systemen gelten, auch weiterhin verbindlich sein werden und entsprechend zu beachten sind.**

Ethische KI

(26) Die Verwirklichung einer vertrauenswürdigen KI erfordert nicht nur das Einhalten von Gesetzen, denn dies ist nur eine ihrer drei Komponenten. Gesetze halten nicht immer mit technologischen Entwicklungen Schritt. Sie stehen nicht notwendigerweise immer im Verhältnis mit ethischen Normen oder sind möglicherweise schlicht nicht dafür geeignet, bestimmte Probleme zu lösen. Damit KI-Systeme vertrauenswürdig sind, müssen sie also außerdem ethisch sein und unbedingt mit ethischen Normen in Einklang stehen.

Robuste KI

(27) Selbst wenn gewährleistet ist, dass KI-Systeme einem ethischen Zweck dienen, müssen die Bürgerinnen und Bürger und die Gesellschaft außerdem darauf vertrauen können, dass sie keinen unbeabsichtigten Schaden anrichten. Solche Systeme sollten ihre Funktionen auf abgesicherte und zuverlässige Weise erfüllen. Sicherheitsmaßnahmen sollten im Vorfeld getroffen werden, um unerwünschte negative Auswirkungen zu verhindern. Es ist deshalb unbedingt dafür zu sorgen, dass KI-Systeme robust sind. Diese Notwendigkeit ergibt sich sowohl aus technischer Perspektive (Gewährleistung der technischen Robustheit des Systems in einem gegebenen Kontext, z. B. im jeweiligen Anwendungsgebiet oder der jeweiligen Phase des Lebenszyklus) als

auch aus sozialer Perspektive (unter angemessener Berücksichtigung des Kontextes und der Umgebung, in der das System operiert). Die Konzepte der ethischen und robusten KI sind also eng miteinander verbunden und ergänzen sich gegenseitig. Die in Kapitel I dargelegten Prinzipien und die daraus abgeleiteten Anforderungen in Kapitel II beziehen sich auf beide Komponenten.

Der Rahmen

(28) Die Hilfestellung im vorliegenden Dokument gliedert sich in drei Abstraktionsebenen, und zwar von den am stärksten abstrahierten Überlegungen in Kapitel I bis hin zu konkreten Hinweisen in Kapitel III:

I) Fundamente einer vertrauenswürdigen KI. In Kapitel I werden die Fundamente einer vertrauenswürdigen KI dargelegt und ihr auf den Grundrechten[12] beruhender Ansatz erläutert. Es werden die ethischen Grundsätze benannt und erläutert, die zur Gewährleistung einer ethischen und robusten KI befolgt werden müssen.

II) Verwirklichung einer vertrauenswürdigen KI. In Kapitel II werden diese ethischen Grundsätze in sieben Anforderungen übersetzt, die KI-Systeme während ihres gesamten Lebenszyklus umsetzen und erfüllen sollten. Darüber hinaus werden sowohl technische als auch nicht-technische Methoden angeboten, die bei der Umsetzung angewendet werden können.

III) Bewertung einer vertrauenswürdigen KI. Akteure im Bereich der KI erwarten konkrete Leitlinien. Kapitel III enthält deshalb eine vorläufige und nicht erschöpfende Bewertungsliste für vertrauenswürdige KI, die dazu dienen soll, die in Kapitel II genannten Anforderungen zu konkretisieren. Diese Bewertung sollte auf die Anwendung des jeweiligen Systems zugeschnitten sein.

(29) Der letzte Abschnitt des Dokuments beschäftigt sich mit Chancen und von KI-Systemen aufgeworfenen kritischen Fragen, die zu untersuchen sind und über die wir eine weitere Debatte anregen möchten.

(30) Die Struktur der Leitlinien wird in der folgenden *Abbildung 1* dargestellt.

[12] Die Grundrechte bilden das Fundament sowohl für internationale als auch für EU-Menschenrechtsbestimmungen und sie stützen die auf dem Rechtsweg durchsetzbaren und durch die EU-Verträge und die EU-Grundrechtecharta garantierten Rechte. Die Einhaltung der rechtsverbindlichen Grundrechte fällt somit unter die erste Komponente einer vertrauenswürdigen KI („rechtmäßige KI"). Grundrechte können allerdings auch so verstanden werden, dass sich in ihnen besondere moralische Ansprüche aller Bürgerinnen und Bürger niederschlagen, die sich allein aus ihrer Menschlichkeit ergeben, und zwar ungeachtet ihres rechtsverbindlichen Charakters. In diesem Sinne sind sie auch Bestandteil der zweiten Komponente einer vertrauenswürdigen KI („ethische KI").

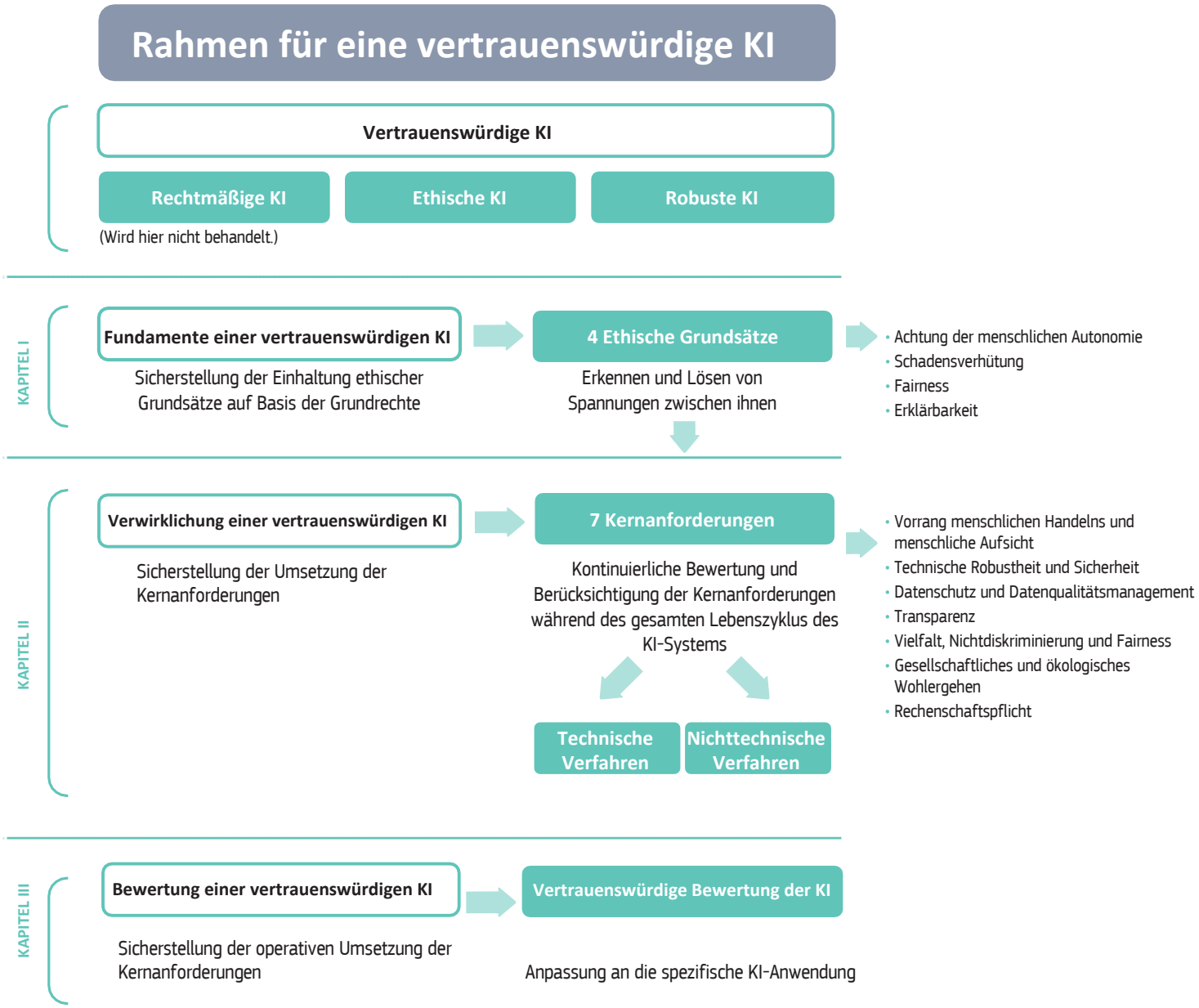

Abbildung 1: Die Leitlinien als Rahmen für eine vertrauenswürdige KI

I. Kapitel I: Fundamente einer vertrauenswürdigen KI

(31) In diesem Kapitel werden die Fundamente einer vertrauenswürdigen KI beschrieben, die sich aus den Grundrechten herleiten und die sich in vier ethischen Grundsätzen niederschlagen, welche zur Gewährleistung einer ethischen und robusten KI zu befolgen sind. Dieses Kapitel nimmt sehr stark auf den Fachbereich der Ethik Bezug.

(32) Die KI-Ethik ist ein Teilbereich der angewandten Ethik und beschäftigt sich mit den ethischen Fragen, die durch die Entwicklung, Einführung und Nutzung von KI aufgeworfen werden. Von zentraler Bedeutung ist dabei die Frage, inwiefern die KI das Leben von Bürgerinnen und Bürgern verbessern kann bzw. welche Bedenken dabei aufgeworfen werden, ob im Hinblick auf die Lebensqualität oder die für eine demokratische Gesellschaft notwendige Autonomie und Freiheit des Menschen.

(33) Die ethische Reflexion über KI-Technologie kann unterschiedlichen Zwecken dienen. Erstens kann sie Überlegungen hinsichtlich der Notwendigkeit anregen, Personen und Gruppen grundsätzlich zu schützen. Zweitens kann sie neue Arten von Innovationen stimulieren, die ethische Werte fördern sollen, z. B. mit dem Zweck, die Ziele für nachhaltige Entwicklung der Vereinten Nationen[13] zu erreichen, welche fester Bestandteil der kommenden Agenda 2030 der EU sein werden[14]. Während sich dieses Dokument hauptsächlich mit dem ersten genannten Zweck beschäftigt, sollte die Bedeutung von Ethik für den zweiten nicht unterschätzt werden. Eine vertrauenswürdige KI kann das Wohl des einzelnen Menschen und das Gemeinwohl durch Schaffung und Mehrung von Wohlstand und durch Wertschöpfung verbessern. Sie kann zum Entstehen einer fairen Gesellschaft beitragen, indem sie dabei hilft, Gesundheit und Wohlergehen von Bürgerinnen und Bürgern so zu verbessern, dass wirtschaftliche, soziale und politische Chancengleichheit gefördert wird.

(34) Wir müssen deshalb unbedingt verstehen, wie wir die Entwicklung, Einführung und Nutzung von KI am besten fördern, damit gewährleistet ist, dass sich jeder Einzelne in einer KI-gestützten Welt selbst verwirklichen kann und dass wir eine bessere Zukunft schaffen, während wir gleichzeitig global wettbewerbsfähig bleiben. Wie jede einflussreiche Technologie stellt uns auch die Nutzung von KI-Systemen in unserer Gesellschaft vor verschiedene ethische Herausforderungen, z. B. im Hinblick darauf, wie sie sich auf Mensch und Gesellschaft, auf die Sicherheit und auf die Fähigkeit auswirkt, Entscheidungen zu treffen. Wenn wir in immer stärkerem Maße die Hilfe von KI-Systemen zum Treffen von Entscheidungen in Anspruch nehmen oder Entscheidungen sogar an sie delegieren, müssen wir sicherstellen, dass diese Systeme sich nicht in unfairer Weise auf das Leben von Menschen auswirken, dass sie mit unabdingbaren Werten im Einklang stehen und entsprechend handeln können und dass geeignete Rechenschaftsprozesse dies gewährleisten können.

(35) Europa muss definieren, welche normative Vision einer von KI durchdrungenen Zukunft es verwirklichen möchte und folglich verstehen, welche Vorstellung von KI in Europa erforscht, entwickelt, eingeführt und genutzt werden sollte, um diese Vision zu erreichen. Mit dem vorliegenden Dokument wollen wir zu diesem Unterfangen beitragen, indem wir die Vorstellung einer vertrauenswürdigen KI einführen, die unserer Meinung nach den richtigen Weg in eine Zukunft mit KI darstellt. Eine Zukunft, in der Demokratie, Rechtsstaatlichkeit und Grundrechte die Basis für KI-Systeme bilden und in der solche Systeme die demokratische Kultur ständig verbessern und verteidigen, kann auch eine Umgebung schaffen, in der Innovation und Wettbewerbsfähigkeit gedeihen.

(36) Ein fachspezifischer Ethikkodex, so einheitlich, hochentwickelt und exakt dieser auch in Zukunft sein mag, kann niemals ein Ersatz für die ethische Vernunft an sich sein; letztere muss stets Einzelheiten im bestehenden Kontext aufgreifen, die sich nicht in allgemeinen Richtlinien erfassen lassen. Die Entwicklung eines Regelwerks reicht nicht aus, wenn wir eine vertrauenswürdige KI gewährleisten wollen. Dazu müssen wir des Weiteren durch öffentliche Debatten, Bildung und praktisches Lernen eine ethische Kultur und Einstellung aufbauen und bewahren.

[13] https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_en

[14] https://sustainabledevelopment.un.org/?menu=1300

1. Grundrechte als moralischer und rechtlicher Anspruch

(37) Wir glauben an eine Herangehensweise an KI-Ethik, die auf den in den EU-Verträgen[15], in der Charta der Grundrechte der Europäischen Union (EU-Grundrechtecharta) und in den internationalen Menschenrechten verankerten Grundrechten beruht[16]. Die Achtung der Grundrechte innerhalb eines Rahmens der Demokratie und Rechtsstaatlichkeit bildet die aussichtsreichste Grundlage für die Bestimmung abstrakter ethischer Grundsätze und Werte, die im Kontext der KI konkretisiert werden können.

(38) Die EU-Verträge und die EU-Grundrechtecharta schreiben eine Reihe von Grundrechten vor, deren Beachtung durch die EU-Mitgliedstaaten und die EU-Organe bei der Umsetzung von EU-Recht rechtlich bindend ist. Diese Rechte werden in der EU-Grundrechtecharta unter Bezugnahme auf Würde, Freiheiten, Gleichheit und Solidarität, Bürgerrechte und Gerechtigkeit beschrieben. Die gemeinsame Grundlage dieser Rechte ist die Achtung der Würde des Menschen. Es handelt sich also um einen „menschenzentrierten Ansatz", wobei der Mensch eine einzigartige und unveräußerliche moralische Vorrangstellung in den Bereichen Zivilgesellschaft, Politik, Wirtschaft und Soziales einnimmt.[17]

(39) Obwohl die in der EU-Grundrechtecharta verankerten Rechte rechtsverbindlich sind, muss unbedingt berücksichtigt werden*[18]*, dass die Grundrechte nicht in jedem Fall einen umfassenden rechtlichen Schutz bieten. Im Hinblick auf die EU-Grundrechtecharta ist beispielsweise zu unterstreichen, dass ihre Anwendung auf Bereiche des EU-Rechts beschränkt ist. Internationale Menschenrechtsgesetze und insbesondere die Europäische Menschenrechtskonvention sind für EU-Mitgliedstaaten auch in Bereichen, die außerhalb des Geltungsbereichs von EU-Recht liegen, rechtsverbindlich. Gleichzeitig ist zu unterstreichen, dass die Grundrechte einzelnen Personen und (bis zu einem gewissen Grad) auch Gruppen aufgrund ihrer moralischen Eigenschaft als Menschen zuteilwerden, unabhängig von ihrer Rechtskraft. Als durchsetzbare Rechte verstanden, fallen Grundrechte daher unter die erste Komponente der vertrauenswürdigen KI (rechtmäßige KI), welche für die Einhaltung von Recht und Gesetz sorgt. Als Rechte eines jeden Menschen verstanden, die in der inhärenten moralischen Eigenschaft des Menschen an sich verwurzelt ist, untermauern sie außerdem die zweite Komponente der vertrauenswürdigen KI (ethische KI), wobei es um ethische Normen geht, die nicht notwendigerweise rechtsverbindlich, aber dennoch unabdingbar für das Gewährleisten von Vertrauenswürdigkeit sind. Da dieses Dokument nicht als Leitfaden zur erstgenannten Komponente dienen soll, beziehen sich Verweise auf Grundrechte im Sinne der vorliegenden nicht-verbindlichen Leitlinien auf letztere Komponente.

2. Von Grundrechten zu ethischen Grundsätzen

2.1 Grundrechte als Basis für eine vertrauenswürdige KI

(40) Unter den umfassenden, unteilbaren Rechten, die durch internationale Menschenrechte, die EU-Verträge und die EU-Grundrechtecharta festgelegt sind, eignen sich die folgenden Kategorien von Grundrechten besonders gut für die Anwendung auf KI-Systeme. Viele dieser Rechte sind unter bestimmten Umständen rechtlich in der EU durchsetzbar, sodass die Einhaltung ihrer Bestimmungen rechtsverbindlich ist. Aber selbst wenn die Einhaltung rechtlich durchsetzbarer Grundrechte verwirklicht worden ist, kann eine ethische Reflexion uns dabei zu verstehen helfen, wie Entwicklung und Nutzung von KI möglicherweise Grundrechte und ihre zugrunde liegenden Werte implizieren, und sie kann dabei helfen, exaktere Leitlinien aufzustellen, wenn es darum geht festzustellen, was wir tun *sollten*, anstatt was wir (derzeit) mit der Technologie tun *können*.

---

[15] Die EU basiert auf der verfassungsmäßigen Verpflichtung zum Schutz der fundamentalen und unteilbaren Menschenrechte, zur Achtung von Rechtsstaatlichkeit, zur Förderung demokratischer Freiheit und des Gemeinwohls. Diese Rechte spiegeln sich in den Artikeln 2 und 3 des Vertrags über die Europäische Union und in der EU-Grundrechtecharta wider.

[16] In anderen Rechtsvorschriften werden diese Verpflichtungen aufgegriffen und weiter konkretisiert, so z. B. in der Europäischen Sozialcharta des Europarats oder in einschlägigen Rechtsvorschriften wie der Datenschutz-Grundverordnung der EU.

[17] Es sei darauf hingewiesen, dass ein Bekenntnis zu einer menschenzentrierten KI und ihre Verankerung in den Grundrechten kollektive gesellschaftliche und konstitutionelle Grundlagen erfordert, in deren Rahmen die individuelle Freiheit und die Achtung der Menschenrechte sowohl praktisch möglich als auch sinnvoll sind, anstatt ein über die Maßen individualistisches Menschenbild zu implizieren.

[18] Gemäß Artikel 51 der Charta gilt sie für EU-Organe und EU-Mitgliedstaaten bei der Umsetzung von EU-Recht.

(41) ***Achtung der Menschenwürde.*** Die menschliche Würde basiert auf dem Gedanken, dass jeder Mensch einen „inhärenten Wert" besitzt, der niemals durch Andere – oder durch neue Technologien wie KI-Systeme geschmälert, kompromittiert oder unterdrückt werden darf.[19] Im Kontext der KI gebietet es die Achtung der Würde des Menschen, dass alle Menschen mit Respekt zu behandeln sind, da es sich um moralische *Subjekte* und nicht um bloße *Objekte* handelt, die es zu sieben, zu sortieren, zu bewerten, zu gruppieren, zu konditionieren oder zu manipulieren gilt. KI-Systeme sollten daher so entwickelt werden, dass die körperliche und geistige Unversehrtheit des Menschen, seine persönliche und kulturelle Identität und die Erfüllung seiner Grundbedürfnisse geachtet, gefördert und geschützt werden.[20]

(42) ***Freiheit des Einzelnen****.* Die Menschen sollten die Freiheit besitzen, eigene Lebensentscheidungen zu treffen. Das bedeutet einerseits Freiheit von staatlichen Eingriffen, erfordert aber andererseits Maßnahmen durch die Regierung und Nichtregierungsorganisationen, die gewährleisten, dass von der Gefahr der Ausgrenzung betroffene Bürgerinnen und Bürger gleichermaßen Zugang zu den Vorteilen und Möglichkeiten der KI haben. In einem KI-bezogenen Kontext gebietet die Freiheit des Einzelnen die Eindämmung von (in)direktem unrechtmäßigem Zwang, von Bedrohungen für die geistige Selbstbestimmung und Gesundheit, von ungerechtfertigter Überwachung, Täuschung und unfairer Manipulation. In der Tat stellt die Freiheit des Einzelnen eine Verpflichtung dar, den Menschen die Fähigkeit zu geben, eine noch stärkere Kontrolle über das eigene Leben auszuüben. Bestandteile dessen sind (neben anderen Rechten) der Schutz der unternehmerischen Freiheit, der Freiheit der Kunst und der Wissenschaft, der Meinungsfreiheit, des Rechts auf Privatleben und Privatsphäre sowie des Versammlungs- und Vereinigungsrechts**.**

(43) ***Achtung von Demokratie, Gerechtigkeit und Rechtsstaatlichkeit.*** Alle Regierungsgewalt in konstitutionellen Demokratien muss gesetzlich geregelt und durch geltendes Recht begrenzt sein. KI-Systeme sollten dazu dienen, demokratische Prozesse zu bewahren und zu fördern sowie die Pluralität der individuellen Werte und Lebensentscheidungen respektieren. KI-Systeme dürfen demokratische Prozesse, menschliche Entscheidungen oder demokratische Abstimmungssysteme nicht beeinträchtigen. In KI-Systemen muss außerdem eine Verpflichtung eingebettet sein, die gewährleistet, dass sie keinesfalls in einer Weise operieren, welche die Grundlagen der Rechtsstaatlichkeit sowie Gesetze und Bestimmungen beeinträchtigt, und dass faire Verfahren und die Gleichheit vor dem Gesetz sichergestellt sind.

(44) ***Gleichheit, Nichtdiskriminierung und Solidarität*** *– einschließlich der Rechte von Personen mit Exklusionsrisiko.* Allen Menschen muss die gleiche Achtung vor ihrem moralischen Wert und ihrer Würde zuteilwerden. Diese Forderung geht über Nichtdiskriminierung hinaus, die Unterscheidungen zwischen unähnlichen Situationen auf der Basis objektiver Rechtfertigungen toleriert. In einem KI-bezogenen Kontext bedeutet Gleichheit, dass das System keine auf unfaire Weise verzerrten Ergebnisse liefern darf (z. B. sollten die zur Justierung der KI-Systeme verwendeten Daten so inklusiv wie möglich sein und verschiedene Bevölkerungsgruppen repräsentieren). Dies erfordert außerdem angemessenen Respekt für möglicherweise gefährdete Personen und Gruppen[21], z. B. Arbeitnehmer, Frauen, Menschen mit Behinderungen, ethnische Minderheiten, Kinder, Verbraucherinnen und Verbraucher oder andere Gruppen mit Exklusionsrisiko.

