ETHIK-LEITLINIEN FÜR KI

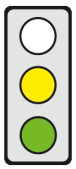
cep**Analyse** Nr. 16/2019

Centrum für Europäische Politik

KERNPUNKTE

Ziel der Mitteilung: Die Kommission stellt – unverbindliche – ethische Grundsätze und Kernanforderungen für künstliche Intelligenz (KI) vor, an die sich Entwickler und Nutzer von KI EU-weit halten sollen.

Betroffene: KI-Entwickler, Unternehmen und natürliche Personen, die KI nutzen oder von KI betroffen sind.



Pro: (1) Die Leitlinien sind ein sinnvoller erster Schritt hin zu einem EU-einheitlichen Konzept „ethisch vertretbarer" KI.

(2) Eine „Bewertungsliste für vertrauenswürdige KI" dient als Orientierungshilfe bei der Umsetzung der ethischen Anforderungen.

Contra: Konkrete Beispiele sollten ergänzt werden, um die Umsetzung der Leitlinien weiter zu erleichtern.

Die wichtigsten Passagen im Text sind durch einen Seitenstrich gekennzeichnet.

INHALT

Titel

Mitteilung COM(2019) 168 vom 8. April 2019: **Schaffung von Vertrauen in eine auf den Menschen ausgerichtete KI,** „**Ethik-Leitlinien für eine vertrauenswürdige KI**" der „Hochrangigen Expertengruppe für KI" vom 8. April 2019

Hinweis: Seitenangaben ohne weitere Angabe beziehen sich auf die Mitteilung COM(2019) 168, solche mit dem Verweis „EL" auf die „Ethik-Leitlinien für eine vertrauenswürdige KI", auf die die Mitteilung Bezug nimmt.

Kurzdarstellung

- **Allgemeiner Hintergrund**
  - Systeme der Künstliche Intelligenz (KI) sind von Menschen entwickelte Softwaresysteme, die [EL S. 47]
    - ihre Umgebung durch die Erfassung von Daten wahrnehmen,
    - die gesammelten Daten interpretieren,
    - Schlussfolgerungen daraus ziehen oder die aus den Daten abgeleiteten Informationen verarbeiten und
    - über die am besten geeigneten Maßnahmen zur Erreichung eines vorgegebenen komplexen Ziels entscheiden.
  - KI hat das Potenzial, unsere Welt zu verbessern, u.a. die Gesundheitsversorgung zu stärken, den Energieverbrauch zu senken und Autos sicherer zu machen; Landwirte können natürliche Ressourcen effizienter nutzen [S. 1].
  - Künstliche Intelligenz (KI) wirft ethische und rechtliche Fragen auf, u.a. weil Maschinen eigenständig „lernen" und ohne menschliches Zutun automatisierte Entscheidungen treffen können. Diese können von Cyber-Angreifern manipuliert werden oder falsch sein, z.B. aus unvollständigen oder verzerrten – d.h. nicht repräsentativen – Datensätzen resultieren [S. 3].

