European Commission

HOCHRANGIGE EXPERTENGRUPPE

der Europäischen Kommission

FÜR KÜNSTLICHE INTELLIGENZ

ENTWURF

ETHIK-LEITLINIEN FÜR EINE VERTRAUENSWÜRDIGE KI

ZUSAMMENFASSUNG

Arbeitsunterlage für die Konsultation der Interessenträger

ENTWURF DER ETHIK-LEITLINIEN

FÜR EINE VERTRAUENSWÜRDIGE KI

Hochrangige Expertengruppe für künstliche Intelligenz
Entwurf der Ethik-Leitlinien für eine vertrauenswürdige KI

Europäische Kommission
Generaldirektion Kommunikation

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

Europäische Kommission
1049 Brüssel (Belgien)

Veröffentlichung des Dokumentes am 18. Dezember 2018 auf Englisch

Diese Arbeitsunterlage wurde von der Hochrangigen Expertengruppe für künstliche Intelligenz (HEG-KI) erstellt. Sie lässt die individuellen Positionen ihrer Mitglieder zu bestimmten Einzelpunkten unberührt und greift ihrer endgültigen Fassung nicht vor. An dieser Unterlage wird noch weiter gearbeitet. Im Anschluss an die Konsultation der Interessenträger im Rahmen der Europäischen KI-Allianz wird im März 2019 eine endgültige Fassung vorgelegt.

Weder die Europäische Kommission noch Personen, die im Namen der Kommission handeln, sind für die Verwendung der nachstehenden Informationen verantwortlich. Für den Inhalt dieser Arbeitsunterlage ist allein die Hochrangige Expertengruppe für künstliche Intelligenz (HEG-KI) verantwortlich. Auch wenn die Ausarbeitung dieser Leitlinien von den Dienststellen der Kommissionsdienststellen gefördert wurde, spiegeln die darin geäußerten Ansichten den Standpunkt der HEG-KI wider und stellen keinesfalls den offiziellen Standpunkt der Europäischen Kommission dar. Dieser Entwurf ist das erste Arbeitsergebnis der HEG-KI. Die endgültige Fassung wird der Kommission im März 2019 vorgelegt. Eine endgültige Fassung des zweiten Arbeitsergebnisses – nämlich der KI-Politik- und -Investitionsempfehlungen – wird Mitte 2019 vorgelegt.

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).

ZUSAMMENFASSUNG

Dieses Arbeitspapier ist ein Entwurf der Ethik-Leitlinien für die KI, der von der Hochrangigen Expertengruppe der Europäischen Kommission für künstliche Intelligenz (HEG-KI) erstellt wurde. Die endgültige Fassung der Leitlinien soll im März 2019 vorgelegt werden.

Die künstliche Intelligenz (KI) gehört zu den revolutionärsten Kräften unserer Zeit und dürfte das Gefüge unserer Gesellschaften verändern. Sie bietet großartige Chancen, Wohlstand und Wachstum zu steigern, die Europa nicht ungenutzt lassen darf. Dank der Verfügbarkeit riesiger Mengen an digitalen Daten, leistungsfähiger Rechnerarchitekturen und der Fortschritte in KI-Technologien wie dem maschinellen Lernen sind im Laufe des letzten Jahrzehnts beträchtliche Fortschritte erzielt worden. Unsere alltägliche Lebensqualität verbessert sich dank wichtiger KI-gestützter Entwicklungen auf Gebieten wie autonomes Fahren, Gesundheitswesen, Heim- und Dienstleistungsroboter, Bildung oder Cybersicherheit. Darüber hinaus ist die KI entscheidend für die Bewältigung vieler großer Herausforderungen, vor denen die Welt heute steht, ob es um Gesundheit und Wohlergehen in der Welt, um Klimawandel und verlässliche rechtliche und demokratische Systeme oder um andere Fragen geht, die in den Zielen der Vereinten Nationen für nachhaltige Entwicklung zum Ausdruck kommen.