(45) ***Bürgerrechte.*** Die Bürgerinnen und Bürger genießen zahlreiche Rechte, darunter das Wahlrecht, das Recht auf eine gute Verwaltung oder den Zugang zu öffentlichen Dokumenten sowie das Petitionsrecht gegenüber der Regierung. Durch KI-Systeme ergibt sich ein erhebliches Potenzial zur Verbesserung der Leistungsfähigkeit und Effizienz von Regierungen bei der Bereitstellung öffentlicher Güter und Dienstleistungen an die Gesellschaft. Gleichzeitig könnten sich KI-Anwendungen aber auch negativ auf die Bürgerrechte auswirken, weshalb diese geschützt werden sollten. Wenn hier der Begriff „Bürgerrechte" verwendet wird, sollen damit keineswegs die Rechte von Staatsangehörigen von Drittstaaten oder von sich unrechtmäßig (oder illegal) in der EU aufhaltenden Personen geleugnet oder negiert werden, denen ebenfalls Rechte unter internationalem Recht –

[19] C. McCrudden, *Human Dignity and Judicial Interpretation of Human Rights*, EJIL, 19(4), 2008.

[20] Ein tieferes Verständnis der „Würde des Menschen" in diesem Zusammenhang vermittelt der Aufsatz von E. Hilgendorf, *Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity*, in: D. Grimm, A. Kemmerer, C. Möllers (Hrg.), *Human Dignity in Context. Explorations of a Contested Concept*, 2018, S. 325 ff.

[21] Eine Erläuterung des Begriffs, wie er im vorliegenden Dokument verwendet wird, finden Sie im Glossar.

und daher auch im Bereich der KI – zustehen.

2.2 Ethische Grundsätze im Kontext von KI-Systemen[22]

(46) Viele öffentliche, private und zivile Organisationen haben sich bei der Erstellung eines ethischen Rahmen für KI von den Grundrechten inspirieren lassen.[23] In der EU hat die Europäische Gruppe für Ethik der Naturwissenschaften und der Neuen Technologien („EGE“) insgesamt neun Grundsätze vorgeschlagen, die auf den Grundwerten der EU-Verträge und der EU-Grundrechtecharta beruhen.[24] Wir bauen auf diese Arbeit weiter auf und erkennen die meisten der bislang von verschiedenen Gruppen vertretenen Grundsätze an, während wir die Ziele klarstellen, die sämtliche Grundsätze fördern und unterstützen sollten. Diese ethischen Grundsätze können ein Ausgangspunkt für neue und spezifische Regulierungsinstrumente sein, können dabei helfen, Grundrechte im Zuge der Entwicklung unseres sozio-technischen Umfelds im Laufe der Zeit zu interpretieren, und sie können die Logik der Entwicklung, Nutzung und Umsetzung von KI-Systemen leiten – und sich analog zur dynamischen Entwicklung der Gesellschaft selbst anpassen.

(47) KI-Systeme sollten das Wohl des einzelnen Menschen und das Gemeinwohl fördern. In diesem Abschnitt werden **vier ethische Grundsätze** angeführt, die in den Grundrechten verwurzelt sind und beachtet werden müssen, um die Entwicklung, Einführung und Nutzung von KI-Systemen auf vertrauenswürdige Art und Weise zu gewährleisten. Sie werden als **ethische Imperative** formuliert, sodass KI-Akteure stets bestrebt sein sollten, sie zu befolgen. Wir zählen die Grundsätze im Folgenden ohne eine hierarchische Struktur auf. Die Reihenfolge ihrer Nennung richtet sich nach dem Erscheinen der Grundrechte, auf denen sie beruhen, in der EU-Grundrechtecharta[25].

(48) Dies sind die Grundsätze:

i) Achtung der menschlichen Autonomie
ii) Schadensverhütung
iii) Fairness
iv) Erklärbarkeit

(49) Viele dieser Grundsätze sind bereits zu einem großen Teil in den bestehenden und zu beachtenden Rechtsvorschriften berücksichtigt und fallen daher auch unter den Geltungsbereich der „rechtmäßigen KI“, die die erste Komponente der vertrauenswürdigen KI darstellt.[26] Obwohl jedoch, wie gerade erläutert, viele gesetzliche Verpflichtungen ethische Grundsätze widerspiegeln, geht die Einhaltung ethischer Grundsätze trotzdem über die Einhaltung geltender Gesetze hinaus.[27]

- Der Grundsatz der Achtung der menschlichen Autonomie

(50) Die Grundrechte, auf welche sich die EU gründet, sollen die Achtung der Freiheit und Autonomie des Menschen gewährleisten. Wenn Menschen mit KI-Systemen interagieren, müssen sie in der Lage sein, die Selbstbestimmung über die eigene Person in vollem Umfang und wirksam auszuüben und es muss ihnen

[22] Diese Grundsätze gelten auch für die Entwicklung, Einführung und Nutzung anderer Technologien und sind deshalb nicht spezifisch auf KI-Systeme zugeschnitten. Im Folgenden haben wir versucht, ihre Bedeutung insbesondere in einem KI-bezogenen Kontext darzulegen.

[23] Vertrauen in die Grundrechte hilft außerdem dabei, regulatorische Unsicherheiten zu begrenzen, da es möglich ist, auf Jahrzehnte der praktischen Erfahrung mit dem Schutz der Grundrechte in der EU zurückzugreifen, was zu Klarheit, Lesbarkeit und Vorhersagbarkeit führt.

[24] Kürzlich hat die Taskforce „AI4People“ die bereits erwähnten EGE-Grundsätze sowie 36 andere bislang vorgeschlagene ethische Grundsätze untersucht und sie unter vier übergeordneten Grundsätzen zusammengefasst: L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), *„AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations“*, Minds and Machines 28(4): 689-707.

[25] Die Achtung der menschlichen Autonomie ist mit dem Recht auf menschliche Würde und Freiheit (verankert in den Artikeln 1 und 6 der Charta) eng verbunden. Die Schadensverhütung ist mit dem Schutz der körperlichen oder geistigen Unversehrtheit (verankert in Artikel 3) eng verbunden. Fairness ist mit dem Recht auf Nichtdiskriminierung, Solidarität und Gerechtigkeit (verankert in Artikel 21 ff.) eng verbunden. Erklärbarkeit und Verantwortung sind mit den Rechten, die mit Gerechtigkeit in Zusammenhang stehen (verankert in Artikel 47) eng verbunden.

[26] Hier seien als Beispiel die DSGVO oder die Verbraucherschutzbestimmungen der EU genannt.

[27] Vertiefende Literatur zum Thema findet sich z. B. in L. Floridi, *Soft Ethics and the Governance of the Digital, Philosophy & Technology*, März 2018, Band 31, Ausgabe 1, S. 1–8.

möglich sein, am demokratischen Prozess teilzuhaben. KI-Systeme sollten Menschen nicht auf ungerechtfertigte Weise unterordnen, nötigen, täuschen, manipulieren, konditionieren oder in eine Gruppe drängen. KI-Systeme sollten vielmehr dazu dienen, die kognitiven, sozialen und kulturellen Fähigkeiten des Menschen zu stärken, zu ergänzen und zu fördern. Die Zuweisung von Funktionen zwischen Menschen und KI-Systemen sollte nach menschenzentrierten Entwicklungsgrundsätzen erfolgen und sinnvolle Spielräume für menschliche Entscheidungen lassen. Dies bedeutet die Sicherstellung der menschlichen Aufsicht[28] und Kontrolle über Arbeitsprozesse in KI-Systemen. KI-Systeme können auch die Arbeitswelt fundamental verändern. Sie sollten den Menschen in seiner Arbeitsumgebung unterstützen und der Schaffung sinnvoller Arbeit dienen.

- Der Grundsatz der Schadensverhütung

(51) KI-Systeme sollten Schäden weder verursachen noch verschärfen oder [29] sich auf andere Art und Weise [30]auf Menschen negativ auswirken. Hierzu gehört der Schutz der Menschenwürde sowie der geistigen und körperlichen Unversehrtheit. Die KI-Systeme und die Umgebungen, in denen sie operieren, müssen sicher und geschützt sein. Sie müssen technisch robust sein und es muss gewährleistet sein, dass sie nicht für Missbrauch anfällig sind. Auf schutzbedürftige Personen sollte in besonderem Maße geachtet werden und sie sollten in die Entwicklung und Einführung von KI-Systemen einbezogen werden. Ein besonderes Augenmerk ist auch auf Situationen zu legen, in denen KI-Systeme durch ungleiche Macht- oder Informationsverteilung negative Auswirkungen verursachen oder verschärfen können, etwa zwischen Arbeitgebern und Arbeitnehmern, Unternehmen und Verbrauchern oder Regierungen und Bürgerinnen und Bürgern. Zur Schadensverhütung gehört außerdem die Berücksichtigung der natürlichen Umwelt und aller Lebewesen.

- Der Grundsatz der Fairness

(52) Die Entwicklung, Einführung und Nutzung von KI-Systemen muss fair sein. Wir sind uns zwar bewusst, dass es viele verschiedene Interpretationen von Fairness gibt, glauben aber, dass Fairness sowohl eine substanzielle als auch eine verfahrenstechnische Dimension hat. Die substanzielle Dimension impliziert eine Verpflichtung zur Gewährleistung einer gleichen und gerechten Verteilung von Vorteilen und Kosten und die Gewährleistung, dass Personen und Gruppen vor unfairer Verzerrung, Diskriminierung und Stigmatisierung geschützt werden. Wenn unfaire Verzerrungen vermieden werden können, könnten KI-Systeme die gesellschaftliche Fairness sogar verbessern. Chancengleichheit beim Zugang zu Bildung, Gütern, Dienstleistungen und Technologie sollte ebenfalls gefördert werden. Des Weiteren sollte der Einsatz von KI-Systemen niemals dazu führen, dass (End-)Nutzer getäuscht oder in ihrer Wahlfreiheit beeinträchtigt werden. Darüber hinaus gebietet es die Fairness, dass KI-Akteure den Grundsatz der Verhältnismäßigkeit zwischen Mittel und Zweck beachten und sorgfältig abwägen, wie sich konkurrierende Interessen und Ziele ins Gleichgewicht bringen lassen.[31] Zur verfahrenstechnischen Dimension der Fairness gehört die Möglichkeit, sich gegen Entscheidungen der KI-Systeme und der sie betreibenden Menschen zu wehren und einen wirksamen Rechtsbehelf einzulegen.[32] Dazu müssen die für die Entscheidung verantwortliche Stelle identifizierbar und der Entscheidungsfindungsprozess erklärbar sein.

---

[28] Das Konzept der menschlichen Aufsicht wird im folgenden Absatz 65 näher erläutert.

[29] Schäden können sich auf den Einzelnen oder die Gemeinschaft auswirken und sie können auch als immaterielle Schäden in sozialen, kulturellen und politischen Umgebungen auftreten.

[30] Dies umfasst auch die Lebensweise von Personen und sozialen Gruppen, etwa die Vermeidung kultureller Schäden.

[31] Dies ist mit dem Grundsatz der Verhältnismäßigkeit verbunden (verankert in der Maxime, dass man nicht „mit Kanonen auf Spatzen schießen" soll). Zum Erreichen eines Ziels ergriffene Maßnahmen (z. B. Datenextraktionsmaßnahmen zur Umsetzung der KI-Optimierungsfunktion) sollten sich auf das absolut Notwendige beschränken. Das hat auch zur Folge, dass, wenn sich mehrere Maßnahmen zum Erreichen eines Ziels anbieten, diejenige vorzuziehen ist, die die Grundrechte und ethischen Normen am wenigsten beeinträchtigt (z. B. sollten KI-Entwickler stets eher auf Daten des öffentlichen Sektors als auf persönliche Daten zurückgreifen). Es sei ebenfalls auf die Verhältnismäßigkeit zwischen Benutzer und Betreiber hingewiesen, wobei die Rechte der Unternehmen einerseits (z. B. geistiges Eigentum und Vertraulichkeit) und die Rechte der Benutzer andererseits zu berücksichtigen sind.

[32] Dazu zählen in der Arbeitswelt auch die Vereinigungsfreiheit und das Recht auf die Mitgliedschaft in einer Gewerkschaft gemäß Artikel 12 der EU-Grundrechtecharta.

- Der Grundsatz der Erklärbarkeit

(53) Erklärbarkeit ist unabdingbar, wenn beim Benutzer dauerhaftes Vertrauen in KI-Systeme entstehen soll. Das bedeutet, dass Prozesse transparent sein müssen, dass die Fähigkeiten und der Zweck von KI-Systemen offen zu kommunizieren sind und dass Entscheidungen – im größtmöglichen Umfang – den direkt und indirekt davon betroffenen Personen erklärbar sein müssen. Ohne diese Informationen kann eine Entscheidung nicht ordnungsgemäß angefochten werden. Eine Erklärung, warum ein Modell ein bestimmtes Ergebnis oder eine bestimmte Entscheidung erzeugt hat (und welche Kombination aus Eingabefaktoren dazu geführt hat) ist nicht immer möglich. Diese Fälle werden als „Blackbox"-Algorithmen bezeichnet und erfordern besondere Beachtung. Unter diesen Umständen sind möglicherweise andere Erklärbarkeitsmaßnahmen notwendig (z. B. Rückverfolgbarkeit, Nachprüfbarkeit und transparente Kommunikation über die Fähigkeiten des Systems), solange das System als Ganzes Grundrechte achtet. Bis zu welchem Grad Erklärbarkeit notwendig ist, hängt sehr stark vom Kontext und der Tragweite der Konsequenzen eines fehlerhaften oder anderweitig unzutreffenden Ergebnisses ab.[33]

2.3 Spannungen zwischen den Grundsätzen

(54) Zwischen den genannten Grundsätzen können Spannungen entstehen, die sich nicht mit einer bestimmten Lösung ausräumen lassen. Da die EU ganz grundsätzlich dem demokratischen Engagement, dem ordnungsgemäßen Verfahren und der offenen politischen Teilhabe verpflichtet ist, sollten Methoden einer verantwortungsvollen Beratung für den Umgang mit solchen Spannungen festgelegt werden. So stehen in verschiedenen Anwendungsbereichen die Grundsätze der *Schadensverhütung* und der *menschlichen Autonomie* möglicherweise miteinander in Konflikt. Betrachten wir als Beispiel den Einsatz von KI-Systemen im Bereich ‚vorausschauende Polizeiarbeit', der möglicherweise bei der Verbrechensbekämpfung hilft, aber Methoden nutzt, die Überwachungsmaßnahmen erfordern und sich daher negativ auf die individuellen Freiheits- und Datenschutzrechte auswirken. Des Weiteren sollten die Vorteile von KI-Systemen insgesamt die vorhersehbaren individuellen Risiken erheblich überwiegen. Diese Grundsätze weisen zwar auf Lösungsansätze hin, bleiben jedoch abstrakte ethische Vorschriften. Von KI-Akteuren kann deshalb nicht erwartet werden, dass sie anhand der genannten Grundsätze die richtige Lösung finden; sie sollten jedoch an ethische Dilemmata und Kompromisse mit vernünftiger, auf Fakten gestützter Reflexion und nicht mit Intuition oder nach Gutdünken herangehen. Es kann jedoch zu Situationen kommen, in denen keine akzeptablen Kompromisse gefunden werden können. Bestimmte Grundrechte und mit ihnen verbundene Prinzipien sind absolut und können nicht gegen andere aufgewogen werden (z. B. die Würde des Menschen).

Wichtige Leitlinien aus Kapitel I:

- ✓ Die Entwicklung, Einführung und Nutzung von KI-Systemen muss so erfolgen, dass die folgenden ethischen Grundsätze eingehalten werden: *Achtung der menschlichen Autonomie, Schadensverhütung, Fairness und Erklärbarkeit.* Die möglichen Spannungen zwischen diesen Grundsätzen müssen zur Kenntnis genommen und gelöst werden.
- ✓ Ein besonderes Augenmerk muss dabei auf Situationen gelegt werden, in denen besonders schutzbedürftige Gruppen wie Kinder, Menschen mit Behinderungen und Minderheiten betroffen sind, die bereits in der Vergangenheit Benachteiligung erfahren haben oder die einem besonders hohen Exklusionsrisiko ausgesetzt sind. Gleiches gilt für Situationen, die sich durch ungleiche Macht- oder Informationsverteilung auszeichnen,[34] etwa zwischen Arbeitgebern und Arbeitnehmern oder Unternehmen und Verbrauchern.
- ✓ Es gilt anzuerkennen und zu berücksichtigen, dass KI-Systeme zwar das Potenzial haben, dem einzelnen Menschen und der Gesellschaft einen erheblichen Nutzen zu bringen, manche KI-Anwendungen sich

[33] Beispielsweise werden unpassende Einkaufsempfehlungen, die von einem KI-System erzeugt werden, kaum ethische Bedenken hervorrufen, ganz im Gegensatz zu KI-Systemen, die bewerten, ob ein Straftäter auf Bewährung freigelassen werden sollte.

[34] Siehe Artikel 24 bis 27 der EU-Charta über die Rechte des Kindes und älterer Menschen, die Integration von Menschen mit Behinderungen sowie die Arbeitnehmerrechte. Siehe auch Artikel 38 über Verbraucherschutz.

jedoch gleichzeitig möglicherweise negativ auswirken, wobei manche negative Auswirkungen möglicherweise schwer absehbar, erkennbar oder messbar sind (z. B. im Hinblick auf Demokratie, Rechtsstaatlichkeit, Verteilungsgerechtigkeit oder den menschlichen Geist als solchen). Zur Abwendung dieser Gefahren müssen gegebenenfalls angemessene und in Anbetracht der Höhe des Risikos verhältnismäßige Maßnahmen getroffen werden.

II. Kapitel II: Verwirklichung einer vertrauenswürdigen KI

(55) Dieses Kapitel bietet Hilfestellung für die Umsetzung und Verwirklichung einer vertrauenswürdigen KI mithilfe einer Liste mit sieben zu erfüllenden Anforderungen, aufbauend auf den in Kapitel I genannten Grundsätzen. Des Weiteren werden derzeit verfügbare technische und nicht-technische Methoden für die Umsetzung dieser Anforderungen während des gesamten Lebenszyklus des KI-Systems eingeführt.

1. Anforderungen an eine vertrauenswürdige KI

(56) Die in Kapitel I genannten Grundsätze müssen in konkrete Anforderungen zur Verwirklichung einer vertrauenswürdigen KI übersetzt werden. Diese Anforderungen gelten für verschiedene Beteiligte, die in den Lebenszyklus eines KI-Systems involviert sind: Entwickler, Betreiber und Endnutzer sowie die breitere Gesellschaft. Als Entwickler bezeichnen wir diejenigen, die KI-Systeme erforschen, entwerfen und/oder entwickeln. Als Betreiber bezeichnen wir öffentliche oder private Organisationen und Unternehmen, die KI-Systeme als Teil ihrer Geschäftsprozesse oder zum Bereitstellen von Produkten und Dienstleistungen an Dritte verwenden. Endnutzer sind Personen, die mit KI-Systemen direkt oder indirekt interagieren. Die breitere Gesellschaft schließlich besteht aus allen anderen Menschen, die direkt oder indirekt von KI-Systemen betroffen sind.

(57) Verschiedene Kategorien von Beteiligten nehmen unterschiedliche Rollen ein, wenn es darum geht, die Einhaltung der Anforderungen zu gewährleisten:

a. Die Entwickler sollten die Anforderungen umsetzen und auf die Konzeption und die Entwicklung von Prozessen anwenden;
b. die Betreiber sollten sicherstellen, dass die von ihnen eingesetzten Systeme sowie die angebotenen Produkte und Dienstleistungen die Anforderungen erfüllen;
c. die Endnutzer und die breitere Gesellschaft sollten über die Anforderungen informiert werden und in der Lage sein, auf ihre Einhaltung zu bestehen.

(58) Die folgende Liste mit Anforderungen erhebt keinen Anspruch auf Vollständigkeit[35]. Sie enthält systemische, individuelle und gesellschaftliche Aspekte:

1 Vorrang menschlichen Handelns und menschliche Aufsicht
z. B. Grundrechte, Vorrang menschlichen Handelns und menschliche Aufsicht

2 Technische Robustheit und Sicherheit
z. B. Widerstandsfähigkeit gegen Angriffe und Sicherheitsverletzungen, Auffangplan und allgemeine Sicherheit, Präzision, Zuverlässigkeit und Reproduzierbarkeit

3 Schutz der Privatsphäre und Datenqualitätsmanagement
z. B. Achtung der Privatsphäre, Qualität und Integrität der Daten sowie Datenzugriff

4 Transparenz
z. B. Nachverfolgbarkeit, Erklärbarkeit und Kommunikation

5 Vielfalt, Nichtdiskriminierung und Fairness
z. B. Vermeidung unfairer Verzerrungen, Zugänglichkeit und universeller Entwurf sowie Beteiligung der

[35] Wir zählen die Grundsätze im Folgenden ohne eine hierarchische Struktur auf. Die Reihenfolge ihrer Nennung richtet sich nach dem Erscheinen der Grundsätze und Grundrechte, auf denen sie beruhen, in der EU-Grundrechtecharta.

Interessenträger

6 **Gesellschaftliches und ökologisches Wohlergehen**

z. B. Nachhaltigkeit und Umweltschutz, soziale Auswirkungen, Gesellschaft und Demokratie

7 **Rechenschaftspflicht**

z. B. Nachprüfbarkeit, Minimierung und Meldung von negativen Auswirkungen, Kompromisse und Rechtsbehelfe.

