Europäische Gesetzesanforderungen für vertrauenswürdige Künstliche Intelligenz: Herausforderungen und Chancen

H. Pham, F. Hartmann, F. Beer

Abstract

Das Themengebiet der Künstlichen Intelligenz (KI) ist ein altes und zugleich auch immer wieder neues digitales Grenzland. Es ist eine treibende Kraft für Innovation und verspricht Annehmlichkeiten in Form von smarter Funktionalität, welche einen Mehrwert liefert. Teilweise medial angeheizt und als Marketinginstrument gewinnbringend ausgeschöpft, rückt dieses Narrativ nicht zuletzt wegen der verblüffenden Resultate von ChatGPT zusehends in unser Bewusstsein. Aber auch weniger spektakuläre Anwendungsszenarien verdeutlichen, dass Technologie basierend auf KI auf dem Vormarsch ist und Einzug in das digitale Leben hält. So vergrößern sich auch rasant die Berührungspunkte mit dem Menschen – mit subtilen, stellenweise allerdings auch erheblichen Auswirkungen, wie Zwischenfälle aus der jüngeren Vergangenheit offenbaren. Diese Umstände bieten genügend Anlass, Eckpfeiler in Form gesetzlicher Rahmenwerke zu manifestieren. Als weltweiter Vorreiter legte die Europäische Union (EU) im Jahr 2021 einen Verordnungsvorschlag zur Festlegung harmonisierter Vorschriften für vertrauenswürdige KI vor. Mit der Bestrebung, eine einheitliche begriffliche Abgrenzung für KI zu etablieren und sektorübergreifend angemessene Rahmenbedingungen bezüglich der Einhaltung europäischer Werte aufzustellen, steht dieser Gesetzesentwurf kurz vor seiner Verabschiedung. Ein besonderer Fokus liegt dabei auf der sogenannten Hochrisiko-KI, von der eine beträchtliche Gefährdung für Gesundheit und Sicherheit, sowie für Grundrechte von Personen ausgehen kann. Bei Verstößen gegen die gesetzlichen Bestimmungen drohen horrende administrative Strafen. Diese sind wirksam, verhältnismäßig und abschreckend und verleihen der richtungsweisenden Bedeutung des Verordnungsvorschlags seitens der EU Nachdruck. Dieses Whitepaper befasst sich inhaltlich mit den Kernpunkten dieser geplanten Regulatorik. Ausgehend von der Definition eines KI-Systems und der zugrundeliegenden Risikomodellierung, werden die essenziellen Anforderungen, die an solche Systeme gestellt werden, formuliert. Auf dieser Grundlage werden Herausforderungen für Unternehmen und Organisationen samt ihren Auswirkungen kritisch beleuchtet. Mit diesen Hürden gehen allerdings auch Chancen einher. Sie werden ebenfalls konkret benannt und zeigen, dass mit einigen Korrekturen im Gesetzesentwurf und einem strategischen Umdenken von Unternehmen und Organisationen das Potenzial von KI nachhaltig ausgeschöpft werden kann.

infodas

Gesellschaft für Systementwicklung
und Informationsverarbeitung mbH

Rhonestraße 2
50765 Köln

vertrieb@infodas.de
www.infodas.de

1 Ausgangslage

Nicht zuletzt durch den Erfolg in der US-Amerikanischen Quizsendung „Jeopardy!" im Jahr 2011, den historischen Sieg im Brettspiel „Go" über den südkoreanischen Meister Lee Sedol im Jahr 2016 oder die aktuell beeindruckenden Ergebnisse von ChatGPT wird das Potenzial von „Künstlicher Intelligenz" (KI) offenkundig. Im englischen Sprachgebrauch auch unter dem Begriff Artificial-Intelligence (AI) bekannt, hat sie längst Einzug in weite Teile der digitalen Gesellschaft gehalten und beschert uns Annehmlichkeiten. Ob Produktempfehlungsdienste, Sprachassistenzsysteme oder Gesichtserkennung – Technologie basierend auf KI begleitet uns seit vielen Jahren und ist nicht mehr aus dem Alltag wegzudenken. Auch zukünftig wird sie präsent sein und sich mit nahezu allen Lebensbereichen verzahnen. Laut Analysten ist damit zu rechnen, dass der Anteil des Bruttoinlandprodukts in Europa bis zum Jahr 2030 allein durch den Einfluss von KI sich auf 2,5 Billionen Euro belaufen wird [1]. Aufgrund des immensen Potentials wird KI auch als Megatrend bzw. Schlüsseltechnologie betitelt. Was sich allerdings genau hinter der Begrifflichkeit verbirgt, ist häufig unklar und es konnte sich bis dato kein allgemeines Verständnis darüber in der Gesellschaft etablieren. So verbindet man für gewöhnlich mit der Wortschöpfung „KI" vor allem eines: Smarte Funktionalität, die einen Mehrwert verspricht. Dieses Narrativ wird nicht zuletzt als Marketinginstrument von vielen Unternehmen geschürt, um sich mit neuen und vermeintlich innovativen Produkten positiv am Markt abzuheben.

KI ist jedoch alles andere als neu. Zur Klärung lohnt ein Blick in die Historie. Bereits im Jahr 1956 fand die Begrifflichkeit auf einer mehrwöchigen wissenschaftlichen Konferenz in Dartmouth, New Hampshire, erstmalig Erwähnung. Seit der Geburtsstunde dieser wissenschaftlichen Disziplin wurde sie durch intensive Forschung sukzessive weiterentwickelt, in welcher auf Grundlage von Beispieldaten versteckte Muster und Strukturen in diesen mittels immer ausgefeilteren Lernstrategien extrahiert werden können. In der Regel wird dazu allerdings ein erhöhtes Volumen an Basisdaten benötigt, so dass KI es lange nicht schaffte, außerhalb des akademischen Kontextes Fuß zu fassen. Erst mit der aufkommenden Digitalisierung in den 2010er Jahren wurden die erforderlichen Rahmenbedingungen etabliert. Breitbandiges Internet, leistungsfähigere Hardware und hohe Datenverfügbarkeit waren dabei wichtige Katalysatoren, die den Einsatz von KI in der Praxis begünstigten. Seitdem entstand eine schier unüberschaubare Menge an lukrativen Business-Cases, die nicht nur darauf abzielen, inhärente Muster zu gewinnen, um Daten abstrakt zu beschreiben, sondern dieses abgeleitete Wissen auch auf ungesehene Daten anzuwenden. Träger dieses sprichwörtlich trainierten Wissens werden auch fachsprachlich als KI-Modelle bezeichnet.