- **Ethik-Richtlinien für „vertrauenswürdige KI"**
  - KI muss den Menschen dienen und zum Ziel haben, das menschliche Wohlbefinden zu steigern. Sie muss „vertrauenswürdig" und „menschzentriert" sein, d.h. so entwickelt werden, dass der Mensch im Mittelpunkt steht" [S. 1ff.].
  - Deshalb hat eine von der Kommission eingesetzte „Unabhängige Hochrangige Expertengruppe für KI" [„Gruppe"] „Ethik-Leitlinien für eine vertrauenswürdige KI" erarbeitet [nachfolgend: „Leitlinien", S. 3, EL S. 4].
  - Die Leitlinien
    - sind unverbindlich und sollen innerhalb der EU „gleiche Wettbewerbsbedingungen" schaffen und die EU weltweit zu einem Vorreiter für „vertrauenswürdige KI" machen [S. 3, 10, EL S. 6],
    - richten sich an alle Beteiligten, unter anderem an Entwickler und Betreiber von KI – das sind private oder öffentliche Nutzer von KI, die Produkte und Dienstleistungen für andere anbieten –, aber auch an Endnutzer und alle anderen, die direkt oder indirekt von KI betroffen sind [EL S. 7, 17].
  - Nach den Leitlinien muss KI, um „vertrauenswürdig" zu sein, während ihres gesamten Lebenszyklus [S. 4, EL S. 6-12]
    - rechtmäßig sein, d.h. das geltende Recht einhalten und insbesondere die Grundrechte achten,
    - robust sein, insbesondere in technischer Hinsicht, z.B. resistent gegen Cyberangriffe sein, und
    - ethisch sein, d.h. ethischen Werten entsprechen, auch wo der technische Fortschritt das Recht überholt; obwohl unverbindlich, können diese Werte bei der Bestimmung helfen, was mit KI getan werden sollte – nicht, was damit getan werden kann – und wie KI Grundrechten und zugrunde liegenden Werten entsprechen kann.
  - Die Leitlinien befassen sich nicht mit „rechtmäßiger" KI, sie bieten nur Hilfestellung für die Schaffung „ethischer" und „robuster" KI. Dennoch sind die Grundrechte auch für „ethische KI" relevant, da sie auf Werten basieren und Individuen aufgrund ihrer moralischen Eigenschaft als Menschen zuteilwerden. [EL S. 8,12]

Autoren: Alessandro Gasparotti und Dr. Anja Hoffmann, LL.M. Eur. | hoffmann@cep.eu
cep | Kaiser-Joseph-Straße 266 | 79098 Freiburg | Telefon +49 (0)761 38693-107 | www.cep.eu

– Die Leitlinien [EL S. 11, 17ff.]
  - skizzieren die „Fundamente“ einer „vertrauenswürdigen KI“, die sich aus den Grundrechten ableiten und sich in vier „ethischen Grundsätzen“ widerspiegeln, und
  - legen auf der Basis dieser theoretischen Prinzipien sieben Kernanforderungen für „vertrauenswürdige KI“ fest.

► **Grundrechte und vier ethische Grundsätze als „Fundamente” „vertrauenswürdiger KI”**

– „Vertrauenswürdige KI“ muss auf den Grundrechten und -werten beruhen, die in den EU-Verträgen, der Charta der Grundrechte der EU und dem Völkerrecht [S. 2, EL S. 11ff.] verankert sind. Die Grundrechte umfassen unter anderem folgende Rechte, die für KI von besonderer Bedeutung sind:
  - Achtung der Menschenwürde, die auf der Vorstellung basiert, dass jeder Mensch einen „inhärenten Wert“ besitzt,
  - Freiheit des Einzelnen, z. B. Meinungs- und Versammlungsfreiheit, Recht auf Privatleben und Datenschutz, und
  - Gleichheit, Nichtdiskriminierung und Solidarität.
– KI muss vier ethische Grundsätze einhalten. Diese Grundsätze spiegeln die Grundrechte wider [EL S. 14-16]:
  - Achtung der menschlichen Autonomie: Menschen sollten die volle Selbstbestimmung über die eigene Person und Kontrolle über KI haben; KI sollte Menschen nicht auf ungerechtfertigte Weise täuschen oder manipulieren;
  - Schadensverhütung: KI sollte keinen Schaden verursachen oder sich sonst auf Menschen negativ auswirken;
  - Fairness: Vorteile und Kosten von KI sollten gesellschaftlich gerecht verteilt, Diskriminierung und „unfaire Verzerrungen“ vermieden werden und effektive Rechtsbehelfe gegen KI-basierte Entscheidungen verfügbar sein;
  - Erklärbarkeit: Der Zweck von KI-Systemen sollte kommuniziert sowie Prozesse transparent und Entscheidungen – je nach Inhalt und Tragweite der Konsequenzen einer Fehlentscheidung – in größtmöglichem Umfang erklärbar gemacht werden.
– Diese Grundsätze können kollidieren, z.B. kann Gesichtserkennungstechnologie Kriminalität reduzieren (Schaden verhüten) und zugleich Privatsphäre und Freiheitsrechte (menschliche Autonomie) einschränken [EL S. 16].
– Solche Konflikte sollten anerkannt und durch „vernünftige Reflexion“ gelöst werden, wobei die Menschenwürde nicht gegen andere Rechte abgewogen werden darf [EL S. 16].
– Die Leitlinien leiten aus den vier ethischen Grundsätzen eine nicht abschließende Liste von sieben Kernanforderungen ab, die „vertrauenswürdige KI“ erfüllen sollte [S. 4-7, EL S. 17-25]:

► **Sieben Kernanforderungen für die Verwirklichung „vertrauenswürdiger KI”**

– **Vorrang menschlichen Handelns und menschlicher Aufsicht:** Um die menschliche Autonomie und Entscheidungsfindung zu sichern, sollte KI menschliches Handeln unterstützen und menschliche Aufsicht ermöglichen [S. 5, EL S. 19]:
  - KI sollte den Menschen helfen, bessere, fundiertere Entscheidungen zu treffen, und so ihr Handeln unterstützen.
  - KI sollte eine menschliche Aufsicht ermöglichen, etwa durch Lenkungs- und Kontrollmechanismen, z.B. indem
    - menschliche Entscheidungsspielräume und die Möglichkeit vorgesehen werden, KI-basierte Entscheidungen außer Kraft zu setzen, und die Entscheidung ermöglicht wird, KI in bestimmten Situationen nicht zu verwenden.
  - KI-Aktivitäten sollten überwacht werden; je weniger Aufsicht möglich ist, desto intensiver muss KI getestet werden.
  - Aufsichtsbehörden müssen in der Lage sein, die KI-Nutzung im Einklang mit ihrem Mandat zu beaufsichtigen.
  - Negative Wirkungen von KI auf Grundrechte sollten vorab geprüft und reduziert oder gerechtfertigt werden.
– **Technische Robustheit und Sicherheit**: Um Schäden zu vermeiden, muss KI technisch robust und sicher sein, d.h. [S. 5f., EL S. 20f.]
  - ein Sicherheitsniveau aufweisen, das in einem angemessenen Verhältnis zur Höhe des jeweiligen Risikos steht,
  - zuverlässig, sicher und widerstandsfähig gegen Angriffe, z.B. Hacking und Manipulationen, sein,
  - gleiche, d.h. „reproduzierbare“ Ergebnisse unter gleichen Bedingungen erzeugen, so dass das Verhalten der KI beschrieben werden kann,
  - genau sein und die Nutzer über die Wahrscheinlichkeit von Fehlern (z.B. begrenzte Genauigkeit) informieren und
  - Sicherheitsvorkehrungen vorsehen, die bei Problemen einen Rückfallplan aktivieren, z.B. einen menschlichen Bediener einbeziehen.
– **Privatsphäre und Datenqualitätsmanagement**: Um Schäden für die Privatsphäre zu vermeiden, ist ein angemessenes Datenqualitätsmanagement erforderlich [S. 6, EL S. 21]:
  - Die Qualität, Integrität und Relevanz der in ein KI-System eingespeisten Daten muss gewährleistet sein, um Verzerrungen und Fehler zu vermeiden.
  - Menschen müssen die volle Kontrolle über ihre für oder durch KI erhobenen Daten haben, und diese Daten dürfen nicht rechtswidrig verwendet werden.
– **Transparenz**: Um „erklärbare“ KI zu fördern, müssen ihre Elemente – Daten, Systeme und Geschäftsmodelle – transparent sein. Insbesondere gilt [S. 6, EL S. 22]:
  - KI-Systeme müssen rückverfolgbar sein, z.B. durch Dokumentation ihrer Entscheidungen und des zugrunde liegenden Prozesses (einschließlich der Daten).
  - Algorithmische Entscheidungsprozesse müssen so weit wie möglich erklärbar sein. Die Erklärbarkeit sollte gegen eine mit ihr verbundene mögliche Einschränkung der Präzision abgewogen werden und unter der Voraussetzung stehen, dass die fragliche KI „das Leben von Menschen erheblich beeinflusst“.

- Nutzer müssen wissen, dass sie mit einem KI-System interagieren, über die Beschränkungen des Systems informiert sein und die Möglichkeit haben, sich für die Einschaltung eines Menschen zu entscheiden, wo dies „erforderlich" ist, um die „Einhaltung der Grundrechte" zu gewährleisten.