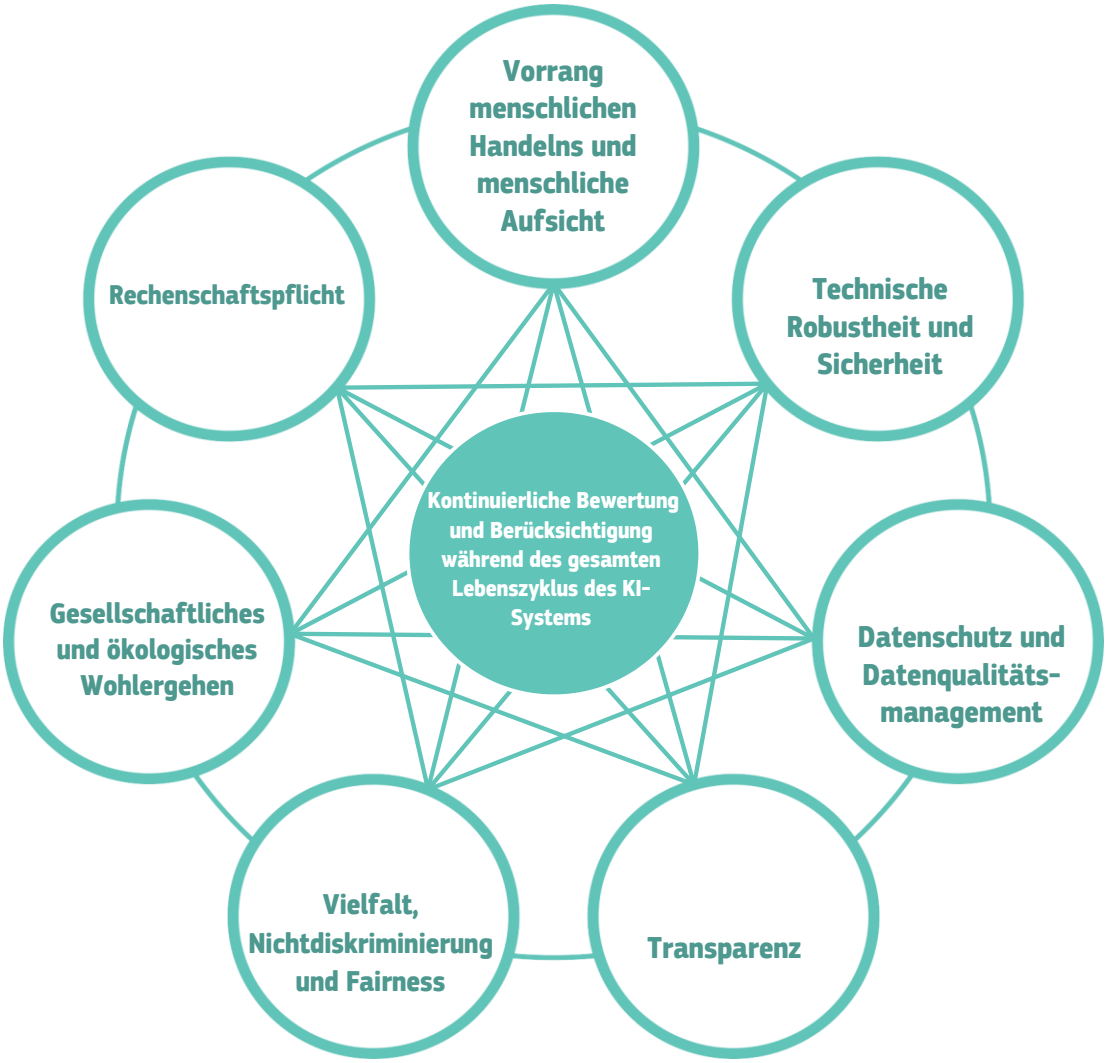

Abbildung 2: Beziehung zwischen den sieben Anforderungen: Alle sind in Bezug auf ihre Bedeutung gleichrangig, unterstützen sich gegenseitig und sollten während des gesamten Lebenszyklus eines KI-Systems umgesetzt und bewertet werden.

(59) Während alle Anforderungen in Bezug auf ihre Bedeutung gleichrangig sind, müssen ihr Kontext und die möglichen Spannungen zwischen ihnen bei der fachbereichs- und branchenübergreifenden Anwendung berücksichtigt werden. Die Umsetzung dieser Anwendungen sollte für den gesamten Lebenszyklus eines KI-Systems erfolgen und sich nach der spezifischen Anwendung richten. Zwar treffen die meisten Anforderungen auf sämtliche KI-Systeme zu, ein besonderes Augenmerk muss jedoch auf diejenigen gelegt werden, welche direkt oder indirekt Menschen betreffen. Für manche Anwendungen (z. B. in einer industriellen Umgebung) haben diese Anforderungen deshalb möglicherweise eine geringe Bedeutung.

(60) Die genannten Anforderungen umfassen Elemente, die in manchen Fällen bereits Bestandteil geltender Gesetze sind. Wir weisen nochmals mit Nachdruck darauf hin, dass es – gemäß der ersten Komponente der

vertrauenswürdigen KI – in der Verantwortung der Entwickler und Betreiber von KI-Systemen liegt, die Einhaltung ihrer gesetzlichen Verpflichtungen sowohl im Hinblick auf horizontal geltende Vorschriften als auch in Bezug auf fachspezifische Bestimmungen sicherzustellen.

(61) In den folgenden Abschnitten wird auf jede Anforderung näher eingegangen.

1. Vorrang menschlichen Handelns und menschliche Aufsicht

(62) KI-Systeme sollten die menschliche Autonomie und Entscheidungsfindung, wie es der Grundsatz der *Achtung der menschlichen Autonomie* vorsieht, unterstützen. Dies erfordert, dass KI-Systeme sowohl einer demokratischen, florierenden und gerechten Gesellschaft dienen, indem sie das menschliche Handeln und die Wahrung der Grundrechte fördern, als auch die menschliche Aufsicht ermöglichen.

(63) ***Grundrechte.*** Wie viele andere Technologien können KI-Systeme für Grundrechte entweder förderlich oder hinderlich sein. Sie können Menschen dadurch Nutzen bringen, dass sie ihnen z. B. dabei helfen, den Überblick über ihre persönlichen Daten zu behalten, oder einen besseren Zugang zu Bildung zu haben, wodurch ihr Recht auf Bildung gestärkt wird. Gerade im Hinblick auf die Reichweite und die Fähigkeit von KI-Systemen kann es aber auch zu negativen Auswirkungen auf die Grundrechte kommen. In solchen risikobehafteten Situationen sollte eine Folgenabschätzung der Auswirkungen auf die Grundrechte vorgenommen werden. Diese Maßnahme sollte vor der Entwicklung stehen und eine Beurteilung darüber enthalten, ob sich diese Risiken verringern lassen bzw. ob sie als in einer demokratischen Gesellschaft notwendig gerechtfertigt sind, da auch die Rechte und Freiheiten Dritter zur respektieren sind. Des Weiteren sollten Mechanismen zur Entgegennahme externer Rückmeldungen über möglicherweise Grundrechte beeinträchtigende KI-Systeme eingerichtet werden.

(64) ***Vorrang menschlichen Handelns.*** Die Benutzer sollten in der Lage sein, informierte Entscheidungen in Bezug auf KI-Systeme zu treffen. Ihnen sollte das notwendige Wissen und die Mittel zur Verfügung gestellt werden, um die KI-Systeme hinreichend zu verstehen und mit ihnen interagieren zu können, und sie sollten möglichst in die Lage versetzt werden, das System angemessen bewerten oder sich ihm entgegenstellen zu können. KI-Systeme sollten die einzelnen Menschen dabei unterstützen, im Einklang mit ihren eigenen Zielen bessere, fundiertere Entscheidungen zu treffen. KI-Systeme können manchmal so eingesetzt werden, dass sie menschliches Verhalten durch möglicherweise schwer zu erkennende Mechanismen gestalten und beeinflussen, da sie sich z. B. unterbewusste Prozesse zunutze machen, einschließlich unterschiedlicher Formen unfairer Manipulation, Täuschung, Bedrängung oder Konditionierung, die ausnahmslos die menschliche Autonomie bedrohen können. Die Richtschnur für das Funktionieren von KI-Systemen muss der Grundsatz der Selbstbestimmung des Nutzers sein. Entscheidend ist in diesem Zusammenhang das Recht einer Person, nicht einer ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt[36].

(65) ***Menschliche Aufsicht.*** Die menschliche Aufsicht hilft, dafür zu sorgen, dass ein KI-System die menschliche Autonomie nicht untergräbt oder sich sonst nachteilig auswirkt. Die Aufsicht kann durch Lenkungs- und Kontrollmechanismen wie die Gewährleistung der interaktiven Einbindung eines Menschen („*Human-in-the-Loop*“), der Überprüfung und Kontrolle durch einen Menschen („*Human-on-the-Loop*“) oder der Gesamtsteuerung durch einen Menschen („*Human-in-Command*“) erreicht werden. „*Human-in-the-Loop*“ (HITL) bezieht sich auf die Fähigkeit des Menschen, in jeden Entscheidungszyklus des Systems einzugreifen, was in vielen Fällen weder möglich noch wünschenswert wäre. „*Human-on-the-Loop*“ (HOTL) bedeutet die Möglichkeit des Menschen, in den Entwurfszyklus des Systems einzugreifen und den Systembetrieb zu überwachen. „*Human-in-Command*“ (HIC) bedeutet die Möglichkeit, den Gesamtbetrieb des KI-Systems (einschließlich seiner weiteren wirtschaftlichen, gesellschaftlichen, rechtlichen und ethischen Auswirkungen) zu beaufsichtigen, sowie die Fähigkeit zu entscheiden, wann und wie das System in einer bestimmten Situation

[36] Es kann hier auf Artikel 22 der DSGVO Bezug genommen werden, wo dieses Recht verankert ist.

eingesetzt werden soll. Dies beinhaltet auch die Entscheidungsmöglichkeit, ein KI-System in einer bestimmten Situation nicht einzusetzen, beim Einsatz des Systems ein bestimmtes Maß an menschlichem Ermessen zuzulassen oder sicherzustellen, dass eine vom System getroffene Entscheidung außer Kraft gesetzt werden kann. Des Weiteren muss sichergestellt werden, dass die Durchsetzungsbehörden stets in der Lage sind, die Aufsicht im Einklang mit ihrem jeweiligen Auftrag auszuüben. Aufsichtsmechanismen sind möglicherweise in unterschiedlichen Graden notwendig, um andere Sicherheits- und Kontrollmaßnahmen je nach Anwendungsbereich und potenziellem Risiko des KI-Systems zu unterstützen. Für ein ansonsten gleiches System muss gelten: Je weniger Aufsicht ein Mensch über ein KI-System ausüben kann, desto ausführlicher muss es zuvor getestet werden und desto strenger muss die Lenkung und Kontrolle sein.

2. Technische Robustheit und Sicherheit

(66) Eine entscheidende Komponente bei der Verwirklichung einer vertrauenswürdigen KI ist die technische Robustheit, die mit dem *Grundsatz der Schadensverhütung* eng verbunden ist. Die technische Robustheit macht es erforderlich, dass KI-Systeme mit einem präventiven Herangehen an Risiken und so entwickelt werden, dass sie sich zuverlässig gemäß ihrer Bestimmung verhalten, bei gleichzeitiger Minimierung von unbeabsichtigtem und unerwartetem Schaden und unter Verhinderung von inakzeptablem Schaden. Dies sollte auch für potenziell veränderte Betriebsumgebungen oder die Anwesenheit von anderen (menschlichen oder künstlichen) wirkenden Kräften gelten, die gegebenenfalls in feindlicher Art und Weise mit dem System interagieren. Darüber hinaus sollte die körperliche und geistige Unversehrtheit von Menschen gewährleistet sein.

(67) ***Widerstandsfähigkeit gegen Angriffe und Sicherheitsverletzungen.*** KI-Systeme sollten wie alle Softwaresysteme vor Sicherheitslücken geschützt werden, die einen Missbrauch durch Angreifer (z. B. durch Hacking) ermöglichen. Die Angriffe richten sich möglicherweise gegen den Datenbestand (*Data Poisoning*), das Modell (*Model Leakage*) oder die zugrunde liegende Infrastruktur, wobei sowohl Software als auch Hardware betroffen sein kann. Wird ein KI-System angegriffen, z. B. im Rahmen feindseliger Angriffe, werden die Daten sowie das Systemverhalten möglicherweise geändert, was dazu führt, dass das System andere Entscheidungen trifft oder sich gänzlich abschaltet. Systeme und Daten können auch durch böswillige Absicht oder durch Konfrontation mit unerwarteten Situationen beschädigt werden. Unzureichende Sicherheitsprozesse können sich außerdem in fehlerhaften Entscheidungen oder sogar körperlichem Schaden niederschlagen. Damit KI-Systeme als sicher gelten können[37], sollten zweckentfremdete Anwendungen der KI (z. B. Mehrfachanwendungen) und ein möglicher Missbrauch eines KI-Systems durch böswillige Angreifer berücksichtigt werden und es sollten Vorkehrungen getroffen werden, damit das verhindert und begrenzt wird.[38]

(68) ***Auffangplan und allgemeine Sicherheit.*** KI-Systeme sollten für den Problemfall über Sicherheitsvorkehrungen verfügen, die einen Auffangplan aktivieren. Dies kann bedeuten, dass KI-Systeme von einem statistischen auf ein regelbasiertes Verfahren umschalten oder dass sie einen menschlichen Bediener anfordern, bevor sie einen Vorgang weiter ausführen.[39] Es muss gewährleistet sein, dass das System bestimmungsgemäß agiert, ohne Lebewesen oder der Umwelt Schaden zuzufügen. Dies schließt das Minimieren unbeabsichtigter Folgen oder Fehler ein. Darüber hinaus sollten Verfahren zur Klärung und Bewertung potenzieller Risiken im Zusammenhang mit dem Einsatz von KI-Systemen in verschiedenen Anwendungsbereichen eingerichtet werden. Das Niveau der erforderlichen Sicherheitsmaßnahmen hängt vom Ausmaß des von einem KI-System ausgehenden Risikos ab, welches wiederum von den Fähigkeiten des Systems bestimmt wird. Sofern absehbar

[37] Siehe z. B. die Erwägungen unter 2.7 des koordinierten Plans der Europäischen Union für künstliche Intelligenz.

[38] Für die Sicherheit von KI-Systemen ist es in hohem Maß geboten, einen positiven Kreislauf in der Forschung und Entwicklung zwischen dem Verstehen von Angriffen, der Entwicklung angemessener Schutzmaßnahmen und der Verbesserung von Evaluierungsmethoden entwickeln zu können. Zu diesem Zweck sollte eine Konvergenz zwischen der KI- Gemeinschaft und der IT-Sicherheitsgemeinschaft gefördert werden. Darüber hinaus liegt es in der Verantwortung aller Beteiligten, gemeinsame grenzübergreifende Schutz- und Sicherheitsnormen sowie ein Umfeld gegenseitigen Vertrauens zu schaffen, die die internationale Zusammenarbeit fördern. Für mögliche Maßnahmen siehe *Malicious Use of AI* (Avin S., Brundage M., et. al., 2018).

[39] Szenarien, in denen ein Mensch nicht sofort eingreifen kann, sollten ebenfalls berücksichtigt werden.

ist, dass der Entwicklungsprozess oder das System an sich außerordentlich hohe Risiken darstellen, ist es unerlässlich, Sicherheitsmaßnahmen proaktiv zu entwickeln und zu testen.

(69) ***Präzision.*** Präzision bezieht sich auf die Fähigkeit eines KI-Systems, Sachverhalte richtig zu beurteilen, z. B. das richtige Einordnen von Informationen in bestimmte Kategorien oder das Treffen richtiger Vorhersagen, Empfehlungen oder Entscheidungen auf der Grundlage von Daten oder Modellen. Ein expliziter und gut ausgestalteter Entwicklungs- und Bewertungsprozess kann unerwünschte Risiken, die sich aus falschen Vorhersagen ergeben, unterstützen, abwenden und korrigieren. Wenn sich gelegentliche ungenaue Vorhersagen nicht vermeiden lassen, sollte das System unbedingt anzeigen können, mit welcher Wahrscheinlichkeit es zu Fehlern kommt. Ein hohes Maß an Präzision ist insbesondere dann unerlässlich, wenn sich KI-Systeme direkt auf das Leben von Menschen auswirken.

(70) ***Zuverlässigkeit und Reproduzierbarkeit.*** Es ist unbedingt erforderlich, dass die von KI-Systemen erzeugten Ergebnisse gleichermaßen reproduzierbar und zuverlässig sind. Ein System ist dann zuverlässig, wenn es mit einer Reihe von Eingaben und in verschiedenen Situationen einwandfrei funktioniert. Dies ist erforderlich, um ein KI-System zu überprüfen und unerwünschte Schäden zu vermeiden. Wiederholbarkeit beschreibt, ob ein KI-Experiment das gleiche Verhalten aufweist, wenn es unter gleichen Bedingungen wiederholt wird. Auf diese Weise können Wissenschaftler und politische Entscheidungsträger genau beschreiben, was KI-Systeme tun. Sogenannte „*Replication Files*“[40] können den Prozess der Prüfung und Reproduktion von Verhaltensweisen erleichtern.

3. Schutz der Privatsphäre und Datenqualitätsmanagement

(71) Eng verbunden mit dem *Grundsatz der Schadensverhütung* ist der Schutz der Privatsphäre, ein Grundrecht, auf das sich KI-Systeme insbesondere auswirken. Die Schadensverhütung in Bezug auf den Schutz der Privatsphäre erfordert außerdem ein angemessenes Datenqualitätsmanagement, wozu auch die Qualität und Integrität der verwendeten Daten, ihre Relevanz gegenüber dem Bereich, in dem die KI-Systeme eingesetzt werden, die Zugangsprotokolle sowie die Fähigkeit zur Datenverarbeitung unter Wahrung des Datenschutzes gehören.

(72) ***Schutz der Privatsphäre und Datenschutz.*** KI-Systeme müssen den Schutz der Privatsphäre und den Datenschutz in allen Phasen des Lebenszyklus eines Systems gewährleisten.[41] Dies umfasst auch anfänglich vom Benutzer angegebene Informationen sowie die im Verlauf der Interaktion mit dem System über den Benutzer erzeugten Informationen (z. B. Ergebnisse, die das KI-System für bestimmte Benutzer erzeugt oder die Reaktion von Benutzern auf bestimmte Empfehlungen). Aus digitalen Aufzeichnungen über das menschliche Verhalten können KI-Systeme nicht nur auf persönliche Vorlieben einzelner Menschen, sondern auch auf sexuelle Ausrichtung, Alter und Geschlecht sowie religiöse oder politische Ansichten schließen. Damit die Menschen Vertrauen in die Datenverarbeitung haben können, muss sichergestellt sein, dass die über sie gesammelten Daten nicht dazu verwendet werden, sie unrechtmäßig oder unfair zu diskriminieren.

(73) ***Qualität und Integrität der Daten.*** Die Qualität der verwendeten Datensätze ist für die Leistungsfähigkeit von KI-Systemen von entscheidender Bedeutung. Bei der Erfassung der Daten können sozial bedingte Verzerrungen, Ungenauigkeiten, Fehler und andere Mängel auftreten. Solche Probleme müssen vor der Ausbildung mit einem bestimmten Datensatz behoben werden. Darüber hinaus muss die Integrität der Daten gewährleistet sein. Die Eingabe schädlicher Daten in ein KI-System kann sein Verhalten verändern, insbesondere bei selbstlernenden Systemen. Die verwendeten Prozesse und Datensätze müssen in allen Schritten wie Planung, Ausbildung, Erprobung und Einsatz getestet und dokumentiert werden. Dies sollte auch für KI-Systeme gelten, die nicht intern entwickelt, sondern von außerhalb erworben werden.

(74) ***Datenzugriff.*** In jeder Organisation, die mit den Daten von Bürgerinnen und Bürgern umgeht (ungeachtet dessen, ob die betreffenden Personen Nutzer des Systems sind oder nicht) sollten Datenprotokolle

[40] Dies betrifft Dateien, die jeden Schritt des Entwicklungsprozesses des KI-Systems von der Forschung und Ersterhebung der Daten bis zu den Ergebnissen reproduzieren.

[41] Hier sei auf bestehende Datenschutzgesetze wie die DSGVO oder die bevorstehende e-Datenschutz-Verordnung verwiesen.

eingerichtet werden, die den Datenzugriff regeln. In diesen Protokollen sollte festgelegt werden, wer unter welchen Umständen auf die Daten zugreifen kann. Nur entsprechend qualifiziertes Personal sollte auf die Daten von Bürgerinnen und Bürgern zugreifen dürfen, wenn die Kompetenz und die Notwendigkeit dafür gegeben sind.

4. Transparenz

(75) Diese Anforderung ist eng mit dem *Grundsatz der Erklärbarkeit* verbunden und bezieht sich auf die Transparenz der für ein KI-System relevanten Komponenten: die Daten, das System und die Geschäftsmodelle.

(76) ***Rückverfolgbarkeit.*** Die Datensätze und Prozesse, die zu der Entscheidung des KI-Systems geführt haben, einschließlich der Datenerfassung und -kennzeichnung sowie der verwendeten Algorithmen, sollten so gut wie möglich dokumentiert werden, um deren Rückverfolgbarkeit sicherzustellen und die Transparenz zu erhöhen. Dies gilt auch für die vom KI-System getroffenen Entscheidungen. So können die Gründe für eine fehlerhafte KI-Entscheidung ermittelt werden, was wiederum zur Vermeidung zukünftiger Fehler beitragen kann. Rückverfolgbarkeit erleichtert somit die Nachprüfbarkeit und Erklärbarkeit.

(77) ***Erklärbarkeit.*** Erklärbarkeit bezieht sich auf die Möglichkeit, sowohl die technischen Prozesse eines KI-Systems als auch die damit verbundenen menschlichen Entscheidungen (z. B. Anwendungsbereiche eines KI-Systems) zu erklären. Technische Erklärbarkeit setzt voraus, dass die von einem KI-System getroffenen Entscheidungen vom Menschen verstanden und rückverfolgt werden können. Darüber hinaus müssen möglicherweise Kompromisse zwischen einer verbesserten Erklärbarkeit eines Systems (was die Präzision beeinträchtigen kann) und mehr Präzision (auf Kosten der Erklärbarkeit) eingegangen werden. Wann immer ein KI-System das Leben von Menschen entscheidend beeinflusst, muss es möglich sein, eine geeignete Erklärung für den Entscheidungsprozess des KI-Systems zu erhalten. Eine solche Erklärung sollte rechtzeitig erfolgen und auf die jeweilige Sachkenntnis des betroffenen Interessenträgers (z. B. Laie, Regulierungsbehörde oder Forscher) zugeschnitten sein. Darüber hinaus sollten Erläuterungen darüber vorliegen, inwieweit ein KI-System die Entscheidungsprozesse einer Organisation beeinflusst und gestaltet, aber auch über die Entwurfsentscheidungen und die Gründe für die Einführung des Systems (zur Gewährleistung der Transparenz des Geschäftsmodells).

(78) ***Kommunikation.*** KI-Systeme dürfen gegenüber den Nutzern nicht als Menschen auftreten. Menschen haben das Recht, darüber informiert zu werden, dass sie mit einem KI-System interagieren. Dies bedeutet, dass KI-Systeme als solche erkennbar sein müssen. Zur Gewährleistung der Einhaltung der Grundrechte sollte darüber hinaus bei Bedarf die Möglichkeit bestehen, sich gegen diese Interaktion und zugunsten einer zwischenmenschlichen Interaktion zu entscheiden. Darüber hinaus sollten die Fähigkeiten und Einschränkungen des KI-Systems den Anwendern der KI und den Endnutzern in einer dem jeweiligen Anwendungsfall angemessenen Weise mitgeteilt werden. Das könnte Informationen zur Präzision des KI-Systems sowie seiner Grenzen umfassen.

5. Vielfalt, Nichtdiskriminierung und Fairness

(79) Zur Schaffung einer vertrauenswürdigen KI müssen Inklusion und Vielfalt während des gesamten Lebenszyklus des KI-Systems garantiert sein. Neben der Berücksichtigung und Einbindung aller betroffenen Interessenträger in den gesamten Prozess setzt dies auch die Sicherstellung eines gleichberechtigten Zugangs durch inklusive Gestaltungsprozesse sowie Gleichbehandlung voraus. Diese Anforderung ist eng mit dem *Grundsatz der Fairness* verbunden.

