Vertrauenswürdige KI
Reihenherausgeber Jordan Pötsch

Jordan Pötsch
Rainer Bernnat

Regulierung von Künstlicher Intelligenz in der EU

Praxisbezogene Lösungsansätze für die Sicherheit von KI-Anwendungen

Springer Vieweg

Vertrauenswürdige KI

Reihe herausgegeben von

Jordan Pötsch, Hattingen, Deutschland

English

Trustworthy AI is essential for ensuring ethical, transparent, and secure decision-making that fosters societal trust and acceptance. As AI becomes more widely deployed in business, implementing trustworthy AI often faces numerous challenges, such as high costs, complex compliance requirements, skill shortages, and competitive pressure. This book series thoroughly examines the challenges and solutions related to trustworthy AI. It analyzes AI-relevant legislation, such as the EU AI Act, and compares regional differences in AI regulation in terms of cultural, political, and economic motivations. While compliance requirements may initially lead to additional effort, organizations should view them as an opportunity to drive the upcoming AI transformation purposefully and build upon a solid governance structure. To effectively implement organizational aspects such as AI governance coordination, risk management, and compliance according to the Three Lines of Defense model, standards like ISO/IEC 42001 are analyzed, compared, and useful best practices for implementation are provided. A key challenge is the technical implementation of trustworthiness dimensions throughout the entire lifecycle, such as fairness, transparency, explainability, reliability, data protection, and sustainability. The series addresses the operationalization of application-specific trustworthiness requirements, the technical quality, and the practical implementation of the desired system characteristics. It also presents methods for auditing and certifying AI products. Our authors are leading experts from academia and industry, providing not only theoretical foundational knowledge but also practical guidelines and case studies. The target audience for the series includes students, AI specialists, and C-level decision-makers.

Deutsch

Vertrauenswürdige KI ist entscheidend, um ethische, transparente und sichere Entscheidungsprozesse zu gewährleisten, die gesellschaftliches Vertrauen und Akzeptanz fördern. Mit der zunehmenden Verbreitung von KI im geschäftlichen Umfeld stehen Unternehmen jedoch häufig vor Herausforderungen bei der Umsetzung vertrauenswürdiger KI. Dazu zählen hohe Kosten, komplexe Compliance-Anforderungen, Fachkräftemangel und Wettbewerbsdruck. Diese Buchreihe untersucht ausführlich die Herausforderungen und Lösungsansätze im Zusammenhang mit vertrauenswürdiger KI. Sie analysiert relevante Gesetzgebungen wie den EU AI Act und vergleicht regionale Unterschiede in der KI-Regulierung hinsichtlich kultureller, politischer und wirtschaftlicher Beweggründe. Auch wenn Compliance-Anforderungen anfänglich zusätzlichen Aufwand bedeuten, sollten Unternehmen diese als Chance betrachten, die bevorstehende KI-Transformation gezielt voranzutreiben und auf einer soliden Governance-Struktur aufzubauen. Um organisatorische Aspekte wie die Koordination der KI-Governance, das Risikomanagement und die Compliance gemäß dem „Three Lines of Defense“-Modell effektiv umzusetzen, werden Standards wie ISO/IEC 42001 analysiert, verglichen und nützliche Best Practices zur Implementierung vorgestellt. Eine zentrale Herausforderung ist die technische Umsetzung der Vertrauenswürdigkeitsdimensionen über den gesamten Lebenszyklus hinweg, wie Fairness, Transparenz, Erklärbarkeit, Zuverlässigkeit, Datenschutz und Nachhaltigkeit. Die Reihe behandelt die Operationalisierung anwendungsspezifischer Vertrauenswürdigkeitsanforderungen, die technische Qualität und die praktische Umsetzung der gewünschten Systemeigenschaften. Darüber hinaus werden Methoden zur Prüfung und Zertifizierung von KI-Produkten vorgestellt. Unsere Autoren sind führende Experten aus Wissenschaft und Industrie und bieten nicht nur theoretisches Grundlagenwissen, sondern auch praxisnahe Leitlinien und Fallstudien. Die Zielgruppe dieser Buchreihe umfasst Studierende, KI-Spezialisten und Entscheidungsträger auf C-Level.

Jordan Pötsch • Rainer Bernnat

Regulierung von Künstlicher Intelligenz in der EU

Praxisbezogene Lösungsansätze für die Sicherheit von KI-Anwendungen

Springer Vieweg

Jordan Pötsch
Hattingen, Deutschland

Rainer Bernnat
Königstein im Taunus, Deutschland

ISSN 3059-2518 ISSN 3059-2526 (electronic)
Vertrauenswürdige KI
ISBN 978-3-658-46748-7 ISBN 978-3-658-46749-4 (eBook)
https://doi.org/10.1007/978-3-658-46749-4

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://portal.dnb.de abrufbar.

Planung/Lektorat: Petra Steinmueller
Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Vorwort

Liebe Leserschaft,

mit dem ersten Band „Regulierung von Künstlicher Intelligenz in der EU – Praxisbezogene Lösungsansätze für die Sicherheit von KI-Anwendungen“ der Buchreihe „Vertrauenswürdige KI“ wird gemäß dem idealtypischen Leibniz'sche Leitbild „Theoria cum praxi“ fundiertes Grundlagenwissen mit praxisorientierten Lösungsansätzen zur Regulierung von KI verbunden.

Erstmals werden KI-relevante Gesetze und Standards nach einer eigens entwickelten Systematik in Beziehung zueinander gesetzt. Diese Systematik ordnet Standards anhand definierter Kriterien und bewertet sie nach ihrer Bedeutung für die Umsetzung der Vorgaben des EU AI Acts. Institutionen stehen bei der Operationalisierung von vertrauenswürdiger KI sektorübergreifend vor vielen Herausforderungen, die maßgeblich den Erfolg eines KI-Projekts beeinflussen. Die vorgestellte Systematik dient hierbei als roter Leitfaden entlang der drei Ebenen einer KI-Governance-Struktur einer Institution: von verbindlichen Compliance-Vorgaben, über organisatorische Aspekte wie Koordination der KI-Governance, dem Risikomanagement und der Compliance entlang dem Three-Lines-of-Defence-Modell, bis hin zur technischen Kontrolle entlang des KI-Lebenszyklus. Somit richtet sich das Buch vor allem an Führungskräfte in Institutionen, insbesondere Entscheidungsträger im Risikomanagement wie der CISO, aber auch an KI-Governance Verantwortliche, Cybersecurity-Experten und Studierende.

Der EU AI Act verpflichtet die Mitgliedstaaten, bis August 2025 nationale Aufsichtsbehörden zu benennen. Dieses Buch bietet Vorschläge zur Umsetzung und Konkretisierung der Vorgaben des AI Acts durch das BSI. Da es nicht darum geht, das Rad neu zu erfinden, liegt der Fokus auf der Ergänzung und Erweiterung bestehender Frameworks wie dem BSI-IT-Grundschutz. Um Synergieeffekte zu erzielen, wird das Risikomanagement von KI sinnvollerweise auf bestehende Prozesse der ISO/IEC 27001 aufgebaut. Die CISO-Organisation stellt hierbei ein zentraler Ausgangspunkt dar, um KI sicher in Unternehmen einzuführen. Erstmals werden in diesem Buch drei Kategorien definiert, die die Rolle von KI in der Cybersecurity umfassend beleuchten. Darüber hinaus wird die Schnittstelle zwischen den Managementsystemen ISMS nach ISO/IEC 27001 und AIMS nach ISO/IEC 42001 konkretisiert. Das Buch veranschaulicht anhand einer Fallstudie, wie eine

CISO-KI-Initiative gestaltet werden kann. Es zeigt auf, welche Anpassungen die CISO-Organisation an ihren Services vornehmen sollte, um das Zusammenspiel von Künstlicher Intelligenz und IT-Sicherheit umfassend zu berücksichtigen.