Dass diese erlernten Modelle wiederum nur unter gewissen Umständen ausreichend gute Aussagen für neue Daten erlauben oder deren Marktreife kritisch zu hinterfragen ist, demonstrieren eine Reihe von Zwischenfällen aus der jüngeren Vergangenheit: So erregte im Jahr 2015 eine neue KI-gestützte Anwendung des Suchmaschinenkonzerns Google, welche Fotos automatisiert Rubriken zuordnet, starkes mediales Aufsehen, denn sie sortierte das Portrait eines afroamerikanischen Pärchens fälschlicherweise der Kategorie „Gorillas" zu [2]. Im selben Jahr erkannten Datenwissenschaftler des Online-Warenhändlers Amazon, dass ihre interne Software zur Selektion von Kandidaten im Bewerbungsprozess keine gender-neutralen Entscheidungen trifft und Frauen gezielt benachteiligt [3]. Beide moralisch verwerflichen Beispiele stehen stellvertretend dafür, dass KI-

Modelle strukturelle Verzerrungen (vgl. Bias) innerhalb der Basisdaten ggf. mitlernen oder die zugrundeliegende Algorithmik per se eine unberechtigte Voreingenommenheit begünstigen kann. Insofern kommt der Fairness von KI speziell im Umgang mit Menschen eine entscheidende Rolle zu, die andernfalls zu Diskriminierung einzelner Personen bis hin zu kategorischer Chancenungleichheit oder Rassismus führen kann. Dass KI noch fatalere Folgen für Menschen mit sich bringen kann, zeigt ein Fall aus dem Jahr 2018. In dieser Zeit führte das Beförderungsunternehmen Uber Testfahrten selbstfahrender Kraftfahrzeuge im US-Bundestaat Arizona durch. Bei einer dieser Fahrten kam es zu einem Verkehrsunfall, bei dem eine Fußgängerin zu Tode kam [4]. Im Zusammenhang mit diesem Unglück bleibt aufgrund ihrer Undurchsichtigkeit rätselhaft, warum die KI-basierte Software des autonomen Fahrzeugs die Passantin nicht rechtzeitig erkannte und sie ungebremst beim Überqueren einer öffentlichen Straße erfasste. Ohne solche Tragödien beschönigen zu wollen, sei an dieser Stelle erlaubt, auftretende Zwischenfälle selbstfahrender Automobile in Relation zu den Millionen gefahrenen Testkilometer zu setzten. Gleiches lässt sich aber auch für weniger kritischere Anwendungsfälle überlegen. Hierbei fällt auf, dass für KI-Modelle ähnliche oder wiederkehrende Situationen, die sie in der Lern- bzw. Trainingsphase gesehen haben, eher unproblematisch sind. Es zeigt sich aber auch, dass gerade neue Situationen zu beliebig schlechten Entscheidungen führen können. Wissenschaftliche Studien zeigen dies deutlich. Beispielsweise sind durch gewissen Störeinflüsse aktuelle Verfahren des Deep-Learning – einem Teilgebiet der KI – nicht in der Lage, eine zuverlässige Einschätzung abzugeben (vgl. [5], [6]). So könnte in der Analogie zum Straßenverkehrsszenario das Anbringen eines Stickers auf einem Stoppschild leicht zu Irritationen seitens der KI führen. Dieses könnte fälschlicherweise als Tempolimit interpretiert werden und demnach fatale Folgen mit sich bringen. In diesem Zusammenhang sind auch gezielte Sabotagen der Umgebung durch Angreifer denkbar, um solche Schwachstellen der KI konsequent auszunutzen und zu ihren Gunsten zu instrumentalisieren.

Die dargelegten Ausführungen und viele weitere Zwischenfälle skizzieren deutlich, dass trotz nahezu 70 Jahren der intensiven Beforschung KI alles andere als perfekt ist. Ein besonderes Augenmerk muss vor allem dann auf sie gerichtet werden, wenn die zugrundeliegenden Anwendungsfälle eine Schnittmenge mit dem Menschen bilden und getroffene Entscheidungen irreversible Konsequenzen mit sich bringen können. Vielen Prognosen zur Folge wird diese Schnittmenge größer und so ist es unerlässlich, dass KI kein Narrativ bleibt, sondern zu einem zuverlässigen Begleiter für den Menschen wird. Hierzu sind gesetzliche Rahmenbedingungen anzustreben, die sowohl die Begrifflichkeit schärfen als auch verhältnismäßige Anforderungen an die Qualität solcher Systeme stellen. Mit dem Verordnungsvorschlag zur Festlegung harmonisierter Vorschriften für KI (kurz AI-Act) [7] und den einschlägigen ANNEXES [8] übernimmt die Europäische Union (EU) eine weltweite Vorreiterrolle, diese Bestrebungen umzusetzen, dessen Geltungsbereich auch weit über die EU-Grenzen reichen soll. Dieses Whitepaper befasst sich inhaltlich mit den wichtigsten Aspekten dieser Regulierung (Kapitel 2), die sich aktuell im Entwurfsstadium befindet und voraussichtlich in diesem Jahr oder 2024 in Kraft treten wird. Darüber hinaus werden Herausforderungen aber auch Chancen aufgezeigt (Kapitel 3), die mit der Einführung der Regulatorik einher gehen. Das Whitepaper schließt mit einem kurzen Fazit, welches abgeleitete Erkenntnisse ebenso benennt (Kapitel 4). Aufgrund der kontinuierlichen Anpassungen des AI-Acts seitens

der EU beziehen sich die hier dargelegten Inhalte auf die ursprüngliche Entwurfsversion sofern nicht anderweitig gekennzeichnet.

