Fraunhofer IAIS

FRAUNHOFER-INSTITUT FÜR INTELLIGENTE ANALYSE- UND INFORMATIONSSYSTEME IAIS

VERTRAUENSWÜRDIGER EINSATZ VON KÜNSTLICHER INTELLIGENZ

HANDLUNGSFELDER AUS PHILOSOPHISCHER, ETHISCHER, RECHTLICHER UND TECHNOLOGISCHER SICHT ALS GRUNDLAGE FÜR EINE ZERTIFIZIERUNG VON KÜNSTLICHER INTELLIGENZ

In Kooperation mit

UNIVERSITÄT BONN

Universität zu Köln

Powered by

KI NRW

Gefördert durch

Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen

Fraunhofer IAIS

FRAUNHOFER-INSTITUT FÜR INTELLIGENTE ANALYSE- UND INFORMATIONSSYSTEME IAIS

VERTRAUENSWÜRDIGER EINSATZ VON KÜNSTLICHER INTELLIGENZ

HANDLUNGSFELDER AUS PHILOSOPHISCHER, ETHISCHER, RECHTLICHER UND TECHNOLOGISCHER SICHT ALS GRUNDLAGE FÜR EINE ZERTIFIZIERUNG VON KÜNSTLICHER INTELLIGENZ

Autorinnen und Autoren

Prof. Dr. Armin B. Cremers | Fraunhofer IAIS
Dr. Alex Englander | Universität Bonn
Prof. Dr. Markus Gabriel | Universität Bonn
Dr. Dirk Hecker | Fraunhofer IAIS
PD Dr. Michael Mock | Fraunhofer IAIS
Dr. Maximilian Poretschkin (Projektleitung) | Fraunhofer IAIS
Julia Rosenzweig | Fraunhofer IAIS
Prof. Dr. Dr. Frauke Rostalski (Projektleitung) | Universität zu Köln
Joachim Sicking | Fraunhofer IAIS
Dr. Julia Volmer | Fraunhofer IAIS
Jan Voosholz (Projektleitung) | Universität Bonn
Dr. Angelika Voss | Fraunhofer IAIS
Prof. Dr. Stefan Wrobel | Fraunhofer IAIS

In cooperation with

UNIVERSITÄT BONN

Universität zu Köln

Powered by

KI NRW

Sponsored by

Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen

INHALT

GRUSSWORT

Liebe Leserinnen und Leser,

wir leben im Zeitalter der Digitalisierung. Daten sind der Treibstoff und Künstliche Intelligenz (KI) ist der Motor, um die Ressource Daten zum Nutzen für Wirtschaft und Gesellschaft einzusetzen. Alle Studien sind sich einig, dass Künstliche Intelligenz in signifikantem Umfang das weltweite Wirtschaftswachstum erhöhen wird. Darüber hinaus hat sie das Potenzial, einen Beitrag zur Bewältigung der großen gesellschaftlichen Herausforderungen wie Klimawandel, Mobilität und Gesundheit zu leisten. Dieses Potenzial gilt es zu erschließen. Wichtig ist dabei, den Menschen in den Mittelpunkt zu stellen. Für dieses große Zukunftsthema brauchen wir einen gesellschaftlichen Dialog über das Verhältnis von Mensch und Maschine sowie ein verlässliches ethisch-rechtliches Fundament.

Nordrhein-Westfalen hat bereits heute eine führende Position in der Entwicklung und Anwendung von Künstlicher Intelligenz. Wir verfügen im Land über viele erstklassige Hochschulen und außeruniversitäre Forschungseinrichtungen, an denen KI-Spitzenforschung mit internationaler Sichtbarkeit durchgeführt wird. In der Wirtschaft haben einige große Unternehmen aus NRW heraus umfassende Kompetenzen im Bereich der Künstlichen Intelligenz aufgebaut und den Weg der digitalen Transformation beschritten. Und nicht zuletzt sind die Rahmenbedingungen für Start-ups in NRW sehr günstig, und es haben sich bereits viele neue, auf Künstlicher Intelligenz basierende Geschäftsmodelle etabliert. Mit der von uns initiierten Kompetenzplattform KI.NRW vernetzen wir die Akteure im Bereich der Künstlichen Intelligenz und stärken den Technologietransfer von der Forschung in die Praxis sowie die berufliche Qualifizierung. Künstliche Intelligenz erhöht aber nur dann den Wohlstand und die Lebensqualität der Menschen, wenn Werten wie Selbstbestimmung, Diskriminierungsfreiheit, Datenschutz und Sicherheit Rechnung getragen wird.

Mit der von uns ins Leben gerufenen Zertifizierung von Künstlicher Intelligenz wollen wir aus Nordrhein-Westfalen heraus die Qualitätsmarke »KI Made in Germany« weiter etablieren, indem sie zuverlässige, sichere Technologien erkennbar macht und nachhaltig schützt. Die Zertifizierung fördert den freien Wettbewerb unterschiedlicher Anbieter und leistet einen Beitrag zur Akzeptanz von Künstlicher Intelligenz in der Gesellschaft.

Die Zertifizierung wird federführend von hochrangigen Expertinnen und Experten aus den Bereichen Maschinelles Lernen, Rechtswissenschaften, Philosophie, Ethik und IT-Sicherheit entwickelt. In einem offen gestalteten Prozess mit einer breiten Beteiligung von Akteuren aus Wirtschaft, Forschung und Gesellschaft sollen die Grundprinzipien für eine technisch zuverlässige und ethisch verantwortungsvolle Künstliche Intelligenz entwickelt werden. Wir freuen uns sehr, dass wir diese Initiative mit Strahlkraft über Deutschland hinaus aus NRW heraus initiieren und fördern dürfen.

Die vorliegende Veröffentlichung bildet die Grundlage für die Entwicklung der KI-Zertifizierung. Sie erläutert die Handlungsfelder, entlang derer der vertrauensvolle Einsatz von Künstlicher Intelligenz erfolgen muss. Gleichzeitig möchte sie auch zum gesellschaftlichen Diskurs zu dieser Zukunftstechnologie anregen, die wir in Nordrhein-Westfalen gemeinsam im Dialog mit Ihnen gestalten wollen.

Mit herzlichen Grüßen,

Prof. Dr. Andreas Pinkwart
Minister für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen

VORWORT

Liebe Leserinnen und Leser,

Künstliche Intelligenz (KI) verändert Gesellschaft, Wirtschaft und unseren Alltag in grundlegender Weise und eröffnet große Chancen für unser Zusammenleben. Sie hilft zum Beispiel Ärzten, Röntgenbilder besser und oftmals auch exakter auszuwerten, beantwortet in Form von Chatbots Fragen zu Versicherungspolicen und anderen Produkten und wird in absehbarer Zeit Autos immer selbstständiger fahren lassen. Gleichzeitig wird stets deutlicher, dass eine sorgfältige Gestaltung solcher Anwendungen notwendig ist, damit wir die Chancen der Künstlichen Intelligenz im Einklang mit unseren gesellschaftlichen Werten und Vorstellungen nutzen können.

Künstliche Intelligenz hat das Potenzial, die Fähigkeiten des Menschen zu erweitern und hilft uns, neue Erkenntnisse zu gewinnen. Entscheidungen von ihren durch Maschinelles Lernen automatisierten oder teilautomatisierten Ergebnissen abhängig zu machen, stellt uns aber auch vor grundlegend neue Herausforderungen. Neben Fragen der technischen Eignung sind dies grundsätzliche philosophisch-ethische Erwägungen, aber auch rechtliche Fragestellungen. So wirft die Möglichkeit »autonom« reagierender intelligenter Maschinen ein neues Licht auf die individuelle Haftung und Verantwortung von Personen und damit auf Grund und Kriterien von »Zurechnung«. Um sicherzustellen, dass der Mensch stets im Mittelpunkt dieser Entwicklung steht, ist daher ein enger Austausch über Künstliche Intelligenz zwischen Informatik, Philosophie und Rechtswissenschaften notwendig.

Angesichts des schnellen Vordringens von Künstlicher Intelligenz in nahezu jedweden gesellschaftlichen Bereich, haben wir uns zum Ziel gesetzt, im interdisziplinären Austausch eine Zertifizierung von Künstlicher Intelligenz zu entwickeln. Die vorliegende Publikation bildet den Auftakt hierzu und diskutiert aktuelle Herausforderungen der Künstlichen Intelligenz aus der Sicht von Informatik, Philosophie und Rechtswissenschaften. Aufbauend auf diesem interdisziplinären Austausch formuliert sie KI-spezifische Handlungsfelder für den vertrauenswürdigen Einsatz von Künstlicher Intelligenz.

Faires Verhalten der KI-Anwendung gegenüber allen Beteiligten, die Anpassung an die Bedürfnisse der Nutzerinnen und Nutzer, eine verständliche, verlässliche und sichere Funktionsweise, sowie der Schutz sensibler Daten sind zentrale Kriterien, die beim vertrauenswürdigen Einsatz einer KI-Anwendung zu erfüllen sind.

Die hier vorgestellten Handlungsfelder bilden die Grundlage für einen KI-Prüfkatalog, den wir aktuell parallel erarbeiten und mithilfe dessen neutrale Prüfer KI-Anwendungen auf ihren vertrauenswürdigen Einsatz hin überprüfen können. Als wichtiger Partner für die Erarbeitung dieses Prüfkatalogs ist zudem das Bundesamt für Sicherheit in der Informationstechnik (BSI) mit seiner langen Erfahrung in der Entwicklung von sicheren IT-Standards beteiligt. Mit der Zertifizierung wollen wir wesentlich dazu beitragen, Qualitätsstandards für eine Künstliche Intelligenz »Made in Europe« zu setzen, den verantwortungsvollen Umgang mit der Technologie zu sichern und einen fairen Wettbewerb verschiedener Anbieter zu befördern.