(80) ***Vermeidung unfairer Verzerrungen***. Die von KI-Systemen (sowohl zur Ausbildung als auch für den Einsatz) verwendeten Datensätze können unbeabsichtigte historische Verzerrungen und schlechte Lenkungs- und Kontrollmodelle aufweisen oder unvollständig sein. Die Fortschreibung solcher Verzerrungen könnte

(in)direkte Vorurteile und die Diskriminierung[42] bestimmter Gruppen oder Personen zur Folge haben und Vorurteile und Marginalisierung potenziell verschärfen. Schaden kann aber auch aus einer beabsichtigten Ausnutzung von Vorurteilen (der Verbraucher) oder durch unlauteren Wettbewerb entstehen, wie z. B. durch die Vereinheitlichung der Preise durch geheime Absprachen oder einen undurchsichtigen Markt.[43] Erkennbare diskriminierende Verzerrungen sollten nach Möglichkeit in der Phase der Datenerhebung beseitigt werden. Auch die Art und Weise, wie KI-Systeme entwickelt werden (z. B. wie der Programmcode eines Algorithmus geschrieben wird), kann durch unfaire Einflüsse beeinträchtigt werden. Dem könnte durch die Einführung von Aufsichtsverfahren entgegengewirkt werden, mit deren Hilfe der Zweck, die Einschränkungen, Anforderungen und Entscheidungen des Systems klar und transparent analysiert und angegangen werden könnten. Darüber hinaus sollte die Beschäftigung von Personen mit unterschiedlichen kulturellen Hintergründen, die aus verschiedenen Fachrichtungen stammen, gefördert werden, da dies zur Gewährleistung der Meinungsvielfalt beitragen könnte.

(81) ***Barrierefreiheit und universeller Entwurf***. Insbesondere in der Beziehung zwischen Unternehmen und Verbrauchern sollten die Systeme benutzerorientiert und so gestaltet sein, dass alle Menschen unabhängig von ihrem Alter, Geschlecht, ihren Fähigkeiten oder Merkmalen KI-Produkte oder -dienstleistungen nutzen können. Die barrierefreie Zugänglichkeit dieser Technologie für Menschen mit Behinderungen, die in allen gesellschaftlichen Gruppen präsent sind, ist von besonderer Bedeutung. KI-Systeme sollten keinen Einheitsansatz verfolgen und die Grundsätze eines universellen Entwurfs[44] sollten einen möglichst breiten Nutzerkreis ansprechen und sich an die einschlägigen Barrierefreiheitsnormen halten.[45] So wird ein gerechter Zugang und eine aktive Beteiligung aller Menschen an bestehenden und neu hinzukommenden computergestützten menschlichen Tätigkeiten, auch im Hinblick auf assistierende Technologien, möglich sein.[46]

(82) ***Beteiligung der Interessenträger***. Zur Entwicklung vertrauenswürdiger KI-Systeme ist eine Konsultation der Interessenträger ratsam, die möglicherweise während des gesamten Lebenszyklus des Systems direkt oder indirekt von diesem betroffen sind. Es ist von Vorteil, auch nach der Einführung eines Systems regelmäßige Rückmeldungen einzuholen und längerfristige Vorkehrungen zur Beteiligung der Interessenträger zu schaffen, beispielsweise durch die Gewährleistung der Schulung, Anhörung und Beteiligung der Arbeitnehmerinnen und Arbeitnehmer während des gesamten Prozesses der Implementierung von KI-Systemen in einem Unternehmen.

6. Gesellschaftliches und ökologisches Wohlergehen

(83) Im Einklang mit den *Grundsätzen der Fairness* und *Schadensverhütung* sollten während des gesamten KI-Lebenszyklus auch die breitere Gesellschaft, andere fühlende Wesen und die Umwelt als Akteure berücksichtigt werden. Die Nachhaltigkeit und ökologische Verantwortung der KI-Systeme sollten gefördert werden und die Erforschung von KI-Lösungen in Bezug auf globale Belange, wie z. B. die Ziele für eine nachhaltige Entwicklung, sollte ausgebaut werden. Im Idealfall sollte die KI zum Wohle aller Menschen, auch künftiger Generationen, eingesetzt werden.

(84) ***Nachhaltige und umweltfreundliche KI.*** KI-Systeme versprechen, einen Beitrag zur Bewältigung einiger der drängendsten gesellschaftlichen Probleme zu leisten, doch muss sichergestellt sein, dass dies auf möglichst umweltfreundliche Art und Weise geschieht. Der Prozess der Entwicklung, Einführung und Nutzung des

[42] Siehe die Definition von direkter und indirekter Diskriminierung, beispielsweise in Artikel 2 der Richtlinie 2000/78/EG des Rates vom 27. November 2000 zur Festlegung eines allgemeinen Rahmens für die Verwirklichung der Gleichbehandlung in Beschäftigung und Beruf. Siehe auch Artikel 21 der EU-Grundrechtecharta.

[43] Vgl. das Papier der Agentur der Europäischen Union für Grundrechte: *„BigData: Discrimination in data-supported decision making (2018)"* http://fra.europa.eu/en/publication/2018/big-data-discrimination.

[44] Artikel 42 der Richtlinie über die öffentliche Auftragsvergabe verlangt, dass Barrierefreiheit und ein „Design für alle" in den technischen Spezifikationen berücksichtigt werden müssen.

[45] Zum Beispiel EN 301 549.

[46] Diese Anforderung knüpft an das Übereinkommen der Vereinten Nationen über die Rechte von Menschen mit Behinderungen an.

Systems sowie dessen gesamte Lieferkette sollten diesbezüglich bewertet werden, z. B. anhand einer kritischen Untersuchung des Verbrauchs von Ressourcen und Energie während der Ausbildung, wobei weniger umweltschädliche Optionen gewählt werden sollten. Maßnahmen zur Sicherung der Umweltverträglichkeit der gesamten Lieferkette des KI-Systems sollten gefördert werden.

(85) ***Soziale Auswirkungen***. Die Omnipräsenz sozialer KI-Systeme[47] in allen Lebensbereichen (ob in Bildung, Arbeit, Pflege oder Unterhaltung) kann unsere Vorstellung von sozialer Handlungsfähigkeit verändern oder unsere sozialen Beziehungen und Bindungen beeinflussen. So wie KI-Systeme zur Verbesserung sozialer Kompetenzen eingesetzt werden können[48], können sie auch zu deren Verschlechterung beitragen. Dies könnte sich ebenso auf das körperliche und geistige Wohlergehen der Menschen auswirken. Die Auswirkungen dieser Systeme müssen daher sorgfältig überwacht und berücksichtigt werden.

(86) ***Gesellschaft und Demokratie.*** Neben der Bewertung der Auswirkungen der Entwicklung, Einführung und Nutzung eines KI-Systems auf den Einzelnen sollten die Systeme auch aus gesellschaftlicher Sicht unter Berücksichtigung der Auswirkungen auf Institutionen, Demokratie und die Gesellschaft insgesamt bewertet werden. Der Einsatz von KI-Systemen sollte insbesondere im Zusammenhang mit dem demokratischen Prozess – einschließlich der politischen Entscheidungsfindung und in Bezug auf Wahlen – sorgfältig geprüft werden.

7. Rechenschaftspflicht

(87) Mit der Rechenschaftspflicht werden die oben genannten Anforderungen, die eng mit dem *Grundsatz der Fairness* verbunden sind, ergänzt. Zu diesem Zweck müssen Vorkehrungen getroffen werden, die die Verantwortlichkeit und Rechenschaftspflicht für KI-Systeme und deren Ergebnisse vor und nach deren Umsetzung gewährleisten.

(88) ***Nachprüfbarkeit.*** Nachprüfbarkeit bedeutet, dass Algorithmen, Daten und das Entwurfsverfahren einer Bewertung unterzogen werden können. Das heißt nicht unbedingt, dass Informationen über Geschäftsmodelle und geistiges Eigentum im Zusammenhang mit dem KI-System immer öffentlich verfügbar sein müssen. Die Bewertung von KI-Systemen durch interne und externe Prüfer und das Vorliegen solcher Bewertungsberichte kann beträchtlich zur Vertrauenswürdigkeit der Technik beitragen. Die externe Nachprüfbarkeit sollte insbesondere bei Anwendungen sichergestellt sein, die sich auf die Grundrechte auswirken, sowie bei sicherheitskritischen Anwendungen.

(89) ***Minimierung negativer Auswirkungen und Meldung.*** Sowohl die Möglichkeit der Berichterstattung über Handlungen oder Entscheidungen, die zu einem bestimmten Systemergebnis beitragen als auch die Reaktionsfähigkeit auf die Folgen eines solchen Ergebnisses müssen gewährleistet sein. Die Ermittlung, Bewertung, Berichterstattung und Minimierung potenziell negativer Auswirkungen von KI-Systemen ist für (in)direkt Betroffene besonders wichtig. Ein angemessener Schutz für Informanten, Nichtregierungsorganisationen, Gewerkschaften und andere Stellen, die berechtigte Bedenken hinsichtlich eines KI-gestützten Systems äußern, muss gewährleistet sein. Der Einsatz von Folgenabschätzungen (z. B. „*Red Teaming*“ oder Formen der Algorithmen-Folgenabschätzungen) sowohl vor als auch während der Entwicklung, Einführung und Nutzung von KI-Systemen kann hilfreich sein, um negative Folgen möglichst gering zu halten. Diese Bewertungen müssen in einem angemessenen Verhältnis zu dem Risiko stehen, das die KI-Systeme darstellen.

(90) ***Kompromisse.*** Bei der Umsetzung der oben genannten Anforderungen können Spannungen zwischen ihnen

[47] Das bezieht sich auf KI-Systeme, die mit Menschen kommunizieren und interagieren und die bei der Interaktion zwischen Mensch und Roboter (eingebettete KI) oder als Avatare in einer virtuellen Realität Sozialität simulieren. Auf diese Weise haben diese Systeme das Potenzial, unsere soziokulturellen Handlungsmuster und das soziale Gefüge zu verändern.

[48] Siehe zum Beispiel das von der EU geförderte Projekt zur Entwicklung einer KI-gestützten Software, die es Robotern ermöglicht, effektiver mit autistischen Kindern im Rahmen von durch Menschen geleitete Therapie zu interagieren und so zur Verbesserung der Sozial- und Kommunikationsfähigkeit der betroffenen Kinder beizutragen: http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968.

auftreten, die Kompromisse erforderlich machen. Solche Kompromisse sollten nach dem neuesten Stand der Technik rational und methodisch angegangen werden. Dies bedeutet, dass relevante Interessen und Werte, die vom KI-System betroffen sind, ermittelt werden sollten und dass im Falle von Konflikten zwischen diesen ausdrückliche Kompromisse eingegangen werden müssen, die im Hinblick auf ihr Risiko für ethische Grundsätze, einschließlich der Grundrechte, bewertet werden müssen. In Situationen, in denen keine ethisch vertretbaren Kompromisse möglich sind, sollte das KI-System in einer anderen Form entwickelt, eingeführt und genutzt werden. Jede Entscheidung darüber, welcher Kompromiss einzugehen ist, sollte begründet und ordnungsgemäß dokumentiert werden. Der Entscheidungsträger ist für die Art und Weise des Zustandekommens eines angemessenen Kompromisses rechenschaftspflichtig und er sollte die Angemessenheit der daraus resultierenden Entscheidung kontinuierlich überprüfen, um sicherzustellen, dass bei Bedarf notwendige Änderungen am System vorgenommen werden können. [49]

(91) ***Rechtsmittel.*** Sollte es schließlich doch zu ungerechten und nachteiligen Auswirkungen kommen, sollten Vorkehrungen für einen angemessenen Rechtsschutz getroffen werden[50]. Zu wissen, dass Rechtsbehelfe möglich sind, wenn etwas schief geht, ist der Schlüssel zur Vertrauensbildung. Schutzbedürftigen Personen oder Gruppen sollte besondere Aufmerksamkeit geschenkt werden.

2. Technische und nicht-technische Methoden zur Schaffung einer vertrauenswürdigen KI

(92) Zur Umsetzung der oben genannten Anforderungen können sowohl technische als auch nicht-technische Methoden zum Einsatz kommen. Diese umfassen alle Phasen des Lebenszyklus eines KI-Systems. Die zur Umsetzung der Anforderungen eingesetzten Verfahren, die Berichterstattung über Änderungen des Umsetzungsprozesses und deren Rechtfertigung[51] sollten fortlaufend bewertet werden. Da KI-Systeme sich in einem dynamischen Umfeld ständig weiterentwickeln, ist die Schaffung einer vertrauenswürdigen KI ein kontinuierlicher Prozess, der in Abbildung 3 unten dargestellt ist.

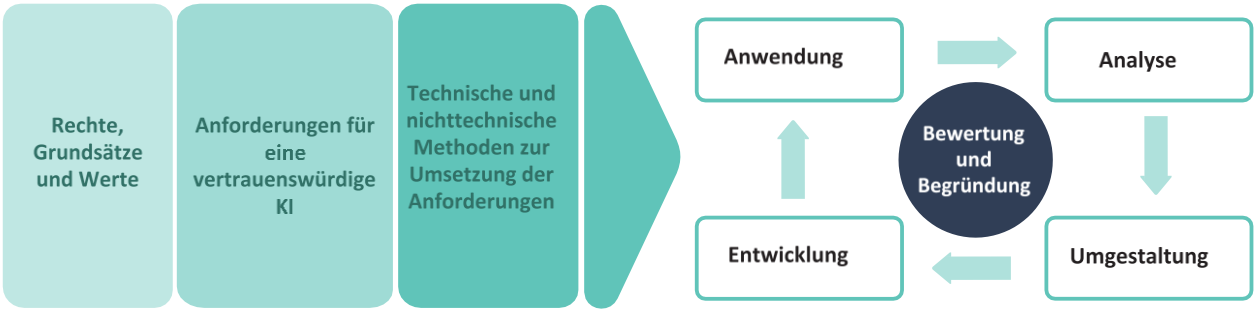

Abbildung 3: Schaffung einer vertrauenswürdigen KI während des gesamten Lebenszyklus des Systems

(93) Die folgenden Verfahrensweisen können als komplementär oder alternativ angesehen werden, da unterschiedliche Anforderungen – und andersartige Empfindlichkeiten – einen Bedarf an verschiedenartigen Umsetzungsmethoden schaffen. Diese Übersicht ist weder umfassend oder vollständig noch verbindlich. Vielmehr wird damit eine Liste mit Vorgehensweisen vorgeschlagen, die bei der Einführung einer vertrauenswürdigen KI hilfreich sein können.

---

[49] Verschiedene Lenkungs- und Kontrollmodelle können dabei behilflich sein. So könnte z. B. die Anwesenheit eines internen und/oder externen ethischen (und sektorspezifischen) Experten oder Gremiums nützlich sein, um potenzielle Konfliktbereiche aufzuzeigen und Vorschläge zu machen, wie ein bestimmter Konflikt am besten gelöst werden kann. Ebenfalls nützlich sind sinnvolle Konsultationen und Gespräche mit Interessenträgern, einschließlich denjenigen, die Gefahr laufen, von den unerwünschten Folgen eines KI-Systems beeinträchtigt zu werden. Die europäischen Universitäten sollten eine führende Rolle bei der Ausbildung der erforderlichen Ethikexperten übernehmen.

[50] Siehe auch die Stellungnahme der Agentur der Europäischen Union für Grundrechte zum Thema *„Improving access to remedy in the area of business and human rights at the EU level"* (2017), https://fra.europa.eu/en/opinion/2017/business-human-rights.

[51] Zur Berücksichtigung der oben genannten Anforderungen sind beispielsweise die Entscheidungen, die beim Entwurf, der Entwicklung und Einführung des Systems getroffen wurden, zu begründen.

1. Technische Verfahren

(94) Dieser Abschnitt beschreibt technische Verfahren zur Gewährleistung einer vertrauenswürdigen KI, die in die Phasen des Entwurfs, der Entwicklung und der Nutzung eines KI-Systems einbezogen werden können. Die nachfolgend aufgeführten Methoden sind unterschiedlich ausgereift[52].

- *Architekturen für eine vertrauenswürdige KI*

(95) Die Anforderungen an eine vertrauenswürdige KI sollten in Verfahrensweisen und/oder Verfahrenseinschränkungen, die in der Architektur des KI-Systems verankert sein sollten, „übersetzt" werden. Dies könnte durch eine Reihe von Regeln einer „weißen Liste" (Verhaltensweisen oder Zustände), die das System immer befolgen sollte, und durch Einschränkungen des Verhaltens oder der Zustände, die das System niemals überschreiten darf und die auf einer „schwarzen Liste" aufgeführt werden, sowie durch Mischungen dieser beiden oder komplexere, nachweisbare Garantien bezüglich des Verhaltens des Systems erreicht werden. Die Überwachung der Einhaltung dieser Einschränkungen durch das System während des Einsatzes kann mithilfe eines separaten Verfahrens erfolgen.

(96) Lernfähige KI-Systeme, die in der Lage sind, ihr Verhalten dynamisch anzupassen, können als ein nichtdeterministisches System verstanden werden, das möglicherweise ein unerwartetes Verhalten an den Tag legen kann. Aus theoretischer Sicht werden diese Systeme oft im Sinne eines „*sense-plan-act*"-Zyklus (dt. erkennen-planen-handeln) verstanden. Die Anpassung dieser Architektur zur Sicherstellung einer vertrauenswürdigen KI erfordert die Einbeziehung der oben aufgeführten Anforderungen in allen drei Phasen des Zyklus: i) Während der Phase des Erkennens („*sense*") sollte das System so entwickelt werden, dass es alle Umgebungselemente erkennt, die notwendig sind, um die Einhaltung der Anforderungen zu gewährleisten. ii) Während der Planungsphase („*plan*") sollte das System nur solche Pläne berücksichtigen, die den Anforderungen entsprechen. iii) Während der Handlungsphase („*act*") sollten die Handlungen des Systems auf Verhaltensweisen beschränkt sein, die die Anforderungen verwirklichen.

(97) Die oben skizzierte Architektur ist unspezifisch und beschreibt die meisten KI-Systeme nur unvollständig. Dennoch gibt sie Anhaltspunkte für Einschränkungen und Strategien, die sich in spezifischen Modulen widerspiegeln sollten, um ein umfassendes System zu gestalten, das vertrauenswürdig ist und auch als solches wahrgenommen wird.

- *Konzeptuell integrierte Ethik und Rechtsstaatlichkeit (X-by-Design)*

(98) Methoden zur Sicherstellung konzeptuell integrierter Werte bieten präzise und explizite Verbindungen zwischen den abstrakten Prinzipien, an die sich das System halten muss, und den spezifischen Entscheidungen bei der Umsetzung. Die Vorstellung, dass die Einhaltung von Vorschriften in das Konzept des KI-Systems einbezogen werden kann, ist der Schlüssel zu dieser Methode. Die Unternehmen sind dafür verantwortlich, die Auswirkungen ihrer KI-Systeme von Anfang an zu erkennen, ebenso wie die Vorschriften, die ihr KI-System einhalten muss, damit negative Auswirkungen verhindert werden. Verschiedene Gestaltungsprinzipien sind bereits weitverbreitet, z. B. *konzeptuell integrierter Datenschutz* und *konzeptuell integrierte Sicherheit.* Wie bereits erwähnt, müssen die Prozesse, Daten und Ergebnisse einer vertrauenswürdigen KI sicher sein und KI-Systeme sollten so konzipiert sein, dass sie robust gegenüber feindlichen Daten und Angriffen sind. KI-Systeme sollten über einen Mechanismus für eine ausfallsichere Abschaltung verfügen und die Betriebsaufnahme nach einer Zwangsabschaltung (z. B. einem Angriff) sollte möglich sein.

- *Erklärungsmethoden*

(99) Damit ein System vertrauenswürdig ist, muss nachvollziehbar sein, warum es sich auf eine bestimmte Art und

[52] Während einige dieser Verfahrensweisen bereits heute verfügbar sind, müssen andere weiter erforscht werden. Die Bereiche, in denen ein weiterer Forschungsbedarf besteht, werden auch Gegenstand des zweiten Berichts der HEG-KI, d. h. der Empfehlungen für Richtlinien und Investitionen, sein.

Weise verhalten hat und warum es eine bestimmte Interpretation hervorgebracht hat. Ein ganzer Forschungsbereich – die erklärbare KI (XAI) – versucht, dieses Problem anzugehen, um die zugrunde liegenden Mechanismen des Systems besser zu verstehen und Lösungen zu finden. Bei KI-Systemen, die auf der Basis neuronaler Netze funktionieren, stehen wir auch heute noch vor dieser Herausforderung. Ausbildungsprozesse mit neuronalen Netzen können dazu führen, dass Netzwerkparameter auf numerische Werte gesetzt werden, die nur schwer mit den Ergebnissen in Einklang gebracht werden können. Darüber hinaus können manchmal kleine Änderungen in den Daten zu dramatischen Änderungen bei deren Interpretation führen, was dazu führen kann, dass das System z. B. einen Schulbus mit einem Strauß verwechselt. Diese Schwachstelle kann auch bei Angriffen auf das System ausgenutzt werden. Verfahren, die XAI-Forschung einbeziehen, sind nicht nur wichtig, um den Nutzern das Verhalten des Systems zu erklären, sondern auch, um zuverlässige Technologien einzuführen.

- *Erproben und prüfen*

(100) Aufgrund des nichtdeterministischen und kontextspezifischen Charakters der KI-Systeme reicht traditionelles Testen nicht aus. Fehlschläge der vom System verwendeten Konzepte und Repräsentationen werden nur dann offensichtlich, wenn ein Programm mit hinreichend realistischen Daten verwendet wird. Zur Prüfung und Validierung der Datenverarbeitung muss das zugrunde liegende Modell daher sowohl während der Ausbildung als auch während des Einsatzes auf seine Stabilität, Robustheit und Funktionsfähigkeit innerhalb gut verständlicher und vorhersehbarer Grenzen sorgfältig überwacht werden. Es muss sichergestellt sein, dass das Planungsergebnis mit der Eingabe übereinstimmt und dass die Entscheidungen so getroffen werden, dass der zugrunde liegende Prozess validiert werden kann.

(101) Das Erproben und Überprüfen des Systems sollte so früh wie möglich einsetzen, damit sichergestellt ist, dass sich das System während des gesamten Lebenszyklus und insbesondere nach der Einführung wie vorgesehen verhält. Es sollten alle Komponenten eines KI-Systems einbezogen werden, einschließlich der Daten, im Voraus geschulter Modelle, Umgebungen und des Verhaltens des Systems als Ganzes. Es sollte von einer möglichst vielfältig zusammengesetzten Personengruppe entworfen und entwickelt werden. Für die Kategorien, die aus verschiedenen Perspektiven getestet werden, sollten vielfältige Metriken entwickelt werden. Kontradiktorische Tests durch vertrauenswürdige, gemischte „rote Teams“, die bewusst versuchen, das System zu „knacken“, um Schwachstellen zu finden, und „Bug-Bounty“-Programme, die Außenstehende dazu anregen, Systemfehler und -schwächen zu erkennen und verantwortungsbewusst zu melden, sollten in Betracht gezogen werden. Schließlich muss sichergestellt sein, dass die Ergebnisse bzw. Handlungen mit den Ergebnissen der vorangegangenen Prozesse übereinstimmen, indem sie zur Gewährleistung ihrer Konformität mit den zuvor definierten Zielsetzungen verglichen werden.