Die Autorenschaft ist sich der hochdynamischen Entwicklungen im Bereich KI bewusst. Um die Halbwertszeit der im Buch enthaltenen Informationen zu maximieren, basiert die entwickelte Systematik auf allgemein gültigen Prinzipien vertrauenswürdiger KI und ist jederzeit erweiterbar.

Sollten Ihnen bei aller Sorgfalt unsererseits (tatsächlich wie vermeintlich) fehlerhafte oder ungenaue Abbildungen, Formulierungen oder Inhalte auffallen, so sind wir allen Leserinnen und Lesern für Hinweise sehr dankbar. Für Fragen, Anregungen und sonstige Rückmeldungen freuen wir uns über Ihre Nachricht per E-Mail an Jordan.Poetsch@web.de.

Unser Dank gilt dem Springer-Verlag sowie allen Personen, die mit wertvollen Hinweisen und Anregungen den Mehrwert dieses Buches gesteigert haben.

Das Autorenteam wünscht Ihnen eine angenehme und bereichernde Lektüre.

Düsseldorf, Deutschland — Jordan Pötsch
Frankfurt, Deutschland — Prof. Dr. Rainer Bernnat
im Juli 2024

Inhaltsverzeichnis

Über die Autoren

Jordan Pötsch ist Unternehmensberater und Auditor im Bereich Cyber Security & Privacy bei PricewaterhouseCoopers Deutschland. Zu seinen Schwerpunkten zählen die Regulierung von Künstlicher Intelligenz (KI) und deren Anwendung in der Cybersicherheit. Darüber hinaus lehrt er Informationssicherheit an der FernUniversität Hagen sowie an weiteren renommierten Studieninstituten. Jordan Pötsch ist Autor zahlreicher Publikationen und Herausgeber der Buchreihe „Vertrauenswürdige KI“ in Springer Nature.

Prof. Dr. Rainer Bernnat ist Senior Partner bei Strategy& und leitet den Bereich Government & Public Services für die PwC Deutschland. Er ist spezialisiert auf strategische Transformationsprogramme und hat maßgeblich zur Einführung der elektronischen Gesundheitskarte in Deutschland beigetragen. Zudem ist er Honorarprofessor an der Wirtschaftswissenschaftlichen Fakultät an der Universität Augsburg. Bernnat war Mitglied des Vorstands der Initiative D21 und Vizepräsident der Special Olympics Deutschland.

Künstliche Intelligenz – gekommen, um zu bleiben

1

Künstliche Intelligenz (KI) etabliert sich zunehmend sowohl in der Privatwirtschaft als auch im öffentlichen Dienst. Der Trend zu einem verstärkten Einsatz dieser Technologien setzt sich rasant fort. So wuchs der weltweite Markt für künstliche Intelligenz bis 2024 auf über 184 Mrd. US-Dollar, ein Sprung von fast 50 Mrd. gegenüber dem Vorjahr. Bis 2030 wird ein weiterer Anstieg auf über 826 Mrd. US-Dollar erwartet.[1] Besonders bemerkenswert ist das Wachstum im Bereich der generativen KI, deren Marktvolumen sich im Jahr 2023 im Vergleich zu 2022 auf knapp 45 Mrd. US-Doller verdoppelte.[2]

Künstliche Intelligenz könnte das weltweite Bruttoinlandsprodukt (BIP) innerhalb der nächsten 10 bis 15 Jahre um mehrere Billionen US-Dollar anheben. Um dieses Potenzial zu realisieren, ist jedoch eine umfassende Transformation der Arbeitswelt erforderlich. Dies umfasst die Anpassung von Fähigkeiten und Arbeitsmodellen an die neuen Technologien. Folgendes Beispiel aus dem Financial Service (Sektor) beschreibt, welches Potenzial generativer KI zugeschrieben wird. Abb. 1.1 veranschaulicht die durch generative KI getriebene Weiterentwicklung des Geschäftsmodells von einem traditionellen Modell hin zu einem nachhaltigen zukünftigen Zustand. Dieser Wandel verläuft über einen Zwischenzustand, in dem digitale Transformation und Automatisierung eine zentrale Rolle spielen.

Im Zuge der Transformation und der Einführung neuer Arbeitsweisen wird die Finanzfunktion deutlich effizienter. Freiwerdende Kapazitäten werden für „wertvollere" Tätigkeiten genutzt, sowie um aktiv zur Unternehmensleistung beizutragen und echten Mehrwert zu schaffen (Abb. 1.2).

[1] Vgl. Artificial intelligence (AI) market size worldwide from 2020 to 2030, Statista, Juni 2024.

[2] Vgl. Generative artificial intelligence (AI) market size worldwide from 2020 to 2030, Statista, Februar 2024.

J. Pötsch, R. Bernnat, *Regulierung von Künstlicher Intelligenz in der EU*, Vertrauenswürdige KI, https://doi.org/10.1007/978-3-658-46749-4_1

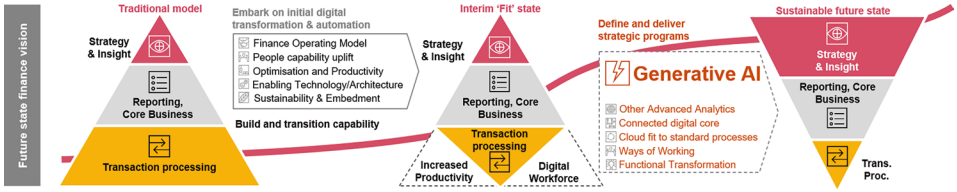

Abb. 1.1 Generative KI beschleunigt die Evolution der Geschäftsfunktion

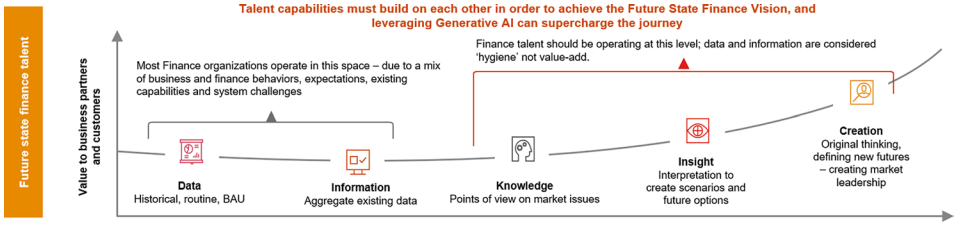

Abb. 1.2 Veränderte Anforderungen durch den Einsatz generativer KI im FS-Sektor

Branchenübergreifend gilt: generative KI beschleunigt die daten- und KI-gestützte Neugestaltung von Unternehmen erheblich. Die Einführung von KI erfordert nicht nur technologische Implementierung, sondern auch einen (ggf. zeitintensiven) kulturellen Wandel innerhalb des Unternehmens. Um jedoch ihr volles Potenzial auszuschöpfen, müssen Führungskräfte sie als mehr als nur ein Werkzeug zur Prozessoptimierung und Kostensenkung betrachten. Sie muss als eine Chance zur Wertschöpfung für Unternehmen, Menschen und die Gesellschaft gesehen werden. Nur eine menschenzentrierte und verantwortungsbewusste Einführung von KI kann nachhaltiges Wachstum fördern. Ein wesentlicher Faktor zum Erfolg liegt in der Weiterentwicklung der Belegschaft. Dabei ist es wichtig, auch soziale Kompetenzen zu stärken, um Vertrauen aufzubauen und die Akzeptanz zu fördern. Führungskräfte sollten eine langfristige, menschenzentrierte Perspektive einnehmen, Arbeitsabläufe neu denken und eine dynamische Belegschaft fördern.