2 AI-Act: Kurz und bündig

Aus den politischen Leitlinien [9] für ein koordiniertes europäisches Konzept, welches die menschlichen und ethischen Implikationen von KI korrespondierend berücksichtigt, entstanden bereits während der Kandidatur der heutigen Präsidentin Ursula von der Leyen im Jahr 2019 entscheidende Impulse für einen entsprechenden Gesetzesentwurf. Der AI-Act ist der erste Versuch einer horizontalen Regelung, welcher durch die europäische Kommission im Frühjahr 2021 angestoßen wurde. Diese angestrebte Regulatorik legt nicht nur sektorübergreifend die Rahmenbedingungen für Hersteller und Nutzer von KI fest. Weiträumig adressiert sie Akteure entlang der gesamten KI-Wertschöpfungskette wie Anbieter, Händler und Importeure und richtet ihren Geltungsbereich extraterritorial aus. Ähnlich zu den Bestimmungen der europäischen Datenschutz-Grundverordnung – der General-Data-Protection-Regulation (GDPR) [10] – reiht sie sich somit nahtlos in die Gruppe von EU-Gesetzestexten ein, die ebenfalls außerhalb des Hoheitsgebiets der EU Anwendung findet. Damit allerdings solch ein Rechtsrahmen etabliert werden kann, bedarf es gleichermaßen einer einheitlich scharfen Definition für KI: Als legalen Begriff deklariert der AI-Act demnach all jene Software als KI-Systeme, welche durch Techniken und Konzepte entwickelt wurden, die im begleitenden ANNEX I aufgelistet sind. Darunter fallen insbesondere maschinelle Lernverfahren[1], logik- und wissensbasierte Methoden, sowie statistische Ansätze. Ferner ist diese Software, im Hinblick auf vom Menschen festgelegte Ziele, in der Lage, Ergebnisse wie Inhalte, Vorhersagen, Empfehlungen oder sonstige Entscheidungen zu produzieren, so dass ihre Umgebung, mit der sie interagiert, beeinflusst wird.

Neben dieser breiten Definition basiert der AI-Act auf dem Fundament eines vier-stufigen Risikomodells, welches KI-Systeme in unterschiedliche Klassen einteilt. Von der vierten und damit niedrigsten Risikostufe werden all solche KI-Systeme erfasst, die ein minimales oder gar kein Risiko bezüglich der Schädigung der Gesundheit, der Beeinträchtigung der Sicherheit oder nachteilige Auswirkungen auf die Grundrechte von Personen darstellt. Dazu gehört beispielsweise KI in Videospielen oder Spamfilter. In die dritte Stufe fallen solche KI-Systeme, die ein limitiertes Risiko mit sich bringen. Repräsentanten dieser Stufe sind u.a. Chatbots (vgl. textgenerierende Dialogsysteme) oder Deepfakes (vgl. realistisch wirkende Medieninhalte). Begründung findet diese eingeräumte Risikostufe deswegen, da mit diesen Systemen interagierende bzw. konsumierende Nutzer nicht immer zweifelsfrei beurteilen können, ob es sich um authentische Inhalte oder um solche handelt, die automatisiert erzeugt wurden. Somit unterliegen KI-Systeme in dieser Stufe einer Transparenzanforderung in Form einer Kennzeichnungspflicht. Sie soll offenlegen, dass Nutzer mit synthetischen Inhalten in Berührung kommen und so für etwaigen Manipulationen sensibilisiert werden. In die zweite Stufe sind sog. Hochrisiko-KI-Systeme einzuordnen. Von Systemen in dieser Stufe gehen erhebliche Risiken für die Gesundheit und Sicherheit von natürlichen Personen ebenso wie deren Grundrechte aus. Dementsprechend legt der AI-Act enge Leitplanken in Form von spezifischen Vorschriften für KI-Systeme dieser Art zu Grunde, welche in der Folge dieses Kapitels dediziert, thematisiert werden. Die

[1] An dieser Stelle sei angemerkt, dass alle geschilderten Vorfälle aus Kapitel 1 diesen Verfahren zuzuordnen werden können und folglich unter die Definition von KI des AI-Acts fallen.

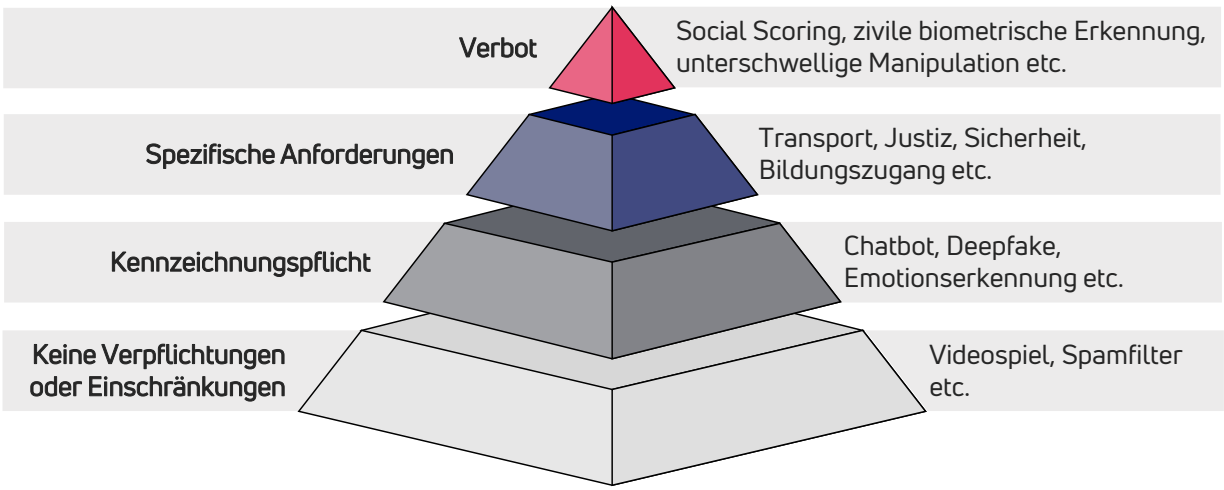

Abbildung 1: Vorgesehene Modellierung von KI-Systemen nach Risikostufen

erste und letzte Stufe des Risikomodells umfasst KI-Systeme und -Praktiken mit unannehmbarem Risiko. Dazu zählen konkret Techniken der unterschwelligen Einflussnahme und solche, die Schwächen bestimmter schutzbedürftiger Gruppen (insbesondere Kinder u. Personen mit Behinderung) ausnutzen. Ferner gehören zu dieser Stufe behördlich angeordnete Bewertungen von natürlichen Personen durch Analysen von sozialem Verhalten oder persönlichen Merkmalen (vgl. Social-Scoring) und die kategorische biometrische Überwachung im öffentlichen Raum zum Zwecke der Strafverfolgung. Eine grafische Darstellung zu diesen Risikostufen ist in Abbildung 1 dargelegt.