Das vorliegende Whitepaper soll zum gesellschaftlichen Diskurs über den Einsatz von Künstlicher Intelligenz beitragen. Denn es ist an uns allen mitzuentscheiden, wie die Welt aussehen soll, in der wir morgen leben.

In diesem Sinne wünschen wir eine spannende und erkenntnisreiche Lektüre.

Prof. Dr. Markus Gabriel
Professor für Philosophie
Universität Bonn

Prof. Dr. Dr. Frauke Rostalski
Professorin für Rechtswissenschaft
Universität Köln

Prof. Dr. Stefan Wrobel
Institutsleiter Fraunhofer IAIS &
Professor für Informatik
Universität Bonn

VORBEMERKUNGEN

Executive Summary

Die vorliegende Publikation dient als Grundlage für die interdisziplinäre Entwicklung einer Zertifizierung von Künstlicher Intelligenz. Angesichts der rasanten Entwicklung von Künstlicher Intelligenz mit disruptiven und nachhaltigen Folgen für Wirtschaft, Gesellschaft und Alltagsleben verdeutlicht sie, dass sich die hieraus ergebenden Herausforderungen nur im interdisziplinären Dialog von Informatik, Rechtswissenschaften, Philosophie und Ethik bewältigen lassen. Als Ergebnis dieses interdisziplinären Austauschs definiert sie zudem sechs KI-spezifische Handlungsfelder für den vertrauensvollen Einsatz von Künstlicher Intelligenz: Sie umfassen Fairness, Transparenz, Autonomie und Kontrolle, Datenschutz sowie Sicherheit und Verlässlichkeit und adressieren dabei ethische und rechtliche Anforderungen. Letztere werden mit dem Ziel der Operationalisierbarkeit weiter konkretisiert.

Aufbau des Whitepapers

Die interdisziplinäre Betrachtung des Themas spiegelt sich in der Kapitelstruktur dieses Whitepapers wider. Kapitel 1 gibt eine Einleitung in die Thematik und motiviert die Notwendigkeit einer Zertifizierung von Künstlicher Intelligenz. In Abschnitt 2.1 wird ein grundlegendes Verständnis der Funktionsweise, Möglichkeiten und Beschränkungen der zugrundeliegenden Technik entwickelt. Die philosophisch-ethische Sichtweise auf das Problem, insbesondere die Rolle der ethischen Konzepte von Autonomie, Freiheit und Selbstbestimmung des Menschen, wird in Abschnitt 2.2 beleuchtet. Die Grundlagen der sich daraus ergebenden rechtlichen Anforderungen werden in Abschnitt 2.3 besprochen, mit besonderem Fokus auf Verantwortlichkeit, Nachvollziehbarkeit und Haftung für KI-Anwendungen. Abschnitt 2.4 stellt die Auswirkungen der unterschiedlichen interdisziplinären Perspektiven, insbesondere im Hinblick auf die Gestaltung konkreter KI-Anwendungen dar. In Kapitel 3 werden dann die konkreten fundamentalen Handlungsfelder motiviert und in eigenen Abschnitten erläutert, von Autonomie und Kontrolle in Abschnitt 3.1 über Fairness, Transparenz, Verlässlichkeit und Sicherheit bis zu Datenschutz im Abschnitt 3.6. In Kapitel 4 geben wir schließlich einen Ausblick auf die geplanten weiteren Schritte bei der Entwicklung einer Zertifizierung.

Kontext

Das vorliegende Whitepaper ist das erste Ergebnis eines interdisziplinären Projekts der Kompetenzplattform KI.NRW mit dem Ziel, eine Zertifizierung für KI-Anwendungen zu entwickeln, die neben der Absicherung der technischen Zuverlässigkeit auch einen verantwortungsvollen Umgang aus ethisch-rechtlicher Perspektive prüft.

1 EINLEITUNG

Jede Zeit hält ihre Herausforderungen bereit. Wir leben im Zeitalter der Digitalisierung. Die neuen Technologien verändern unser Miteinander gravierend. Sie durchdringen nahezu jedweden gesellschaftlichen Bereich – sei es die Arbeitswelt, den Straßenverkehr, den Gesundheitssektor oder schlicht die Art und Weise, wie wir Menschen miteinander kommunizieren. Auch wenn sich vieles davon im Stillen oder als schleichender Prozess vollzieht, ist die Geschwindigkeit im Vergleich zu früheren gesellschaftlichen Veränderungen beispiellos und hätte unsere Vorfahren zu Zeiten der industriellen Revolution im 18. und 19. Jahrhundert in Angst und Schrecken versetzt.

Eine zentrale Antriebsfeder der Digitalisierung ist die rasante Entwicklung der Künstlichen Intelligenz (KI), die ausgelöst wurde durch Durchbrüche in sogenannten tiefen künstlichen neuronalen Netzen auf hochleistungsfähigen Rechnern. In Spezialgebieten wie der Bilderkennung oder komplexen Strategiespielen können KI-Anwendungen sogar die besten menschlichen Experten schlagen. Künstliche Intelligenz eröffnet große Chancen für neue technische Anwendungen, digitale Geschäftsmodelle und praktische Erleichterungen im Alltagsleben. Ihre Anwendungen verbreiten sich unaufhaltsam in vielfältigen Bereichen. Automatisierte Übersetzungshilfen, Sprachassistenten in den Wohnungen oder selbstfahrende Autos sind nur einige bekannte Beispiele. Künstliche Intelligenz besitzt ein disruptives Potenzial: Die wissenschaftlichen und wirtschaftlichen Anwendungsmöglichkeiten sind derart weitreichend, dass derzeit kaum abzusehen ist, wie unsere Erkenntnis- und Handlungsweisen durch Künstliche Intelligenz verändert werden. Außerdem werden Problemkontexte entstehen, auf die wir mit unseren traditionellen rechtlichen, politischen, ethischen und sozialen Mitteln nicht ausreichend reagieren können. Die KI-Forschung verbessert die Generalisierbarkeit und Übertragbarkeit von Anwendungen auf neue Kontexte und verdrängt so sukzessive ältere Technologien. Herkömmliche Wertschöpfungsketten werden disruptiv verändert. Die gesteigerte Produktivität geht gleichzeitig einher mit einer Entlastung der Menschen, die in bestimmten Bereichen weniger monotone oder schwere Arbeiten verrichten müssen.

Allgemein wird erwartet, dass die Zahl der KI-Anwendungen in den nächsten Jahren exponentiell wachsen wird. McKinsey prognostiziert bis 2030 global bis zu 13 Billionen Dollar zusätzliche Wertsteigerung durch Künstliche Intelligenz[1]. Weiterhin wird prognostiziert, dass Künstliche Intelligenz 1,2 Prozentpunkte zum jährlichen Wachstum des globalen Bruttoinlandprodukts beiträgt. Somit sind die Auswirkungen mindestens vergleichbar mit dem Produktivitätswachstum der vorangegangenen industriellen Revolutionen, wie der Dampfmaschine (0,3 Prozentpunkte), den Industrierobotern (0,4 Prozentpunkte) oder der Verbreitung der Informationstechnologie (0,6 Prozentpunkte). Dieses beeindruckende Wachstum beruht auf immer mehr verfügbaren und verknüpfbaren Daten, einer höheren Vernetzung und einer immer besseren Rechenleistung, die einen größeren Grad an Automatisierung und Individualisierung von Produkten und Dienstleistungen erlauben. Die Individualisierung ist hierbei umso erfolgreicher, je mehr Informationen über Nutzer[2] und Kunden bekannt sind.

Dabei liegt auf der Hand, dass der Einsatz von KI-Anwendungen in kurzer Zeit Auswirkungen auf das gesamte gesellschaftliche Miteinander haben wird. Dies wird am Beispiel der Überwachungssysteme besonders augenfällig. So wurde Personenerkennung in Pilotprojekten auch in Deutschland bereits erprobt, wie zum Beispiel am Bahnhof Berlin Südkreuz, wobei die Ergebnisse unter anderem allerdings als zu fehlerhaft bewertet wurden. Dies zeigt zum einen, dass die Frage nach der Zuverlässigkeit bei KI-Anwendungen im Gegensatz zu herkömmlicher Software neue Herausforderungen stellt, zum anderen ist es jedoch aus technischer Sicht nur noch eine Frage der Zeit und des

1 Notes from the frontier modelling the impact of AI on the world economy, Discussion Paper, McKinsey Global Institute, September 2018, www.mckinsey.com/mgi

2 Im Interesse einer besseren Lesbarkeit wird nicht ausdrücklich in geschlechtsspezifischen Personenbezeichnungen differenziert. Die gewählte männliche Form schließt eine adäquate weibliche Form gleichberechtigt ein.

Aufwands, bis eine hinreichende Zuverlässigkeit – zumindest für den hier betrachteten Fall von Personenerkennungen auf Überwachungssystemen – erreicht werden kann. Prinzipiell ließe sich auch eine KI-basierte Intentionserkennung mit der Personenerkennung kombinieren, so dass es eventuell sogar möglich wäre, gezielt Alarm zu schlagen, wenn sich Personen im öffentlichen Raum auffällig verhielten. Es stellt sich unmittelbar die Frage, inwieweit eine solche Überwachung – selbst bei optimaler Funktionsweise – nach geltendem Recht zulässig wäre oder ob bzw. wie das Recht hierfür geändert werden müsste. Auf diese Weise entstehen neue ethische Fragestellungen, da gesellschaftlicher Klärungsbedarf besteht, welche KI-Anwendungen wir überhaupt zulassen sollten. Recht und Ethik müssen in diesen neuen Handlungssituationen kooperieren.