Innerhalb dieser Transformation differenziert sich die Wettbewerbsfähigkeit von Unternehmen nach der Verfügbarkeit und dem sicheren Zugriff der KI auf die internen respektive proprietären Unternehmensdaten. Große Sprachmodelle, engl. Large Language Models (LLMs), konkurrieren um die höchste Qualität der Ausgaben, Geschwindigkeit und Ressourceneffizienz. Durch den Einsatz von Techniken wie retrieval-augmented Generation (RAG) ist eine Individualisierung des Outputs möglich. Diese Technik kombiniert die Fähigkeit von KI-Modellen, Wissen aus einer großen Menge von Textdaten abzurufen (Retrieval), mit der Fähigkeit, natürlichsprachliche Antworten zu generieren (Generation). RAG wird insbesondere in Anwendungen eingesetzt, die präzise und kontextuelle Informationen erfordern. Foundation Models sind große, vortrainierte LLMs, die auf einer breiten Basis von Daten trainiert wurden und für eine Vielzahl von Aufgaben anpassbar sind. RAG kann diese Modelle verbessern, indem es spezifische Informationen aus einer Datenbank oder externen Quellen abruft und in die generierten

Antworten einfließen lässt. Dabei bleiben die Daten in einer sicheren Umgebung, ohne dass sie direkt in die Modelle eingearbeitet werden müssen. Die Marktperspektive zeigt: Während sich die Qualität der Ausgaben großer Sprachmodelle annähert, ergibt sich die Differenzierung durch die Daten, auf die die Modelle zugreifen können. Unternehmen müssen Wege finden, ihre Daten sicher (secure) in die generative KI einzuspeisen, um nicht ins Hintertreffen zu geraten.

Der Einsatz von KI bringt spezifische Compliance-Anforderungen und Risiken mit sich. Dennoch ist der aktuelle Trend zur Entwicklung und Nutzung generativer KI gut begründet. Unternehmen, die in diesem Bereich zurückhaltend agieren, riskieren, den Anschluss an den Wettbewerb zu verlieren. Mitarbeiter werden weiterhin KI für ihre Aufgaben einsetzen, unabhängig von der Unternehmensrichtlinie, und es wird immer Wettbewerber geben, die bereit sind, Risiken einzugehen, um einen Vorteil zu erlangen. Unternehmen müssen KI implementieren, um langfristig wettbewerbsfähig zu bleiben. Vertrauenswürdige und sichere KI (eine genaue Definition der Vertrauenswürdigkeit von KI erfolgt in Abschn. 2.2), entweder als Anforderungen aus einem Gesetz wie dem EU AI Act oder als selbstauferlegte Best Practice, wird im ersten Schritt mit Kosten verbunden. Es zeichnet sich jedoch ab, dass vertrauenswürdige KI ein Differenziator im Geschäftsmodell ist, z. B. wenn durch Erfüllung von Transparenzanforderungen von KI das Vertrauen der Kunden in die Geschäftsbeziehung gestärkt wird. Aus der Risikoperspektive ist es ein Fehler, den Beschäftigten keine KI-Tool zur Verfügung zu stellen, indem beispielsweise der Zugriff auf GenAI-Tools über die Webproxies blockiert sind. Werden offizielle KI-gestützte Werkzeuge nicht bereitgestellt, greifen Mitarbeiter möglicherweise auf private, unsichere IT-Lösungen zurück, um Zugriffsbeschränkungen zu umgehen. Dieses Phänomen, bekannt als Schatten-KI (shadow AI), birgt erhebliche Sicherheitsrisiken. Mitarbeiter könnten sensible Daten auf private Geräte übertragen, was ein eigenes Risikopotenzial birgt und die Datenintegrität gefährden kann. Statt KI-Tools den Mitarbeitern schlichtweg zu verbieten, sollten im ersten Schritt nachweislich vertrauenswürdige KI-Tools zum Experimentieren/Arbeiten aufgezeigt werden. Somit kann die Nutzung von KI anhand von „Do's and Don'ts" zielgerichtet kanalisiert, Raum für Experimente geschaffen und die Gefahr einer Umgehung der Regeln minimiert werden. Durch die Entwicklung von Sicherheitsstrategien, die die KI-Nutzung unterstützen, statt sie einzuschränken, können Unternehmen nicht nur ihre Sicherheit verbessern, sondern auch ihre Effizienz steigern und Innovation vorantreiben. Dies fördert einen signifikanten Wettbewerbsvorteil und stellt sicher, dass sowohl Daten als auch Systeme umfassend geschützt sind. Um Synergieeffekte zu erzeugen und den Aufwand zu minimieren, empfiehlt es sich, bestehende Sicherheitsstrukturen zu nutzen und diese um KI-spezifische Aspekte zu ergänzen.

Die obigen Ausführungen zeigen auf, warum eine nachhaltige und vertrauenswürdige Einführung von KI in Organisationen sowohl aus der Businessperspektive als auch aus der Risikoperspektive ein unumgänglicher Schritt ist. Ist einmal der Entschluss gefasst, die KI-Transformation anzustoßen, gibt es viele Herausforderungen und Risiken vertrauenswürdiger KI, die letztlich über den Erfolg der Transformation entscheiden können. Auf diesen Aspekt geht das nachfolgende Kapitel ein.

1.1 Herausforderungen vertrauenswürdiger KI

Diese rasanten Entwicklungen zwingen Gesetzgeber dazu, den KI-Markt zu regulieren. Die Europäische Union hat auf diese Dynamik reagiert, indem sie den EU AI Act im August 2024 in Kraft gesetzt hat. Dieses Gesetz ist Teil des New Legislative Framework (NLF). Bereits bestehende Gesetze wie die EU-DSGVO (Datenschutz-Grundverordnung) sind für KI einzuhalten. Auch andere Regionen der Welt wie die USA bringen momentan Gesetze zur Regulierung von KI auf den Weg. Europa und die USA entwickeln sich in Bezug auf Innovation und Regulierung asynchron, was ohne Wertung betrachtet werden kann. Während die USA durch eine hohe Innovationsgeschwindigkeit voranschreiten, nimmt sich Europa mehr Zeit für regulierte und ethisch abgewogene Fortschritte. Spannend wird es zu sehen, wie sich diese unterschiedlichen Ansätze langfristig auf unsere Wettbewerbsfähigkeit auswirken. Der EU AI Act könnte sich als Blaupause für globale Standards erweisen, da Europa nicht nur Responsible AI fördert, sondern auch eine nachhaltige Integration in den Markt ermöglicht. Unternehmen, die auf unregulierte Märkte wie Saudi-Arabien setzen, mögen kurzfristig Vorteile genießen, doch langfristig wird Nachhaltigkeit und globale Skalierbarkeit entscheidend sein – insbesondere im Bereich der Künstlichen Intelligenz. In diesem Buch werden die KI-relevanten Gesetze und Standards analysiert, ein Vergleich mit den Vorgaben aus den USA gezogen und inhaltliche Überlappungen aufgezeigt.