- *Dienstqualitätsparameter*

(102) Zur Gewährleistung, dass ein KI-System unter Berücksichtigung der entsprechenden Sicherheitsvorkehrungen erprobt und entwickelt wurde, können angemessene Indikatoren für die Dienstqualität definiert werden. Diese Indikatoren könnten Maßnahmen zur Bewertung der Tests und der Ausbildung der Algorithmen sowie traditionelle Softwaremetriken bezüglich Funktionalität, Leistung, Benutzerfreundlichkeit, Zuverlässigkeit, Sicherheit und Wartbarkeit umfassen.

2. Nichttechnische Verfahren

(103) Dieser Abschnitt beschreibt eine Vielzahl nichttechnischer Verfahren, die eine wichtige Rolle bei der Sicherung und Erhaltung einer vertrauenswürdigen KI spielen können. Auch diese sollten **fortlaufend** bewertet werden.

- *Regulierung*

(104) Wie bereits erwähnt, gibt es bereits heute Regelungen zur Stärkung der Vertrauenswürdigkeit der KI, beispielsweise die Rechtsvorschriften zur Produktsicherheit und zur Produkthaftung. Soweit wir der Ansicht sind, dass eine Regulierung sowohl zum Schutz als auch zur Ermöglichung der KI überarbeitet, angepasst oder eingeführt werden sollte, werden wir dies in unserem zweiten Bericht mit Empfehlungen für Richtlinien und

Investitionen im Zusammenhang mit der KI behandeln.

- *Verhaltenskodizes*

(105) Zur Unterstützung einer vertrauenswürdigen KI können Unternehmen und Interessenträger sich den Leitlinien anschließen und ihre Grundsätze zur sozialen Verantwortung, ihre wesentlichen Leistungsindikatoren (KPI), Verhaltenskodizes oder internen Strategiepapiere anpassen. Eine Organisation, die an einem KI-System arbeitet, kann, ganz allgemein, ihre Absichten dokumentieren und mit Standards für bestimmte wünschenswerte Werte wie Grundrechte, Transparenz und Schadensvermeidung versehen.

- *Standarisierung*

(106) Normen, z. B. für Konzeption, Fertigung und Geschäftspraktiken können für KI-Anwender, Verbraucher, Organisationen, Forschungseinrichtungen und Regierungen als Qualitätsmanagementsystem dienen, da sie die Möglichkeit bieten, ethisches Verhalten bei der Kaufentscheidung zu erkennen und zu fördern. Über die herkömmlichen Standards hinaus gibt es Ansätze zur Koregulierung: Akkreditierungssysteme, Verhaltenskodizes auf dem Gebiet der Berufsethik oder Standards für ein mit den Grundrechten konformes Konzept. Aktuelle Beispiele sind unter anderem die ISO-Normen oder die Normenreihe IEEE P7000. Zukünftig könnte auch ein mögliches Gütezeichen „Vertrauenswürdige KI" geeignet sein, das unter Bezugnahme auf spezifische technische Normen bestätigt, dass das System beispielsweise die Anforderungen an Sicherheit, technische Robustheit und Erklärbarkeit einhält.

- *Zertifizierung*

(107) Da nicht davon ausgegangen werden kann, dass alle Menschen in der Lage sind, die Funktionsweise und die Auswirkungen von KI-Systemen vollständig zu verstehen, sollten Organisationen in Betracht gezogen werden, die der breiten Öffentlichkeit gegenüber bestätigen können, dass ein KI-System transparent, rechenschaftspflichtig und fair ist[53]. Diese Zertifizierungen würden auf Normen beruhen, die für verschiedene Anwendungsbereiche und KI-Techniken entwickelt wurden und auf die entsprechenden industriellen und gesellschaftlichen Vorgaben des jeweiligen Kontextes ausgerichtet sind. Zertifizierungen sind jedoch kein Ersatz für Verantwortung. Sie sollten daher durch Rahmenbedingungen für die Rechenschaftspflicht, einschließlich Haftungsausschlüsse sowie Verfahren zur Überprüfung und Mängelbeseitigung ergänzt werden[54].

- *Rechenschaftspflicht durch Rahmenbedingungen für die Lenkung und Kontrolle*

(108) Unternehmen sollten interne und externe Rahmenbedingungen für die Lenkung und Kontrolle von KI-Systemen einführen und somit die Rechenschaftspflicht bezüglich der ethischen Aspekte der Entscheidungsfindung im Zusammenhang mit der Entwicklung, Einführung und Nutzung von KI gewährleisten. Das kann beispielsweise durch die Ernennung eines Ethikbeauftragten für KI oder eines internen/externen Ethikgremiums oder -rates erfolgen. Aufsicht und Beratung sind mögliche Aufgabenbereiche solcher Personen, Gremien oder Räte. Wie oben dargelegt, können in diesem Zusammenhang auch Zertifizierungsspezifikationen und/oder -stellen eine Rolle spielen. Zum Austausch bewährter Verfahren, Besprechen von Problemen oder zur Meldung neu auftretender ethischer Bedenken sollten Kommunikationskanäle mit Industrie und/oder Aufsichtsbehörden gewährleistet sein. Solche Maßnahmen können die Rechtsaufsicht zwar ergänzen, (z. B. in Form der Bestellung eines Datenschutzbeauftragten oder gleichwertiger, datenschutzrechtlich vorgeschriebener Maßnahmen), aber nicht ersetzen.

- *Bildung und Bewusstsein zur Förderung einer ethischen Mentalität*

(109) Eine vertrauenswürdige KI beruht auf der sachkundigen Beteiligung aller Interessenträger. Kommunikation

[53] Zum Beispiel im Einklang mit der Empfehlung der Initiative der IEEE „*Ethically Aligned Design*": https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] Weitere Informationen zu den Grenzen der Zertifizierung finden Sie auf: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

und Aus- und Weiterbildung spielen eine wichtige Rolle, um die Verbreitung der Kenntnisse über die potenziellen Auswirkungen von KI-Systemen unter der Bevölkerung sicherzustellen und um die Menschen darauf aufmerksam zu machen, dass sie an der Gestaltung der gesellschaftlichen Entwicklung teilnehmen können. Das betrifft alle Interessenträger, z. B. die an der Produktfertigung Beteiligten (Designer und Entwickler), die Nutzer (Unternehmen oder Einzelpersonen) und weitere betroffene Gruppen (Personen, die kein KI-System erwerben oder verwenden, die jedoch von den Entscheidungen eines KI-Systems betroffen sind) und die Gesellschaft im Allgemeinen. Grundlegende KI-Kompetenzen sollten in der gesamten Gesellschaft gefördert werden. Eine Voraussetzung für die Aufklärung der Öffentlichkeit ist, dass Ethiker in diesem Bereich ausgebildet werden, damit sie über die angemessenen Kompetenzen verfügen.

- *Beteiligung der Interessenträger und sozialer Dialog*

(110) Die KI hat vielfältige Vorteile und in Europa muss sichergestellt werden, dass diese auch für alle zugänglich sind. Zu diesem Zweck ist eine offene Debatte und die Beteiligung der Sozialpartner, Interessenträger und der Öffentlichkeit im Allgemeinen erforderlich. Viele Unternehmen setzen bereits auf Arbeitsgruppen, deren Mitglieder die Nutzung von KI-Systemen und die Datenanalyse besprechen. Diese Gremien haben verschiedene Mitglieder: z. B. Experten für Recht und Technik, Ethiker, Verbrauchervertreter, Mitarbeiterinnen und Mitarbeiter. Die Bewertung der Ergebnisse und Ansätze beruht auch auf der aktiven Suche nach Beteiligung und dem Dialog über die Verwendung und die Auswirkungen von KI-Systemen und kann insbesondere in komplexen Fällen hilfreich sein.

- *Vielfalt und inklusive Entwurfsteams*

(111) Vielfalt und Integration spielen eine wesentliche Rolle bei der Entwicklung von KI-Systemen, die in der realen Welt eingesetzt werden. Es ist wichtig, dass die Teams, die nunmehr immer autonomere KI-Systeme entwerfen, entwickeln, erproben, warten, bereitstellen und/oder beschaffen, die Vielfalt der Nutzer und die Gesellschaft im Allgemeinen widerspiegeln. So wird ein Beitrag zur Objektivität und Berücksichtigung unterschiedlicher Perspektiven, Bedürfnisse und Ziele geleistet. Im Idealfall sollten die Teams nicht nur in Bezug auf Geschlecht, Kultur und Alter, sondern auch im Hinblick auf ihren beruflichen Hintergrund und ihre Kompetenzen unterschiedlich zusammengesetzt sein.

Wichtige Leitlinien aus Kapitel II:

- ✓ Gewährleistung, dass ein KI-System während seines gesamten Lebenszyklus den Anforderungen an eine vertrauenswürdige KI entspricht: 1) Vorrang menschlichen Handelns und menschliche Aufsicht, 2) technische Robustheit und Sicherheit, 3) Datenschutz und Datenqualitätsmanagement, 4) Transparenz, 5) Vielfalt, Nichtdiskriminierung und Fairness, 6) gesellschaftliches und ökologisches Wohlergehen und 7) Rechenschaftspflicht.
- ✓ Berücksichtigung technischer und nichttechnischer Methoden, um die Umsetzung dieser Anforderungen sicherzustellen.
- ✓ Förderung von Forschung und Innovation zur Unterstützung der Bewertung von KI-Systemen und der Erfüllung der Anforderungen; Verbreitung von Ergebnissen und offenen Fragen in der breiten Öffentlichkeit und systematische Schulung einer neuen Generation von KI-Ethikexperten.
- ✓ Klare und proaktive Informationsübermittlung an betroffene Kreise über die Fähigkeiten und Grenzen der KI-Systeme, die realistische Erwartungen ermöglichen, sowie über die Art und Weise der Implementierung der Anforderungen. Für die Anwender muss klar erkennbar sein, dass sie es mit einem KI-System zu tun haben.
- ✓ Erleichterung der Rückverfolgbarkeit und Nachprüfbarkeit der KI-Systeme, insbesondere in kritischen Kontexten und Situationen.
- ✓ Beteiligung der Interessenträger während des gesamten Lebenszyklus des KI-Systems. Schulungs- und Ausbildungsförderung mit dem Ziel, allen Interessenträgern Kompetenzen auf dem Gebiet der vertrauenswürdigen KI zu vermitteln.

- ✓ Zwischen den verschiedenen Grundsätzen und Anforderungen können möglicherweise wesentliche Spannungen auftreten. Diesbezügliche Kompromisse und Lösungen müssen kontinuierlich ermittelt, bewertet, dokumentiert und mitgeteilt werden.

III. Kapitel III: Bewertung einer vertrauenswürdigen KI

(112) Auf der Grundlage der in Kapitel II dargelegten Kernanforderungen wird in diesem Kapitel eine nicht erschöpfende **Bewertungsliste für vertrauenswürdige KI** (Pilotversion) zur **praktischen Anwendung vertrauenswürdiger KI** vorgestellt. Sie gilt insbesondere für KI-Systeme, die direkt mit den Anwendern interagieren, und richtet sich in erster Linie an Entwickler und Betreiber von KI-Systemen (unabhängig davon, ob die Systeme selbst entwickelt oder von Dritten erworben wurden). Diese Bewertungsliste befasst sich nicht mit der praktischen Umsetzung der ersten Komponente für eine vertrauenswürdige KI (rechtmäßige KI). Die Einhaltung dieser Bewertungsliste ist kein Nachweis der Gesetzeskonformität und sie ist auch nicht als Leitfaden zur Einhaltung des geltenden Rechts gedacht. Angesichts der Anwendungsspezifität von KI-Systemen muss die Bewertungsliste auf die spezifischen Anwendungsfälle und Kontexte, in denen die Systeme zum Einsatz kommen, zugeschnitten werden. Darüber hinaus enthält dieses Kapitel eine allgemeine Empfehlung zur Umsetzung der Bewertungsliste für vertrauenswürdige KI auf der Grundlage einer Lenkungs- und Kontrollstruktur, die sowohl die operative als auch die Managementebene umfasst.

(113) Die Bewertungsliste sowie die Lenkungs- und Kontrollstruktur werden in enger Zusammenarbeit mit Interessenträgern aus dem öffentlichen und privaten Sektor entwickelt. Eine Pilotphase ermöglicht umfassende Rückmeldungen aus zwei parallel verlaufenden Prozessen:

a. einem qualitativen Prozess zur Sicherstellung der Darstellbarkeit, an dem sich eine kleine Auswahl an Unternehmen, Organisationen und Institutionen (aus verschiedenen Sektoren und unterschiedlicher Größe) beteiligt, um die Bewertungsliste und die Lenkungs- und Kontrollstruktur in der Praxis zu erproben und umfassende Rückmeldungen zu geben, und

b. einem quantitativen Prozess, an dem sich alle interessierten Akteure beteiligen können, um die Bewertungsliste in der Praxis zu erproben und im Rahmen einer offenen Konsultation eine Rückmeldung zu geben.

(114) Nach der Pilotphase werden die Ergebnisse der Rückmeldungen in die Bewertungsliste eingearbeitet und Anfang 2020 wird eine überarbeitete Fassung erstellt. Ziel ist es, einen Rahmen zu schaffen, der bereichsübergreifend für alle Anwendungen genutzt werden kann und damit die Grundlage für eine vertrauenswürdige KI in allen Bereichen bildet. Sobald eine solche Grundlage geschaffen ist, können sektorale oder anwendungsspezifische Rahmenbedingungen entwickelt werden.

Lenkung und Kontrolle

(115) Unternehmen, Organisationen und Institutionen möchten möglicherweise prüfen, wie die Bewertungsliste zur vertrauenswürdigen KI in ihrer Organisation umgesetzt werden kann. Das kann durch die Einbeziehung des Bewertungsprozesses in die bestehenden Lenkungs- und Kontrollmechanismen oder durch die Einführung neuer Prozesse erfolgen. Die Entscheidung hängt von der internen Struktur der jeweiligen Organisation sowie von ihrer Größe und den verfügbaren Ressourcen ab.

(116) Forschungsergebnisse[55] weisen darauf hin, dass das Engagement auf höchster Führungsebene unerlässlich für die Herbeiführung von Veränderungen ist. Sie zeigen auch, dass die Einbeziehung aller Akteure eines Unternehmens, einer Organisation oder einer Institution die Akzeptanz und Relevanz der Einführung eines

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation

neuen Prozesses (egal, ob technologisch oder nicht)[56] erhöht. Daher ist es empfehlenswert, sowohl die operative als auch die oberste Führungsebene an der Prozessimplementierung zu beteiligen.

Ebene	**Relevante Aufgaben (organisationsabhängig)**
Management und Vorstand	Auf oberster Führungsebene werden die Entwicklung und Einführung bzw. die Beschaffung von KI-Systemen besprochen und bewertet. Falls kritische Bedenken aufkommen, dient diese Ebene als Eskalationsgremium für die Bewertung aller KI-Innovationen und -nutzungen. Die von der möglichen Einführung eines KI-Systems betroffenen Personen (z. B. Arbeitnehmerinnen und Arbeitnehmer) und deren Vertreter werden mittels Informations-, Konsultations- und Beteiligungsverfahren in den gesamten Prozess einbezogen.
Qualitätsbeurteilung/Rechtsabteilung/Abteilung für Unternehmensverantwortung	Die Abteilung für Unternehmensverantwortung überwacht die Umsetzung der Bewertungsliste und deren Weiterentwicklung, die erforderlich ist, um die Beurteilung an technische oder regulatorische Änderungen anzupassen. Sie aktualisiert die Vorschriften oder internen Richtlinien für KI-Systeme und stellt sicher, dass die Nutzung der KI-System im Einklang mit den bestehenden Rechts- und Verwaltungsvorschriften sowie den Unternehmenswerten erfolgt.
Produkt- und Dienstleistungsentwicklung oder Gleichwertiges	Die Abteilung für Produkt- und Dienstleistungsentwicklung verwendet die Beurteilungsliste zur Bewertung KI-gestützter Produkte und Dienstleistungen und protokolliert alle Ergebnisse. Diese werden auf der Führungsebene besprochen, die schließlich die neuen oder überarbeiteten KI-gestützten Anwendungen genehmigt.
Qualitätssicherung	Die Abteilung für Qualitätssicherung (oder eine gleichwertige) überprüft die Ergebnisse der Arbeit mit der Bewertungsliste und ergreift Maßnahmen um ein Problem weiter oben zu eskalieren, wenn ein Ergebnis nicht zufriedenstellend ist oder wenn unvorhergesehene Ergebnisse festgestellt worden sind.
Personalabteilung	Die Personalabteilung sorgt für eine angemessene Mischung an Kompetenzen und vielfältige Profile im Bereich der KI-System-Entwickler. Sie stellt sicher, dass die Mitarbeiterinnen und Mitarbeiter im Hinblick auf eine vertrauenswürdige KI im Unternehmen angemessen geschult werden.
Beschaffung	Die Beschaffungsabteilung stellt sicher, dass der Prozess zur Beschaffung von KI-gestützten Produkten oder Dienstleistungen anhand der Bewertungsliste für

[56] Siehe z. B.: A. Bryson, E. Barth und H. Dale-Olsen, *The Effects of Organisational Change on Worker Well-Being and the Moderating Role of Trade Unions*, *ILRReview*, 66(4), Juli 2013; Jirjahn, U. und Smith, S.C. (2006). '*What Factors Lead Management to Support or Oppose Employee Participation—With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations*, 45(4), 650–680; Michie, J. und Sheehan, M. (2003). '*Labour market deregulation, "Flexibility" and Innovation', Cambridge Journal of Economics*, 27(1), 123–143.

	vertrauenswürdige KI überprüft wird.
Alltagsbetrieb	Entwickler/innen und Projektleiter/innen beziehen die Bewertungsliste in ihre tägliche Arbeit ein und dokumentieren die Ergebnisse der Bewertung.

Verwendung der Bewertungsliste für vertrauenswürdige KI

(117) Bei der praktischen Anwendung der Bewertungsliste ist es empfehlenswert, nicht nur auf die wichtigsten Bereiche zu achten, sondern auch auf die Fragen, die sich nicht (so leicht) beantworten lassen. Ein mögliches Problem könnte der Mangel an vielfältigen Fertigkeiten und Kompetenzen im Team sein, das für die Entwicklung und Erprobung des KI-Systems zuständig ist, was die Einbeziehung weiterer interner oder externer Akteure erforderlich machen könnte. Es wird dringend empfohlen, alle Ergebnisse – sowohl technische als auch betriebswirtschaftliche – zu protokollieren, um sicherzustellen, dass die Problemlösung auf allen Ebenen der Leitungs- und Kontrollstruktur verständlich ist.

(118) Diese Bewertungsliste soll KI-Akteuren bei der Entwicklung, Einführung und Verwendung von vertrauenswürdiger KI unterstützen. Die Bewertung sollte angemessen auf den konkreten Anwendungsfall zugeschnitten sein. Während der Pilotphase werden möglicherweise spezifische sensible Bereiche aufgedeckt. Die Notwendigkeit weiterer Vorgaben für solche Fällen wird dann im nächsten Schritt bewertet. Diese Bewertungsliste gibt zwar keine konkreten Antworten auf die aufgeworfenen Fragen, ermutigt aber zum Nachdenken über die Schritte, die dazu beitragen könnten, die Vertrauenswürdigkeit von KI-Systemen zu gewährleisten, und die Maßnahmen, die diesbezüglich ergriffen werden sollten.

Bezug zu bestehenden Gesetzen und Prozessen

(119) Die an der Entwicklung, Einführung und Nutzung von KI Beteiligten sollten wissen, dass bestimmte Verfahrensweisen gesetzlich vorgeschrieben und bestimmte Ergebnisse gesetzlich verboten sind, und dass diese sich mit einigen der in der Bewertungsliste aufgeführten Maßnahmen überschneiden und damit übereinstimmen können. So legt das Datenschutzrecht beispielsweise eine Reihe von gesetzlichen Anforderungen fest, die von denjenigen erfüllt werden müssen, die an der Erhebung und Verarbeitung personenbezogener Daten beteiligt sind. Da eine vertrauenswürdige KI aber auch den ethischen Umgang mit Daten erforderlich macht, können interne Verfahren und Strategien, die die Einhaltung der Datenschutzrichtlinien gewährleisten, auch einen ethischen Umgang mit Daten erleichtern und somit geltende rechtliche Verfahren ergänzen. Die Einhaltung dieser Bewertungsliste ist *kein* Nachweis der Gesetzeskonformität und ist nicht als Leitfaden zur Einhaltung des geltenden Rechts gedacht. Vielmehr soll mit dieser Bewertungsliste den Adressaten eine Reihe spezifischer Fragen an die Hand gegeben werden, mit deren Hilfe sie sicherstellen können, dass ihr Ansatz bei der Entwicklung und Einführung von KI auf die Gewährleistung einer vertrauenswürdigen KI ausgerichtet ist.

(120) Viele KI-Akteure verfügen bereits über bestehende Bewertungsinstrumente und Verfahren zur Softwareentwicklung, die der Einhaltung auch nicht-rechtlicher Vorgaben dienen. Die nachstehende Bewertungsliste muss nicht unbedingt einzeln angewendet werden, sie kann vielmehr in bestehende Verfahren einbezogen werden.

BEWERTUNGSLISTE FÜR VERTRAUENSWÜRDIGE KI (PILOTVERSION)

1. Vorrang menschlichen Handelns und menschliche Aufsicht

Grundrechte:

- ✓ Haben Sie für die Anwendungsfälle, bei denen die Möglichkeit einer Beeinträchtigung der Grundrechte besteht, eine Folgenabschätzung in Bezug auf die Grundrechte durchgeführt? Haben Sie mögliche Kompromisse zwischen den verschiedenen Grundsätzen und Rechten erkannt und dokumentiert?

- ✓ Beeinflusst das KI-System die Entscheidungsfindung menschlicher Endnutzer (z. B. Empfehlungen für Handlungen oder Entscheidungen, Vorschlag verschiedener Möglichkeiten)?

  - Besteht in diesen Fällen die Gefahr, dass das KI-System die menschliche Autonomie beeinträchtigt, indem es unbeabsichtigt in den Entscheidungsprozess des Endverbrauchers eingreift?

  - Haben Sie in Betracht gezogen, ob das KI-System den Nutzern mitteilen sollte, dass eine Entscheidung, ein Inhalt, eine Empfehlung oder ein Ergebnis die Folge einer auf Algorithmen basierenden Entscheidung ist?

  - Falls das KI-System über ein Chatbot oder ein Konversationssystem verfügt, werden die menschlichen Endnutzer auf die Tatsache aufmerksam gemacht, dass sie mit einem nicht-menschlichen Wesen, also einer Maschine, interagieren?