Das Szenario verdeutlicht, dass sich das prognostizierte wirtschaftliche Wachstum auf Dauer nur dann realisieren lässt, wenn ein ausreichendes Vertrauen in die KI-Technologie vorliegt. Um Vertrauen herzustellen, muss eine KI-Anwendung überprüfbar so konstruiert werden, dass sie sicher und zuverlässig funktioniert sowie ethischen und rechtlichen Rahmenbedingungen genügt. Dazu muss neben der technischen Absicherung auch geklärt werden, unter welchen Voraussetzungen der Einsatz ethisch vertretbar ist und welche Anforderungen sich insbesondere aus rechtlicher Sicht ergeben. Die damit verbundenen Herausforderungen berühren Grundsatzfragen, die sich nur in einem interdisziplinären Austausch zwischen Informatik, Philosophie und Rechtswissenschaften angehen lassen. Da Künstliche Intelligenz in nahezu alle gesellschaftlichen Sphären vordringt, sind davon rechtlich schutzwürdige Interessen einer Vielzahl an Akteuren betroffen. Eventuell müssen hier rechtliche Rahmenbedingungen konkretisiert oder neu geschaffen werden. Umgekehrt muss jedoch vermieden werden, dass eine Überregulierung innovationshemmend wirkt oder aufgrund der Dynamik des technologischen Fortschritts zu schnell veraltet und somit gar nicht anwendbar ist. Denn die Ethik steht nicht ein- für allemal fest, weshalb es angesichts gesellschaftlicher und technologischer Umbrüche stets die Möglichkeit ethischen Fort- und Rückschritts gibt.

Entwicklung einer Zertifizierung von KI-Anwendungen

Da KI-Anwendungen oft auf besonders großen Datenmengen und dem Einsatz hochkomplexer Modelle beruhen, ist es für Anwender in der Praxis schwierig zu überprüfen, inwiefern die zugesicherten Eigenschaften erfüllt werden. Eine Zertifizierung von KI-Anwendungen, die auf einer sachkundigen und neutralen Prüfung beruht, kann hier Vertrauen und Akzeptanz schaffen – sowohl bei Unternehmen als auch bei Nutzern und gesellschaftlichen Akteuren.

Angesichts der dargestellten Herausforderungen beim Einsatz von Künstlicher Intelligenz hat sich die Kompetenzplattform KI.NRW das Ziel gesetzt, eine durch akkreditierte Prüfer operativ durchführbare Zertifizierung für KI-Anwendungen zu entwickeln, die neben der Absicherung der technischen Zuverlässigkeit auch einen verantwortungsvollen Umgang aus ethisch-rechtlicher Perspektive prüft. Das Zertifikat soll einen Qualitätsstandard bescheinigen, der es den Anbietern erlaubt, KI-Anwendungen überprüfbar rechtskonform und ethisch akzeptabel zu gestalten und der es zudem ermöglicht, KI-Anwendungen unterschiedlicher Anbieter zu vergleichen und so den freien Wettbewerb in der Künstlichen Intelligenz zu fördern.

Neben der Forderung, dass eine KI-Anwendung ethischen und rechtlichen Grundsätzen entsprechen muss, wurden im interdisziplinären Dialog sechs KI-spezifische Handlungsfelder identifiziert und so zugeschnitten, dass sie möglichst von jeweils verschiedenen Fachexperten geprüft werden können. Dabei werden die Anforderungen dieser Handlungsfelder aus bestehenden ethischen, philosophischen und rechtlichen Grundsätzen (wie zum Beispiel dem allgemeinen Gleichbehandlungsgrundsatz) abgeleitet. Sie umfassen die Bereiche Fairness, Transparenz, Autonomie und Kontrolle, Datenschutz sowie Sicherheit und Verlässlichkeit. Während die Sicherheit die üblichen Aspekte der Betriebssicherheit umfasst, betrifft die Verlässlichkeit die besonderen Prüfungsherausforderungen von komplexen KI-Modellen wie tiefen neuronalen Netzen.

Die Frage, wie KI-Anwendungen verantwortlich und zuverlässig eingesetzt werden können, ist bereits seit einiger Zeit Gegenstand intensiver gesellschaftlicher und wissenschaftlicher Diskussionen im internationalen Raum. Auf europäischer Ebene hat die EU-Kommission eine sogenannte HLEG (High-Level Expert Group) für Künstliche Intelligenz ins Leben gerufen. Sie hat im April 2019 Empfehlungen dafür formuliert, welche Aspekte bei der Entwicklung und Anwendung von Künstlicher Intelligenz zu berücksichtigen sind[3]. Das vorliegende Whitepaper nimmt diese Empfehlungen auf, differenziert sie aus und geht an einigen Stellen über sie hinaus. Dies ist deswegen geboten, da die Empfehlungen der HLEG vorrangig allgemeiner Natur sind und bisher weder rechtliche Aspekte – insbesondere nicht die Spezifika der jeweiligen nationalen Rechtssysteme –, noch operationalisierbare ethische Vorgaben mit dem klaren Ziel der Zertifizierung in den Blick nehmen. Insoweit geht die vorliegende Publikation im Vergleich mit den Vorschlägen der HLEG sowohl in die Breite als auch in die Tiefe: Es nimmt neben der philosophischen Ethik das Recht in den Blick und setzt beide zueinander in Beziehung. Um den Anforderungen der Operationalisierbarkeit zu genügen, sind die auf diese Weise erarbeiteten Handlungsfelder an vielen Stellen außerdem spezifischer und tiefer ausgeführt als die Kategorien der HLEG.

3 https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

2 INFORMATISCHE, PHILOSOPHISCHE UND RECHTLICHE PERSPEKTIVEN

2.1 Ausgangslage und Fragen aus der Informatik

1956 entstand Künstliche Intelligenz als Teilgebiet der Informatik mit dem Ziel, intelligentes Verhalten zu automatisieren. Inspiriert durch Kybernetik, Kognitions- und Neurowissenschaften wurden vielfältige Techniken entwickelt. Dazu zählen intelligente Agenten, die über Sensoren und Aktuatoren mit der Umgebung oder auch untereinander interagieren, die Kombination von Logiksystemen mit Heuristiken, Methoden zur symbolischen Wissensrepräsentation und -auswertung sowie das Maschinelle Lernen (ML) mit statistischen Verfahren und Optimierung, die gerade in der jüngeren Entwicklung einen enormen Zuwachs verzeichnet haben. Bald nach Entstehen der Disziplin hat man sich auch schon kritisch mit einem verantwortungsvollen Umgang auseinandergesetzt[4].

Viele KI-Techniken beruhen auf der Anwendung von Modellen, die Wissen und Erfahrung über spezifische Aufgaben enthalten. Beim Maschinellen Lernen erzeugen Lernalgorithmen das Modell aus vielen Beispielen, den sogenannten Trainingsdaten. Zu jeder Art von Modell gibt es Berechnungs- oder »Inferenzverfahren«, die für eine Eingabe eine Ausgabe erzeugen. Damit kann das Modell anschließend auf neue, potenziell unbekannte Daten derselben Art angewendet werden. Immer wenn Prozesse zu kompliziert sind, um sie analytisch zu beschreiben, aber genügend viele Beispieldaten – etwa Sensordaten, Bilder oder Texte – verfügbar sind, bietet sich Maschinelles Lernen an. Mit den gelernten Modellen können Vorhersagen getroffen oder Empfehlungen und Entscheidungen generiert werden – ganz ohne im Vorhinein festgelegte Regeln oder Berechnungsvorschriften.

Eine wichtige und große Klasse von ML-Modellen bilden die tiefen neuronalen Netze. Sie bestehen aus einer Vielzahl durch Software realisierter sogenannter künstlicher Neuronen, die durch gewichtete Verbindungen miteinander verknüpft sind. Ein solches Netz enthält Tausende bis Millionen offener Parameter, die auf die Trainingsdaten optimiert werden.

Aufbau einer KI-Anwendung

Die Funktion einer KI-Anwendung[5] wird wesentlich bestimmt durch die antrainierten ML-Modelle mit ihren Berechnungsverfahren und möglicherweise weiteren Vor- und Nachverarbeitungsprozeduren. Dieser Kern einer jeden KI-Anwendung wird im Folgenden als »KI-Komponente« bezeichnet. Die KI-Komponente ist immer in weitere Software-Module der KI-Anwendung eingebettet. Die Module aktivieren die KI-Komponente und verarbeiten ihre Ergebnisse weiter. Sie realisieren letztendlich das nach außen sichtbare »Verhalten« der KI-Anwendung und sind für die Interaktion mit dem Nutzer zuständig. Insbesondere obliegt es ihnen, ein Versagen der KI-Komponente festzustellen und abzufangen sowie auf Störungen und Notfälle zu reagieren. Eine KI-Anwendung kann selbstständig sein oder Teil eines Systems. Zum Beispiel kann die Fußgängererkennung als eine KI-Anwendung in einem autonomen Auto, in einer Drohne oder in einem Geländeüberwachungssystem integriert sein. Bei der Diskussion und Beurteilung einer KI-Anwendung besteht ein erster wichtiger Schritt darin, die Grenzen der KI-Anwendung im Gesamtsystem zu definieren und die KI-Komponente in der KI-Anwendung abzugrenzen.

4 Joseph Weizenbaum. Die Macht der Computer und die Ohnmacht der Vernunft. Suhrkamp Verlag, 1. Aufl. (1977).
Armin B. Cremers et al. (Hsg. i.A. des Vereins Deutscher Ingenieure). Künstliche Intelligenz. Leitvorstellungen und Verantwortbarkeit. VDI-Report 17, 189 S. 2. Aufl. (1993), VDI-Report 21, 121 S. (1994).