Standards, die im Gegensatz zu Gesetzen nicht verbindlich sind, bieten technische Konkretisierung und sind essenziell, um den geforderten Stand der Technik zu erfüllen. In den Jahren 2023 und 2024 wurden bereits eine Vielzahl neuer Standards veröffentlicht. Zu den geläufigsten Frameworks für die Sicherheit von KI gehören das NIST AI Risk Management, ENISAs Multilayer Framework for Good Cybersecurity Practices for AI, die Veröffentlichungen des UK National Cyber Security Center und die Richtlinien der US Cybersecurity and Infrastructure Security Agency für die sichere Entwicklung von KI-Systemen, OWASP AI Exchange, ISO/IEC 42001 und der MITRE ATLAS. Die Liste verdeutlicht, dass es zu einer Flut von Standards und Veröffentlichungen kam, die sich mit der Regulierung von KI befassen. Die zentrale Leitfrage ist, welche Standards zur Compliance mit gesetzlichen Vorgaben wie dem EU AI Act berücksichtigt werden sollten. Die wesentliche Herausforderung bei der Auswahl und Umsetzung von KI-Standards ist ihre Inhomogenität. Beispielsweise decken Standards unterschiedliche Geltungsbereiche (international, europäisch, national) ab, beziehen sich auf eine oder auch mehrere Phasen entlang des KI-Lebenszyklus, beschränken sich auf bestimme Dimensionen und betrachten organisatorische oder systemische Aspekte gleichzeitig respektive getrennt voneinander. Aber auch die hohe Entwicklungsdynamik am Markt hat einen Einfluss auf die Auswahl von Standards. Beispielsweise sind Standards mit einem Veröffentlichungsdatum vor dem Durchbruch generativer KI Ende 2022 meist auf Risiken klassischer KI eingegangen. Wie in Kap. 3 aufgezeigt, sind Begrifflichkeiten in Standards nicht einheitlich definiert und es fehlt global an einem gemeinsamen Verständnis, was unter vertrauenswürdiger KI zu verstehen ist. Daraus resultiert, dass eine international agierende Firma, die global

KI-Anforderungen in ihrer Governance definiert, vor der Herausforderung steht, welche Standards zur Compliance umzusetzen sind und in welchen Fällen Mehraufwand aufgrund von Überlappungen vermieden werden kann. Eine weitere Herausforderung besteht darin, diejenigen Standards zu identifizieren, mit deren Hilfe die highlevel Vorgaben eines Gesetzes in technisch konkrete Vorgaben übersetzen kann. Neben der hohen Marktdynamik der KI-Technologie selbst besitzt ihre Regulierung ebenfalls eine hohe Dynamik. Die bereits hohe Anzahl an Frameworks steigt stetig weiter. Somit existiert die Wahl zwischen verschiedenen Ansätzen, was jedoch zu der Herausforderung führt, den Überblick in einem hochdynamischen regulatorischen Umfeld und einer sich schnell entwickelnden Technologie wie KI zu behalten.

Der Einsatz von KI führt zu Produktivitätssteigerungen. Trotz dieser Vorteile sorgen sich Führungskräfte, insbesondere Chief Information Security Officers (CISOs), um die Sicherheit und den Datenschutz bei der Einführung dieser Technologien. Viele CISOs fühlen sich unsicher, da es an spezifischen Leitlinien für den sicheren Einsatz von KI mangelt. Diese Unsicherheit führt oft dazu, dass Innovationen gebremst werden. Auch besteht Unsicherheit, wie verlässlich KI-Tools für die Cybersicherheit sind oder wie effiziente Maßnahmen gegen Schäden durch Verwendung von KI-Tools bei Angriffen aussehen sollten.

Bestehende Rahmenwerke Frameworks wie NIST und AIGA teilen ähnliche Grundprinzipien zur Definition von KI-Governance, bieten jedoch oft keine klaren Anweisungen zur Umsetzung eines Governance-Prozesses. In diesem Buch werden praktische Ansätze aufzuzeigen, wie ein KI-Governance-Prozess in einer Organisation eingeführt werden kann.

1.2 Ziel und Eingrenzung des Buches

Das Buch dient dazu, Entscheidungsträgern in Organisationen einen Überblick über die wichtigsten Herausforderungen zur Realisierung von vertrauenswürdiger KI zu geben. Damit die KI-Transformation zum Erfolg führt, wird ein businessnaher Leitfaden zur Umsetzung von KI-Governance, Risk und Complaince (GRC) entlang des Three-Lines-of-Defence-Modell gegeben. Die in diesem Buch erarbeitete Systematik charakterisiert KI-Standards anhand verschiedener Kriterien. Um Synergien zu erzeugen, wird ausgehend von einem Informationssicherheitsmanagementsystem (ISMS) nach ISO/IEC 27001 bestehende Risikoprozesse um KI-Spezifika ergänzt oder um neue Prozesse erweitert.

Das Buch definiert Begrifflichkeiten im Kontext der Regulierung von KI. Es wird ein sinnvoller Geltungsbereich festgelegt, in dem sowohl systemische als auch organisatorische Aspekte der Vertrauenswürdigkeit von KI vollumfänglich betrachtet werden. Ein klar definierter Geltungsberiech ist eine wichtige Voraussetzung, um Risken zu identifizieren und geeignete Maßnahmen zu ergreifen. Aus diesem Grund vermittelt das Buch inhaltliche und konzeptuelle Grundlagen hinsichtlich des formalen Aufbaus einer KI-Anwendung und dessen Einsatzumgebung. Ferner wird erläutert, was Vertrauenswürdigkeit von KI bedeutet, weshalb regionale Unterschiede bestehen, wie ein Vergleich mit dem

europäischen Ansatz ausfällt und wie die KI-Compliance global eingehalten werden kann. Weiterhin ist es insbesondere für Risikoexperten (z. B. ein Chief Information Security Officer, kurz CISO) von besonderer Bedeutung, worin die Unterschiede zur klassischen IT/OT liegen und welchen Einfluss dies auf die Governance der Organisation hat.

Ferner wird anhand eines Shared-Responsibilty-Models erläutert, wie sich Verantwortlichkeiten für KI-Sicherheit entlang des KI-Lebenszyklus zwischen den Beteiligten aufgliedern. Dieses Modell hilft, Lücken von GRC-Kontrollen zu verhindern, Unklarheiten auszuräumen und somit insgesamt Risken zu mindern. Herausforderungen hinsichtlich der Verantwortlichkeiten bestehen nicht nur zwischen externen Parteien entlang des Lebenszyklus, sondern auch von Parteien (CISO, Chief Compliance Officer, Chief Data Officer, CAIO falls vorhanden) innerhalb einer Organisation. Auch zu dieser Herausforderung werden Best Practice vermittelt, indem beispielsweise die Rolle des CISOs oder aber auch aufkommende Rollen wie der Chief AI Officer (CAIO) eingeordnet werden. Für die jeweiligen Rollen (für KI-Verantwortliche wie der CISO) werden Leitfäden zur Verfügung gestellt. Es folgt eine Anleitung zur strukturierten Dokumentation von technischen und organisatorischen Maßnahmen entlang des Lebenszyklus einer KI-Anwendung, die dem aktuellen Stand der Technik entsprechen und durch deren Umsetzung mögliche Risiken abgeschwächt werden können.