– **Vielfalt, Nichtdiskriminierung und Fairness**: „Faire" KI muss so gestaltet sein, dass alle den gleichen Zugang zur Nutzung des Produkts oder der Dienstleistung haben. Alle Betroffenen sollten in Designprozesse einbezogen und unfaire Verzerrungen in Datensätzen vermieden werden, da sie zu Diskriminierung führen können. [S. 7, EL S. 22f.]
– **Gesellschaftliches und ökologisches Wohlergehen**: Um fair zu sein und Schäden zu vermeiden, sollte KI nachhaltig, ökologisch und sozial verträglich sein; ihre Auswirkungen auf Umwelt, Mensch, Gesellschaft und Demokratie sollten berücksichtigt werden [S. 7].
– **Rechenschaftspflicht**: „Faire" KI sollte nachprüfbar sein, vor allem, wenn ihre Verwendung in Grundrechte eingreift, ohne zwingend z.B. als geistiges Eigentum geschützte Informationen offenlegen zu müssen. Negative Auswirkungen sollten gemeldet und minimiert und angemessene Rechtsbehelfe vorgesehen werden. [S. 7, EL S. 24]
– Die Anforderungen können umgesetzt werden durch [EL S. 25-29]:
  - „technische Methoden", z.B. durch Einführung von Verfahren, die KI befolgen muss oder nicht befolgen darf, und integrierte Ethik „by design", durch die die Einhaltung ethischer Normen von Anfang an sichergestellt wird;
  - „nicht-technische Methoden" wie Regulierung, Verhaltenskodizes, Standardisierung und Zertifizierung.

► **Bewertungsliste und Pilotphase**
– Um die Umsetzung der Leitlinien in der Praxis zu gewährleisten, hat die „Gruppe" auf Grundlage der Anforderungen eine „Bewertungsliste für vertrauenswürdige KI" erstellt, die in erster Linie KI-Entwickler und -Betreiber bei der Schaffung „vertrauenswürdiger KI" unterstützen soll [EL S. 32-41].
– In einer Pilotphase bis Dezember 2019 können betroffene Akteure die Liste testen und bewerten. Die Gruppe wird die Leitlinien Anfang 2020 aktualisieren. [S. 7, EL S. 24]

Politischer Kontext

Die Kommission veröffentlichte 2018 eine KI-Strategie [COM(2018) 237] und einen „Koordinierten Plan" [COM(2018) 795], die u.a. darauf abzielen, angemessene rechtliche und ethische Regeln für KI zu schaffen [vgl. cep**Analyse** 13/2019, s. auch cep**Analysen** 10/2019 und 12/2019]. Das Europäische Parlament schlug 2017 einen Verhaltenskodex vor [s. Entschließung] und forderte 2019 einen „Ethik-Leitrahmen" für eine „menschzentrierte KI" [s. Entschließung]. Der Rat betonte die Wichtigkeit von Ethikleitlinien für KI in der EU und auf globaler Ebene [s. Schlussfolgerungen 6331/19]. Im Mai 2019 veröffentlichte die OECD eigene Ethik-Leitlinien für KI, die von den G20 gebilligt wurden [s. cep**Input** Nr. 07/2019].

Politische Einflussmöglichkeiten

Generaldirektionen:	GD Kommunikationsnetze, Inhalte und Technologien
Ausschüsse des Europäischen Parlaments:	Industrie, Forschung und Energie (federführend)
Bundesministerien:	Bundesministerium des Innern, für Bau und Heimat, Justiz und Verbraucherschutz (federführend); Datenethikkommission der Bundesregierung
Ausschüsse des Deutschen Bundestags:	Bildung, Forschung und Technikfolgenabschätzung (federführend); Enquete-Kommission „Künstliche Intelligenz", Vorsitz: Daniela Kolbe (SPD)