5 Die folgende Diskussion setzt den Fokus auf das Maschinelle Lernen als Schlüsseltechnologie zur Realisierung von Künstlicher Intelligenz.

Herausforderungen beim Einsatz von ML-Modellen

Abhängigkeit von den Trainingsdaten

Eine konkrete Anfrage an eine KI-Komponente liefert die Eingabe für das ML-Modell, aus dessen Berechnungsergebnis die KI-Komponente eine Antwort oder Reaktion erzeugt. Da KI-Anwendungen ihr »Verhalten« aus der Verallgemeinerung von Beispieldaten »lernen«, hängt die Qualität der KI-Anwendung erheblich von der Güte und den Eigenschaften der verwendeten Datenbestände ab. Wenn die Trainingsdaten statistisch nicht repräsentativ für die Daten sind, die im Betrieb anfallen, kann es passieren, dass die Ergebnisse in eine Richtung »verzerrt« werden. Darum muss im Betrieb regelmäßig kontrolliert werden, wie gut die Datenverteilungen zueinander passen und ob sie auseinanderlaufen.

Probabilistischer Charakter

Aufgrund der statistischen Natur des Modells sowie qualitativen Unsicherheiten von Eingabe- und Lerndaten sind die Ergebnisse approximativ und mit mehr oder weniger Unsicherheit behaftet. Oft gibt es faktisch auch gar keine eindeutig richtige oder falsche Antwort. Die KI-Komponente könnte die besten Alternativen ausgeben, zusammen mit einer Unsicherheitsangabe. Bei der Interpretation solcher Ergebnisse muss der Mensch innerhalb seines Ermessensspielraums entscheiden. Bei einer vollautomatisierten Anwendung müssen entsprechende Vorkehrungen im umgebenden Gesamtsystem getroffen werden, zu dem wiederum der Mensch wesentlich hinzugehört.

Verständlichkeit und Transparenz des ML-Modells und seiner Ergebnisse

Viele ML-Modelle sind sogenannte »Black Boxes«. Unter einer Black Box versteht man hierbei Systeme, bei denen nur das äußere Verhalten betrachtet werden kann, die inneren Funktionsmechanismen aufgrund von Komplexität oder fehlendem Wissen aber unzugänglich sind. Es ist somit oftmals nicht möglich, das Zustandekommen von Antworten nachzuvollziehen. Bei manchen Anwendungen kann es deshalb angezeigt sein, auf bestimmte Arten von ML-Modellen zu verzichten. Man kann das ML-Modell aber auch um ein weiteres, das sogenannte »Erklärmodell« ergänzen, das ermittelt, welche Teile der Eingabe ausschlaggebend für ein bestimmtes Ergebnis waren. So hat man zum Beispiel herausgefunden, dass eine KI-Anwendung in einer Bilddatenbank Pferde an einem Wasserzeichen, also einem Artefakt in den Bildern, erkannt hat, statt etwa an der Form der Tiere.

Testen von ML-Modellen

Klassische Softwaretestmethoden scheitern, weil sich die Modelle nicht in immer kleinere, separat prüfbare Einheiten zerlegen lassen. Es ist im Allgemeinen nicht einmal möglich, eine Formel zu finden, um zulässige Eingaben zu charakterisieren. Dies wurde eindrücklich demonstriert, als eine automatische Verkehrszeichenerkennung durch unauffällige Aufkleber auf den Schildern völlig in die Irre geführt wurde. An Stelle des modularen Testens tritt hier das quantitative Testen des Modells anhand von separaten Testdaten, die dieselbe statistische Verteilung wie Trainingsdaten und Einsatzdaten haben sollten.

Selbstlernen im Betrieb

Prinzipiell können ML-Modelle im laufenden Betrieb automatisch weiterlernen, zum Beispiel, indem das Nutzerfeedback mit herangezogen wird. Das ML-Modell unterliegt dann einer kontinuierlichen Veränderung. Ein bekanntes Beispiel ist der Chatbot Tay von Microsoft, der innerhalb eines Tages von seinen Nutzern viele rassistische Äußerungen lernte und daraufhin aus dem Verkehr gezogen wurde. Da es ausgesprochen schwierig ist, »Leitplanken« zu konstruieren, innerhalb derer eine KI-Komponente weiterlernen darf, stellt der kontrollierte Einsatz von solchen KI-Anwendungen zurzeit noch eine ungelöste Herausforderung dar. Die derzeit beste Absicherung in diesem Fall ist die engmaschige Überwachung der KI-Anwendung durch Menschen.

2.2 Ausgangslage und Fragen aus der Philosophie

An die Philosophie, besonders ihre Subdisziplin Ethik, wird aktuell der Wunsch herangetragen, eine Ethik der Künstlichen Intelligenz zu liefern, um dem disruptiven Potenzial dieser Technologie zu begegnen. Unter »Ethik der Künstlichen Intelligenz« versteht man einen allgemeingültigen Anspruch daran, wie die Anwendungskontexte (der Einsatzbereich inklusive der Mensch-Maschine-Interaktion), die eingesetzten Techniken und die Schnittstellen der Anwendungskontexte zum Rest der sozialen und digitalen Sphäre gestaltet sein müssen. Das Ziel ist dabei, dass alle Beteiligten nach ihren jeweiligen moralischen Überzeugungen gut handeln bzw. sich gut verhalten können und niemand in Rechten, Autonomie oder Freiheit unzulässig beschnitten wird. Die Zertifizierung von KI-Anwendungen in ihren konkreten Anwendungskontexten ist ein erster wichtiger Schritt zu einer allgemeinen Ethik der KI.

Dabei müssen zwei Missverständnisse vermieden werden: Erstens bezieht sich Ethik der Künstlichen Intelligenz hier auf konkrete KI-Anwendungen für spezifische Aufgaben. Damit sind Fragen ausgeschlossen wie: Welche moralischen Pflichten und welche Verantwortung haben wir gegenüber intelligenten Maschinen? Sollten wir vor diesem Hintergrund überhaupt versuchen, Künstliche Intelligenz mit genereller Intelligenz zu bauen? Wann kann eine KI-Anwendung als moralischer Agent zählen und besitzt sie Freiheit und Rechte? Diese Fragen berühren nicht die Zertifizierung konkreter KI-Anwendungen, um die es derzeit faktisch geht.

Zweitens lässt sich eine Ethik der Künstlichen Intelligenz nicht als Code umsetzen, der bei jeder gegebenen Frage aus einem konkreten Problemkontext heraus eine binäre Ja-Nein-Antwort produziert. Die Frage, »welches moralische System ist so programmierbar oder modellierbar, dass damit KI-Anwendungen künftig ausgestattet werden sollen?«, ist verfehlt, insofern weder Ethik abschließend programmierbar, da prinzipiell offen für Veränderungen ist, noch ohne Schwierigkeiten Einigkeit über das korrekte moralische System herstellbar ist. Denn Ethik verdankt sich historisch variablen Erfahrungen des Menschen. Gesellschaftliche Transformationen wie die Digitalisierung werfen bisher unbekannte ethische Probleme auf, sodass man durch Erforschung der konkreten Mensch-Maschine-Interaktion zu allererst neue Richtlinien erarbeiten muss, die mit dem universalen Wertesystem der humanen Lebensform (den Menschenrechten als Rahmenbedingungen von Recht und Ethik) vereinbar sind.

Der zentrale Beitrag von Philosophie und Ethik bei der Entwicklung von Standards für Künstliche Intelligenz sind demnach neu festzulegende Leitlinien für den Umgang mit unseren derzeit existenten KI-Techniken. Diese Leitlinien müssen im Einklang mit fundamentalen ethischen Grundprinzipien wie der Menschenwürde, der Autonomie und der individuellen sowie der demokratischen Freiheit stehen. Sie geben den Rahmen vor, in dem sich KI-Anwendungen in ihrem Anwendungskontext bewegen müssen, um nicht ethischen Grundprinzipien wie Fairness oder Transparenz zu widersprechen. Dazu muss man sowohl die KI-Anwendung selbst als auch ihre Schnittstelle zur sozialen Sphäre in den Blick nehmen, was nur gelingt, wenn man den Menschen als KI-Anwender ins Zentrum rückt.

2.3 Ausgangslage und Fragen aus dem Recht

Für das Recht gehen mit den Techniken der Künstlichen Intelligenz eine Vielzahl an Herausforderungen einher, mit denen wir uns als Gesellschaft befassen müssen. Hierzu zählt die Frage, inwieweit Maschinelles Lernen ein neues Licht auf die individuelle Haftung bzw. Verantwortung von Personen und damit Grund und Kriterien von »Zurechnung« wirft. Systeme, die durch maschinell gelernte Modelle gesteuert werden, können Fehler aufweisen, die sich insbesondere in Form von Vorurteilen negativ auf den Einzelnen auswirken können. Hinzu tritt die Schwierigkeit, dass Transparenz im Hinblick auf lernende Systeme oftmals nur eingeschränkt möglich ist. Ob und ggfs. in welchem Umfang entsprechende Techniken in sensiblen gesellschaftlichen Bereichen verwendet werden sollten, bedarf daher einer Klärung.