Vorrang menschlichen Handelns:

- ✓ Falls das KI-System Teil des Arbeitsprozesses ist, haben Sie bei der Aufgabenverteilung zwischen dem KI-System und menschlichen Beschäftigten eine sinnvolle Interaktion und eine angemessene menschliche Aufsicht und Kontrolle berücksichtigt?

  - Verbessert oder erweitert das KI-System die menschlichen Fähigkeiten?

  - Haben Sie Vorkehrungen getroffen, um im Rahmen der Arbeitsprozesse einem übermäßigen Vertrauen in das KI-System vorzubeugen?

Menschliche Aufsicht:

- ✓ Haben Sie das angemessene Maß an menschlicher Kontrolle für das jeweilige KI-System und den spezifischen Anwendungsfall überprüft?

  - Können Sie gegebenenfalls das Ausmaß an menschlicher Kontrolle bzw. menschlicher Teilnahme beschreiben? Welche Person ist mit der Kontrolle beauftragt und welche Werkzeuge stehen für den menschlichen Eingriff zur Verfügung?

  - Haben Sie Mechanismen und Maßnahmen eingeführt, um eine solche potenzielle menschliche Kontrolle oder Aufsicht zu gewährleisten oder sicherzustellen, dass Entscheidungen unter der Gesamtverantwortung eines Menschen getroffen werden?

  - Haben Sie Maßnahmen zur Nachprüfbarkeit und Verhinderung von Problemen im Zusammenhang mit der Steuerung und Kontrolle der Autonomie der KI ergriffen?

- ✓ Haben Sie für den Fall eines selbstlernenden oder autonomen KI-Systems oder Anwendungsfalls spezifischere Vorkehrungen zur Kontrolle und Aufsicht eingeführt?

  - Welche Art von Erkennungs- und Reaktionsmechanismen haben Sie etabliert, um zu beurteilen, ob etwas schief gehen könnte?

  - Haben Sie eine „Stopptaste“ oder gegebenenfalls ein Verfahren zum sicheren Abbrechen eines Vorgangs sichergestellt? Wird dabei der Prozess ganz oder teilweise abgebrochen oder die

Kontrolle an einen Menschen übergeben?

2. <u>Technische Robustheit und Sicherheit</u>

Schutz gegen Angriffe und Sicherheit:

- ✓ Haben Sie mögliche Angriffsformen eingeschätzt, für die das KI-System anfällig sein könnte?
  - Haben Sie insbesondere verschiedene Arten möglicher Schwachstellen wie Datenverunreinigung, physische Infrastruktur, Cyberangriffe berücksichtigt?
- ✓ Haben Sie Maßnahmen oder Systeme eingeführt, um die Integrität und Belastbarkeit des KI-Systems gegenüber möglichen Angriffen zu gewährleisten?
- ✓ Haben Sie das Verhalten Ihres Systems in unerwarteten Situationen und Umgebungen bewertet?
- ✓ Haben Sie in Erwägung gezogen, ob und inwieweit Ihr System einen doppelten Verwendungszweck (*Dual-Use*) haben könnte? Wenn ja, haben Sie geeignete vorbeugende Maßnahmen für diesen Fall ergriffen (z. B. Verzicht auf die Veröffentlichung der Forschung oder Verzicht auf den Einsatz des Systems)?

Auffangplan und allgemeine Sicherheit:

- ✓ Haben Sie sichergestellt, dass Ihr System über einen befriedigenden Auffangplan verfügt, falls es feindlichen Angriffen oder anderen unerwarteten Situationen ausgesetzt ist (z. B. technische Umschaltverfahren oder das Heranziehen eines menschlichen Bedieners, bevor die Arbeit fortgesetzt wird)?
- ✓ Haben Sie das vom KI-System verursachte Risiko in diesem speziellen Anwendungsfall berücksichtigt?
  - Haben Sie ein Verfahren zur Messung und Bewertung von Risiken und Sicherheit eingeführt?
  - Haben Sie die notwendigen Informationen für den Fall der Bedrohung der körperlichen Unversehrtheit von Menschen bereitgestellt?
  - Haben Sie eine Versicherung in Betracht gezogen, die mögliche, vom KI-System verursachte Schäden deckt?
  - Haben Sie die potenziellen Sicherheitsrisiken infolge der vorhersehbaren (und anderer) Anwendungen der Technologie, einschließlich eines versehentlichen oder böswilligen Missbrauchs, ermittelt? Gibt es einen Plan zur Begrenzung oder Minderung dieser Gefahren?
- ✓ Haben Sie beurteilt, ob die Möglichkeit besteht, dass das KI-System den Anwendern oder Dritten Schaden zufügt? Wenn ja, haben Sie die Wahrscheinlichkeit, den potenziellen Schaden, den betroffenen Personenkreis und die Schwere des Schadens eingeschätzt?
  - Falls die Gefahr besteht, dass das KI-System Schäden verursacht, haben Sie die Haftungs- und Verbraucherschutzbestimmungen berücksichtigt? In welcher Art und Weise haben Sie dies getan?
  - Haben Sie die möglichen Auswirkungen oder Sicherheitsrisiken für Umwelt oder Tiere in Betracht gezogen?
  - Haben Sie bei Ihrer Risikoanalyse berücksichtigt, ob Sicherheits- oder Netzwerkprobleme (z. B. Bedrohungen der Cybersicherheit) ein Sicherheitsrisiko oder Schäden durch ein unbeabsichtigtes

Verhalten des KI-Systems Schäden verursachen könnten?

- ✓ Haben Sie die möglichen Auswirkungen eines Ausfalls Ihres KI-Systems eingeschätzt und ob dieser zu falschen Ergebnissen, zur Nichtverfügbarkeit Ihres Systems oder zu gesellschaftlich inakzeptablen Ergebnissen (z. B. diskriminierende Praktiken) führen könnte?
  - Haben Sie Schwellenwerte sowie Maßnahmen zur Lenkung und Kontrolle für die oben genannten Szenarien festgelegt, damit alternative oder Auffangpläne aktiviert werden?
  - Haben Sie Auffangpläne festgelegt und erprobt?

Präzision

- ✓ Haben Sie beurteilt, welcher Grad an Präzision und welche Definition von Genauigkeit im Zusammenhang mit dem KI-System und dem jeweiligen Anwendungsfall erforderlich sind?
  - Haben Sie beurteilt, wie die Präzision gemessen und gewährleistet wird?
  - Haben Sie Maßnahmen ergriffen, um sicherzustellen, dass die verwendeten Daten umfassend und aktuell sind?
  - Haben Sie Maßnahmen ergriffen, um zu beurteilen, ob zusätzliche Daten erforderlich sind, z. B. um die Präzision zu verbessern oder Verzerrungen zu vermeiden?
- ✓ Haben Sie den Schaden bewertet, der entstehen würde, wenn das KI-System ungenaue Vorhersagen machen würde?
- ✓ Haben Sie Methoden eingeführt, um zu messen, ob Ihr System eine inakzeptable Anzahl nicht präziser Vorhersagen ausgibt?
- ✓ Wenn ungenaue Vorhersagen gemacht werden, haben Sie die nötigen Schritte zur Lösung des Problems veranlasst?

Zuverlässigkeit und Wiederholbarkeit:

- ✓ Haben Sie eine Strategie zur Überwachung und Erprobung eingeführt, die Ihnen zu erkennen hilft, ob Ihr KI-System die Ziele, Zwecke und vorgesehenen Anwendungen erfüllt?
  - Haben Sie getestet, ob bestimmte Kontexte oder Bedingungen berücksichtigt werden müssen, damit die Wiederholbarkeit gewährleistet ist?
  - Haben Sie Prüfverfahren oder -methoden zur Messung und Sicherstellung verschiedener Aspekte der Zuverlässigkeit und Wiederholbarkeit eingeführt?
  - Haben Sie Verfahren eingeführt, die beschreiben, wann ein KI-System bei bestimmten Einstellungen ausfällt?
  - Haben Sie diese Prozesse zur Erprobung und Prüfung der Zuverlässigkeit von KI-Systemen eindeutig dokumentiert und operativ umgesetzt?

    Haben Sie Vorkehrungen getroffen oder ein Kommunikationssystem eingerichtet, mit dem Sie den (End-)Nutzern die Zuverlässigkeit des KI-Systems garantieren können?

3. <u>Schutz der Privatsphäre und Datenqualitätsmanagement</u>

Achtung der Privatsphäre und Gewährleistung des Datenschutzes:

- ✓ Haben Sie je nach Anwendungsfall einen Mechanismus eingerichtet, der es anderen ermöglicht, Fragen der Privatsphäre oder des Datenschutzes im Zusammenhang mit den Prozessen der Datenerfassung (für Ausbildung und Einsatz) und Datenverarbeitung des KI-Systems zu kennzeichnen?
- ✓ Haben Sie die Art und den Umfang der in Ihren Datensätzen enthaltenen Daten beurteilt (z. B. ob personenbezogene Daten enthalten sind)?
- ✓ Haben Sie Methoden in Betracht gezogen, wie Sie das KI-System entwickeln oder das Modell ausbilden können, wobei keine oder eine minimale Anzahl von potenziell sensiblen oder personenbezogenen Daten verwendet werden?
- ✓ Haben Sie je nach Anwendungsfall Vorkehrungen zur Kenntnisnahme und Kontrolle personenbezogener Daten getroffen (z. B. eine gültige Einwilligungserklärung und ggf. eine Widerrufsmöglichkeit)?
- ✓ Haben Sie Maßnahmen zur Verstärkung der Privatsphäre ergriffen, z. B. durch Verschlüsselung, Anonymisierung und Aggregierung?
- ✓ Falls es einen Datenschutzbeauftragten (DSB) gibt, haben Sie diese Person frühzeitig in den Prozess einbezogen?

Qualität und Integrität der Daten:

- ✓ Haben Sie Ihr System auf potenziell relevante Normen (z. B. ISO, IEEE) oder weitverbreitete Protokolle für Ihr tägliches Datenqualitätsmanagement abgestimmt?
- ✓ Haben Sie Kontrollmechanismen für die Erhebung, Speicherung, Verarbeitung und Nutzung der Daten eingerichtet?
- ✓ Haben Sie beurteilt, inwieweit Sie die Qualität der verwendeten externen Datenquellen unter Kontrolle haben?
- ✓ Haben Sie Prozesse zur Gewährleistung der Qualität und Integrität Ihrer Daten eingeführt? Haben Sie weitere Prozesse in Betracht gezogen? Wie stellen Sie sicher, dass Ihre Datensätze weder beeinträchtigt noch gehackt wurden?

Datenzugriff:

- ✓ An welche Protokolle, Prozesse und Verfahren haben Sie sich gehalten, um ein ordnungsgemäßes Datenqualitätsmanagement sicherzustellen?
  - Haben Sie festgesetzt, wer unter welchen Umständen auf die Benutzerdaten zugreifen darf?
  - Haben Sie sichergestellt, dass diese Personen dazu berechtigt sind und es erforderlich ist, dass sie auf die Daten zugreifen, und dass sie über die notwendigen Kompetenzen verfügen, um die Einzelheiten der Datenschutzvorschriften zu verstehen?
  - Haben Sie einen Überwachungsmechanismus eingerichtet, mit dem protokolliert wird, wann, wo, wie, von wem und zu welchem Zweck ein Datenzugriff erfolgte?

4. <u>Transparenz</u>

Rückverfolgbarkeit:

- ✓ Haben Sie Maßnahmen zur Gewährleistung der Rückverfolgbarkeit ergriffen? Diese könnten folgende Dokumentation umfassen:
    - ➢ Verwendete Methoden für den Entwurf und die Entwicklung des algorithmischen Systems:
        - o Im Falle eines regelbasierten KI-Systems sollte die Art der Programmierung oder die Art und Weise, wie das Modell erstellt wurde, dokumentiert werden.
        - o Im Falle eines lernbasierten KI-Systems sollte die Methode zur Ausbildung des Algorithmus, einschließlich der Angabe, welche Eingabedaten erhoben und ausgewählt wurden und wie dies geschah, dokumentiert werden.
    - ➢ Verwendete Methoden für die Erprobung und Validierung des algorithmischen Systems:
        - o Im Falle eines regelbasierten KI-Systems sollten die Szenarien oder Fälle, die zur Erprobung und Validierung verwendet wurden, dokumentiert werden.
        - o Im Falle eines lernbasierten Modells sollten Angaben zu den zur Erprobung und Validierung verwendeten Daten dokumentiert werden.
    - ➢ Ergebnisse des algorithmischen Systems:
        - o Die Ergebnisse des Algorithmus oder die vom Algorithmus getroffenen Entscheidungen sowie mögliche andere Entscheidungen, die sich aus anderen Fällen (z. B. für andere Benutzeruntergruppen) ergeben könnten, sollten dokumentiert werden.

Erklärbarkeit:

- ✓ Haben Sie beurteilt, inwieweit die Entscheidungen und damit das Ergebnis des KI-Systems nachvollziehbar sind?
- ✓ Haben Sie sichergestellt, dass eine Erklärung der Gründe, warum ein System eine bestimmte Wahl getroffen hat, die zu einem bestimmten Ergebnis führt, für alle Benutzer, die eine Erklärung wünschen, verständlich gemacht werden kann?
- ✓ Haben Sie beurteilt, inwieweit die Entscheidung des Systems die Entscheidungsprozesse der Organisation beeinflusst?
- ✓ Haben Sie beurteilt, warum dieses spezielle System in diesen speziellen Bereich eingeführt wurde?
- ✓ Haben Sie das Geschäftsmodell im Zusammenhang mit diesem System bewertet (z. B. auf welche Weise es Wert für die Organisation schafft)?
- ✓ Haben Sie das KI-System von Anfang an auf die Möglichkeit der Interpretierbarkeit ausgelegt?
    - ▪ Haben Sie recherchiert und versucht, das einfachste und am besten interpretierbare Modell für die jeweilige Anwendung zu verwenden?
    - ▪ Haben Sie geprüft, ob Sie die Daten, die Sie für die Ausbildung und Erprobung verwenden, analysieren können? Können Sie das im Laufe der Zeit ändern und aktualisieren?
    - ▪ Haben Sie beurteilt, ob Sie nach der Ausbildung und Entwicklung des Modells über Möglichkeiten zur Überprüfung der Interpretierbarkeit verfügen oder ob Sie Zugang zum internen Workflow des Modells haben?

Kommunikation:

- ✓ Haben Sie den (End-)Nutzern – mit einem Haftungsausschluss oder anderweitig – mitgeteilt, dass sie mit einem KI-System und nicht mit einem anderen Menschen interagieren? Haben Sie Ihr KI-System

als solches ausgewiesen?

- ✓ Haben Sie Mechanismen eingerichtet, mit denen Sie die Nutzer über die Gründe und Kriterien, die den Ergebnissen des KI-Systems zugrunde liegen, informieren?
  - Wird dies den angesprochenen Nutzern klar und verständlich mitgeteilt?
  - Haben Sie Prozesse eingeleitet, die die Rückmeldungen der Benutzer berücksichtigen und auf deren Grundlage das System angepasst wird?
  - Haben Sie auch über potenzielle oder tatsächlich erkannte Risiken, wie z. B. Verzerrungen, informiert?
  - Haben Sie je nach Anwendungsfall auch den Informationsaustausch und die Transparenz gegenüber anderen Zielgruppen, Dritten oder der Öffentlichkeit erwogen?
- ✓ Haben Sie klargestellt, worin der Zweck des KI-Systems besteht und wer oder was von dem Produkt/der Dienstleistung profitieren kann?
  - Wurden die Anwendungsszenarien für das Produkt angegeben und klar verständlich mitgeteilt? Wurden dabei auch alternative Kommunikationsformen in Betracht gezogen, um sicherzustellen, dass sie für die angesprochenen Benutzergruppen verständlich und angemessen sind?
  - Haben Sie je nach Anwendungsfall die menschliche Psychologie und mögliche Einschränkungen wie Verwechslungsgefahr, Bestätigungsfehler oder geistige Ermüdung in Betracht gezogen?
- ✓ Haben Sie eindeutig die Merkmale, Grenzen und potenziellen Mängel des KI-Systems folgenden Personen mitgeteilt?
  - Bei der Entwicklung: denjenigen Personen, die das KI-System in ein Produkt oder eine Dienstleistung einbauen?
  - Bei der Einführung: den Endnutzern oder Verbrauchern?

5. <u>Vielfalt, Nichtdiskriminierung und Fairness</u>

Vermeidung unfairer Verzerrungen:

- ✓ Haben Sie eine Strategie oder eine Reihe von Verfahren vorgesehen, um die Entstehung oder Verstärkung unfairer Verzerrungen im KI-System hinsichtlich sowohl der Verwendung der Eingangsdaten als auch des Entwurfs der Algorithmen zu vermeiden?
  - Haben Sie mögliche Einschränkungen, die sich aus der Zusammensetzung der verwendeten Datensätze ergeben könnten, bewertet und zur Kenntnis genommen?
  - Haben Sie beachtet, dass die Daten die Vielfalt der Nutzer widerspiegeln und repräsentativ sind? Haben Sie das KI-System auf bestimmte Anwendergruppen oder problematische Anwendungsfälle getestet?
  - Haben Sie die verfügbaren technischen Hilfsmittel recherchiert und eingesetzt, um ein besseres Verständnis von Daten, Modell und Leistung zu erzielen?
  - Haben Sie während der Phasen der Entwicklung, Einführung und Nutzung des Systems Prozesse eingerichtet, mit denen Sie das System auf mögliche Verzerrungen untersuchen und überwachen

können?

- ✓ Haben Sie abhängig vom Anwendungsfall einen Mechanismus vorgesehen, der es anderen ermöglicht, Probleme im Zusammenhang mit Verzerrungen, Diskriminierung oder schlechter Leistung des KI-Systems zu kennzeichnen?
  - Haben Sie klare Schritte und Kommunikationswege darüber in Betracht gezogen, wie und an wen solche Themen herangetragen werden sollten?
  - Haben Sie außer den (End-)Nutzern auch weitere, möglicherweise indirekt vom KI-System Betroffene berücksichtigt?
- ✓ Haben Sie untersucht, ob unter den gleichen Bedingungen die Entscheidungen bestimmten Schwankungen unterliegen könnten?
  - Wenn ja, haben Sie über die möglichen Ursachen nachgedacht?
  - Haben Sie in Bezug auf solche Schwankungen Vorkehrungen zur Messung oder Bewertung möglicher Auswirkungen auf die Grundrechte getroffen?
- ✓ Haben Sie eine angemessene Arbeitsdefinition von „Fairness“ festgelegt, die Sie bei der Gestaltung von KI-Systemen anwenden?
  - Wird Ihre Definition üblicherweise verwendet? Haben Sie andere Definitionen in Betracht gezogen, bevor Sie sich für diese entschieden haben?
  - Haben Sie eine quantitative Analyse oder Metriken zur Messung und Prüfung der angewandten Definition von Fairness vorgesehen?
  - Haben Sie Mechanismen eingeführt, mit denen Sie gewährleisten können, dass Sie faire KI-Systeme einsetzen? Haben Sie andere mögliche Vorkehrungen in Betracht gezogen?

Barrierefreiheit und universeller Entwurf:

- ✓ Haben Sie sichergestellt, dass das KI-System eine Vielzahl individueller Vorlieben und Fertigkeiten berücksichtigt?
  - Haben Sie geprüft, ob das KI-System für Menschen mit besonderen Bedürfnissen oder Behinderungen oder für von Ausgrenzung bedrohten Menschen geeignet ist? Wie wurden diese Gesichtspunkte in das System integriert und wie werden sie überprüft?
  - Haben Sie sichergestellt, dass die Informationen über das KI-System auch für Nutzer assistierender Technologien zugänglich sind?
  - Haben Sie diese Gruppe in die Entwicklungsphase des KI-Systems einbezogen oder konsultiert?
- ✓ Haben Sie die Auswirkungen Ihres KI-Systems auf ihre potenzielle Zielgruppe berücksichtigt?
  - Ist das am Aufbau des KI-Systems beteiligte Team für Ihre Zielgruppe repräsentativ? Repräsentiert das System unter Einbeziehung möglicherweise indirekt betroffener Gruppen die allgemeine Bevölkerung?
  - Haben Sie beurteilt, ob es Personen oder Gruppen geben kann, die überproportional von unerwünschten Auswirkungen betroffen sein könnten?
  - Haben Sie Rückmeldungen von anderen Teams oder Gruppen erhalten, die unterschiedliche

Hintergründe und Erfahrungen repräsentieren?

Beteiligung der Interessenträger:

- ✓ Haben Sie Vorkehrungen getroffen, um die Beteiligung verschiedener Interessenträger an der Entwicklung und Nutzung des KI-Systems sicherzustellen?
- ✓ Haben Sie den Weg für die Einführung des KI-Systems in Ihre Organisation geebnet, indem Sie die betroffenen Arbeitnehmerinnen und Arbeitnehmer sowie deren Vertreter im Voraus informiert und in den Prozess einbezogen haben?

6. Gesellschaftliches und ökologisches Wohlergehen

Nachhaltige und umweltfreundliche KI:

- ✓ Haben Sie während der Entwicklung, der Einführung und Nutzung des KI-Systems Maßnahmen zur Messung der Umweltauswirkungen eingeführt (z. B. Energieverbrauch des Rechenzentrums, Art der von den Rechenzentren verwendeten Energie usw.)?
- ✓ Haben Sie Maßnahmen zur Reduzierung der durch Ihr KI-System während seines gesamten Lebenszyklus verursachten Umweltbelastung eingeführt?

Soziale Auswirkungen:

- ✓ Falls das KI-System direkt mit Menschen interagiert:
  - Haben Sie beurteilt, ob das KI-System den Menschen ermutigt, sich mit dem System verbunden zu fühlen und Empathie zu entwickeln?
  - Haben Sie sichergestellt, dass das KI-System deutlich zu verstehen gibt, dass es soziale Interaktionen nur simuliert und dass es weder „verstehen“ noch „fühlen“ kann?
- ✓ Haben Sie ein gutes Verständnis der sozialen Auswirkungen des KI-Systems sichergestellt? Haben Sie beispielsweise das Risiko drohender Arbeitsplatzverluste und der Dequalifizierung von Arbeitnehmerinnen und Arbeitnehmer beurteilt? Welche Maßnahmen wurden ergriffen, um diesen Risiken entgegenzuwirken?

Gesellschaft und Demokratie:

- ✓ Haben Sie die weiter reichenden gesellschaftlichen Auswirkungen der Nutzung des KI-Systems über die einzelnen (End-)Nutzer hinausgehend bewertet und dabei beispielsweise andere, möglicherweise indirekt betroffene Akteure berücksichtigt?

7. Rechenschaftspflicht

Nachprüfbarkeit:

- ✓ Haben Sie Mechanismen eingeführt, die die Nachprüfbarkeit des Systems durch interne und/oder externe Prüfer erleichtern? Haben Sie z. B. die Rückverfolgbarkeit und Protokollierung der Prozesse und Ergebnisse des KI-Systems sichergestellt?