Die KI kann gewaltige Vorteile für den Einzelnen und für die Gesellschaft bringen, sie birgt aber auch bestimmte Risiken, mit denen angemessen umzugehen ist. Ausgehend davon, dass die Vorteile der KI ihre Risiken insgesamt überwiegen, müssen wir sicher sein, dass wir einen Weg beschreiten, auf dem wir **den größtmöglichen Nutzen aus der KI erzielen und gleichzeitig die geringstmöglichen Risiken eingehen**. Um sicherzugehen, dass wir auf dem richtigen Weg bleiben, **brauchen wir einen auf den Menschen ausgerichteten („menschenzentrierten") Ansatz für die KI**, der uns stets daran erinnert, dass die Entwicklung und Nutzung der KI kein Selbstzweck ist, sondern dem Wohlergehen der Menschen dienen muss. Eine **vertrauenswürdige KI ist dabei unser Ziel**, denn die Menschen werden die Vorteile der KI nur dann zuversichtlich und umfassend nutzen können, wenn sie der Technik vertrauen können.

Vertrauenswürdige KI hat **zwei Komponenten**: 1) Sie muss die Grundrechte, das geltende Recht und die zentralen Grundsätze und Werte achten, d. h. einem „**ethischen Zweck**" dienen, und 2) sie muss **technisch robust** und zuverlässig sein, denn selbst mit dem besten Willen kann eine mangelnde technische Beherrschung zu unbeabsichtigtem Schaden führen.

In diesen Leitlinien wird daher ein **Rahmen für vertrauenswürdige KI** festgelegt:

- In **Kapitel I** geht es um die **ethische Zweckbestimmung der KI**. Hierzu werden die Grundrechte, Grundsätze und Werte festgelegt, denen sie entsprechen soll.
- Aus diesen Grundsätzen werden in **Kapitel II** sodann **Leitlinien für die Verwirklichung** einer vertrauenswürdigen KI abgeleitet, die sowohl einem ethischen Zweck dient als auch technische Robustheit garantiert. Dazu werden die Anforderungen an eine vertrauenswürdige KI aufgelistet, und es wird ein Überblick über die technischen und nichttechnischen Methoden gegeben, die zu ihrer Umsetzung angewandt werden können.
- In **Kapitel III** wird die Einhaltung der Anforderungen auf ein **operatives Fundament** gestellt, indem eine konkrete, aber nicht erschöpfende Prüfliste zur Bewertung vertrauenswürdiger KI vorgestellt wird. Diese Liste wird anschließend an bestimmte Anwendungsfälle angepasst.

Im Gegensatz zu anderen Papieren, die sich mit ethischer KI befassen, soll in diesen Leitlinien daher keine weitere Liste zentraler Werte und Grundsätze für die KI aufgestellt werden, sondern es soll vielmehr eine Richtschnur für die konkrete Umsetzung und praktische Anwendung dieser Werte und Grundsätze in KI-Systemen gegeben werden. Diese Richtschnur hat drei Abstraktionsstufen – von der abstraktesten in Kapitel I (Grundrechte, Grundsätze und Werte) – bis zur konkretesten in Kapitel III (Bewertungsliste).

Die Leitlinien richten sich an alle **einschlägigen Beteiligten, die sich mit der Entwicklung, Einführung oder Nutzung von KI befassen**, also an Unternehmen, Organisationen, Forscherinnen und Forscher, öffentliche Dienste, Institutionen, Einzelpersonen oder andere Stellen. Die endgültige Fassung dieser

Leitlinien wird einen Mechanismus vorsehen, dem sich alle Beteiligten dann freiwillig werden anschließen können.

Es sei ausdrücklich darauf hingewiesen, dass diese Leitlinien keineswegs als Ersatz für politische Entscheidungen oder gesetzliche Regelungen gedacht sind (darum wird es im zweiten Arbeitsergebnis der HEG-KI gehen, nämlich in den im Mai 2019 fälligen KI-Politik- und -Investitionsempfehlungen) und auch nicht darauf abzielen, solche Entscheidungen oder Regelungen zu verhindern. Darüber hinaus sollten die Leitlinien als dynamisches Arbeitspapier betrachtet werden, das im Laufe der Zeit regelmäßig zu überarbeiten sein wird, damit es mit der Entwicklung der Technik und unseres diesbezüglichen Wissens Schritt halten kann. Diese Leitlinien sollen daher einen Ausgangspunkt für die Diskussion über „**vertrauenswürdige KI *made in Europe***“ bilden.