Ein wesentlicher Schwerpunkt dieses Buches liegt auf der Untersuchung und Einordnung KI-relevanter Gesetze sowie Standards. Zunächst werden KI-Regulierungsinitiativen vorgestellt, mit dem Ziel, dem Leser ein übergeordnetes Verständnis für die von verschiedenen politischen und wirtschaftlichen Interessieren getriebenen Regulierungsinitiativen zu geben. Es erfolgt ein Vergleich der europäischen Ansätze mit den Regulierungsinitiativen der USA. Es werden die Rechtsakte EU AI Act, EU-DSGVO, EU Data Act und der CyberSecurity Act hinsichtlich ihrer Bedeutung für KI analysiert. Die bereits angesprochene Systematik setzt Gesetze und KI-Standards in Bezug zueinander, sodass sich draus ein Leitfaden zur Herstellung von KI-Compliancen für Unternehmen ergibt. Es wird das Zusammenspiel verschiedener Standards erläutert. Das Rad muss nicht neu erfunden werden, wie anhand folgenden Beispiels erläutert: durch die Highlevel Structur (HLS) der ISO-Standards ist ein integriertes Managementsystem (IMS) einführbar, sodass beispielsweise Unternehmensprozesse des Artificial Intelligence Management Systems (AIMS) auf den Prozessen des ISMS nach ISO/IEC 27001 aufgebaut werden können.

Der Ausgangspunkt in diesem Buch ist die CISO-Organisation. Die Gründe dafür sind vielfältig. Wie bereits erwähnt, bietet es sich an, auf bestehende Risikomanagementprozesse des ISMS aufzubauen. Sicherheit (Security) ist eine Dimension, die als kleinster gemeinsamer Nenner unterschiedlicher KI-Regulierungsinitiativen. Außerdem existieren in der CISO-Organisation viele Use Cases zur Nutzung von KI zur Erhöhung der Cybersicherheit. Aus oben genannten Gründen liegt in diesem Buch der Fokus auf dem Zusammenspiel von Cybersecurity und KI. Es existieren neue Bedrohungsszenarien durch KI, aber auch neue Abwehrmöglichkeiten mittels KI. Dazu folgt die Definition von drei Hauptkategorien und eine Übersicht über sinnvolle Anpassungen der CISO-Organisation im Rahmen der KI-Transformation.

Die Ausgestaltung der KI-Governance variiert je nach Unternehmenskontext, Zielsetzung und Risikobereitschaft. Dennoch existieren grundlegende Fragestellungen, die bei der Einführung von KI allgemein relevant sind und in diesem Buch ausführlich behandelt werden. Der Fokus liegt auf der Regulatorik, wie sie in der EU anzufinden ist. Die KI-Governance ist integraler Bestandteil bestehender Prozesse und den drei LoD. Das AIMS fungiert als operatives Werkzeug, das innerhalb des Rahmens der KI-Governance arbeitet.

Das Buch geht auf die fachlich-inhaltlich sinnvollen Handlungsmöglichkeiten des Bundesamtes für Sicherheit in der Informationstechnik (BSI) ein, um KI unter Berücksichtigung des BSI-IT-Grundschutzes und bestehender Personenkompetenzen abzusichern. Dazu folgenden in dem Buch eine detaillierte Analyse und Lösungsvorschläge. Es wird ein VS-KI-Grobkonzept vorgestellt.

In dem Buch werden sektorspezifische Lösungen für die KI-Sicherheit analysiert, die sich aus bestimmten Use Cases oder einer sektorspezifischen Regulatorik ergeben.

Eingrenzung
Dieses Werk richtet sich an ein breites Publikum und setzt keine juristischen Fachkenntnisse voraus. Obwohl der EU AI Act erläutert wird, erfolgt keine vertiefte Behandlung des Themas, wie es in juristischer Fachliteratur üblich wäre. Das Buch ist somit für ein nicht spezialisiertes Publikum leicht zugänglich, ohne dabei auf juristische Fachsprache zurückzugreifen. Es erfolgt keine detaillierte Betrachtung, welche Risiken die Dimensionen inhaltlich abdecken, mit Ausnahme der Dimension Sicherheit (Security). Es wird nicht erläutert, wie sich eine systemische KI-Prüfung respektive Zertifizierung konkret gestalten kann. Mit diesem Buch wurde kein neuer KI-Standard entwickelt, sondern es werden Best Practices zu KI Governance, Cybersecurity und KI sowie mit dem Umgang von Standards aufgezeigt, um AI Act Compliance sicherzustellen.

1.3 Inhaltlicher Aufbau

Das Buch gliedert sich in elf Kapitel. Diese lassen sich inhaltlich die folgenden Abschnitte differenzieren:

- **Abschn. 1 (Kap. 2 und 3) – Grundlagen der Vertrauenswürdigkeit von KI –** gibt einen kompakten Überblick über die wichtigsten KI-Systembestandteile sowie Begrifflichkeiten im Regulierungskontext, beschreibt die wesentlichen Merkmale und Prinzipien für eine vertrauenswürdige KI, geht auf die Unterschiede zu IT/OT hinsichtlich Sicherheit ein, zeigt eine Vorgehensweise zur Risikoanalyse und beleuchtet die Risikotaxonomie, die von unregulierten KI-Systemen ausgehen.

- **Abschn. 2 (Kap. 4 und 5) – Aktuelle Regulierungsinitiativen und Gesetzgebung in Bezug auf KI-Systeme –** beschreibt internationale, europäische Regulierungsinitiativen und untersucht die bestehenden Gesetze, die den Einsatz von KI in der EU regeln. Es wird ein Vergleich mit den gesetzlichen Vorgaben in den USA gezogen.

- **Abschn. 3 (Kap. 6) – Standards und Frameworks –** in diesem Kapitel wird eine Übersicht über bestehende Standards und Frameworks gegeben. Diese werden anhand einer Systematik aufgegliedert und in ein Verhältnis mit dem EU AI Act gesetzt.

- **Abschn. 4 (Kap. 7) – Das Zusammenspiel von KI und Cybersecurity –** definiert die drei Kategorien, die das Zusammenspiel von KI und Cybersecurity ausmachen.
- **Abschn. 5 (Kap. 8 bis 10) – Best Practices –** bietet einen detaillierten Leitfaden für Unternehmen, um zielgerichtet durch die neue Regulatorik zu navigieren und vertrauenswürdige KI einzuführen. Es wird erläutert, wie die wesentlichen KI-Akteure in einem mehrschichtigen Ansatz zusammenarbeiten, um KI-Risiken zu managen und gleichzeitig das Potenzial der KI-Systeme für die Gesellschaft zu nutzen. Darüber hinaus erfolgen spezifische Untersuchungen zur Sicherheit von Künstlicher Intelligenz in den jeweiligen Sektoren.

Innerhalb der jeweiligen Abschnitte bietet es sich an, die Lesereihenfolge einzuhalten, um aufeinander aufbauende Inhalte nachvollziehen zu können.

Grundlagen der Vertrauenswürdigkeit von KI

2

In diesem Kapitel erfolgt eine fundierte Einführung in die zentralen Aspekte der KI-Sicherheit. Ziel ist es, Leserinnen und Lesern ohne Vorkenntnisse zur Regulierung von künstlicher Intelligenz eine fundierte Basis zu vermitteln, um die weiterführenden Inhalte des Buches zu verstehen. Das Kapitel erörtert umfassend die Begrifflichkeiten und regulatorischen Rahmenbedingungen im Bereich der künstlichen Intelligenz. Es folgt eine kurze und prägnante Einführung in die Taxonomie der KI sowie Arten und Eigenschaften von Machine Learning Algorithmen. Es legt einen klaren Geltungsbereich fest, der sowohl systemische als auch organisatorische Aspekte der KI-Vertrauenswürdigkeit berücksichtigt. Dies ist essenziell, um Risiken effektiv zu identifizieren und adäquate Maßnahmen zu entwickeln. Das Kapitel bietet eine solide Grundlage in Bezug auf den formalen Aufbau und die Einsatzumgebungen von KI-Anwendungen. Es erklärt die Bedeutung der Vertrauenswürdigkeit von KI, diskutiert regionale Unterschiede und vergleicht diese mit dem europäischen Ansatz. Darüber hinaus werden die unterschiedlichen Sicherheitseigenschaften von KI im Vergleich zu Operational Technology (OT) und Informationstechnologie (IT) beleuchtet. Es wird eine Differenzierung zwischen klassischer und generativer KI aus einer Sicherheitsperspektive vorgenommen, um ein tiefgreifendes Verständnis für die jeweiligen Risiken und Schutzmaßnahmen zu fördern. Das Kapitel schließt mit einem Überblick verschiedener Governance-Modelle entlang des Three-Lines-of-Defence-Modell.