Minimierung und Meldung negativer Auswirkungen:

- ✓ Haben Sie eine Risiko- oder Folgenabschätzung für das KI-System unter Berücksichtigung verschiedener, direkt oder indirekt betroffener Interessenträger durchgeführt?

- ✓ Haben Sie zur Entwicklung von Verfahren zur Rechenschaftspflicht einen Rahmen für Aus- und Weiterbildungsmaßnahmen geschaffen?
  - Welche Mitarbeiterinnen und Mitarbeiter oder Abteilungen sind daran beteiligt? Gehen diese Maßnahmen über die Entwicklungsphase hinaus?
  - Werden im Rahmen dieser Schulungen auch die möglicherweise für das KI-System geltenden rechtlichen Rahmenbedingungen vermittelt?
  - Haben Sie die Einrichtung eines „Ethikausschusses“ oder etwas ähnliches in Betracht gezogen, um die umfassende Rechenschaftspflicht sowie ethische Verfahrensweisen einschließlich potenziell unklarer Grauzonen zu erörtern?
- ✓ Gibt es neben internen Initiativen oder Rahmenbedingungen zur Überwachung von Ethik und Rechenschaftspflicht auch externe Leitlinien oder wurden zusätzliche Prüfverfahren eingeführt?
- ✓ Sind für Drittpersonen (z. B. Lieferanten, Verbraucher, Händler/Anbieter) oder Mitarbeiterinnen und Mitarbeiter Verfahren zur Meldung potenzieller Schwachstellen, Risiken oder Verzerrungen des KI-Systems/der Anwendung vorgesehen?

Dokumentation von Kompromissen:

- ✓ Haben Sie Vorkehrungen zum Erkennen relevanter Interessen und Werte im Zusammenhang mit dem KI-System und zum Feststellen möglicherweise erforderlicher Kompromisse getroffen?
- ✓ Auf welche Prozesse stützen Sie sich, um Entscheidungen bezüglich solcher Kompromisse zu treffen? Haben Sie die Dokumentation einer solchen Kompromisslösung sichergestellt?

Rechtsschutz:

- ✓ Haben Sie angemessene Vorkehrungen zum Schadensausgleich im Falle eines Schadens oder nachteiliger Auswirkungen des KI-Systems getroffen?
- ✓ Haben Sie Vorkehrungen getroffen, um die (End-)Nutzer/Dritte über ihre Möglichkeiten zum Einlegen von Rechtsmitteln zu informieren?

Wir fordern alle Beteiligten auf, diese Bewertungsliste in der Praxis zu erproben und uns eine Rückmeldung über deren Umsetzbarkeit, Vollständigkeit und Relevanz für die spezifische KI-Anwendung oder den Anwendungsbereich sowie über Überschneidungen oder Entsprechungen mit bestehenden Konformitäts- oder Bewertungsverfahren zu geben. Auf der Grundlage dieser Rückmeldungen wird der Kommission Anfang 2020 eine überarbeitete Fassung der Bewertungsliste für vertrauenswürdige KI unterbreitet.

Wichtige Leitlinien aus Kapitel III:

- ✓ Verwenden Sie bei der Entwicklung, Einführung oder Verwendung eines KI-Systems die **Bewertungsliste** für vertrauenswürdige KI und passen Sie sie an den spezifischen Anwendungsfall an, auf den das System ausgerichtet ist.
- ✓ Denken Sie daran, dass eine solche Bewertungsliste **niemals vollständig** sein wird. Bei der Gewährleistung einer vertrauenswürdigen KI geht es nicht darum, Kästchen anzukreuzen, sondern darum, kontinuierlich Anforderungen zu erkennen, Lösungen zu bewerten, die Ergebnisse während des gesamten Lebenszyklus des KI-Systems laufend zu verbessern und die Akteure in diesen Prozess einzubeziehen.

C. BEISPIELE FÜR DIE MÖGLICHKEITEN, DIE DIE KI BIETET, UND KRITISCHE ERWÄGUNGEN

(121) Im Folgenden finden Sie Beispiele für die Entwicklung und Nutzung von KI, die gefördert werden sollten, sowie Beispiele dafür, wie die Entwicklung, Einführung oder Nutzung von KI unseren Werten zuwiderlaufen und besondere Bedenken aufwerfen könnte. Es muss ein Gleichgewicht gefunden werden zwischen dem, was mit KI getan werden sollte und dem, was mit KI getan werden kann, und es muss mit der gebotenen Sorgfalt darauf geachtet werden, was mit KI nicht getan werden sollte.

1. Beispiele für die Chancen einer vertrauenswürdigen KI

(122) Eine vertrauenswürdige KI kann bei der Linderung dringender gesellschaftlicher Herausforderungen wie Überalterung, wachsende soziale Ungleichheit und Umweltverschmutzung von großem Nutzen sein. Dieses Potenzial spiegelt sich auch weltweit wider, beispielsweise in den nachhaltigen Entwicklungszielen der Vereinten Nationen[57]. Der folgende Abschnitt befasst sich mit der Frage, wie eine europäische KI-Strategie, die einige dieser Herausforderungen angeht, gefördert werden kann.

a. Klimaschutz und nachhaltige Infrastruktur

(123) Obwohl die Bekämpfung des Klimawandels für politische Entscheidungsträger weltweit an erster Stelle stehen sollte, bergen der digitale Wandel und die vertrauenswürdige KI ein großes Potenzial, die Belastung der Umwelt durch den Menschen zu verringern und eine effiziente und effektive Nutzung von Energie und natürlichen Ressourcen zu ermöglichen[58]. Eine vertrauenswürdige KI kann beispielsweise mit großen Datenmengen (Big Data) verknüpft werden, um den Energiebedarf genauer zu erfassen, was eine effizientere Energieinfrastruktur und einen sparsameren Energieverbrauch zur Folge hat[59].

(124) In Sektoren wie dem öffentlichen Verkehrswesen können KI-Systeme für intelligente Verkehrssysteme[60] eingesetzt werden, um Warteschlangen zu reduzieren, die Streckenführung zu optimieren, sehbehinderten Menschen mehr Unabhängigkeit zu ermöglichen[61], energieeffiziente Motoren zu optimieren und so die Bemühungen zur Dekarbonisierung zu stärken und den Umweltfußabdruck zugunsten einer umweltfreundlicheren Gesellschaft zu verringern. Derzeit stirbt weltweit alle 23 Sekunden eine Person bei einem Autounfall[62]. KI-Systeme könnten dazu beitragen, die Zahl der Todesopfer deutlich herabzusetzen, zum Beispiel durch verbesserte Reaktionszeiten und eine bessere Einhaltung der Verkehrsregeln[63].

b. Gesundheit und Wohlergehen

(125) Vertrauenswürdige KI-Technologien können eingesetzt werden – und werden bereits eingesetzt –, um Behandlungen intelligenter und zielgerichteter zu gestalten und lebensbedrohliche Krankheiten zu

[57] https://sustainabledevelopment.un.org/?menu=1300

[58] Eine Reihe von EU-Projekten hat die Entwicklung von intelligenten Netzen und Energiespeichern zum Ziel, die zusammen mit KI-gestützten und anderen digitalen Lösungen einen Beitrag zu einem erfolgreichen digital gestützten Energiewandel leisten können. Zur Ergänzung der Arbeit dieser einzelnen Projekte hat die Kommission die BRIDGE-Initiative gestartet, die es den laufenden Projekten zu intelligenten Netzen und Energiespeichern im Rahmen von Horizont 2020 ermöglicht, eine gemeinsame Sichtweise für übergreifende Themen zu entwickeln: https://www.h2020-bridge.eu/.

[59] Siehe z. B. das Projekt Encompass: http://www.encompass-project.eu/.

[60] Neue KI-gestützte Lösungen helfen den Städten, sich auf die Zukunft der Mobilität vorzubereiten. Siehe beispielsweise das von der EU geförderte Projekt Fabulos: https://fabulos.eu/.

[61] Siehe beispielsweise das Projekt PRO4VIP, das Teil der europäischen Strategie Vision 2020 zur Bekämpfung vermeidbarer Blindheit, insbesondere im Alter, ist. Mobilität und Orientierung waren Schwerpunkte des Projekts.

[62] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[63] Das europäische Projekt UP-Drive hat beispielsweise zum Ziel, die beschriebenen Verkehrsprobleme durch die Förderung einer schrittweisen Automatisierung und Zusammenarbeit zwischen den Fahrzeugen und somit eines sichereren, integrativeren und erschwinglicheren Verkehrssystems anzugehen. https://up-drive.eu/.

verhindern[64]. Ärzte und medizinisches Fachpersonal haben die Möglichkeit, eine präzisere und detailliertere Analyse der komplexen Gesundheitsdaten eines Patienten durchzuführen, noch bevor eine Person erkrankt, und maßgeschneiderte vorbeugende Maßnahmen anzubieten[65]. Vor dem Hintergrund der Bevölkerungsalterung in Europa können KI und Robotik wertvolle Werkzeuge zur Unterstützung der Pflegekräfte, der Altenpflege[66] und zur Überwachung des Zustands der Patienten in Echtzeit sein, wodurch Leben gerettet werden[67].

(126) Vertrauenswürdige KI kann auch eine Hilfe auf einer breiteren Basis sein. Sie kann beispielsweise allgemeine Trends im Gesundheitswesen und dem therapeutischen Bereich analysieren und erkennen[68] und damit eine frühere Krankheitserkennung, eine effizientere Arzneimittelentwicklung und zielgerichtetere Therapien ermöglichen[69], was letztlich mehr Leben rettet.

c. Hochwertige Bildung und digitaler Wandel

(127) Neue technologische, wirtschaftliche und ökologische Veränderungen erfordern eine stärker proaktive Gesellschaft. Regierungen, Branchenführer, Bildungseinrichtungen und Gewerkschaften stehen in der Verantwortung, Bürgerinnen und Bürger in das neue digitale Zeitalter zu führen und sicherzustellen, dass diese über die angemessenen Kompetenzen verfügen, um die künftigen Arbeitsplätze zu besetzen. Vertrauenswürdige KI-Technologien könnten dazu beitragen, genauer vorherzusagen, welche Arbeitsplätze und Berufe durch die Technologie zerstört werden, welche neuen Aufgaben geschaffen und welche Kompetenzen benötigt werden. Sie könnten Regierungen, Gewerkschaften und Industrie bei der Planung von (Um)schulungsmaßnahmen für Arbeitnehmerinnen und Arbeitnehmer unterstützen. Sie könnten auch von Entlassungen bedrohten Bürgerinnen und Bürgern bei ihrer Vorbereitung auf eine neue Tätigkeit helfen.

(128) Darüber hinaus kann die KI ein großartiges Instrument zur Bekämpfung von Ungleichheiten im Bildungsbereich und zur Schaffung personalisierter und anpassungsfähiger Bildungsprogramme sein, die allen Menschen beim Erwerb neuer Qualifikationen, Fähigkeiten und Kompetenzen entsprechend ihrer Lernfähigkeit helfen können[70]. Von der Grundschule bis zur Universität könnte die KI sowohl die Lerngeschwindigkeit als auch die Qualität der Bildung erhöhen.

---

[64] Siehe z. B. das Projekt REVOLVER (*Repeated Evolution of Cancer*): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/ oder das Murab-Projekt, in dessen Rahmen präzisere Biopsien durchgeführt werden und eine schnellere Diagnose von Krebs und anderen Krankheiten anstrebt wird:https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

[65] Siehe beispielsweise das Projekt Live INCITE: www.karolinska.se/en/live-incite. Dieser Zusammenschluss von Leistungserbringern im Gesundheitswesen fordert die Industrie heraus, intelligente KI- und andere IKT-Lösungen zu entwickeln, die Interventionen in den Lebensstil während des perioperativen Prozesses ermöglichen. Angestrebt werden neue innovative eHealth-Lösungen, die Patientinnen und Patienten personalisiert ansprechen können, damit diese vor und nach einer Operation die notwendigen Maßnahmen für einen das Therapieergebnis optimierenden Lebensstil ergreifen.

[66] Das von der EU geförderte Projekt CARESSES beschäftigt sich mit Robotern für die Altenpflege und konzentriert sich dabei auf deren kulturelles Einfühlungsvermögen: Die Roboter passen ihre Handlungs- und Sprechweise an die Kultur und die Gewohnheiten der von ihnen unterstützten älteren Person an: http://caressesrobot.org/en/project/. Siehe auch die KI-Anwendung Alfred, einen virtuellen Assistenten, der älteren Menschen hilft, aktiv zu bleiben: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. Darüber hinaus wird im Rahmen des Projekts EMPATTICS (EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems) untersucht und definiert, wie Gesundheitspersonal und Patienten IKT-Technologien, einschließlich KI- Systeme nutzen können, um Eingriffe bei Patienten zu planen und den Verlauf ihres körperlichen und geistigen Zustands zu überwachen: www.empattics.eu.

[67] Siehe zum Beispiel den MyHealth Avatar (www.myhealthavatar.eu), der den Gesundheitszustand eines Patienten digital darstellt. Das Forschungsprojekt brachte eine App und eine Online-Plattform heraus, die digitalen Langzeit-Gesundheitsdaten sammelt und zugänglich machen. Dies geschieht in Form eines lebenslangen Gesundheitsbegleiters („Avatar"). MyHealthAvatar prognostiziert auch das Risiko einer Person, an Schlaganfall, Diabetes, Herz-Kreislauf-Erkrankungen oder Bluthochdruck zu erkranken.

[68] Siehe beispielsweise die Verwendung von KI durch Sophia Genetics, die statistische Inferenz, Mustererkennung und maschinelles Lernen nutzt, um die Datenwerte aus der Genomik und Radionik zu optimieren: https://www.sophiagenetics.com/home.html.

[69] Siehe z. B. das Projekt ENRICHME (www.enrichme.eu), das sich mit dem fortschreitenden Abbau der kognitiven Leistungsfähigkeit der alternden Bevölkerung befasst. Eine integrierte Plattform für „Umgebungsunterstütztes Leben" (AAL) und ein mobiler Serviceroboter zur Langzeitbeobachtung und -interaktion sollen älteren Menschen dabei helfen, länger unabhängig und aktiv zu bleiben.

[70] Siehe beispielsweise das Projekt MaTHiSiS, das auf der Grundlage hochwertiger technologischer Geräte und Algorithmen Lösungen für affektbasiertes Lernen in einer komfortablen Lernumgebung zum Ziel hat: (http://mathisis-project.eu/). Siehe auch Watson Classroom von IBM oder die Plattform von Century Tech.

2. Beispiele für bedenkliche KI-Anwendungen

(129) Das Nichteinhalten einer der Anforderungen an die vertrauenswürdigen KI kann zu kritischen Situationen führen. Viele der nachfolgend aufgeführten Bedenken fallen bereits in den Geltungsbereich bestehender gesetzlicher Vorschriften, die verbindlich sind und daher eingehalten werden müssen. Doch selbst wenn die gesetzlichen Anforderungen nachweislich eingehalten wurden, bedeutet dies nicht, dass damit auch die gesamte Bandbreite auftretender ethischer Belange berücksichtigt wurde. Da sich unser Verständnis von der Angemessenheit von Regeln und ethischen Grundsätzen ständig weiterentwickelt und im Laufe der Zeit ändern kann, kann die folgende nicht erschöpfende Besorgnisliste in Zukunft verkürzt, erweitert, bearbeitet oder aktualisiert werden.

a. Identifizierung und Ortung von Personen mithilfe von KI

(130) KI ermöglicht eine immer effizientere Identifizierung einzelner Personen durch öffentliche und private Stellen. Bemerkenswerte Beispiele für eine skalierbare KI-Identifikationstechnologie sind die Gesichtserkennung und andere unfreiwillige Erkennungsmethoden anhand von biometrischen Daten (z. B. Lügendetektion, Persönlichkeitsbewertung durch Mikroexpressionen und automatische Spracherkennung). Die Personenerkennung ist manchmal wünschenswert und orientiert sich an ethischen Grundsätzen (z. B. bei der Aufdeckung von Betrug, Geldwäsche oder Terrorismusfinanzierung). Automatische Identifizierung wirft jedoch starke Bedenken sowohl rechtlicher als auch ethischer Art auf, da sie viele unerwartete psychologische und soziokulturelle Auswirkungen haben kann. Zur Wahrung der Autonomie der europäischen Bürgerinnen und Bürger ist ein verhältnismäßiger Einsatz von KI-Überwachungstechniken erforderlich. Eine eindeutige Definition, ob, wann und wie KI zur automatisierten Personenerkennung verwendet werden darf und die Unterscheidung zwischen der Erkennung einer Person gegenüber der Ortung einer Person und zwischen gezielter Überwachung und Massenüberwachung ist für die Schaffung einer vertrauenswürdigen KI in Zukunft entscheidend. Die Anwendung solcher Technologien muss im geltenden Recht eindeutig geregelt werden[71]. Wenn eine solche Anwendung rechtlich auf einer „Einwilligungserklärung" basiert, müssen praktische Mittel[72] entwickelt werden, die es den Betroffenen ermöglichen, eine sinnvolle und verifizierte Zustimmung zur automatischen Erkennung durch KI oder gleichwertige Technologien abzugeben. Dies gilt auch für die Verwendung „anonymer" personenbezogener Daten, die im Nachhinein wieder personalisiert werden können.

b. Verdeckte KI-Systeme

(131) Der Mensch sollte immer wissen, ob er mit einem anderen Menschen oder einer Maschine interagiert, und es liegt in der Verantwortung der KI-Akteure, dies in zuverlässiger Weise zu gewährleisten. Die KI-Akteure sollten daher sicherstellen, dass Menschen auf die Tatsache aufmerksam gemacht werden, dass sie mit einem KI-System interagieren (z. B. durch eindeutige und transparente Haftungsausschlüsse), oder es soll ihnen eine Möglichkeit zur Anfrage und Nachprüfung dieser Tatsache eingeräumt werden. Dabei ist zu beachten, dass es Grenzfälle gibt, die die Sache verkomplizieren (z. B. eine KI-gefilterte, menschliche Stimme). Man darf nicht vergessen, dass die Verwechslung zwischen Mensch und Maschine vielfältige Folgen wie z. B. Bindung, Einflussnahme oder eine verminderte Wertschätzung des Menschen haben kann.[73] Die Entwicklung von menschenähnlichen Robotern[74] sollte daher einer sorgfältigen ethischen Bewertung unterzogen werden.

c. KI-gestützte Bürgerinnen- und Bürgerbewertungen – eine Form der Grundrechtsverletzung

(132) Die Freiheit und Autonomie aller Bürgerinnen und Bürger müssen von der Gesellschaft geschützt werden. Jede Form der Bürgerbewertung kann zum Verlust dieser Autonomie führen und den Grundsatz der Nichtdiskriminierung gefährden. Das Scoring (dt. Bewertung nach einem Punktesystem) sollte nur dann

[71] In diesem Zusammenhang kann auf Artikel 6 der Datenschutz-Grundverordnung verwiesen werden, der unter anderem vorsieht, dass die Verarbeitung von Daten nur rechtmäßig ist, wenn sie auf einer gültigen Rechtsgrundlage beruht.

[72] Wie die aktuellen Vorkehrungen zur Abgabe einer Einwilligung nach Aufklärung im Internet zeigen, geben die Verbraucher in der Regel ihre Zustimmung ohne weitere Erwägungen. Daher können sie eher als unzweckmäßig eingestuft werden.

[73] Madary und Metzinger (2016). *Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. Frontiers in Robotics and AI*, 3(3).

[74] Dies gilt auch für KI-gesteuerte Avatare.

eingesetzt werden, wenn es eindeutig gerechtfertigt ist und die Maßnahmen verhältnismäßig und fair sind. Eine standarisierte Bürgerbewertung (allgemeine Beurteilung der „moralischen Persönlichkeit“ oder der „ethischen Integrität“) durch Behörden oder private Akteure in großem Umfang gefährdet *alle* Aspekte dieser Werte, insbesondere wenn sie nicht unter Achtung der Grundrechte erfolgt und überproportional und ohne einen konkreten und legitimen Zweck, der den Betroffenen mitgeteilt wird, eingesetzt wird.

(133) Bürgerbewertungen in großem oder kleinem Umfang werden schon heute häufig im Rahmen rein deskriptiver und bereichsspezifischer Bewertungen (z. B. Schulsysteme, E-Learning und Führerscheine) eingesetzt. Selbst für diese begrenzten Anwendungszwecke sollte den Bürgerinnen und Bürgern ein völlig transparentes Verfahren zur Verfügung gestellt werden, das Informationen über den Prozess, den Zweck und das methodische Vorgehen der Bewertung informieren sollte. Dabei ist zu bedenken, dass Transparenz weder Diskriminierung verhindern noch Fairness gewährleisten kann und nicht das Allheilmittel gegen das Scoring-Problem ist. Im Idealfall sollte eine Möglichkeit bestehen, dass sich die Betroffenen einem solchen Bewertungsverfahren entziehen können, ohne dass ihnen daraus Nachteile entstehen. Andernfalls müssen Verfahren zur Anfechtung und Berichtigung der Bewertungen bereitgestellt werden. Dies ist besonders wichtig, wenn ein Ungleichgewicht der Kräfteverhältnisse zwischen den Beteiligten besteht. Solche Opt-out-Möglichkeiten sollten bei der Gestaltung der Technologie immer dann gewährleistet sein, wenn andernfalls die Einhaltung der Grundrechte gefährdet wäre und dies für den Erhalt einer demokratischen Gesellschaft notwendig ist.

d. Tödliche autonome Waffensysteme

(134) Derzeit werden in einer unbekannten Anzahl von Ländern und Branchen tödliche autonome Waffensysteme – angefangen bei selbstständig zielsuchenden Raketen, bis hin zu lernfähigen Maschinen mit kognitiven Fähigkeiten – erforscht und entwickelt. Diese Waffensysteme sind in der Lage zu entscheiden, wer, wann und was ohne menschliches Eingreifen angegriffen werden soll. Das wirft grundlegende ethische Bedenken auf, z. B. die Tatsache, dass es zu einem historisch beispiellosen, unkontrollierbaren Wettrüsten führen könnte und militärische Umfelder geschaffen werden könnten, in denen die menschliche Kontrolle fast vollständig aufgegeben wird und die Risiken von Fehlfunktionen nicht berücksichtigt werden. Das Europäische Parlament hat die dringende Entwicklung eines gemeinsamen, rechtsverbindlichen Standpunkts zu ethischen und rechtlichen Fragen bezüglich menschlicher Kontrolle und Aufsicht, Rechenschaftspflicht, der Umsetzung der internationalen Gesetze für Menschenrechte und des humanitären Völkerrechts sowie militärischer Strategien gefordert.[75] Unter Hinweis auf das von der Europäischen Union angestrebte Ziel der Friedensförderung, das in Artikel 3 des Vertrags über die Europäische Union verankert ist, stehen wir hinter der Entschließung des Parlaments vom 12. September 2018 und werden jegliche Bemühungen im Zusammenhang mit autonomen Waffensystemen unterstützen.

e. Mögliche längerfristige Bedenken

(135) Die KI-Entwicklung ist nach wie vor bereichsspezifisch und erfordert gut ausgebildete Humanwissenschaftler und Ingenieure, die die Ziele genau festlegen. Ein forschender Blick in die spätere Zukunft wirft jedoch bestimmte, langfristige Bedenken auf[76]. Ein risikobasierter Ansatz legt nahe, dass diese Bedenken angesichts möglicher unvorhergesehener Ereignisse und „Schwarzer Schwäne“ berücksichtigt werden sollten.[77] Wegen der überaus starken Auswirkungen dieser Belange und der derzeitigen Unsicherheit bezüglich der entsprechenden Entwicklungen sollten diese Angelegenheiten regelmäßig bewertet werden.