Aus dieser Risikomodellierung, welche den fundamentalen Werten der EU unterliegt, lässt sich ableiten, dass KI-Systeme der ersten Stufe ein inakzeptables Risiko darstellen. Als Folge dessen ist deren Inverkehrbringung ausdrücklich untersagt. Systeme der zweiten bis vierten Stufe dürfen mit keinen oder gewissen Auflagen in Verkehr gebracht werden, wobei KI der zweiten Stufe am kritischsten zu bewerten ist. Aus diesem Grund stehen genau diese Systeme im Fokus des AI-Acts. Sie lassen sich dann in die Stufe der Hochrisiko-KI einordnen, wenn sie als Sicherheitskomponente in einem bereits regulierten Produkt (vgl. ANNEX II) verwendet werden, oder selbst ein solches darstellen. Überdies schildert der begleitende ANNEX III weitere Anwendungsfelder für Hochrisiko-KI-Systeme. Darunter fallen der Betrieb und das Management von kritischen Infrastrukturen (KRITIS), die biometrische Erkennung von Personen, das Personal-, Kredit- und Bildungswesen, der Bereich Migration, Asyl und Grenzkontrolle sowie die Justiz, die Strafverfolgung und der Zugang zu essenziellen privaten und öffentlichen Diensten. Ist demnach eine KI als hochriskant anzusehen, greifen die Bestimmungen für Mindestanforderungen des AI-Acts. Thematisch lassen sie sich in sechs wichtige Anforderungskerngruppen (AKGs) einteilen, auf die im Weiteren näher eingegangen wird:

- Risikomanagementsystem
- Daten-Governance u. -Management
- Technische Dokumentation
- Aufzeichnungspflicht
- Transparenz u. menschliche Aufsicht
- Genauigkeit, Robustheit sowie Cybersicherheit

Um sämtliche Risiken über den gesamten Lebenszyklus zu erfassen, bedarf es eines kontinuierlichen iterativen Prozesses – dem sog. *Risikomanagementsystem*. Mittels dieses Sys-

tems werden alle offensichtlichen und abschätzbaren Risiken inventarisiert, bewertet und geeigneten Managementmaßnahmen zugeordnet, die dem aktuellen Stand der Technik entsprechen. Dabei sind neue unbekannte Risiken, die während des Betriebs der KI auftreten, gleichermaßen zu berücksichtigen und systematisch in diesem Risikomanagementsystem zu pflegen. Des Weiteren besteht für Hochrisiko-KI-Systeme das Erfordernis, Trainings-, Validierungs- und Testdatensätze zu verwenden, sofern es sich um Systeme mit antrainierten KI-Modellen handelt. Für sie müssen einschlägige Verfahren des *Daten-Governance und Daten-Management* angewandt werden, mit dem Zweck eine möglichst gute Qualität der Datenbasis von Beginn an zu forcieren. Darunter fällt u.a. die Protokollierung getätigter Datenvorverarbeitungsschritte (vgl. Bereinigung, Anreicherung und Aggregierung), sowie eine Eignungsprüfung der verwendeten Daten, auch im Hinblick auf möglichen Bias oder sonstiger Mängel. Im direkten Zusammenhang damit steht, dass die verwendeten Daten relevant, repräsentativ, vollständig und frei von Fehlern sein müssen. Da auch KI-Systeme ohne Trainingsphase auf Daten basieren, ist ein geeignetes Daten-Governance und Daten-Management ebenfalls vorzusehen. Für sie gelten aber nur eine Teilmenge der Anforderungen. Nebst diesen Bestimmungen zur Sicherung der Datenqualität fordert der AI-Act weiter eine *technische Dokumentation*. Die Vorgabe zu Inhalt und Aufbau dieser AKG sind in ANNEX IV dargelegt, so dass letztlich aus diesem Dokument auditierungsfähig hervorgehen muss, dass alle notwendigen Maßnahmen für Hochrisiko-KI umgesetzt werden. Die *Aufzeichnungspflicht* bildet eine weitere wichtige technische Säule, denn sie legt den Grundstein für eine detaillierte Protokollierungsfunktionalität, die alle KI-Systeme der zweiten Risikostufe umsetzen müssen. So kann eine Überwachung des KI-Systems nach Inbetriebnahme mittels anerkannter Normen oder üblicher Spezifikationen erfolgen, um eine nahtlose Nachvollziehbarkeit des Systemverhaltens zu erwirken, was ebenfalls eine Forderung seitens des AI-Acts darstellt. Hierfür sind geeignete Monitoring-Pläne in der technischen Dokumentation zu hinterlegen. Mit der Nachvollziehbarkeit eng verknüpft ist *Transparenz und menschliche Aufsicht*. Während bei der Transparenz das Hauptaugenmerk auf der Interpretierbarkeit von erzeugten Ergebnissen und der Anfertigung eines Nutzerleitfadens liegt, sind für die menschliche Aufsicht geeignete Maßnahmen (vgl. Mensch-Maschine-Schnittstelle) zu realisieren. Daraus verspricht sich die EU, KI im laufenden Betrieb wirksam beaufsichtigen zu können und etwaige Risiken für die Gesundheit, Sicherheit oder Grundrechte minimieren oder gar verhindern werden können. Hierzu müssen Personen, denen die menschliche Aufsicht übertragen wurden, in die Lage versetzt werden, in den Betrieb eingreifen zu können, um beispielsweise bei kritischem Systemverhalten mit einer sog. „Stopptaste" oder ähnlichen Verfahren den Betrieb zu unterbrechen. In die letzte AKG fallen die Themenbereiche *Genauigkeit, Robustheit sowie Cybersicherheit*. Hier werden erwartungsgemäß bezüglich der Genauigkeit keine konkreten Schwellwerte angegeben, welche das KI-System nicht unterschreiten darf, wohl aber müssen die Genauigkeitskennzahlen im Nutzerleitfaden hinterlegt sein und dem aktuellen Stand der Technik entsprechen. Weiter wird von der KI erwartet, dass sie robust gegenüber Fehlern, Störungen oder Unstimmigkeiten ist. Als Maßnahme zur Gewährleistung der Robustheit führt der AI-Act konkret technische Redundanzen an, was auch Sicherungs- oder Störungssicherheitspläne beinhalten kann. Ebenso muss bei KI, welche einer kontinuierlichen Lernphase auch nach der Inverkehrbringung unterliegt, darauf geachtet werden, geeignete Risikominderungsmaßnahmen vorzusehen. Hinsichtlich Cybersicherheit

Abbildung 2: Hochrisiko-KI - Anwendungsfelder und Anforderungen

ist gefordert, dass das Hochrisiko-KI-System eine Widerstandsfähigkeit gegenüber Angriffsversuchen unbefugter Dritter besitzt, die darauf abzielen, die Zweckbestimmung des Systems durch Ausnutzung von Schwachstellen zu untergraben. Darunter fallen u.a. gängige Angriffsvektoren wie das Data-Poisoning, Adversarial-Attacks oder die Ausnutzung sonstiger in KI-Modellen enthaltener Mängel zu Gunsten des Angreifers. Für einen Überblick über die Anwendungsfelder als auch die zu erfüllenden AKGs seitens Hochrisiko-KI-Systeme sei an dieser Stelle auf Abbildung 2 verwiesen.