Ein Blick in die Zukunft – Drei KI-Typen und ihre Regulierung

- Von der Automatisierung einfacher Aufgaben bis hin zur Theorie von Maschinen, die menschliche Intelligenz übertreffen könnten, wird KI in verschiedene Kategorien eingeteilt, die jeweils spezifische Merkmale und Herausforderungen aufweisen. Diese Kategorien bestimmen nicht nur die technischen Möglichkeiten, sondern auch die ethischen und regulatorischen Rahmenbedingungen, die notwendig sind, um eine sichere

J. Pötsch, R. Bernnat, *Regulierung von Künstlicher Intelligenz in der EU*, Vertrauenswürdige KI, https://doi.org/10.1007/978-3-658-46749-4_2

und vertrauenswürdige Nutzung dieser Technologien zu gewährleisten. Die drei Kategorien sind:**Artificial Narrow Intelligence (ANI):** Auch als schwache KI bekannt, ist ANI auf spezifische Aufgaben spezialisiert. Beispiele hierfür sind virtuelle Assistenten wie Siri und Alexa, selbstfahrende Autos, und Bilderkennungssysteme. ANI kann in einem bestimmten Anwendungskontext effizient arbeiten (je nach Use Case besser als ein Mensch), hat aber keine Fähigkeit, über diesen Bereich hinaus zu lernen. Alle bisher entwickelten KI-Systeme fallen in diese Kategorie. GenAI (Stand August 2024) gehört zur ANI, da es auf spezifische Aufgaben wie Text- oder Bildgenerierung beschränkt ist. Es verfügt über kein Bewusstsein oder Verständnis. ANI-Systeme sind auf enge Anwendungsgebiete spezialisiert.

- **Artificial General Intelligence (AGI)**: Diese bisher unerreichte Form der KI, auch starke KI genannt, würde die Fähigkeit besitzen, in verschiedenen Kontexten zu lernen und zu handeln, ähnlich wie ein Mensch. Maschinen dieser Kategorie entwickeln die kognitiven Fähigkeiten auf menschlichem Niveau. Die Prognosen für die Erreichung von AGI variieren stark unter Experten. Einige optimistische Schätzungen sehen AGI bereits in diesem Jahrzehnt erreicht, während konservativere Schätzungen eher von einem Zeitraum um 2040 bis 2060 ausgehen.
- **Artificial Superintelligence (ASI)**: ASI ist eine hypothetische Form der KI, die weit über die menschliche Intelligenz hinausgeht. Sie könnte eigenständige Bedürfnisse und Überzeugungen entwickeln sowie theoretisch alle intellektuellen Fähigkeiten des Menschen übertreffen.

Die Entwicklung der Regulierung von KI muss sich den spezifischen Herausforderungen der verschiedenen Kategorien anpassen. Mit jedem Schritt zur nächsten Kategorie wächst der Einfluss auf die Gesellschaft. Während ANI bereits heute reguliert wird, stellt AGI zukünftige regulatorische Herausforderungen dar, die einer proaktiven Planung bedürfen. ASI hingegen erfordert ein tiefgehendes Verständnis potenzieller Risiken und die Entwicklung von Mechanismen, die diese Risiken minimieren können. Die hohe Dynamik der KI-Entwicklung machen es notwendig, dass Regulierungen flexibel und zukunftsorientiert gestaltet werden, um den Fortschritt sicher und ethisch verantwortungsvoll zu gestalten. Eine Beurteilung, wie zukunftsorientiert die aktuellen Vorgaben in dem EU AI Act sind, folgt in Kap. 5.

Definition von KI

Die Definition von künstlicher Intelligenz (KI) und die klare Abgrenzung von nicht-KI-basierten maschinellen Systemen ist eine komplexe Herausforderung. Es fehlt eine einheitliche, allgemein akzeptierte Definition, da es ein Kontinuum von Merkmalen gibt, die KI charakterisieren, anstatt einer klaren Trennlinie. Technologien wie die optische Zeichenerkennung, die einst als KI galten, werden heute oft nicht mehr so wahrgenommen,

obwohl sie weiterhin klassische KI-Methoden nutzen. Die rasante Entwicklung und zunehmende Vielfalt von KI-Anwendungen erschweren eine stabile und zeitbeständige Definition. Diese Dynamik zeigt, dass die Definition und Abgrenzung von KI eine kontinuierlich herausfordernde Aufgabe bleibt.

Der aktuelle Diskurs über künstliche Intelligenz (KI) präsentiert eine Vielzahl regionaler, normativer und kontextabhängiger Begriffsdefinitionen. Im Folgenden werden exemplarisch vier zentrale Definitionen aufgeführt:

1. **Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD):** Die OECD hat die Prinzipien der Arbeitsgruppe für vertrauenswürdige KI-Richtlinien aus dem Jahr 2019 aktualisiert. Die neue Definition lautet:
   a. „An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."
   b. Zu Deutsch: „Ein KI-System ist ein maschinenbasiertes System, das für explizite oder implizite Ziele entwickelt wurde. Es zieht aus den ihm zugeführten Daten Rückschlüsse, um Ausgaben wie Vorhersagen, Empfehlungen oder Entscheidungen zu erzeugen, die physische oder virtuelle Umgebungen beeinflussen können. Unterschiedliche KI-Systeme variieren in ihrem Grad der Autonomie und Anpassungsfähigkeit nach der Implementierung."

2. **EU AI Act:** Gemäß Article 3(1) der Verordnung werden die folgenden Definitionen verwendet:
   a. „AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."
   b. In der offizielle deutsche Version des EU AI Acts vom 12.07.2024[1] bezeichnet der Ausdruck „ „KI-System" ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können"

[1] Vgl. https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L_202401689.

   c. In Erwägungsgrund 12 des EU AI Acts wird die klare Definition des Begriffs „KI-System“ gefordert, um rechtliche Sicherheit zu gewährleisten und internationale Abstimmung sowie Anpassungsfähigkeit an technologische Entwicklungen zu ermöglichen.