D. FAZIT

[75] Entschließung des Europäischen Parlaments 2018/2752(RSP).

[76] Während einige der Ansicht sind, dass künstliche Intelligenz im Allgemeinen, künstliches Bewusstsein, künstliche moralische Wesen, Superintelligenz oder transformative KI Beispiele für solche (derzeit nicht vorhandenen) langfristigen Bedenken sein könnten, halten sie viele andere für unrealistisch.

[77] Ein „Schwarzer-Schwan-Ereignis“ ist ein sehr seltenes, aber hochwirksames Ereignis, das so selten eintritt, dass es wahrscheinlich nicht beobachtet worden ist. Daher kann die Eintrittswahrscheinlichkeit normalerweise nur mit hoher Unsicherheit vorhergesagt werden.

(136) Dieses Papier enthält die KI-Ethik-Leitlinien, die von der hochrangigen Expertengruppe für künstliche Intelligenz (HEG-KI) ausgearbeitet wurden.

(137) Wir sind uns der positiven gewerblichen und gesellschaftlichen Auswirkungen bewusst, die KI-Systeme bereits haben und auch weiterhin haben werden. Es ist uns jedoch auch ein Anliegen sicherzustellen, dass mit den Risiken und weiteren unerwünschten Auswirkungen, die mit diesen Technologien verbunden sind, bei der Anwendung der KI angemessen und verhältnismäßig umgegangen wird. Die KI ist eine transformierende, umwälzende Technologie, deren Entwicklung in den letzten Jahren dank der Verfügbarkeit enormer Mengen digitaler Daten, rasanter technologischer Fortschritte bei der Rechenleistung und Speicherkapazität der Computer sowie bedeutender wissenschaftlicher und technischer Innovationen bei den Methoden und Werkzeugen für KI angekurbelt wurde. KI-Systeme werden sich weiterhin auf eine Weise auf die Gesellschaft und die Bürgerinnen und Bürger auswirken, die wir uns noch nicht vorstellen können.

(138) In diesem Zusammenhang ist es wichtig, vertrauenswürdige KI-Systeme zu schaffen, denn nur wenn die Technologie – einschließlich der Prozesse und Menschen hinter der Technologie – vertrauenswürdig ist, kann der Mensch ihr vertrauen und die Vorteile voll ausschöpfen. Beim Entwurf dieser Leitlinien war unser grundlegendes Ziel daher eine vertrauenswürdige KI.

(139) Eine vertrauenswürdige KI hat drei Komponenten: 1) Sie sollte rechtmäßig sein und die Einhaltung aller geltenden Gesetze und Vorschriften sicherstellen. 2) Sie sollte ethisch sein und die Einhaltung ethischer Grundsätze und Werte sicherstellen. 3) Sie sollte sowohl aus technischer als auch aus sozialer Sicht robust sein, damit sichergestellt ist, dass KI-Systeme auch bei guten Absichten keinen unbeabsichtigten Schaden anrichten. Jede einzelne Komponente ist für die Schaffung einer vertrauenswürdigen KI notwendig, aber nicht ausreichend. Idealerweise wirken alle drei Komponenten harmonisch zusammen und überlappen sich in ihrer Funktionsweise. Wo Spannungen auftreten, sollte versucht werden, diese auszugleichen.

(140) In Kapitel I haben wir die Grundrechte und eine entsprechende Reihe ethischer Grundsätze formuliert, die für das Umfeld der KI entscheidend sind. In Kapitel II haben wir sieben Kernanforderungen aufgeführt, die KI-Systeme erfüllen sollten, um vertrauenswürdig zu sein. Wir haben technische und nicht-technische Methoden vorgeschlagen, die bei der Umsetzung hilfreich sein können. Schließlich haben wir in Kapitel III eine Bewertungsliste für eine vertrauenswürdige KI erstellt, die bei der Umsetzung der sieben Anforderungen unterstützend herangezogen werden sollte. In einem letzten Abschnitt haben wir Beispiele für vorteilhafte Einsatzmöglichkeiten und kritische Belange im Zusammenhang mit den KI-Systemen vorgestellt, mit denen wir zu weiteren Debatten anregen möchten.

(141) Europa hat eine einzigartige Perspektive, die darauf ausgerichtet ist, die Bürgerinnen und Bürger in den Mittelpunkt aller Bemühungen zu stellen. Dieser Schwerpunkt ist kraft der Gründungsverträge der Europäischen Union wesenhaft. Das vorliegende Dokument ist Teil einer Vision der Förderung einer vertrauenswürdigen KI, auf deren Grundlage Europa unserer Meinung nach seine Führungsrolle auf dem Gebiet innovativer, hochmoderner KI-Systeme ausbauen kann. Diese ehrgeizige Vision wird zum Wohlergehen der europäischen Bürgerinnen und Bürger sowohl in ihrer Eigenschaft als Einzelpersonen als auch als Kollektiv beitragen. Unser Ziel ist es, eine Kultur der „vertrauenswürdigen KI für Europa“ zu schaffen, in der die Vorteile der KI von allen in einer Weise genutzt werden kann, die die Achtung unserer grundlegenden Werte – Grundrechte, Demokratie und Rechtsstaatlichkeit – gewährleistet.

GLOSSAR

(142) Dieses Glossar ist Bestandteil dieser Leitlinien und dient dem Verständnis der in diesem Papier verwendeten Begriffe.

Künstliche Intelligenz oder KI-Systeme

(143) Künstliche-Intelligenz-(KI)-Systeme sind vom Menschen entwickelte Software- (und möglicherweise auch Hardware-) Systeme[78], die in Bezug auf ein komplexes Ziel auf physischer oder digitaler Ebene agieren, indem sie ihre Umgebung durch Datenerfassung wahrnehmen, die gesammelten strukturierten oder unstrukturierten Daten interpretieren, Schlussfolgerungen daraus ziehen oder die aus diesen Daten abgeleiteten Informationen verarbeiten und über die geeignete(n) Maßnahme(n) zur Erreichung des vorgegebenen Ziels entscheiden. KI-Systeme können entweder symbolische Regeln verwenden oder ein numerisches Modell erlernen, und sie können auch ihr Verhalten anpassen, indem sie analysieren, wie die Umgebung von ihren vorherigen Aktionen beeinflusst wird.

(144) Als wissenschaftliche Disziplin umfasst die KI mehrere Ansätze und Techniken wie z. B. maschinelles Lernen (Beispiele dafür sind „Deep Learning“ und bestärkendes Lernen), maschinelles Denken (es umfasst Planung, Terminierung, Wissensrepräsentation und Schlussfolgerung, Suche und Optimierung) und die Robotik (sie umfasst Steuerung, Wahrnehmung, Sensoren und Aktoren sowie die Einbeziehung aller anderen Techniken in cyber-physische Systeme).

(145) In einem von der HEG-KI separat erarbeiteten und parallel zu dem vorliegenden Dokument veröffentlichten Papier mit dem Titel „Eine Definition der KI: Wichtigste Fähigkeiten und Wissenschaftsgebiete“ wird die Definition des Begriffs *KI-Systeme* für die Zwecke dieser Leitlinien eingehender erläutert.

KI-Akteure

(146) Als KI-Akteure bezeichnen wir alle Personen oder Organisationen, die KI-Systeme entwickeln (einschließlich Forschung, Entwurf und Bereitstellung von Daten), bereitstellen (einschließlich Einführung) oder nutzen, mit Ausnahme derjenigen, die KI-Systeme als Endnutzer oder Verbraucher verwenden.

Lebenszyklus eines KI-Systems

(147) Der Lebenszyklus eines KI-Systems umfasst die Phasen der Entwicklung (einschließlich Forschung, Entwurf, Datenbereitstellung und eingeschränkte Erprobungen), Einführung (einschließlich Umsetzung) und Nutzung.

Nachprüfbarkeit

(148) Nachprüfbarkeit bezieht sich auf die Fähigkeit eines KI-Systems, einer Bewertung der Algorithmen, Daten und der Verfahren zum Entwurf des Systems unterzogen zu werden. Nachprüfbarkeit ist eine der sieben Anforderungen, die ein vertrauenswürdiges KI-System erfüllen sollte. Dies bedeutet nicht unbedingt, dass Informationen über Geschäftsmodelle und geistiges Eigentum im Zusammenhang mit dem KI-System offengelegt werden müssen. Mit der Sicherstellung von Verfahren zur Rückverfolgbarkeit und Berichterstattung bereits in der frühen Entwurfsphase des KI-Systems kann ein Beitrag zur Nachprüfbarkeit des Systems geleistet werden.

Verzerrung

(149) Eine Verzerrung ist eine Neigung zu Vorurteilen gegenüber oder gegen eine Person, ein Objekt oder eine Position. Verzerrungen können in KI-Systemen in vielerlei Hinsicht auftreten. So kann z. B. bei datengesteuerten KI-Systemen, die beispielsweise auf der Grundlage maschinellen Lernens geschaffen werden, eine Verzerrung bei der Datenerfassung und Ausbildung dazu führen, dass ein KI-System eine Verzerrung aufweist. Bei logikbasierter KI, wie z. B. bei regelbasierten Systemen, können infolge der Perspektive, aus der ein Wissensingenieur die Regeln, die für eine bestimmte Umgebung gelten sollen,

[78] Menschen entwerfen KI-Systeme direkt, sie können deren Entwurf aber auch mithilfe von KI-Techniken optimieren.

betrachtet, Verzerrungen auftreten. Verzerrungen können auch durch Online-Lernen und durch Anpassungen infolge von Interaktion entstehen. Sie können auch durch Personalisierung entstehen, wenn z. B. Benutzern Empfehlungen oder Informationsströme präsentiert werden, die auf die jeweiligen Benutzervorlieben zugeschnitten sind. Sie stehen nicht notwendigerweise im Zusammenhang mit Vorurteilen oder der Datenerhebung durch Menschen. Sie können beispielsweise dadurch entstehen, dass ein System in einem begrenzten Kontext verwendet wird und deshalb keine Möglichkeit zur Übertragung der Erkenntnisse auf andere Umgebungen hat. Verzerrungen können gut oder schlecht, absichtlich oder unabsichtlich sein. In bestimmten Fällen kann eine Verzerrung zu diskriminierenden und/oder unlauteren Ergebnissen führen, was in diesem Papier als unfaire Verzerrung bezeichnet wird.

Ethik

(150) Ethik ist eine wissenschaftliche Disziplin und ein Teilgebiet der Philosophie. Im Allgemeinen geht es um Fragen wie „Was ist eine gute Tat?“, „Welchen Wert hat das menschliche Leben?“, „Was ist Gerechtigkeit?“ oder „Was ist ein gutes Leben?“. In der wissenschaftlichen Ethik gibt es vier Hauptforschungsgebiete: i) Metaethik: Sie bezieht sich vor allem auf die Bedeutung und den Bezug normativer Sätze und die Frage, wie deren Wahrheitswerte (falls vorhanden) bestimmt werden können. ii) Normative Ethik: Sie beschäftigt sich mit praktischen Mitteln zur Bestimmung einer moralischen Handlungsweise durch Überprüfung der Normen für richtiges und falsches Handeln und Zuweisung eines Wertes zu bestimmten Handlungen. iii) Deskriptive Ethik: Gegenstand ist die empirische Untersuchung des moralischen Verhaltens und der Überzeugungen der Menschen. iv) Angewandte Ethik: Gegenstand ist das Handeln, zu dem die Menschen verpflichtet sind (oder das ihnen erlaubt ist) in einer bestimmten (oft historisch neuen) Situation oder einem bestimmten Kontext von (oft historisch ungekannten) Handlungsmöglichkeiten. Die angewandte Ethik beschäftigt sich mit realen Situationen, in denen Entscheidungen unter Zeitdruck und oftmals mit begrenzter Rationalität getroffen werden müssen. Die KI-Ethik wird im Allgemeinen als ein Beispiel für angewandte Ethik betrachtet. Sie konzentriert sich auf die normativen Fragen, die sich aus dem Entwurf, der Entwicklung, der Umsetzung und Verwendung von KI ergeben.

(151) In ethischen Debatten werden oft die Begriffe „moralisch“ und „ethisch“ verwendet. Der Begriff „Moral“ bezieht sich auf konkrete, faktische Verhaltensmuster, Bräuche und Sitten, die bei bestimmten Kulturen, Gruppen oder Einzelpersonen zu einem bestimmten Zeitpunkt beobachtet werden können. Der Begriff „ethisch“ bezieht sich auf eine Bewertung solcher konkreten Handlungen und Verhaltensweisen aus einer systematischen, wissenschaftlichen Perspektive.

Ethische KI

(152) In diesem Dokument wird ethische KI in Bezug auf die Entwicklung, Einführung und Verwendung von KI verwendet, die die Einhaltung der ethischen Normen – einschließlich der Grundrechte als besonderer moralischer Ansprüche – der ethischen Grundsätze und der damit verbundenen Grundwerte gewährleistet. Sie ist das zweite der drei Schlüsselelemente, die für die Erreichung einer vertrauenswürdigen KI erforderlich sind.

Menschenzentrierte KI

(153) Der menschenzentrierte Ansatz für die KI soll sicherstellen, dass menschliche Werte im Mittelpunkt der Entwicklung, Einführung, Nutzung und Überwachung der KI-Systeme stehen. Das soll durch die Achtung der Grundrechte gewährleistet werden, einschließlich der in den Verträgen der Europäischen Union und der Charta der Grundrechte der Europäischen Union verankerten Rechte, die durch Bezugnahme auf eine gemeinsame Grundlage miteinander verbunden sind, welche auf der Achtung der Menschenwürde beruht und dem Menschen einen einzigartigen und unveräußerlichen moralischen Status garantiert. Dazu zählt auch die Berücksichtigung der natürlichen Umwelt und anderer Lebewesen, die Teil des menschlichen Ökosystems sind, sowie ein nachhaltiger Ansatz, der das Gedeihen zukünftiger Generationen ermöglicht.

Red Teaming

(154) „*Red Teaming*“ ist ein Verfahren, bei dem ein „rotes Team“ bzw. eine unabhängige Gruppe eine Organisation herausfordert, ihre Effektivität zu verbessern, indem sie eine feindliche Rolle oder Sichtweise einnimmt. Dieses Verfahren wird insbesondere zur Erkennung und Behebung potenzieller Sicherheitsschwachstellen eingesetzt.

Wiederholbarkeit

(155) Wiederholbarkeit beschreibt, ob ein KI-Experiment das gleiche Verhalten aufweist, wenn es unter gleichen Bedingungen wiederholt wird.

Robuste KI

(156) Die Robustheit eines KI-Systems umfasst sowohl seine technische Robustheit (Angemessenheit für einen bestimmten Kontext, wie z. B. den Anwendungsbereich oder die Phase des Lebenszyklus) als auch seine soziale Robustheit (Gewährleistung einer angemessenen Berücksichtigung des Kontexts und der Umgebung, in denen das System eingesetzt wird). Das ist entscheidend, um sicherzustellen, dass auch bei guten Absichten keine unbeabsichtigten Schäden auftreten können. Robustheit ist das dritte der drei Schlüsselelemente, die für die Erreichung einer vertrauenswürdigen KI erforderlich sind.

Interessenträger

(157) Als Interessenträger bezeichnen wir all diejenigen, die KI entwickeln, entwerfen, einsetzen oder nutzen, sowie diejenigen, die (direkt oder indirekt) von KI betroffen sind – einschließlich, aber nicht beschränkt auf Unternehmen, Organisationen, Forscher, öffentliche Dienste, Institutionen, Organisationen der Zivilgesellschaft, Regierungen, Regulierungsbehörden, Sozialpartner, Einzelpersonen, Bürger, Arbeitnehmer und Verbraucher.

Rückverfolgbarkeit

(158) Die Rückverfolgbarkeit eines KI-Systems bezieht sich auf die Möglichkeit, die Daten-, Entwicklungs- und Bereitstellungsprozesse des Systems nachzuvollziehen, in der Regel anhand von dokumentierten Prozessaufzeichnungen.

Vertrauen

(159) Die folgende Definition ist der Literatur entnommen: „Unter Vertrauen wird Folgendes verstanden: 1) eine Reihe spezifischer Überzeugungen, die sich mit Wohlwollen, Kompetenz, Integrität und Berechenbarkeit befassen (vertrauensvolle Überzeugungen); 2) die Bereitschaft einer Partei, in einer riskanten Situation von einer anderen abhängig zu sein (vertrauensvolle Absicht) oder 3) die Kombination all dieser Elemente.“[79] Während „Vertrauen“ in der Regel keine Eigenschaft ist, die Maschinen zugeschrieben wird, wird in diesem Dokument betont, wie wichtig es ist, nicht nur darauf vertrauen zu können, dass KI-Systeme rechtskonform, ethisch einwandfrei und robust sind, sondern auch darauf, dass dieses Vertrauen auch für alle Personen und Prozesse gilt, die am Lebenszyklus des KI-Systems beteiligt sind.

Vertrauenswürdige KI

(160) Eine vertrauenswürdige KI hat drei Komponenten: 1) Sie sollte rechtmäßig sein und somit geltendes Recht und gesetzliche Bestimmungen einhalten, 2) sie sollte ethisch sein und somit die Einhaltung ethischer Grundprinzipien und Werte gewährleisten und 3) sie sollte robust sein, und zwar sowohl aus technischer als auch aus sozialer Sicht, da KI-Systeme unbeabsichtigten Schaden verursachen können, selbst wenn ihnen gute Absichten zugrunde liegen. Die Vertrauenswürdige KI bezieht sich nicht nur auf die Vertrauenswürdigkeit des KI-Systems selbst, sondern umfasst auch die Vertrauenswürdigkeit aller Prozesse und Akteure, die am Lebenszyklus des Systems beteiligt sind.

[79] Siau, K., Wang, W. (2018), *Building Trust in Artificial Intelligence, Machine Learning, and Robotics*, *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), S. 47–53.

Schutzbedürftige Personen und Gruppen

(161) Wegen ihrer heterogenen Zusammensetzung gibt es keine allgemein anerkannte rechtliche Definition für schutzbedürftige Personen. Was eine schutzbedürftige Person oder Gruppe ausmacht, ist oft kontextspezifisch. Kurzzeitige Lebensereignisse (z. B. Kindheit oder Krankheit), Marktfaktoren (z. B. Informationsasymmetrie oder Marktmacht), wirtschaftliche Faktoren (z. B. Armut), Faktoren, die mit der eigenen Identität verbunden sind (z. B. Geschlecht, Religion oder Kultur) und andere Faktoren können eine Rolle spielen. In Artikel 21 der EU-Grundrechtecharta über Nichtdiskriminierung sind folgende Gründe aufgeführt, die unter anderem einen Bezugspunkt darstellen können: Geschlecht, Rasse, Hautfarbe, ethnische oder soziale Herkunft, genetische Merkmale, Sprache, Religion oder Weltanschauung, politische oder sonstige Anschauung, Zugehörigkeit zu einer nationalen Minderheit, Vermögen, Geburt, Behinderung, Alter und sexuelle Orientierung. Zusätzlich zu den obigen Punkten befassen sich andere Rechtsvorschriften mit den Rechten bestimmter Gruppen. Solche Listen sind unvollständig und können im Laufe der Zeit Änderungen unterworfen sein. Eine schutzbedürftige Gruppe ist eine Gruppe von Personen, die ein oder mehrere Merkmale der Schutzbedürftigkeit aufweisen.

Dieses Dokument wurde von den Mitgliedern der hochrangigen Expertengruppe für KI (HEG-KI) erstellt,

die nachstehend in alphabetischer Reihenfolge aufgelistet sind:

Pekka Ala-Pietilä, Vorsitzender der HEG-KI
AI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – Mitberichterstatter
Zalando
Mária Bieliková
Slowakische Technische Universität Bratislava
Cecilia Bonefeld-Dahl – Mitberichterstatterin
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
IEEE Initiative Ethics of Intelligent/Autonomous Systems und Universität Sorbonne
Mark Coeckelbergh
Universität Wien
Virginia Dignum – Mitberichterstatterin
Universität Umea
Luciano Floridi
Universität Oxford
Jean-Francois Gagné – Berichterstatter
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Agentur der Europäischen Union für Grundrechte
Sami Haddadin
Munich School of Robotics and Machine Intelligence

Gry Hasselbalch
The thinkdotank DataEthics und Universität Kopenhagen
Fredrik Heintz
Universität Linköping
Fanny Hidvegi
Access Now
Eric Hilgendorf
Universität Würzburg
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
TU Wien
Robert Kroplewski
Rechtsanwalt und Berater der polnischen Regierung
Elisabeth Ling
RELX
Pierre Lucas
Orgalim – Europe's Technology Industries
Ieva Martinkenaite
Telenor
Thomas Metzinger – Mitberichterstatter
JGU Mainz und European University Association
Catelijne Muller
ALLAI Netherlands und EWSA
Markus Noga
SAP
Barry O'Sullivan, Stellvertretender Vorsitzender der HEG-KI
University College Cork
Ursula Pachl
BEUC
Nicolas Petit – Mitberichterstatter
Universität Lüttich
Christoph Peylo
Bosch

Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
College of Europe Faculty und CEPS
Francesca Rossi
IBM
Cristina San José
Europäischer Bankenverband
George Sharkov
Digital SME Alliance
Philipp Slusallek
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
Françoise Soulié Fogelman
KI-Beraterin
Saskia Steinacker – Mitberichterstatterin
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – Mitberichterstatterin
TU Delft
Thiébaut Weber
EGB
Cecile Wendling
AXA
Karen Yeung – Mitberichterstatterin
Universität Birmingham

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe und Karen Yeung waren Berichterstatterinnen und Berichterstatter für dieses Dokument.

Pekka Ala-Pietilä ist Vorsitzender der HEG-KI. Barry O'Sullivan ist stellvertretender Vorsitzender der HEG-KI und koordiniert den zweiten Bericht der HEG-KI. Nozha Boujemaa, stellvertretende Vorsitzende bis zum 1. Februar 2019, koordinierte den ersten Bericht und trug außerdem zu den Inhalten dieses Dokuments bei.
Nathalie Smuha war für die redaktionelle Unterstützung verantwortlich.

ISBN 978-92-76-11987-6 KK-02-19-841-DE-N