BEWERTUNG

Ökonomische Folgenabschätzung

Das Ziel der Kommission, „vertrauenswürdige KI" zu fördern, kann die Akzeptanz dieser Technologie erleichtern. **Die Leitlinien sind** jedoch **zu allgemein und vage, als dass sie direkt angewandt werden könnten.** Sie erhöhen auch nicht die Rechtssicherheit, da sie Entwicklern nicht dabei helfen, die geltenden Gesetze einzuhalten. Es ist daher fraglich, ob die Leitlinien allein EU-weit gleiche Wettbewerbsbedingungen für KI schaffen und die EU damit weltweit zu einem Vorreiter für vertrauenswürdige KI machen werden. **Dennoch bieten die** sieben **Kernanforderungen einen umfassenden Rahmen für die Weiterentwicklung präziserer** – z.B. branchenspezifischer – **Leitlinien**.
Wenn bei einer KI ethische Grundsätze kollidieren, sollte zusätzlich die Öffentlichkeit darüber informiert werden, wie der Entwickler den ethischen Zielkonflikt gelöst hat. Solche Informationen schaden Unternehmen nicht und belasten sie kaum. Die Leitlinien sollten zudem Lösungsansätze oder zumindest Beispiele zur Lösung solche Zielkonflikte enthalten.
Im Folgenden werden die sieben Kernanforderungen einzeln bewertet:
(1) Es ist sachgerecht, dass KI menschliche Aufsicht und Entscheidungsspielräume ermöglichen muss, z.B. um KI-Entscheidungen außer Kraft zu setzen, die manipuliert wurden oder verzerrt sind. Es wird jedoch nicht näher konkretisiert, wann welcher Entscheidungsspielraum für menschliches Handeln angemessen wäre.
(2) Es ist sachgerecht, dass das Sicherheitsniveau einer KI in einem angemessenen Verhältnis zu dem von der KI ausgehenden Risiko stehen soll. So wären gleiche Standards für einen KI-gesteuerten Kernreaktor und eine KI, die Nutzern

Musik vorschlägt, unverhältnismäßig. Da die Leitlinien jedoch keine Standards vorgeben, haben Unternehmen einen Anreiz, mögliche Risiken ihrer KI herunterzuspielen, um nur selbst gesetzte Minimalstandards einhalten zu müssen.
(3) Die Qualität, Integrität und Relevanz der in ein KI-System eingespeisten Daten sind entscheidend, um das Risiko unzuverlässiger Ergebnisse zu verringern. Die Erfüllung dieser Anforderung wird i.d.R. durch wettbewerbliche Märkte sichergestellt. Dennoch sollten Beispiele als Orientierungshilfe für den Umgang mit Daten aufgenommen werden.
(4) Zur Förderung von Transparenz ist es unerlässlich, dass KI-Entscheidungen rückverfolgbar und soweit wie möglich erklärbar sind. Je besser Nutzer informiert sind, desto besser können sie beurteilen, inwieweit sie sich auf eine KI verlassen wollen. Insbesondere sollten Benutzer in der Lage sein, Informationen, die eine KI über sie produziert und für ihre Entscheidungen nutzt, einzusehen und zu korrigieren, etwa ihr Alter oder ihre Interessen.
(5) Um Nichtdiskriminierung zu gewährleisten, sollte präzisiert werden, wann eine Verzerrung „unfair" ist. Die Einbeziehung aller Betroffenen bereits in der Designphase von KI dürfte kaum realisierbar sein. Im Übrigen haben die Hersteller ein Eigeninteresse an einer umfassenden Einbeziehung, da Rückmeldungen die Qualität von KI verbessern.
(6) Es ist unklar, was „sozial verträglich" konkret bedeutet. Zwar besagt die „Bewertungsliste für vertrauenswürdige KI" z.B., dass KI dem Risiko von Arbeitsplatzverlusten entgegenwirken sollte, um sozial verträglich zu sein. Jedoch ist nicht klar, ob eine KI sozial verträglich ist, die fairere – z.B. nichtdiskriminierende – Vorstellungsgespräche führt als Menschen, wenn dadurch einige Angestellte ihren Arbeitsplatz verlieren.
(7) Um Rechenschaftspflicht zu gewährleisten, ohne Innovationsanreize zu hemmen, sollte bei der Prüfung einer KI-Entscheidung die Veröffentlichung geschützter Informationen so weit möglich vermieden werden. Auch die Schaffung angemessener Rechtsbehelfe ist sachgerecht, um die Sicherheit für Unternehmen und Verbraucher zu erhöhen.