3. **Richtlinien für die Entwicklung sicherer KI-Systeme („Guidelines for secure AI system development“):**[2] Veröffentlicht durch das UK National Cyber Security Centre (NCSC), die US Cybersecurity and Infrastructure Security Agency (CISA) und 20 weitere nationale Regulierungsbehörden, wird KI im Sicherheitskontext wie folgt definiert:
   a. „In this document we use ‘AI’ to refer specifically to machine learning (ML) applications3. All types of ML are in scope. We define ML applications as applications that:

      > involve software components (models) that allow computers to recognize and bring context to patterns in data without the rules having to be explicitly programmed by a human,

      > generate predictions, recommendations, or decisions based on statistical reasoning.“
   b. Eine offizielle deutsche Übersetzung existiert nicht, man kann es sinngemäß wie folgt übersetzen: „In diesem Dokument beziehen wir uns mit ‚KI‘ speziell auf Anwendungen des maschinellen Lernens (ML). Alle Typen des ML sind eingeschlossen. ML-Anwendungen sind solche, die Softwarekomponenten (Modelle) beinhalten, welche Computern ermöglichen, Muster in Daten zu erkennen und zu kontextualisieren, ohne dass die Regeln explizit durch einen Menschen programmiert werden müssen. Diese generieren Vorhersagen, Empfehlungen oder Entscheidungen auf Basis statistischer Schlussfolgerungen.“

4. **ISO-Normen:** Aktuelle ISO-Standards wie die ISO/IEC 42001 verweisen in den Referenzen auf den „ISO/IEC-Standard 22989:2022, Informationstechnologie – Künstliche Intelligenz – Begriffe und Terminologie der künstlichen Intelligenz“, sodass die dortige Definition von KI dokumentenübergreifend Anwendung findet. In Abschn. 3.1 des Standards wird AI wie folgt definiert:
   a. „Künstliche Intelligenz (KI) als Disziplin: Forschung und Entwicklung von Mechanismen und Anwendungen von KI-Systemen. Anmerkung: Forschung und Entwicklung können in beliebigen Bereichen wie Informatik, Datenwissenschaft, Geisteswissenschaften, Mathematik und Naturwissenschaften stattfinden.“

[2] Vgl. https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf.

b. „Ein KI-System ist ein maschinenbasiertes System, das, basierend auf expliziten oder impliziten Zielen, aus den erhaltenen Eingaben Schlüsse zieht, wie es Ausgaben wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erzeugen kann, die physische oder virtuelle Umgebungen beeinflussen. Verschiedene KI-Systeme weisen nach der Implementierung unterschiedliche Grade an Autonomie und Anpassungsfähigkeit auf."

In den weiteren Abschnitten dieses Fachbuchs, insbesondere im Kontext der europäischen Regulierung, wird sich auf die Definition des EU AI Acts gestützt. Komplexe IT-Systeme, die auf maschinellem Lernen basierende Komponenten zur Ausführung bestimmter Aufgaben enthalten, werden im weiteren Verlauf des Buches als KI-Systeme bezeichnet. An dieser Stelle sei darauf verwiesen, dass aufgrund der uneinheitlichen Begriffsdefinitionen Unternehmen ein Augenmerk auf den Geltungsbereich der jeweiligen Regulatorik legen sollten, siehe dazu auch Kap. 8 „Best Practices".

Taxonomie

Die Taxonomie der KI umfasst mehrere zentrale Unterkategorien, die jeweils spezifische Technologien und Methoden zur Nachahmung menschlicher Intelligenz repräsentieren. Diese Unterkategorien sind Künstliche Intelligenz, Maschinelles Lernen, Deep Learning und Generative Künstliche Intelligenz (siehe Abb. 2.1).

- **Künstliche Intelligenz** ist der Oberbegriff für Systeme und Programme, die in der Lage sind, Aufgaben zu bewältigen, die typischerweise menschliche Intelligenz erfordern. Dazu gehören Bereiche wie Sprachverarbeitung, Entscheidungsfindung und Bilderkennung. Künstliche Intelligenz kann regelbasiert sein oder statistische Methoden und Algorithmen einsetzen, um aus Daten zu lernen und sich kontinuierlich zu verbessern.
- **Maschinelles Lernen** stellt eine spezifische Unterkategorie der Künstlichen Intelligenz dar, bei der Algorithmen verwendet werden, um aus Daten zu lernen und Vorhersagen oder Entscheidungen zu treffen, ohne dass explizite Programmierung notwendig ist. Der Prozess des maschinellen Lernens beinhaltet das Einspeisen von Daten in ein Modell, das Trainieren dieses Modells, das Testen sowie die letztendliche Bereitstellung des Modells zur Ausführung von Vorhersageaufgaben. Typische Anwendungen des maschinellen Lernens umfassen Zeitreihenprognosen, Kreditscoring, Textklassifikation und Empfehlungssysteme.
- **Deep Learning** ist eine spezialisierte Form des maschinellen Lernens, die künstliche neuronale Netze mit vielen Schichten nutzt, um komplexe Muster und Zusammenhänge in großen Datenmengen zu erkennen. Der Begriff „Deep" bezieht sich auf die Vielzahl von Schichten, durch die die Daten verarbeitet werden, ähnlich der Arbeitsweise des menschlichen Gehirns. Deep Learning hat signifikante Fortschritte in Bereichen wie autonomem Fahren, Spracherkennung und Bildverarbeitung ermöglicht.

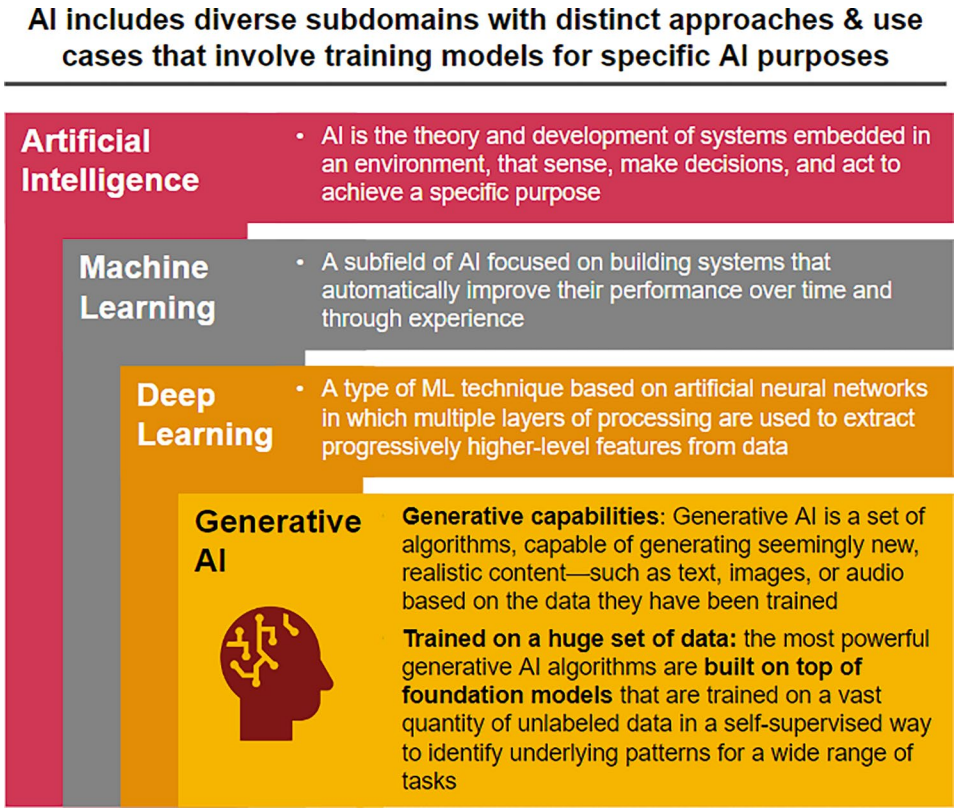

Abb. 2.1 High-Level Taxonomie von KI

- **Generative Künstliche Intelligenz** ist ein Bereich der Künstlichen Intelligenz, der darauf abzielt, neue Inhalte zu generieren, wie Texte, Bilder oder Musik. Diese Technologie verwendet Modelle wie Generative Adversarial Networks (GANs) und Variational Autoencoders (VAEs), um neue, datenähnliche Inhalte zu erstellen. Generative Künstliche Intelligenz wird beispielsweise zur Erzeugung realistischer Bilder oder zur Erstellung natürlicher Sprachtexte eingesetzt. Diese Modelle basieren häufig auf umfangreichen vortrainierten Datensätzen und nutzen Deep Learning-Techniken zur Erfüllung ihrer Aufgaben.

