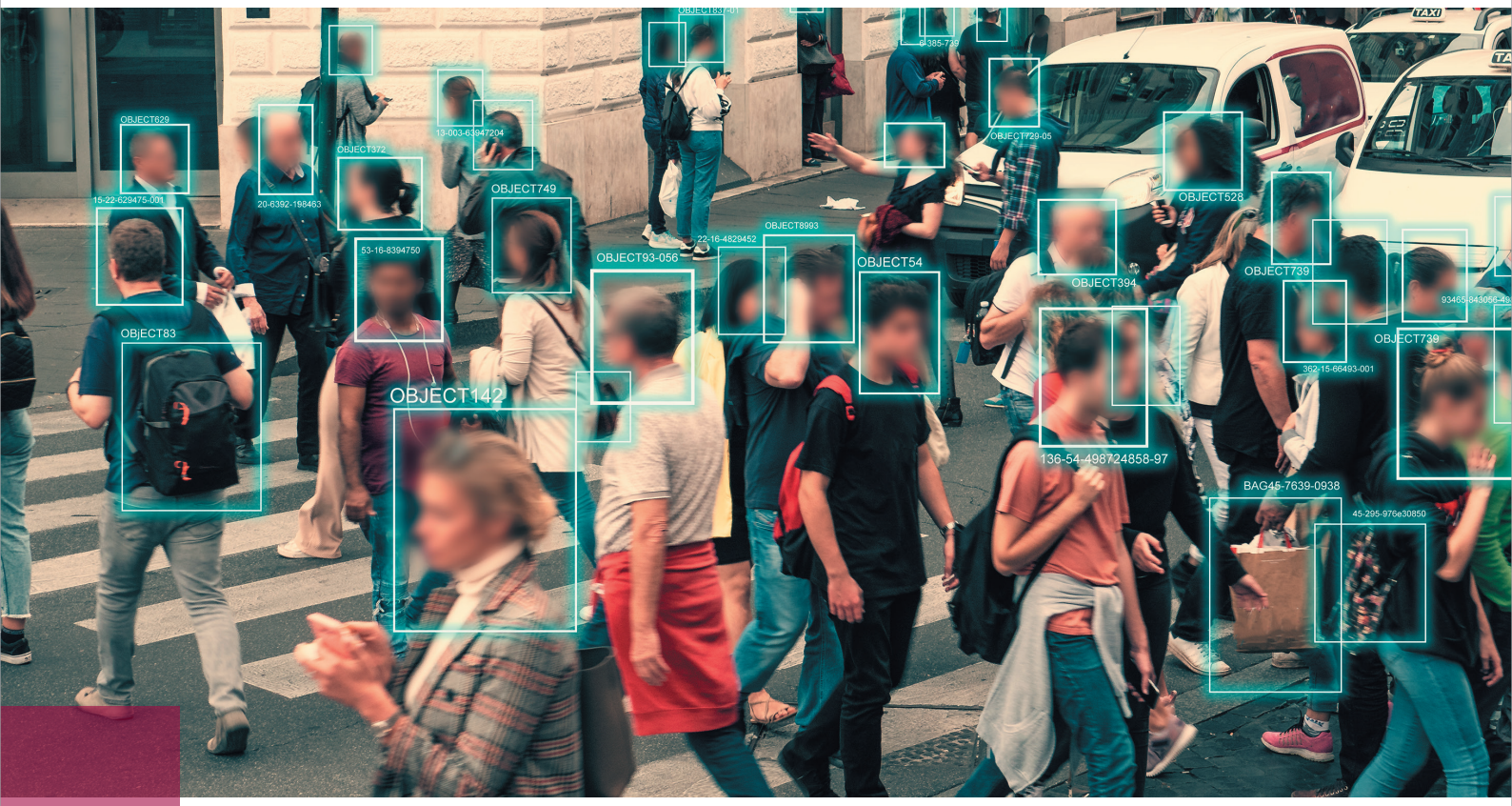
Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ



Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten

Ein notwendiger, aber nicht hinreichender Baustein für Vertrauenswürdigkeit

WHITEPAPER

Jessica Heesen, Jörn Müller-Quade, Stefan Wrobel et al.
AG IT-Sicherheit, Privacy, Recht und Ethik
AG Technologische Wegbereiter und Data Science

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN

Inhalt

Zusammenfassung 3

1 Einleitung 5
1.1 Untersuchungsgegenstand 7
1.2 Begriff der Kritikalität und bestehende Konzepte 9

2. Risiken und Schäden bei KI-Systemen 13
2.1 Akteure, Verwendungsintentionen und Anwendungskontexte 13
2.2 Gefahrenabwehr versus Vorsorge 14
2.3 Verantwortung, Verantwortungsketten und Haftung 15

3. Kritikalitätsbewertung von KI-Systemen vor dem Hintergrund ihrer Anwendungskontexte 18
3.1 Dimensionen zur Bewertung 18
3.2 Von den Dimensionen zur Kritikalität 25