Zu denken ist dabei beispielsweise an den Gesundheitssektor. Hier bietet Künstliche Intelligenz etwa in der Krebsdiagnose oder in Gestalt von sogenannten »Gesundheits-Apps« eine Unterstützung der Arbeit von Ärzten. In der Pflege sollen künftig immer mehr Roboter nicht zuletzt zur Ersetzung des menschlichen Personals herangezogen werden. Damit werden Techniken der Künstlichen Intelligenz den Gesundheitsmarkt im nächsten Jahrzehnt unter Umständen deutlich verändern. Betroffen ist auch der Berufsstand der Juristen, wie es die Fortschritte im Bereich der »Legal Tech« nahelegen. Und so kommen als weiterer Einsatzbereich von KI-Anwendungen nicht zuletzt deutsche Gerichtssäle in Betracht: Hier könnte etwa eine KI-Anwendung herangezogen werden, die Prognosen für die künftige Gefährlichkeit von Straftätern trifft. Dies passiert gegenwärtig in Teilen der USA schon im Bereich der gerichtlichen Bewährungsentscheidungen. Weniger futuristisch, da bereits die Realität deutscher Polizeibeamter prägend, erscheint außerdem der Einsatz digitaler Technologien in der Verbrechensbekämpfung. So verbirgt sich hinter dem Begriff des »Predictive Policing« die datenbasierte Prognose von Straftaten, die zur Polizeieinsatzplanung verwendet wird. Dabei ist das »Predictive Policing« ein Anwendungsbereich, dessen Ausbau in der »Strategie Künstliche Intelligenz der Bundesregierung« beabsichtigt ist.

All diese Entwicklungen betreffen im Kern die Frage, wie wir in unserer Gesellschaft leben wollen. Gibt es ein »Menschenbild der Digitalisierung« – und ist dieses mit einem

freiheitlichen Rechtsstaat in Einklang zu bringen? In dessen Zentrum steht nach wie vor der mit Würde begabte Mensch, wofür Art. 1 Abs. 1 GG die normative Grundlage bietet. Danach darf der Mensch nicht zum bloßen Objekt staatlichen Handels herabgewürdigt werden (S. etwa BVerfGE 9, 89; 27, 1; 28, 386; 117, 71, 89; 131, 268, 286 im Anschluss an G. Dürig, AöR 117 (1956), 127.) – eine Vorgabe, deren Wahrung in Zeiten der Disruption durch Künstliche Intelligenz in besonderem Maße der kritischen Überprüfung bedarf, was eine intensive Kooperation von Recht und Philosophie voraussetzt. Dabei steht eines fest: Technologische Revolutionen sind nicht als »Selbstläufer« zu begreifen. Vielmehr liegt ihr Hergang in der Hand des Menschen als deren maßgeblichem Akteur. Aus juristischer Sicht liegt daher im Zertifizierungsprojekt von KI.NRW auch ein Augenmerk auf den Gestaltungsmöglichkeiten, die im Hinblick auf den Einsatz von Künstlicher Intelligenz zur Verfügung stehen. Ziel ist es, auf diese Weise einen relevanten Beitrag zu leisten zu dem Bild, das die Gesellschaft in Zeiten großer technologischer Fortschritte im Bereich der Künstlicher Intelligenz von sich selbst zeichnen wird.

2.4 Interdisziplinäre Betrachtung

Disruptive Technologien, die wie die Künstliche Intelligenz an den Wurzeln einer Gesellschaft ansetzen und Veränderungen von bislang unbekanntem Ausmaß und ungeahnter Geschwindigkeit herbeiführen können, bedürfen eines holistischen Blickwinkels, um ihnen angemessen Rechnung zu tragen. Für Philosophie, Recht und Technologie steht in einem freiheitlichen Rechtsstaat der Mensch im Mittelpunkt. Die Kooperation der Wissenschaften ist daher nicht nur abstrakt wünschenswert, sondern ein notwendiges Gebot unserer Zeit.

Gestaltung der ethisch-rechtlichen Rahmenbedingungen von Künstlicher Intelligenz

Unsere Gesellschaft und somit jeder Einzelne hat die Möglichkeit (mit-)zu entscheiden, wie die Welt aussehen soll, in der wir künftig mit Künstlicher Intelligenz leben wollen. In dem Diskurs, der hierfür geführt werden muss, spielen Philosophie, Recht und Technologie eine zentrale Rolle. Die technologische Entwicklung generiert die Problemfelder dieses gesellschaftlichen Diskurses. Gleichzeitig zeigt sie auf, was tatsächlich im Bereich des Möglichen ist und was in den Bereich der Science-Fiction gehört. Die Philosophie ordnet zentrale Begriffe der Ethik, wie etwa den moralischen Akteur, im Kontext von Künstlicher Intelligenz neu ein und liefert eine Begründung der universellen Gültigkeit von bestimmten ethischen Prinzipien und Rechtsnormen, wie etwa den Menschenrechten. Hierdurch stabilisiert sie überhaupt erst den Rahmen, innerhalb dessen der gesellschaftliche Diskurs sinnvoll und zielführend stattfinden kann. Bei der Umsetzung des Diskursergebnisses hin zur rechtlich richtigen Lösung nutzt die Rechtswissenschaft ethische Argumente. Dies wird vor allem in Bereichen relevant, in denen das Recht unter Umständen erst noch auf den Plan treten muss. Angesichts der Vielzahl an Veränderungen, die die Anwendungen der Künstlichen Intelligenz in jedweden gesellschaftlichen Bereich trägt, fragt sich aus Sicht der Juristen, ob insoweit ein Regelungsbedarf besteht. Weiterhin geraten rechtliche Grundbegriffe in Konfrontation mit neuen technischen Entwicklungen auf den Prüfstand. Dies betrifft beispielsweise den Begriff der Verantwortung bzw. der »Schuld«. Davon sind auch Grundbegriffe der philosophischen Ethik wie Gerechtigkeit, Gleichheit, Autonomie, Fairness und Transparenz usw. betroffen. Diese müssen für den Zusammenhang von KI-Anwendungen genau gefasst werden, da diese Begriffe hier eine spezifische Bedeutung gewinnen, die sie erst durch die neue Technologie bekommen. Diese Bedeutungen lassen sich nur in trilateraler Zusammenarbeit klären. Es stellt sich im Zusammenhang mit Künstlicher Intelligenz die Frage, ob an den bisherigen Begriffen festgehalten werden kann oder ob sie einer Modifikation bedürfen. Dabei setzt jedwede rechtliche Bewertung ein klares Verständnis der technischen Zusammenhänge voraus. Dies betrifft in erster Linie die tatsächlichen Möglichkeiten von KI-Anwendungen und damit die Frage nach einer faktischen Umsetzbarkeit rechtlicher Vorgaben. Sofern diese nicht besteht, kommt eine Situation auf, in der von rechtlicher Seite etwas verlangt wird, das technisch nicht umgesetzt werden kann (Beispiel: uneingeschränkte Transparenz). Sofern aber eine entsprechende technische Umsetzbarkeit nicht besteht, stellt sich die weitergehende rechtliche Frage, ob gleichwohl eine Zulässigkeit der jeweiligen KI-Anwendung begründet werden kann.

Entwicklung von konkreten KI-Anwendungen

Bei der Konstruktion von KI-Anwendungen innerhalb eines bestehenden ethischen und rechtlichen Rahmens ist es essentiell, die Sichtweise aller drei Disziplinen miteinzubeziehen. Bereits im Design der KI-Anwendung muss geklärt werden, ob die Anwendung ethisch und rechtlich zulässig ist und

falls ja, welche Leitplanken für ihre Ausgestaltung formuliert werden sollten. Ein notwendiges Kriterium ist es hierbei, allen Beteiligten dieselben Möglichkeiten einer moralischen Entscheidung zu geben, welche sie auch im Falle eines Verzichts auf den KI-Einsatz hätten, und ihre Rechte sowie ihre Freiheit zu achten. Viele weitere Folgefragen, die sich hieraus ergeben – beispielsweise was Fairness im Kontext der Anwendung zu bedeuten hat, oder welche Auswirkungen auf die Nutzer, wie etwa emotionale Bindungen zur KI, vertretbar sind – lassen sich aus technologischer Perspektive allein nicht beantworten, sondern bedürfen wiederum eines holistischen Ansatzes.

Ist die grundsätzliche Zulässigkeit der KI-Anwendung sichergestellt, ergeben sich auch für die weitere Entwicklung bis hin zur Veröffentlichung, z. B. als Open Source, interdisziplinäre Fragen. Diese betreffen etwa den Umgang mit unvermeidbaren Konflikten und Trade-offs zwischen den verschiedenen Handlungsfeldern. In unterschiedlichen Kontexten ist jeweils eine andere Balance zwischen den einzelnen Werten gefragt. Widerstreitende Interessen können durch den ethisch-rechtlichen Grundsatz der Verhältnismäßigkeit in eine ausgewogene Beziehung miteinander gebracht werden. Auf diese Weise werden sämtliche Perspektiven der beteiligten Akteure in die notwendige Interessenabwägung einbezogen. Wenngleich also Abwägungsentscheidungen in Einzelfällen nicht auf der Metaebene getroffen werden können, hält das Verhältnismäßigkeitsprinzip ein Instrument bereit, um die Zulässigkeit von spezifischen KI-Anwendungen festzustellen.