Seinen ethischen Ansatz für die KI kann Europa zwar nur dann zur Geltung bringen, wenn es auch weltweit wettbewerbsfähig ist, ein **ethisches Herangehen an die KI ist aber der Schlüssel zu einer verantwortlichen Wettbewerbsfähigkeit**, denn es schafft Vertrauen bei den Nutzern und erleichtert eine breite Einführung und Nutzung der KI. Diese Leitlinien sollen nicht etwa die Innovation im Bereich der KI in Europa im Keim ersticken, sondern stattdessen die Ethik als Inspirationsquelle für die Entwicklung einer einzigartigen KI-Ausprägung verwenden, die sowohl dem Schutz des Einzelnen als auch dem Gemeinwohl dient. Auf diese Weise wird Europa in der Lage sein, sich als Vorreiter einer hochmodernen, sicheren und ethischen KI zu positionieren. Nur wenn die Vertrauenswürdigkeit gesichert ist, werden die europäischen Bürgerinnen und Bürger von den Vorteilen der KI uneingeschränkt profitieren können.

Über Europa hinaus sollen diese Leitlinien schließlich auch die **Reflexion und Diskussion** über einen ethischen Rahmen für die KI **auf weltweiter Ebene** fördern.

ÜBERSICHT ÜBER DIE LEITLINIEN

Jedes Kapitel der Leitlinien enthält Hinweise zur Verwirklichung einer vertrauenswürdigen KI und richtet sich an alle einschlägigen Beteiligten, die sich mit der Entwicklung, Einführung oder Nutzung von KI befassen. Die Leitlinien lassen sich wie folgt zusammenfassen:

Kapitel I: Richtschnur für die Gewährleistung des ethischen Zwecks:

- Gewährleistung, dass die KI **auf den Menschen ausgerichtet** (menschenzentriert) ist: KI sollte mit einem „**ethischen Zweck**“ entwickelt, eingeführt und genutzt werden, der auf Grundrechten, gesellschaftlichen Werten und den folgenden ethischen Grundsätzen beruht und diese widerspiegelt: *Benefizienz* (Gutes tun), *Schadensverhütung* (keinen Schaden zufügen), *Achtung der menschlichen Autonomie*, *Gerechtigkeit* und *Erklärbarkeit*. Diese bilden die Voraussetzung für die Arbeit an einer **vertrauenswürdigen KI**.
- Ausgehend von Grundrechten, ethischen Grundsätzen und Werten erfolgt eine vorausschauende Abschätzung möglicher Folgen der KI für die Menschen und das Gemeinwohl. Das **besondere Augenmerk** liegt dabei auf Situationen, in denen **schutzbedürftige Gruppen** wie Kinder, Menschen mit Behinderungen oder Minderheiten betroffen sind, oder auf Situationen mit **ungleicher Macht- oder Informationsverteilung**, etwa zwischen Arbeitgebern und Arbeitnehmern oder Unternehmen und Verbrauchern.
- Bewusstmachung und Berücksichtigung der Tatsache, dass die künstliche Intelligenz zwar erhebliche Vorteile für den Einzelnen und die Gesellschaft bringt, aber auch negative Auswirkungen haben kann. Es gilt, in wichtigen Problembereichen wachsam zu bleiben.

Kapitel II: Richtschnur für die Verwirklichung einer vertrauenswürdigen KI:

- Einbeziehung der **Anforderungen an eine vertrauenswürdige KI ab der frühesten Entwurfsphase**: Rechenschaftspflicht, Datenqualitätsmanagement, Entwurf für alle, Steuerung der KI-Autonomie (menschliche Aufsicht), Nichtdiskriminierung, Achtung der menschlichen Autonomie, Wahrung der

Privatsphäre, Robustheit, Sicherheit, Transparenz.