4. Aktuelle Fragen und Antworten zum Thema Kritikalität 28

5. Fazit 37

Literatur 38

Über dieses Whitepaper 40

Zusammenfassung

Der Einsatz von KI-Systemen birgt ein großes Nutzenpotenzial. In der Medizin können sie Diagnosen und Behandlungen verbessern, in der Mobilität die Routenplanung optimieren und eine passgenauere Abstimmung zwischen Bedarfen und Angeboten ermöglichen. Sie helfen zudem allgemein dabei, die Zukunftsfähigkeit von Wirtschaft und Gesellschaft zu sichern. Gleichzeitig kann die Anwendung mancher KI-Systeme aber auch ein gewisses Risiko mit sich bringen. Diesem Risiko kann unter anderem mithilfe einer passgenauen Regulierung ausgewählter KI-Systeme vor deren Einsatz (ex ante) im europäischen Binnenmarkt begegnet werden. Ziel ist es, sicherere und vertrauenswürdigere KI-Systeme in die Anwendung zu bringen, also Qualität zu liefern, und gleichzeitig Überregulierung zu vermeiden sowie Innovationen zu fördern. Die Europäische Kommission hat deshalb im April 2021 einen Vorschlag zur Regulierung von KI-Systemen vorgelegt.

Expertinnen und Experten der Arbeitsgruppen IT-Sicherheit, Privacy, Recht und Ethik sowie Technologische Wegbereiter und Data Science setzen sich im vorliegenden Whitepaper mit ausgewählten Inhalten dieses EU-Vorschlags zur Regulierung von KI-Systemen auseinander: nämlich anhand welcher Kriterien festgelegt werden kann, in welchen Fällen der Einsatz von KI-Systemen von vornherein zu regulieren ist und wann dies nicht notwendig ist. Ziel ist es, diejenigen KI-Systeme zu identifizieren, welche vor dem Hintergrund ihres jeweiligen Anwendungskontextes einer Regulierung zur Sicherstellung der Sicherheit und Vertrauenswürdigkeit bedürfen. Dies auch um letztlich eine Balance zwischen Innovationsoffenheit auf der einen und Schutz des Subjekts auf der anderen Seite zu schaffen. Ein vielversprechender Ansatz stellt hierfür das Konzept der Kritikalität da, auf dem auch der EU-Vorschlag zur Regulierung von KI-Systemen beruht.

Zunächst wird der Begriff der Kritikalität im Zusammenhang mit KI-Systemen genauer definiert, worunter allgemein ein Maß für potenzielle Gefahren verstanden wird, die vom Einsatz eines KI-Systems in einem spezifischen Anwendungskontext ausgehen können. Diesem Kritikalitätsverständnis folgt auch der EU-Vorschlag. Es findet sich u. a. bereits in mehreren bestehenden Konzepten zur Kritikalitätseinstufung wieder (Kapitel 1). Die Autorinnen und Autoren betrachten in ihren Überlegungen KI-Systeme mit einem besonderen Fokus auf algorithmische Entscheidungssysteme (sog. ADM-Systeme/Algorithmic Decision Making-Systems) mit lernenden Komponenten, während der EU-Regulierungsvorschlag eine breitere Definition umfasst.

Risiken, die der Einsatz von KI-Systemen mit sich bringen kann, und Schäden, die bei solchen Einsätzen entstehen können, sind vielfältig und folglich nur schwer vorherseh- und abschätzbar (Kapitel 2). Neben der Regulierung anhand der Kritikalität im Vorhinein und dem Schaffen von Verbraucherschutzregimen für mögliche Schadensfälle bildet auch die Aufteilung der Verantwortung des Risikos über Haftungsregelungen einen wichtigen Baustein hin zu vertrauenswürdigen KI-Systemen. Deshalb sollte nach Auffassung der Autorinnen und Autoren bereits vor dem Einsatz dieser Systeme geklärt werden, wie im Falle von Schäden die Verantwortung unter den Akteuren aufgeteilt werden sollte.

Die Betrachtung der Kritikalität sollte daher im Sinne einer Gefahrenvorsorge statt einer Gefahrenabwehr erfolgen, da Risiken nicht nur rein technisch, sondern auch sozio-technisch zu bewerten sind. Anhand eines Modells zur Verantwortungsaufteilung im Fall von Schäden durch KI-Systeme präzisieren die Autorinnen und Autoren die jeweiligen Verantwortlichkeiten.