3 HANDLUNGSFELDER DER ZERTIFIZIERUNG

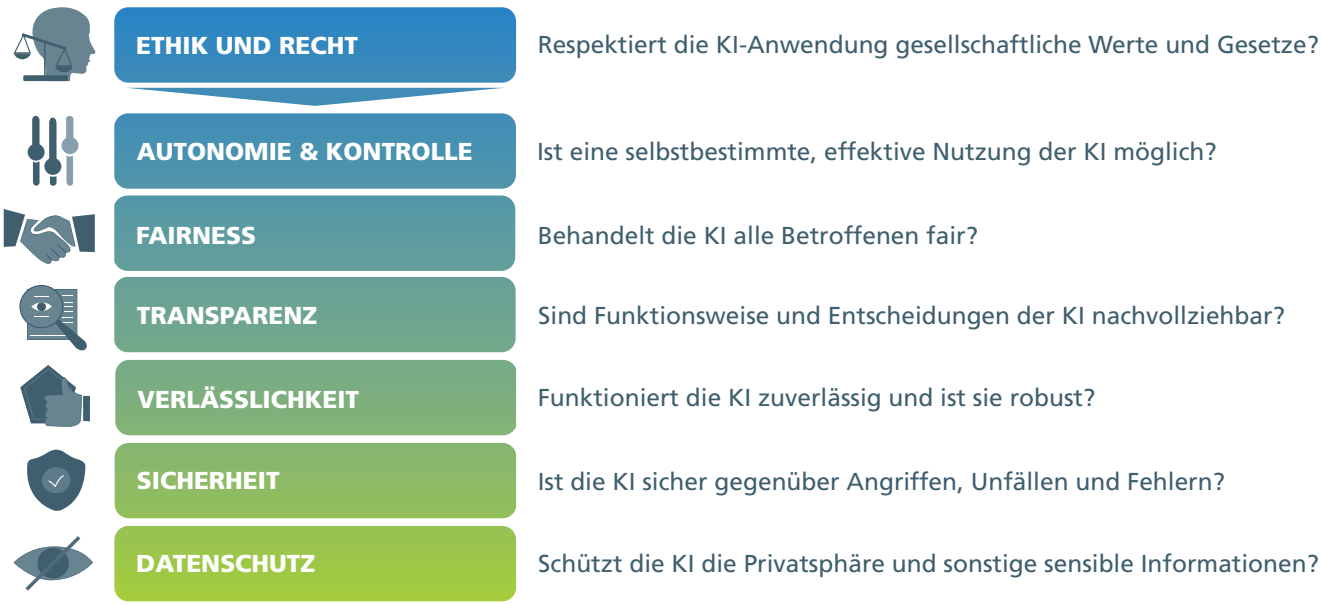
Aufgrund ihres disruptiven Potenzials ist es für KI-Anwendungen in besonderem Maße wichtig, die Übereinstimmung mit philosophischen, ethischen und rechtlichen Rahmenbedingungen zu gewährleisten. Ihre Zertifizierung dient in erster Linie dem Schutz rechtlich bzw. ethisch grundlegender Interessen von Personen. Auf diese Weise soll vermieden werden, dass es zu unzulässigen Beeinträchtigungen Einzelner und von Gruppen kommt. Die KI-Zertifizierung verfolgt insofern den allgemeinen Zweck, Unrecht bzw. ethisch nicht gerechtfertigte Zustände von der Gesellschaft abzuwenden. Dies betrifft neben den Freiheitsrechten des Einzelnen und dem Grundsatz der Gleichbehandlung gerade auch allgemeine gesellschaftliche Interessen wie etwa den Schutz und die Erhaltung der Umwelt sowie des demokratischen Rechtsstaats.

Aus diesen Grundwerten und Prinzipien eines freiheitlich geordneten Gemeinwesens lässt sich unter Berücksichtigung des rechtsstaatlichen Grundsatzes der Verhältnismäßigkeit eine Vielzahl an Konkretisierungen ableiten. Auf diese Weise entstehen auf der Basis von Ethik und Recht sowie informatischen Anforderungen konkrete, für die Zertifizierung maßgebliche Handlungsfelder.

Für die Entwicklung einer KI-Anwendung folgt hieraus insbesondere, dass Anwendungsbereich, Zweck und Umfang sowie Betroffene frühzeitig identifiziert werden müssen. In diesen Prozess sind alle direkt oder indirekt betroffenen Akteure angemessen zu involvieren. Es sollte eine Risikoanalyse durchgeführt werden, die die Möglichkeiten von Missbrauch und Dual Use einschließt, deren Konsequenzen angemessen bei der weiteren Entwicklung miteinbezogen werden müssen. Schließlich sollte die Anwendung »by-design« so konstruiert werden, dass sie in dem festgelegten Umfang auditierbar und prüfbar ist.

3.1 Autonomie und Kontrolle

Autonomie ist sowohl als ethischer als auch als rechtlicher Wert anerkannt. Philosophisch ist Autonomie die Grundlage aller Werte, weil wir uns Werte als menschliche Gemeinschaft selbst geben müssen. Sie ist im Allgemeinen die Fähigkeit zur moralisch relevanten Selbstbestimmung. Begründet ist damit die Freiheit des Einzelnen, selbstbestimmt Entscheidungen zu treffen. Dies umfasst auch all jene Entscheidungen, die seine eigene Rechtsposition berühren und darüber hinaus die Freiheit, die Ziele des eigenen Verhaltens ebenso wie die Wahl der Mittel zur Erreichung dieser Ziele zu bestimmen.



KI-Anwendungen übernehmen immer mehr Routinetätigkeiten und agieren zunehmend selbstständiger. Dabei ist zu beachten, dass oftmals mobile Systeme (z. B. Roboter, Fahrzeuge), die durch die eingesetzten KI-Anwendungen gesteuert werden, ebenfalls als »autonom« bezeichnet werden. Hierbei wird jedoch dem System lediglich die Wahl der Mittel, nicht aber die eigentliche Zielsetzung freigestellt. Von daher spricht man in diesem Kontext irreführend von der »Handlungsautonomie« des Systems, die sich eigentlich aus der menschlichen Zielsetzung ergibt. Deswegen ergibt sich auch in diesem Kontext ein Spannungsfeld zur Autonomie (des Menschen), da solche KI-Anwendungen den Menschen ihrerseits in der Wahl seiner Ziele und Mittel beeinflussen können. Letzteres ist insbesondere der Fall, wenn die KI-Anwendung mit menschlicher Entscheidungsfindung interagiert, indem sie zum Beispiel Entscheidungsvorschläge generiert, Steuerbefehle erzeugt (und evtl. sogar ausführt), mit dem Menschen direkt kommuniziert (Sprachassistenten, Chatbots, …) oder in Arbeitsprozesse integriert ist.

Künstliche Intelligenz darf die Autonomie von Individuen und sozialen Gruppen nicht unverhältnismäßig einschränken. Vor diesem Hintergrund ist es bei der Entwicklung und dem Betrieb einer KI-Anwendung wichtig darzulegen, inwiefern individuelle bzw. kollektive Nutzer übermäßiges Vertrauen in die KI-Anwendung entwickeln, emotionale Bindungen aufbauen oder in ihrer Entscheidungsfindung unzulässig beeinträchtigt beziehungsweise gelenkt werden könnten. Die Aufgabenverteilung und Interaktionsmöglichkeiten zwischen KI-Anwendung und Nutzer müssen daher klar und transparent geregelt sein. Nutzer müssen angemessen mit den möglichen Risiken bzgl. einer eventuellen Beeinträchtigung ihrer Autonomie, mit ihren Rechten, Pflichten und Eingriffs- sowie Beschwerdemöglichkeiten vertraut gemacht werden. Dem Nutzer muss in angemessenem Umfang die Möglichkeit zur Steuerung des Systems verliehen werden. Dies schließt es ein, dass die Zustimmung zur Nutzung einer KI-Anwendung entziehbar sein muss. Dabei sollten den Nutzern keine bloße Ja/Nein-Option, sondern plurale Nutzungsmöglichkeiten eingeräumt werden. Insbesondere muss es auch möglich sein, die Anwendung in Gänze abzuschalten. Nutzer müssen angemessen in der Wahrung ihrer Autonomie unterstützt werden, indem sie die dazu erforderlichen Informationen über das Verhalten der KI-Anwendung im Betrieb erhalten, ohne überfordert zu werden. Letzteres muss gegebenenfalls insbesondere auch auf Personen mit speziellen Bedürfnissen abgestimmt sein. Zudem sind angemessen sichere Eingriffsmöglichkeiten für den Fall bereitzustellen, dass eine Gefährdung der Autonomie der Nutzer erkannt wird.

3.2 Fairness

Als Ausfluss des allgemeinen Gleichbehandlungsgrundsatzes ist sowohl in ethischer als auch in rechtlicher Hinsicht von einer KI-Anwendung die Wahrung des Prinzips der Fairness zu verlangen. Gemeint ist damit das Verbot, gleiche soziale Sachverhalte ungleich oder ungleiche gleich zu behandeln, es sei denn, ein abweichendes Vorgehen wäre sachlich gerechtfertigt. Damit erstreckt sich das Prinzip auf das Verbot einer ungerechtfertigten Ungleichbehandlung in einer KI-Anwendung und schließt unzulässige Diskriminierungen aus. Dies bedeutet insbesondere, dass Individuen nicht aufgrund ihrer Zugehörigkeit zu einer marginalisierten oder diskriminierten Gruppe wiederum im sozialen Ergebnis diskriminiert werden dürfen. Zum Beispiel dürfen nicht Menschen mit bestimmten Familiennamen, einer spezifischen Religionszugehörigkeit oder einem besonderen Geschlecht eine bessere oder schlechtere Bewertung erhalten. Ebenso müssen Sprachsteuerungen auch auf Personen mit besonderen Akzenten oder Soziolekten reagieren können und individuell anpassbar sein. Darüber hinaus darf Gesichtserkennungssoftware grundsätzlich nicht häufiger Fehler bei Menschen mit einer bestimmten Hautfarbe oder anderen phänotypischen Merkmalen machen.