Zur Überprüfung der geschilderten AKGs für Hochrisiko-KI-Systeme schreibt der AI-Act verpflichtend das Durchlaufen einer Konformitätsbewertung für Anbieter und Importeure[2] vor. In Abhängigkeit zum jeweiligen Anwendungsfeld kann im Wesentlichen zwischen zwei Verfahren unterschieden werden: die Selbstkontrolle nach ANNEX VI und die externe Auditierung wie in ANNEX VII dargelegt. Beide Konformitätsbewertungsverfahren basieren auf einem sog. Qualitätsmanagementsystem. Es fungiert als umschließendes Element für die Bewertung und beinhaltet im Kern ein Konzept und Kontrollmechanismen zur Einhaltung der gesetzlichen Vorgaben, so dass diese lückenlos während des gesamten Lebenszyklus des KI-Systems dokumentiert sind. Darüber hinaus verpflichten sich Anbieter, bei der Selbstkontrolle alle in der technischen Dokumentation ausgeführten Informationen gegen die auferlegten Anforderungen ihrer Hochrisiko-KI zu prüfen, und den Entwicklungsprozess sowie die Überwachung im Sinne der Aufzeichnungspflicht nach Inbetriebnahme konform der Dokumentation umzusetzen. Bei der externen Auditierung kommt das Konstrukt der sog. notifizierten Stellen[3] zum Einsatz. Sie werden von der jeweiligen zuständigen nationalen Behörde ermächtigt und bewerten die Konformität einer Hochrisiko-KI unabhängig vom Anbieter. Dazu kann ggf. die Offenlegung von Trainings-, Validierungs- und Testdaten oder sogar des Quellcodes eingefordert werden. Ebenso besteht eine Mitteilungspflicht seitens des Anbieters oder stellvertretender Instanzen über Änderungen am Qualitätsmanagementsystem. Geht aus dem Bewertungsverfahren nach ANNEX VI eine Konformität hervor, erstellt der Anbieter eine bindende EU-Konformitätserklärung. Im Falle einer positiven Bewertung nach ANNEX VII erstellt die notifizierte Stelle fernhin eine Konformitätsbescheinigung. So kann im EU-Binnenmarkt mit dem System frei verkehrt

[2] Speziell ist mit Importeur jede in der Union ansässige natürliche oder juristische Person gemeint, die ein KI-System auf dem EU-Markt platziert oder einsetzt, welches den Namen oder die Marke einer außerhalb der Union ansässigen natürlichen oder juristischen Person trägt.

[3] In Ausnahmefällen übernimmt die zuständige Marktüberwachungsbehörde die Tätigkeit der notifizierten Stelle.

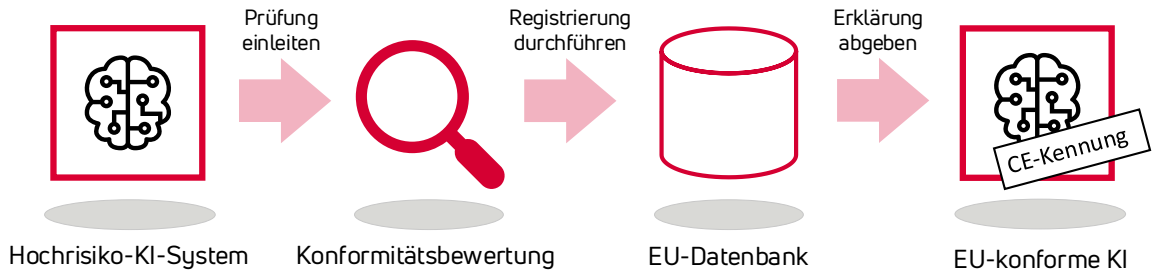

Abbildung 3: Schematische Prozessdarstellung hin zu einem EU-konformen KI-System

werden. In Analogie zur bekannten CE-Kennzeichnung [11], welche auf sämtlichen in der EU zugelassenen Produkten zu finden ist, ist ebenfalls angestrebt, diese Konformitätskennzeichnung für das jeweilige KI-System zu etablieren. Jedoch muss vor jedem Inverkehrbringen oder Inbetriebnahme das Hochrisiko-KI-System in einem unionsweiten, öffentlich zugänglichen Register – der sog. EU-Datenbank – festgehalten werden, mit der EU-Kommission als verantwortliche Stelle. Der Ablauf von einer Hochrisiko-KI hin zu einer EU-konformen und damit vertrauenswürdigen KI im Sinne europäischer Grundsätze ist schematisch durch Abbildung 3 beschrieben.

3 Herausforderungen und Chancen

Mit der Herausgabe des AI-Acts wird der Weg hin zu vertrauenswürdigen KI-Systemen innerhalb der EU geebnet. Doch dafür müssen essenzielle Hürden vor bzw. nach dem Inkrafttreten des AI-Acts überwunden werden. Darüber hinaus ergeben sich auch inhärente Chancen aus dem Gesetzestext.