Die Begriffe **Generative AI, Foundation Models und Large Language Models (LLMs)** bezeichnen unterschiedliche, jedoch miteinander verbundene Bereiche der künstlichen Intelligenz. Generative AI bezieht sich auf Techniken der künstlichen Intelligenz, die aus vorhandenen Daten lernen und diese nutzen, um neue Inhalte zu generieren. Diese Technologie umfasst Systeme, die Texte, Bilder, Musik oder andere Medien erzeugen können. Beispiele für generative AI-Systeme sind Bildgeneratoren wie Stable Diffusion und textgenerierende Modelle wie ChatGPT. Foundation Models sind groß angelegte maschi-

nelle Lernmodelle, die auf breiten Datensätzen trainiert werden und als Grundlage für verschiedene spezifische Anwendungen dienen können. Sie sind darauf ausgelegt, vielseitig und anpassbar zu sein, sodass sie für eine Vielzahl von Aufgaben gefinetuned (feinabgestimmt/spezialisiert/angepasst) werden können. Diese Modelle bilden die Basis (daher der Name „Foundation"), auf denen spezifische Anwendungen entwickelt werden können. Ein Foundation Model könnte zum Beispiel ein Sprachmodell sein, das durch zusätzliche Trainingsdaten für die Nutzung in einem Chatbot spezialisiert wird. Large Language Models (LLMs) sind eine spezielle Art von Foundation Models, die auf Sprachverarbeitung fokussiert sind. Diese Modelle werden mit enormen Mengen an Textdaten trainiert, um menschliche Sprache zu verstehen und zu generieren. Sie lernen Muster und Zusammenhänge in den Daten und können dadurch Texte schreiben, übersetzen oder interpretieren. LLMs wie GPT-4 oder Google's PaLM sind darauf ausgelegt, menschlich klingende Antworten auf textbasierte Eingaben zu geben und sind das Herzstück vieler generativer AI-Anwendungen. Zusammenfassend lässt sich sagen, dass Generative AI die übergeordnete Kategorie ist, die alle Techniken zur Generierung neuer Inhalte umfasst. Foundation Models sind umfassende, vielseitige Modelle, die als Basis für spezialisierte Anwendungen dienen können, und LLMs sind eine spezielle Unterkategorie von Foundation Models, die sich auf die Verarbeitung und Generierung von Sprache konzentrieren.

Ein tiefgehendes Verständnis der Taxonomie und auch Terminologie der Künstlichen Intelligenz ist für die Regulierung entscheidend. Die eindeutige Terminologie schafft die Grundlage für technische Standards und Frameworks. Die EU und die USA verfolgen einen menschenzentrierten Ansatz für KI: Dieser Ansatz erfordert, dass die verwendete Terminologie das menschliche, gesellschaftliche und ökologische Wohlbefinden sowie Rechtsstaatlichkeit, Menschenrechte, demokratische Werte und nachhaltige Entwicklung in den Mittelpunkt stellt. Unterschiedliche terminologische Ansätze spiegeln verschiedene technologische Kulturen wider und offenbaren Lücken, Divergenzen und Inkonsistenzen wider. Für Interoperabilität von Standards und die Zusammenarbeit/Kooperation von Regulierungsbehörden ist eine einheitliche Verwendung von Begriffen entscheidend.

Machine Learning Algorithmen – Arten und Eigenschaften

Für die Betrachtung der Vertrauenswürdigkeit und Sicherheit von KI spielt das zugrunde liegende maschinelle Lernverfahren bzw. das dadurch erstellte KI-Modell eine wesentliche Rolle. Vor dem Hintergrund von Sicherheits- und Vertrauenswürdigkeitsaspekten folgt eine detaillierte Beschreibung der gängigen Lernverfahren.

Machine Learning (ML) Algorithmen lassen sich in verschiedene Kategorien einteilen, wobei die gängigsten Ansätze das überwachte Lernen (**supervised learning**), unüberwachte Lernen (**unsupervised learning**) und verstärkendes Lernen (**reinforcement learning**) umfassen.

Beim überwachten Lernen wird ein Modell mit gekennzeichneten Trainingsdaten trainiert, bei denen Eingabedaten mit den entsprechenden Ausgabedaten (Labels) gepaart sind. Dieser Ansatz ermöglicht es dem Modell, Muster und Beziehungen in den Daten zu

erkennen, die dann auf neue, unbekannte Daten angewendet werden können. Typische Algorithmen in dieser Kategorie sind die lineare Regression, Entscheidungsbäume und neuronale Netze. Anwendungen finden sich in der Spracherkennung, Betrugserkennung und medizinischen Diagnostik. Aus sicherheitstechnischer Sicht ist das überwachte Lernen anfällig für adversariale Angriffe, bei denen böswillig manipulierte Eingaben das Modell in die Irre führen können. Datenschutzprobleme können ebenfalls auftreten, insbesondere wenn sensible persönliche Daten für das Training verwendet werden.

Das unüberwachte Lernen hingegen arbeitet mit unbeschrifteten Daten, das heißt, das Modell versucht, Muster oder Strukturen in den Eingaben zu erkennen, ohne dass spezifische Ausgabedaten vorgegeben sind. Bekannte Algorithmen sind hier das K-Means Clustering und die Hauptkomponentenanalyse (PCA). Unüberwachtes Lernen wird häufig in der Kundensegmentierung oder bei der Anomalie Erkennung eingesetzt. Sicherheitsaspekte umfassen die Gefahr der Missinterpretation von Daten und potenzielle Datenschutzrisiken, da die Daten oft nicht spezifiziert sind und persönliche Informationen enthalten können.

Das verstärkende Lernen ist ein Ansatz, bei dem ein Agent durch Interaktionen mit seiner Umgebung lernt und versucht, Belohnungen zu maximieren. Dieser Lernprozess basiert auf den Prinzipien der Belohnung und Bestrafung, ähnlich dem Lernen durch Versuch und Irrtum. Anwendungen finden sich in der Robotik, bei selbstfahrenden Autos und in dynamischen Preissystemen. Sicherheitsrisiken in diesem Bereich beinhalten die Möglichkeit der Datenmanipulation und des Modell-Drifts, was die Zuverlässigkeit der Entscheidungen beeinträchtigen kann.

An dieser Stelle der Hinweis, dass die Definition der Vertrauenswürdigkeit von KI in Abschn. 2.2.

2.1 Geltungsbereich für die Betrachtung der Vertrauenswürdigkeit

In diesem Unterkapitel wird ein für vertrauenswürdige KI sinnvoller Geltungsbereich festgelegt, in dem sowohl systemische als auch organisatorische Aspekte der Vertrauenswürdigkeit von KI vollumfänglich betrachtet werden. Ein klar definierter Geltungsberiech ist eine wichtige Voraussetzung, um Risken zu identifizieren und geeignete Maßnahmen zu ergreifen. Aus diesem Grund vermittelt dieses Kapitel inhaltliche und konzeptuelle Grundlagen hinsichtlich des formalen Aufbaus einer KI-Anwendung und dessen Einsatzumgebung.

Ein erster Schritt bei der Identifizierung von Risken und Bewertung der Vertrauenswürdigkeit einer KI-Anwendung besteht darin, ihren formalen Aufbau zu spezifizieren und den Geltungsbereich klar abzugrenzen. Ziel dieses Abschnitts ist es, ein einheitliches Verständnis der Begriffe in Bezug auf die Struktur einer KI-Anwendung zu schaffen. KI-Anwendungen sind komplexe Systeme, die oft aus einer Kombination von maschinellen Lernmodellen, Expertensystemen und anderen klassischen Softwarekomponenten beste-