Die „Bewertungsliste für vertrauenswürdige KI" dient als Orientierungshilfe bei der Umsetzung der ethischen Anforderungen. Konkrete Beispiele sollten jedoch **ergänzt werden, um die Umsetzung weiter zu erleichtern.** Zusätzlich zu dem sinnvollen Praxistest dieser Liste durch KI-Entwickler und -Betreiber sollte ein umfassender und transparenter öffentlicher Diskurs über KI aktiver gefördert werden, in den auch Ethikgremien, Universitäten, Parlamente und die Bevölkerung aktiv einzubeziehen sind. Dabei sollte – ggf. über die Leitlinien hinaus – auch der Schutz vor Zukunftsrisiken und die mögliche Festlegung ethischer Grenzen für den Einsatz von KI in der EU näher erörtert werden.

Juristische Bewertung

Kompetenz

Unproblematisch. Es handelt sich nicht um rechtliche, sondern um „ethische" Leitlinien, die zudem unverbindlich sind.

Subsidiarität

Unproblematisch (s.o.).

Verhältnismäßigkeit gegenüber den Mitgliedstaaten

Abhängig von der Ausgestaltung der Folgemaßnahmen.

Sonstige Vereinbarkeit mit EU-Recht

„Vertrauenswürdige" KI muss u.a. das geltende Recht und insbesondere die Grundrechte achten. Da die Leitlinien jedoch keine rechtlichen, sondern „ethische" Anforderungen an KI aufstellen, bestehen keine Zweifel an ihrer Vereinbarkeit mit EU-Recht. Vielmehr geht es um die – zwingend zu erörternde – Frage, ob und unter welchen Voraussetzungen der zunehmende Einsatz von KI ethisch vertretbar oder ggf. sogar geboten ist.

Die Leitlinien sind ein sinnvoller erster Schritt hin zu einem EU-einheitlichen Konzept „ethisch vertretbarer" KI. Ethische Sichtweisen, Wert- und Moralvorstellungen weichen in der EU trotz gemeinsamer Grundwerte voneinander ab, so dass eine Zersplitterung ethischer Anforderungen an KI droht. Die vier ausgewählten Grundsätze – menschliche Autonomie, Schadensverhütung, Fairness und Erklärbarkeit von KI – sind ein geeigneter Ausgangspunkt. Sie werden zu Recht aus den in der EU geltenden Grundrechten und -werten und insbesondere aus der Garantie der Menschenwürde hergeleitet, die oberster Grundwert ist [Art. 2 S. 1 EUV] und sich aus den gemeinsamen Verfassungsüberlieferungen der Mitgliedstaaten ergibt [Meyer-Borowski, Charta der Grundrechte der EU, Art. 1 Rn. 26]. Dadurch kann bei Auslegungsfragen an die grundrechtliche Judikatur angeknüpft und eine zu große Diskrepanz zwischen Ethik und Recht vermieden werden. Da der EU-Wertekanon den Menschen in den Mittelpunkt des EU-Aufbauwerks stellt [Calliess/Ruffert, Art. 2 EUV, Rn. 11], ist die Entscheidung der EU für „menschzentrierte" KI folgerichtig.

Auswirkungen auf das deutsche Recht

Abhängig von der Ausgestaltung der Folgemaßnahmen.

Zusammenfassung der Bewertung

Die Leitlinien sind zu allgemein und vage, als dass sie direkt angewandt werden könnten. Dennoch bieten die Kernanforderungen einen umfassenden Rahmen für die Weiterentwicklung präziserer Leitlinien. Die Leitlinien sind daher ein sinnvoller erster Schritt hin zu einem EU-einheitlichen Konzept „ethisch vertretbarer" KI. Die „Bewertungsliste für vertrauenswürdige KI" dient als Orientierungshilfe bei der Umsetzung der ethischen Anforderungen. Konkrete Beispiele sollten jedoch ergänzt werden, um die Umsetzung weiter zu erleichtern.