Unmittelbare Auswirkungen werden vor allem Entwickler bzw. Hersteller und Drittanbieter von KI-Systemen spüren, da der AI-Act hauptsächlich auf sie abzielt. Äußerst ungünstig ist momentan noch die Definition von „KI" im Gesetzesentwurf. Durch die breitgefächerte Auslegung birgt sie die Gefahr, dass auch fortschrittliche Software, die nach jetzigem Stand nicht auf KI-Algorithmen basiert, trotzdem zu den KI-Systemen gezählt wird [12]. Beispielsweise kann es sich dabei um Berechnungsprogramme in Banken handeln, welche auf einem statistischen Ansatz beruhen und in Echtzeit die Einlagen- und Kreditgeschäfte erledigen. Folglich würden Unternehmen oder Organisationen, bei denen solche Software verankert ist, dadurch auch unter die Regulatorik des AI-Acts fallen. Dies verlangt nach einer sehr aufwändigen Inventarisierung von Bestandssystemen, was offenkundig ungeplante Ressourcen bindet. Vor diesem Hintergrund besteht bereits jetzt der Drang, sich mit dem AI-Act und den darin beschriebenen Anforderungen an KI zu befassen, um nicht den Anschluss zu verlieren, rechtzeitig EU-konform zu sein [13]. Auch hier müssen Aufwände vorgehalten werden, um sich auf ein Konformitätsbewertungsverfahren vorzubereiten und dieses im Anschluss auch zu durchschreiten. Wie hoch diese Aufwände ausfallen werden, ist jedoch schwer greifbar insbesondere im Kontext von Bestandssystemen. Hürden zur Erfüllung dürften hierbei spürbar hoher ausfallen, da die dargelegten Auflagen bereits beim Entwurf hätten berücksichtigt werden müssen, und so nur sehr mühsam eine Konformität erreicht werden kann. Darüber hinaus geht dieser Kraftakt einher mit nicht zu vernachlässigenden Auswirkungen querschnittlicher Natur. Es ist zu erwarten, dass für Bestandssysteme und Neuentwicklungen sowohl Großunternehmen als auch kleine und mittlere

Unternehmen (KMUs) betreffende Prozessstrukturen, welche mitunter tief im Unternehmen verwurzelt sein können, angepasst oder gänzlich neu überdacht werden müssen. Im Sinne einer ganzheitlichen Analyse der Wertschöpfungskette schließt dies Lieferanten, Kunden, Entwickler und andere interne sowie externe Stakeholder gleichermaßen mit ein. Die Prozesslandkarte muss zudem unter dem Aspekt der Nachhaltigkeit betrachtet werden, denn ausgestellte Konformitätsbescheinigungen seitens der notifizierten Stellen haben nur eine Gültigkeit von fünf Jahren.

Neben dieser Neuausrichtung und den damit verknüpften strategischen Überlegungen ist eine der größten Herausforderungen bei der Umsetzung des AI-Acts das Fehlen von geeigneten Prüfmethoden für KI-Systeme. Es existieren jenseits der klassischen Software bis dato nur wenige harmonisierten Normen für KI-Systeme. Dies hat zur Folge, dass KI-Systeme weithin nicht eingesetzt oder weiterentwickelt werden können. Der AI-Act macht keine konkreten Angaben darüber, wie und nach welchen Kriterien auferlegte Anforderungen getestet werden sollen. Somit wird neben dem Gesetzestext eine Grundlage für eine Etablierung generalisierbarer Regeln und anwendungsspezifischer Normen gefordert, die eine Konformitätsbewertung erst ermöglichen [14]. Zu dieser Ungewissheit kommen Strafankündigungen hinzu. Bei Nichtkonformität werden abhängig vom Schweregrad des Vergehens administrative Strafen verhängt, die mitunter sensibel hoch ausfallen können. Es drohen Geldbußen von bis zu 30 Millionen Euro oder von bis zu sechs Prozent des gesamten Jahresumsatzes des vorangegangenen Geschäftsjahres im Falle von Unternehmen. Dies betrifft auch Importeure und Bevollmächtigte von Anbietern. Sie haften genauso für die eingeführten KI-Systeme und müssen auf die Einhaltung der Konformität achten. Auch wenn diese Strafen – so der AI-Act – „wirksam, verhältnismäßig und abschreckend" sein sollen, sorgt die Kombination beider Faktoren für einen lähmenden Schlag in Bezug auf die technologische Wettbewerbsfähigkeit von Europa. Tangiert wird davon gleichermaßen die akademische Forschung, welche maßgebend zu dem Erfolg von KI bis heute beiträgt, was zusätzlich die Innovationsfähigkeit in ihrer Breite schmälern dürfte. Um diesen potenziellen Negativtrend zu antizipieren, stellt der AI-Act das Konzept der KI-Reallabore dem entgegen. Sie sollen die Erprobung innovativer KI-Systeme innerhalb kontrollierter Umgebungen ermöglichen und stehen unter direkter Aufsicht der zuständigen Behörden. Besonders KMUs und Startups würden von einem Zugangsvorrang profitieren, um nicht den Anschluss an großen und etablierten Unternehmen zu verlieren. Die Hauptidee eines solchen Reallabors, welches u.a. in einigen europäischen Ländern erfolgreich zum Einsatz gekommen ist, besteht darin, einen regulatorischen Rahmen für junge und innovative Unternehmen zu schaffen, unter dem eine Entwicklung des Unternehmens hin zu einem vollregulierten Marktteilnehmer möglich wird [15]. Das Problem dabei ist, dass bis heute keine allgemein anerkannte Definition für Reallabore besteht. Ein Großteil der vorhandenen Reallaborprojekte konzentriert sich auf den traditionell sehr stark regulierten Finanzbereich. Für KI im Speziellen hat etwas in dieser Form noch nicht existiert und so gibt es durchaus auch kritische Haltungen hinsichtlich der Etablierung von Reallaboren [16].

Insgesamt eröffnet der AI-Act aber auch neue Türen für Europa. Eine transparente Definition, eine kontextualisierte Einordnung in Risikostufen sowie eine einheitliche Aufklärung über KI werden zu mehr Verständnis und Vertrauen innerhalb der europäischen Staatengemeinschaft beitragen. Das ist äußerst förderlich und ein entscheidender Impuls für die Beschleunigung der digitalen Transformation, denn nach