KI-Anwendungen lernen aus historischen Daten. Diese sind nicht notwendigerweise vorurteilsfrei. Beinhalten die Daten beispielsweise Benachteiligungen von Frauen, so kann die KI-Komponente diese Vorurteile übernehmen. Außerdem können in der Datengrundlage bestimmte Gruppen unterrepräsentiert sein. Man spricht dann von Bias. Bias kann ebenfalls zu Entscheidungen führen, die unfair sind. Als abschreckendes Beispiel bekannt geworden ist die fehlerhafte Klassifikation von dunkelhäutigen Menschen als Gorillas durch Google Fotos. Daher müssen repräsentative Trainingsdaten sichergestellt werden. Darüber hinaus kommt als geeignetes Instrument zur Vermeidung von Bias eine Nachbesserung der Ausgabe des ML-Modells in Betracht.

Um Fairness zu operationalisieren, muss aus technischer Sicht jeweils ein quantifizierbarer Fairnessbegriff entwickelt werden. Dies setzt in einem ersten Schritt voraus, diejenigen Gruppen zu identifizieren, die nicht benachteiligt werden sollen. Diese Gruppen können gesellschaftliche Minderheiten darstellen, aber auch Unternehmen oder allgemein juristische Personen, wie es beispielsweise bei der Preisbildung auf digitalen Marktplätzen der Fall ist. In einem zweiten Schritt muss eine Quantifizierung der gewählten Fairnessdefinition erfolgen. Besonders hervorzuheben ist dabei die Unterscheidung von

Gruppenfairness und individueller Fairness. Bei Gruppenfairness ist zu verlangen, dass die Ergebnisse für alle vorhandenen Gruppen vergleichbar sind, z. B. im Sinne von gleicher »Trefferwahrscheinlichkeit« in allen Gruppen. Bei individueller Fairness wird die gleiche Behandlung von gleichen Individuen als Maßstab gesetzt.

3.3 Transparenz

Die Transparenz einer KI-Anwendung kann für ihre Akzeptanz entscheidend sein. Dabei sind zwei Aspekte zu unterscheiden. Erstens müssen Informationen zum richtigen Umgang mit der KI-Anwendung verfügbar sein. Zum anderen geht es um Anforderungen an die Interpretierbarkeit, Nachverfolgbarkeit und Reproduzierbarkeit von Ergebnissen, die Einsichten in die inneren Prozesse der KI-Anwendung erfordern.

Information zum Umgang mit einer KI-Anwendung

Zu allererst ist zu verlangen, dass in einer Kommunikationssituation grundsätzlich klar sein muss, dass diese mit einer KI-Anwendung stattfindet. Darüber hinaus müssen die Akteure angemessen mit dem Gebrauch der Anwendung vertraut gemacht werden. Dazu gehört ein Verständnis dafür, welchem Zweck die Anwendung dient, was sie leistet, was potenzielle Risiken (auch in Bezug auf andere Handlungsfelder wie z. B. Verlässlichkeit, Sicherheit und Fairness) sind und wer die Zielgruppe der Anwendung ist.

Nachvollziehbarkeit und Interpretierbarkeit des ML-Modells

Aus ethisch-rechtlicher Sicht kann ein Interessenskonflikt zwischen dem Wunsch nach Transparenz für die Nutzer (bzw. für interessierte Gruppen) einerseits und der Wahrung von Geschäftsgeheimnissen bzw. der allgemeinen gesellschaftlichen Sicherheit andererseits bestehen. Daraus ergeben sich konkret die folgenden Anforderungen an die Transparenz einer KI-Anwendung:

- KI-Anwendungen, von denen die Rechte und Interessen Dritter betroffen sind, müssen grundsätzlich transparent sein. Transparenz bedeutet die Nachvollziehbarkeit der Arbeitsweise der KI-Anwendung.

- KI-Anwendungen müssen nicht nach außen transparent gemacht werden. Dies gilt nicht, wenn weit überwiegende gesellschaftliche Interessen an der Verstehbarkeit der KI-Anwendung bestehen.

- KI-Anwendungen, von denen die Rechte und Interessen Dritter betroffen sind, dürfen ausnahmsweise intransparent sein, sofern dies bei Abwägung der widerstreitenden Interessen verhältnismäßig ist.

Diese zweite Art der Transparenz betrifft die inneren Prozesse der KI-Anwendung und speziell des ML-Modells. Dabei geht es um die Fragen von Interpretierbarkeit, Nachverfolgbarkeit und Reproduzierbarkeit von Ergebnissen für verschiedene Akteure und Zwecke. Im Speziellen ist unter anderem zu verlangen:

- Akteure müssen die Ausgabe der KI-Anwendung insoweit nachvollziehen können, als dass sie eine informierte Einwilligung oder Ablehnung geben. Das kann häufig durch das Aufzeigen der entscheidungsrelevanten Passagen in der Eingabe geschehen.

- Für eine informierte Intervention beim Einsatz einer KI-Anwendung im Arbeitsprozess müssen die mitgeteilten Informationen nach der Maßgabe ausgewählt werden, die Nutzer nicht durch irrelevante Details zu überfordern.

- Experten müssen grundsätzlich die Funktionsweise der KI-Anwendung auf technischer Detailebene nachvollziehen können, z. B. zum Zwecke der Verbesserung oder der Klärung von Konfliktfällen. Zwar müssen die Experten nicht jede Ausgabe einer KI-Anwendung vorhersagen können, ihr generelles Verhalten muss jedoch während der Entwicklung und auch später im produktiven Betrieb prinzipiell erklärbar, nachvollziehbar und dokumentiert sein. Hierzu dienen Logging, Dokumentationen bzw. Archivierungen des Designs, der Daten, des Trainings, des Testens/Validieren des Modells, sowie der einbettenden Umgebung.

Aus technischer Sicht ist die Frage der grundsätzlichen Transparenz nicht trivial und das Spannungsfeld zwischen höherer Genauigkeit bzw. Robustheit und der Erklärbarkeit von Modellen ist in der KI-Welt ein altbekanntes Dilemma. »Black Box«-Modelle sind zwar in vielen Fällen genauer bzw. robuster als beispielsweise regelbasierte Modelle, jedoch sind sie nur bedingt interpretierbar. Teilweise kann diese Erklärbarkeit auch durch nachgeschaltete Verfahren, wie z. B. durch das Trainieren von Erklärmodellen oder einer Analyse des Eingabe/Ausgabe-Verhaltens von Modellen (sogenannte LIME Analyse-Local Interpretable Model-agnostic Explanations) erreicht werden. Zurzeit ist die Interpretierbarkeit von Modellen ein

aktives Forschungsfeld und es werden viele Anstrengungen unternommen, die Lernprozesse von »Black Box«-Modellen besser zu verstehen, sowie ihre internen Prozesse zu visualisieren und die resultierenden Entscheidungen erklären zu können.

3.4 Verlässlichkeit

Aus technischer Sicht stellt Verlässlichkeit einen Sammelbegriff dar, der zum Teil deutlich unterschiedliche Aspekte der Güte einer KI-Komponente umfasst: Die Korrektheit der KI-Ausgaben, die Einschätzung der ML-Modellunsicherheiten sowie die Robustheit gegenüber schädlichen Eingaben (z. B. adversarial attacks), Fehlern oder unerwarteten Situationen. Neuartige, für Menschen untypische und damit unerwartete Fehlermodi können zu potenziell kritischen, da nicht eingeübten Situationen führen, insbesondere bei direkter Mensch-Maschine-Interaktion.

Profundes Anwendungswissen ist nötig, um diese Verlässlichkeitsdimensionen für eine konkrete KI-Anwendung zu bewerten und festzulegen, unter welchen Voraussetzungen die Anwendung nach diesen Dimensionen als verlässlich einzustufen ist. Für diese Einstufung sind die bislang bereits erhobenen Anforderungen, die initiale Risikobewertung sowie ethische und rechtliche Rahmenbedingungen vollumfänglich zu berücksichtigen. Die Übersetzung der Anforderungen in quantitative Maße und Zielwerte erfordert Domänenwissen sowie mathematisch-technische Expertise und ist naturgemäß niemals vollständig. Gleiches gilt für die Beschreibung des Anwendungsbereichs der KI-Anwendung. Er ist möglichst genau zu spezifizieren und zu formalisieren, um sicherzustellen, dass die verwendeten Trainings- und Testdaten die Menge der im Betrieb zu erwartenden Eingaben der KI-Anwendung hinreichend abdecken. In jedem Fall sollte die Verlässlichkeit der KI-Anwendung auf das Leistungsvermögen des Menschen abgestimmt werden.

Eine korrekte Implementierung der Trainingsroutinen und des fertig trainierten Modells ist eine notwendige Voraussetzung, um diesen Anforderungen gerecht zu werden. Die dazu durchzuführenden Tests sollten im Bereich des Maschinellen Lernens etabliert und auf die jeweilige Anwendung abgestimmt sein. Werden ML-Modellschwächen aufgedeckt, ist mit geeigneten Korrekturmechanismen bis hin zum Einsatz eines Rückfallplans darauf zu reagieren. Dabei ist die Verlässlichkeit der KI-Anwendung zu jedem Zeitpunkt im Produktivbetrieb zu gewährleisten. Dies impliziert, dass die korrekte Funktionsweise regelmäßig in angemessenen Abständen überprüft werden muss. Um außerdem die Verlässlichkeit schrittweise zu erhöhen, sollten geeignete Maßnahmen etabliert werden, z. B. durch das Abspeichern herausfordernder Szenarien im Produktiveinsatz.