- Berücksichtigung technischer und nichttechnischer Methoden, um die Umsetzung dieser Anforderungen im KI-System sicherzustellen. Auch bei der Zusammenstellung des Teams für die Arbeit an dem System, im System selbst, dem Testumfeld und den möglichen Anwendungen des Systems dürfen diese Anforderungen nicht außer Acht gelassen werden.
- Klare und proaktive **Information der Beteiligten** (Kunden, Mitarbeiter usw.) über die Möglichkeiten und Grenzen des KI-Systems, damit sie realistische Erwartungen haben können. In dieser Hinsicht kommt es entscheidend darauf an, dass die **Rückverfolgbarkeit** des KI-Systems gewährleistet ist.
- Einbeziehung einer vertrauenswürdigen KI als **Teil der Unternehmens- oder Organisationskultur** und Information der Beteiligten und Betroffenen darüber, wie die vertrauenswürdige KI in die Gestaltung und Nutzung von KI-Systemen eingebunden ist. Die vertrauenswürdige KI kann auch in die berufsethischen Regeln oder Verhaltenskodizes der Organisationen aufgenommen werden.
- Gewährleistung der Teilnahme und **Einbeziehung der Interessenträger** in Entwurf und Entwicklung des KI-Systems. Außerdem sollte bei der Aufstellung der Teams, die das Produkt entwickeln, implementieren und testen sollen, auf **Diversität** geachtet werden.
- Bemühung um eine **leichtere Überprüfbarkeit** von KI-Systemen, insbesondere in kritischen Zusammenhängen oder Situationen. Soweit möglich, sollte das System so konzipiert werden, dass individuelle Entscheidungen anhand der verschiedenen Eingaben, Daten und vortrainierten Modelle zurückverfolgt werden können. Außerdem sollten **Erklärungsmethoden** für das KI-System geschaffen werden.
- Gewährleistung eines besonderen Prozesses für die **Gewährleistung der Rechenschaftspflicht**.
- Einplanung von **Schulung und Ausbildung**, damit Manager, Entwickler, Nutzer und Arbeitgeber sich der vertrauenswürdigen KI bewusst werden und lernen, damit umzugehen.
- Berücksichtigung der Tatsache, dass grundlegende Konflikte zwischen verschiedenen Zielen auftreten können (Transparenz kann dem Missbrauch Tür und Tor öffnen, die Erkennung und Berichtigung von Verzerrungen kann gegen den Datenschutz verstoßen). Diese Abwägungen müssen kommuniziert und dokumentiert werden.
- Förderung der Forschung und Innovation, um die Erfüllung der Anforderungen an vertrauenswürdige KI weiter zu verbessern.

<u>Kapitel III: Richtschnur für die Bewertung vertrauenswürdiger KI:</u>

- Aufstellung einer **Bewertungsliste** für vertrauenswürdige KI in Bezug auf die Entwicklung, Einführung oder Nutzung der KI und deren Anpassung an den konkreten Anwendungsfall, in dem das System eingesetzt wird.
- Es gilt zu bedenken, dass eine Bewertungsliste **niemals erschöpfend** ist und dass es bei vertrauenswürdiger KI nicht um das Abhaken von Kästchen auf einer Liste geht, sondern um einen kontinuierlichen Prozess der Ermittlung von Anforderungen, der Bewertung von Lösungen und der Erzielung besserer Ergebnisse über den gesamten Lebenszyklus des KI-Systems.

Diese Leitlinien sind Teil einer Vision, die ein auf den Menschen ausgerichtetes Herangehen an die künstliche Intelligenz aufgreift und Europa die Chance eröffnet, zu einem weltweit führenden Innovator auf dem Gebiet der ethischen, sicheren und hochmodernen KI aufzusteigen. Ziel ist es, eine „**vertrauenswürdige KI *made in Europe***“ zum Wohle der europäischen Bürgerinnen und Bürger zu ermöglichen und zu fördern.