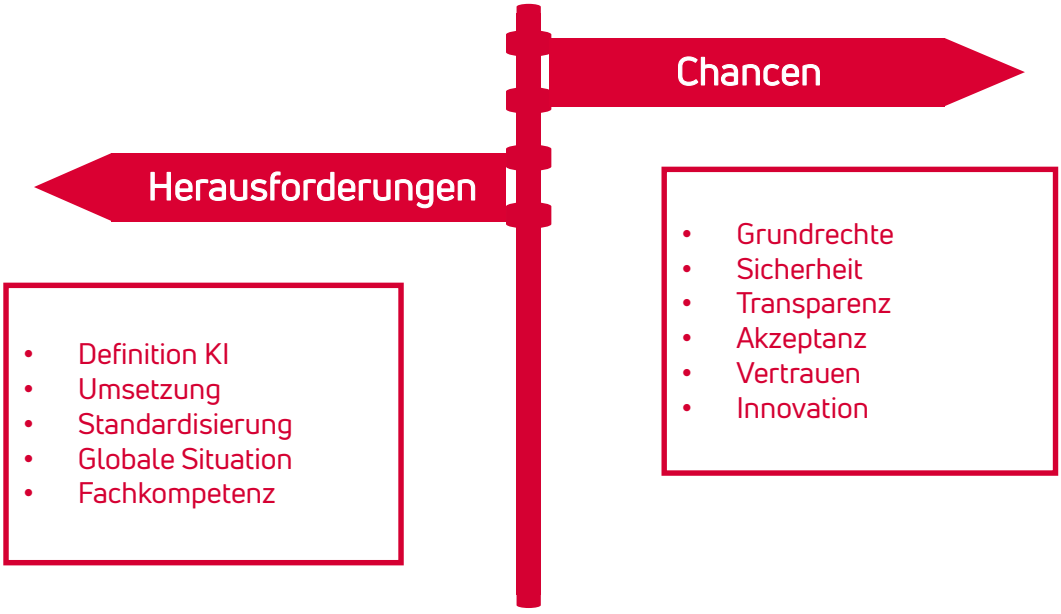

Abbildung 4: Herausforderungen und Chancen durch den AI-Act

einer vom europäischen Parlament in Auftrag gegebenen Studie beschränkte sich der Anteil aller Unternehmen in der EU, welche KI-Technologien aktiv fördern, auf 26 Prozent [17]. In anderen Staaten ist dieser Anteil deutlich höher, wie am Beispiel China deutlich wird. Dort investieren 53 Prozent der Unternehmen laut dieser Studie in KI. Auch für Deutschland gibt es Chancen. Immerhin erkennen laut einer Umfrage des Industrieverbandes bitkom inzwischen ca. 70 Prozent der deutschen Unternehmen die Wichtigkeit von KI als Schlüsseltechnologie für ihre eigene Zukunft [18]. Leider nutzen oder beschäftigen sich aber nur 37 Prozent aller Unternehmen mit KI [19]. Als Folgen der Corona-Pandemie und dem Ukraine-Krieg sahen sich auch viele deutsche Unternehmen gezwungen in den Krisenmodus zu schalten und das Thema KI auf ihrer Prioritätsliste nach unten zu verlagern. Steigende Energiekosten, hohe Inflationsraten und unterbrochene Lieferketten geben nach einer weiteren bitkom-Umfrage wenig Spielraum für neue Technologien und Geschäftsmodelle [20]. Mehr als drei Viertel aller befragten deutschen Unternehmen fordern daher mehr finanzielle Förderungen. Vielen würde zudem eine Kollaboration mit anderen Unternehmen helfen, die schon tiefergehende Erfahrungen im Umgang mit KI gesammelt haben. Wünschenswert bleiben dabei – so die Umfrage – eine bessere Verfügbarkeit von Fachpersonal auf dem Arbeitsmarkt, bessere Informationen über marktfähige KI-Systeme und einfachere Zugriffskanäle auf Daten. Die Initiative des AI-Acts könnte die Erfüllung dieser Faktoren begünstigen, doch es muss darauf geachtet werden, dass zu viele Regulierungen nicht diese Zugänge behindern. Es besteht zudem noch Unklarheit darüber, welche Behörden in Deutschland die Umsetzung des AI-Acts federführend beaufsichtigen und koordinieren werden. Es ist nicht ausgeschlossen, dass dafür möglicherweise sogar eine völlig neue Instanz ins Leben gerufen werden kann.

Diese diskutierten Perspektiven erhebt keinen Anspruch auf Vollständigkeit. Sie geben jedoch wichtige Hinweise darauf, was bei der Umsetzung von vertrauenswürdiger KI auf Unternehmen und Organisationen zu kommen kann und welches Potential damit entsteht. Diese und weitere wichtige Punkte sind in Abbildung 4 hervorgehoben.

4 Fazit und Ausblick

Die Herausgabe des AI-Acts durch die EU, gepaart mit ihrer eigenen Führungsposition in den Themen Robotik oder Fertigung sowie Dienstleistung [21], ist sicherlich ein willkommener Schritt in Richtung der Regulierung für KI-Systeme, um die technologische Unabhängigkeit Europas weiter auszubauen und eine Vertrauensbasis in zukunftsorientierte KI-gestützte Anwendungen zu etablieren. Trotz dieser guten Initiative ist das Fundament allerdings noch nicht stabil genug. An vielen Stellen des AI-Acts besteht noch Verbesserungsbedarf. Der Begriff „KI" umfasst noch zu viele Systeme und Algorithmen, welche bereits weitläufig im Umlauf sind und verwendet werden. Es existiert derzeit noch eine starke Verunsicherung, ob überhaupt, und wenn ja, welche Systeme unter die Regulierung des AI-Acts fallen. Die Frage bleibt daher offen, inwiefern sie einer Konformitätsüberprüfung unterziehen werden müssen, um sich nicht unwissentlich strafbar zu machen. Der AI-Act sollte außerdem noch auf die Nutzungsweise von KI-Systemen tiefgreifender eingehen, wie etwa Kontext und Art des Einsatzes des Systems sowie auf die Verpflichtungen seitens der Nutzer und Kunden, und sich nicht nur auf die Systeme selbst und ihre Anbieter konzentrieren.