3.5 Sicherheit

Sicherheit im Sinne von Schutz vor Angriffen (Security) und Schutz vor Gefährdungen, die von der KI-Anwendung ausgehen (Safety), ist für KI-Anwendungen mindestens von ebenso großer Wichtigkeit wie für andere Informations- und technische Systeme. Beide Sicherheitskonzepte sind auf die gesamte KI-Anwendung anwendbar, in die die KI-Komponente eingebettet ist, und nicht auf die KI-Komponente selbst. Dabei sind KI-spezifische Gefährdungen abzufangen oder geeignet zu behandeln. Diese Gefährdungen können sich als Funktionsausfall oder starke Funktionsänderung der KI-Komponente, sowie unautorisierter Informationsabfluss äußern. Auf Ursachen des Funktionsausfalls, z. B. durch adversarial attacks, und einer starken Funktionsänderung wird, sofern möglich, bereits innerhalb der KI-Komponente reagiert, was in die Verantwortlichkeit des Handlungsfeldes Verlässlichkeit fällt. Ist dies nicht vollumfänglich möglich, greifen die Maßnahmen der umgebenden KI-Anwendung und liegen in der Verantwortlichkeit des Handlungsfeldes Sicherheit.

Die HLEG hat abstrakte Sicherheitsziele für KI-Anwendungen definiert. Diese abstrakten Zielsetzungen (und darüberhinausgehende) sind jedoch weit von einer Operationalisierung etwa durch einen Prüfkatalog oder eine Norm entfernt. Umgekehrt existieren gerade im Bereich Sicherheit eine ganze Reihe von operativ überprüfbaren Spezifikationen und Normen, die jedoch keinen speziellen Bezug auf die Besonderheiten von KI-Anwendungen nehmen. Ziel des Handlungsfeldes Sicherheit ist es, die Anforderungen aus bestehenden Normen, die unerlässlich für den Schutz vor Angriffen und Gefährdungen von KI-Anwendungen sind, zusammenzuführen und mit weiteren spezifischen KI-Anforderungen zu ergänzen.

3.6 Datenschutz

KI-Anwendungen sind geeignet, in eine Vielzahl an Rechtspositionen einzugreifen. Besonders häufig handelt es sich dabei um Eingriffe in die Privatsphäre bzw. das Recht auf informationelle Selbstbestimmung. So verarbeiten KI-Anwendungen oftmals sensible Informationen, wie zum Beispiel Geschäftsgeheimnisse, personenbezogene oder persönliche

Daten, etwa Stimmaufnahmen, Fotos oder Videos. Daher ist sicherzustellen, dass die einschlägigen datenschutzrechtlichen Bestimmungen wie etwa die Datenschutz-Grundverordnung (DSGVO) und das Bundesdatenschutzgesetz (BDSG) eingehalten werden. KI-Anwendungen können nicht nur ein Risiko für die Privatsphäre des Einzelnen darstellen. Darüber hinaus können davon (Geschäfts-)Geheimnisse betroffen sein, die keine personenbezogenen Daten im Sinne der DSGVO darstellen, dennoch aber ethisch sowie rechtlich schutzwürdig sind. Dabei kann es sich beispielsweise um Maschinendaten handeln, die völlig unabhängig von der Frage, welche Person als Maschinenbediener tätig war, Informationen über die Prozessauslastung oder Fehlerquoten beinhalten. Durch die KI-Zertifizierung soll sichergestellt werden, dass Datenschutzrisiken und -maßnahmen der KI-Anwendung, unter Berücksichtigung der besonderen Herausforderungen im Datenschutz durch die Künstliche Intelligenz so ausreichend analysiert und dokumentiert sind, dass der grundsätzlich zu benennende Datenschutzbeauftragte sinnvoll darin unterstützt wird, seine Untersuchung und letztliche Entscheidung bzgl. der Datenschutzfreigabe durchführen zu können.

Die Herausforderungen an den Datenschutz sind in KI-Anwendungen potenziell deutlich höher als in klassischen IT-Systemen, da KI-Anwendungen oft Daten zusammenführen, die bislang nicht verknüpft waren, und erst durch Maschinelles Lernen neue Methoden der Verknüpfung von Daten entstehen. Je mehr Daten verknüpft werden (»data linkage«), umso mehr steigt das Risiko, Personen oder z. B. konkrete Betriebsstätten auch ohne direkte Angabe entsprechender Attribute identifizieren zu können. So ist es zum Beispiel möglich, mit ca. 95 Prozent Verlässlichkeit Personen an der Art und Weise, wie sie eine Computertastatur bedienen, zu re-identifzieren. Gäbe es nun eine öffentliche (oder käufliche) Datenbank mit der Zuordnung von Tastatur-Anschlagsmustern zu Personen, so wird das Anschlagsmuster zu einem sogenannten »Quasi-Identifier«, der einen Personenbezug ermöglicht. Ebenso können mit KI-Methoden potenziell Personenbezüge bei der Verarbeitung von Text, Sprach- und Bilddaten, sowie aus protokollierten Nutzungsdaten erstellt werden. Zusätzlich besteht das Risiko, dass ein trainiertes Modell wieder personenbezogene Rückschlüsse erlaubt, ohne selbst personenbezogene Daten zu beinhalten.

Hieraus ergibt sich, dass die erlangten Informationen sowohl während des Trainings als auch im Betrieb wirksam geschützt werden müssen. KI-Anwendungen dürfen auf personenbezogene Daten ausschließlich mit Einwilligung des Berechtigten Zugriff nehmen. Eine Weiterverarbeitung sowie die Weitergabe an Dritte dürfen – vorbehaltlich weiterer Beschränkungen – ausschließlich mit Zustimmung des Rechtsgutsinhabers erfolgen. Es muss sichergestellt werden, dass keine Schutzlücken bestehen, die einen unberechtigten Zugriff ermöglichen. Dem Einzelnen muss die Möglichkeit der Löschung seiner Daten eingeräumt werden. Zu den erforderlichen Maßnahmen gehören damit u. a. die Information der Betroffenen über Zweck und Einsatz der personenbezogenen Daten oder daraus abgeleiteter Daten, die Bereitstellung ausreichender Einwilligungs-, Auskunfts-, Einspruchs-, und Widerrufsmechanismen bzgl. der Nutzung personenbezogener Daten, die Einhaltung der Grundsätze der Datensparsamkeit und zweckgebundenen Verwendung, sowie eine Risikoanalyse bzgl. der potenziellen Herstellbarkeit eines Personenbezuges, in der möglicherweise eingesetzte Maßnahmen zur Anonymisierung oder Aggregation von Daten mit dem Potenzial einer Re-Identifikation durch Verknüpfung mit Hintergrundwissen abgeglichen wird.

4 AUSBLICK

Das vorliegende Whitepaper ist das erste Ergebnis eines interdisziplinären Projekts der Kompetenzplattform KI.NRW mit dem Ziel, eine Zertifizierung für KI-Anwendungen zu entwickeln, die neben der Absicherung der technischen Zuverlässigkeit auch einen verantwortungsvollen Umgang aus ethisch-rechtlicher Perspektive prüft. Grundlage für diese Zertifizierung ist ein KI-Prüfkatalog, welcher sich aktuell in der Entwicklung befindet und anhand dessen akkreditierte Prüfer KI-Anwendungen sachkundig und neutral beurteilen können.

Es ist geplant, Anfang 2020 eine erste Version des Prüfkatalogs zu veröffentlichen und die ersten KI-Anwendungen zu zertifizieren. Aufgrund der Komplexität des Themas wird die erste Version an einigen Stellen geeignete Einschränkungen bezüglich der Anwendbarkeit machen, wie beispielsweise im Bereich des Weiterlernens im Betrieb oder für die Steuerung von sicherheitskritischen Anwendungen. Eine Reihe unterschiedlicher KI-Anwendungen dient bereits während der Entwicklung des Prüfkatalogs dazu, die Vollständigkeit und Allgemeinheit der Prüfziele und -anforderungen zu testen und die Anwendung des Katalogs zu evaluieren und zu demonstrieren.

Eine besondere Aufgabe wird hierbei der Abgleich der Prüfziele mit existierenden Standards sein und die Abgrenzung gegenüber existierenden Prüfkatalogen und Gesetzen, zum Beispiel für IT-Sicherheit und die Datenschutzgrundverordnung. Dazu kooperiert das Projekt mit dem Bundesamt für Sicherheit in der Informationstechnik (BSI), um dessen langjährige Erfahrung im Bereich der IT-Sicherheit und in der Ausgestaltung und Anerkennung von IT-Prüfstandards miteinzubeziehen.

Die Methoden und Anwendungsmöglichkeiten der Künstlichen Intelligenz werden kontinuierlich und massiv weiterentwickelt. Es ist davon auszugehen, dass sich mit ihnen die gesellschaftliche Vorstellung von Ethik und die Regulierung von Künstlicher Intelligenz prägen wird. Deshalb muss der Prüfkatalog ein lebendes Dokument sein, das stetiger Aktualisierungen aus den drei Bereichen Informatik, Recht und Philosophie bedarf. Parallel hierzu wird der Gültigkeitsbereich des Katalogs schrittweise erweitert und es werden für bestimmte Anwendungsbereiche und Risikoklassen Spezialkataloge ausgearbeitet.

5 IMPRESSUM

Herausgeber
Fraunhofer-Institut für Intelligente Analyse-
und Informationssysteme IAIS
Schloss Birlinghoven
53757 Sankt Augustin
kinrw-pr@iais.fraunhofer.de
www.iais.fraunhofer.de

Kontakt
Dr. Maximilian Poretschkin
Telefon: +49 22 41 14 - 19 84

Titelbild
© mila103, ryzhi, zapp2photo / fotolia.com

Layout und Satz
Svenja Niehues, Fraunhofer-Institut für Intelligente
Analyse- und Informationssysteme IAIS, Sankt Augustin