Es ist essenziell zu erkennen, dass KI spätestens mit dem Erscheinen des AI-Acts nicht wie übliche Software behandelt werden kann, welche bedenkenlos entwickelt wird und funktioniert. Ähnlich wie bei der IT-Sicherheit und der Einführung der GDPR muss KI innerhalb eines Unternehmens oder einer Organisation als Wert aufgefasst werden. Dies erfordert auch, dass entsprechende Rahmenbedingungen in Form einer internen KI-Strategie geschaffen werden, um ein einheitliches Bewusstsein für KI zu entwickeln. Zur Strategie gehören vor allem ein Grundverständnis der Mitarbeiter zur Gesamtthematik sowie die Sensibilisierung im Umgang mit KI – beispielsweise in Form von Awareness-Trainings. Somit ist ein frühzeitiges Handeln unabdingbar. Die Entwicklungen des AI-Acts sollten weiterhin mitverfolgt werden, damit die Chance nicht verpasst wird, sich rechtzeitig und selbstbewusst auf dem KI-Markt zu positionieren.

infodas als verlässlicher Partner bei der Umsetzung vertrauenswürdiger KI

Mit langjähriger Erfahrung in IT Consulting und IT Security Consulting im behördlichen, militärischen sowie wirtschaftlichen Umfeld, bietet infodas GmbH die notwendige Kompetenz, um Sie beim Anforderungsmanagement und bei sicherheitsrelevanten Fragestellungen hinsichtlich eines geplanten oder sich in der Entwicklung befindlichen KI-Systems zu begleiten. Darüber hinaus unterstützen wir Sie mit unserem breiten Technologiewissen über KI, welches unsere Datenwissenschaftler und -ingenieure seit Jahren sehr erfolgreich in datengetriebenen Projekten aufgebaut haben. So können praxisnah die nötigen Maßnahmen ergriffen werden, um vertrauenswürdige KI zu entwickeln und zu etablieren, welche die Erwartungshaltung seitens der EU – und vor allem Ihre – nachhaltig erfüllt.

5 Literaturverzeichnis

[1] A. Rao und G. Verweij, „Sizing the prize - What's the real value of AI for your business and how can you capitalise?," PWC, 2017. [Online]. Available: https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf. [Zugriff am 21. November 2022].

[2] „Google apologises for Photos app's racist blunder," BBC, 1. Juli 2015. [Online]. Available: https://www.bbc.com/news/technology-33347866. [Zugriff am 21. November 2022].

[3] J. Dastin, „Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, 11. Oktober 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. [Zugriff am 21. November 2022].

[4] „Driver Charged in Uber's Fatal 2018 Autonomous Car Crash," The New York Times, 15. September 2020. [Online]. Available: https://www.nytimes.com/2020/09/15/technology/uber-autonomous-crash-driver-charged.html. [Zugriff am 21. November 2022].

[5] X. Yuan, P. He, Q. Zhu und L. Xiaolin, „Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Transactions on Neural Networks and Learning Systems,* pp. 2805-2824, 13. Januar 2019.

[6] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno und D. Song, „Robust Physical-World Attacks on Deep Learning Visual Classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City (UT), USA, 2018.

[7] „Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS," European Commission, 21. April 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. [Zugriff am 21. November 2022].

[8] „The AI Act Annexes," European Commission, 21. April 2021. [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_2&format=PDF. [Zugriff am 21. November 2022].

[9] U. von der Leyen, „A Union that strives for more," 09. Oktober 2019. [Online]. Available: https://ec.europa.eu/info/sites/default/files/political-guidelines-next-commission_en_0.pdf. [Zugriff am 21. November 2022].

[10] „GDPR," EU Parlament und EU Rat, 27. April 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. [Zugriff am 02. Januar 2023].

[11] „Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93," European Union, 09. Juli 2008. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008R0765. [Zugriff am 27. Januar 2023].

[12] Bitkom Bundesverband Informationswirtschaft,

Telekommunikation und Neue Medien e.V., „Position paper: Bitkom principles for the Artificial Intelligence (AI) Act," 04. August 2021. [Online]. Available: https://www.bitkom.org/sites/main/files/2021-08/2021august_bitkomposition_aiact.pdf. [Zugriff am 21. November 2022].

[13] C. Schwerdt und K. Ellert, „Die Kraft von Weltmodellen, der AI Act und die Notwendigkeit von KI-Compliance," Tagesspiegel, 14. November 2022. [Online]. Available: https://background.tagesspiegel.de/digitalisierung/die-kraft-von-weltmodellen-der-ai-act-und-die-notwendigkeit-von-ki-compliance. [Zugriff am 21. November 2022].

[14] Gesellschaft für Informatik e.V. (GI), „Arbeitspapier | KI-Regulierung made in Europe: Positionen zum Gesetzentwurf der Euroäischen Kommission," Berlin, 2021.

[15] N. Eberle, „Die „Regulatory Sandbox" - (K)ein Modell für Deutschland?," 29. Juni 2020. [Online]. Available: https://lrz.legal/images//pdf/Die_Regulatory_Sandbox.pdf. [Zugriff am 16. Januar 2023].

[16] V. Ritter-Döring und J. Gröning, „Innovation trifft Regulierung: Ein Sandkasten für Künstliche Intelligenz (KI)," 17. November 2021. [Online]. Available: https://www.taylorwessing.com/de/interface/2021/ai-act/innovation-meets-regulation-a-sandbox-for-artificial-intelligence-ai. [Zugriff am 16. Januar 2023].

[17] J. Eager, M. Whittle, J. Smit, G. Cacciaguerra und E. Lale-Demoz, „Opportunities of Artificial Intelligence," Juni 2020. [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652713/IPOL_STU(2020)652713_EN.pdf. [Zugriff am 21. November 2022].

[18] A. Streim und M. Uhl, „Künstliche Intelligenz kommt in Unternehmen allmählich voran," Bitkom, 21. April 2022. [Online]. Available: https://www.bitkom.org/Presse/Presseinformation/Kuenstliche-Intelligenz-kommt-in-Unternehmen-allmaehlich-voran. [Zugriff am 21. November 2022].

[19] A. Streim und C. Meinecke, „Dämpfer für die Digitalisierung: Weltlage bremst digitale Transformation der Wirtschaft," Bitkom, 20. Juni 2022. [Online]. Available: https://www.bitkom.org/Presse/Presseinformation/Daempfer-Digitalisierung-Weltlage-bremst-digitale-Transformation-Wirtschaft. [Zugriff am 21. November 2022].

[20] A. Streim und M. Uhl, „KI gilt in der deutschen Wirtschaft als Zukunftstechnologie - wird aber selten genutzt," Bitkom, 13. September 2022. [Online]. Available: https://www.bitkom.org/Presse/Presseinformation/Kuenstliche-Intelligenz-2022. [Zugriff am 25. November 2022].

[21] „Weissbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen," Europäische Kommission, 19. Februar 2020. [Online]. Available: https://commission.europa.eu/document/d2ec4039-c5be-423a-81ef-b9e44e79825b_de. [Zugriff am 21. November 2022].