Das Konzept der Kritikalität ist trotz einiger Schwächen, die es bei der Anwendung auf KI-Systeme aufweist, wie beispielsweise die starke Komplexitätsreduktion oder die mangelnde Vorherseh- und Absehbarkeit von Schäden, nach Auffassung der Autorinnen und Autoren eine vielversprechende Lösung für die Kritikalitätsbewertung und damit ein weiterer, wenn auch noch nicht hinreichender Baustein für die Vertrauenswürdigkeit von KI-Systemen. Deshalb weisen die Autorinnen und Autoren im Folgenden auf Aspekte hin, die ihrer Einschätzung nach bei der Ermittlung der Kritikalität eines KI-Systems in seinem jeweiligen Anwendungskontext beachtet werden müssen, um für diese Klassifizierungsform ein praktikables, wenn auch nicht endgültiges Ergebnis zu erzielen. Entscheidend ist, dass die Regulierung von KI-Systemen immer vor dem Hintergrund des jeweiligen Anwendungskontextes erfolgt und dass KI-Regulierungsvorschläge unter Beachtung bestehender allgemeiner und domänenspezifischer Technikstandards reflektiert werden, denn das gleiche System kann in einem Anwendungskontext unproblematisch und in einem anderen höchst kritisch sein. Anhand dieser Einzelfallbewertung kann die konkrete Operationalisierung des Konzepts präzisiert werden. Für die Ermittlung der Kritikalität sind stets zwei Dimensionen abzufragen: Erstens: Möglichkeit und Ausmaß der Verletzung und Gefährdung von Menschen und weiteren hohen Rechtsgütern sowie zweitens Umfang der Handlungsmöglichkeiten der Individuen. Je höher das Ausmaß der möglichen Verletzung von Menschenleben und weiteren hohen Rechtsgütern und je geringer der Umfang der Handlungsmöglichkeiten des Individuums, desto höher ist die Kritikalität – und daraus abgeleitet der Regulierungsbedarf – und vice versa (Kapitel 3).

Die vorgestellten Ansatzpunkte werden in Kapitel 4 von Expertinnen und Experten der Plattform Lernende Systeme in kurzen Interviews um weitere Aspekte aus ihrer jeweiligen Fachexpertise heraus ergänzt und genauer erörtert.

Das vorliegende Papier zeigt, dass das von der Europäischen Kommission gewählte Konzept der Risiko- oder auch Kritikalitätsbewertung zur Sicherstellung der Qualität von KI-Systemen eine gute Orientierungsfunktion zur Bewertung und Regulierung hat, wenn es um einige Einordnungskriterien ergänzt wird. Es ist jedoch unerlässlich, dass das Konzept erweitert wird sowie auch weitere Kriterien definiert und präzisiert werden. Entscheidend ist, dass die Regulierung von KI-Systemen immer vor dem Hintergrund des jeweiligen Anwendungskontextes erfolgt. Gleichzeitig besteht aber noch weiterer Forschungsbedarf zur Überwindung von Schwachstellen. Das Papier leistet damit einen Beitrag zur Diskussion, wie durch weitere bestimmte Kriterien die Bewertung der Kritikalität von KI-Systemen genauer klassifiziert werden kann, um einerseits sicherzustellen, dass KI-Systeme vertrauenswürdig sind, sowie andererseits Überregulierung zu vermeiden und Innovationen zu fördern.

1. Einleitung

Der Einsatz von KI-Systemen birgt ein großes Nutzenpotenzial. So können KI-Systeme in der Medizin Diagnosen und Behandlungen verbessern, in der Mobilität die Routenplanung optimieren und eine passgenauere Abstimmung zwischen Bedarfen und Angeboten ermöglichen und allgemein dabei helfen, die Zukunftsfähigkeit von Wirtschaft und Gesellschaft zu sichern (siehe Plattform Lernende Systeme 2019a, b und 2020a). Gleichzeitig kann die Anwendung mancher KI-Systeme aber auch ein gewisses Risiko mit sich bringen. Diesem Risiko kann mithilfe einer passgenauen Regulierung ausgewählter KI-Systeme vor deren Einsatz (ex ante) im europäischen Binnenmarkt begegnet werden. So können sicherere und vertrauenswürdigere KI-Systeme in die Anwendung gebracht werden. Im April 2021 hat die Europäische Kommission deshalb einen Vorschlag zur Regulierung von KI vorgelegt (siehe 2021a). Dieser wird nun im Europäischen Parlament und im Rat der Europäischen Union (Ministerrat) diskutiert und anschließend verabschiedet. Vorliegendes Papier setzt sich mit ausgewählten Inhalten des Vorschlags auseinander und möchte so eine Perspektive zur aktuellen politischen Debatte beitragen.

Zentrale Fragestellung

Damit eine Regulierung von KI Qualität sicherstellen, Überregulierung vermeiden und gleichzeitig Innovationen fördern kann, muss eine gute Antwort auf die Frage gefunden werden, anhand welcher Kriterien festgelegt wird, in welchen Fällen der Einsatz von KI-Systemen von vornherein (ex ante) zu regulieren ist und wann dies nicht notwendig ist. Ziel ist es, diejenigen KI-Systeme zu identifizieren, welche vor dem Hintergrund ihres jeweiligen Anwendungskontextes einer Regulierung (siehe Infobox) zur Sicherstellung der Sicherheit und Vertrauenswürdigkeit bedürfen. Ebenso sollen diejenigen KI-Systeme identifiziert werden, bei welchen vor dem Hintergrund ihres Anwendungskontextes kein Regulierungsbedarf vorliegt. Es soll eine Balance zwischen Innovationsoffenheit auf der einen und Schutz des Subjekts auf der anderen Seite geschaffen werden. Das Spektrum der Möglichkeiten der Regulierung reicht von keinerlei Regulierung über freiwillige Regulierung in einigen Fällen bis hin zu verpflichtender Regulierung für alle Systeme in allen Anwendungsgebieten.[1]

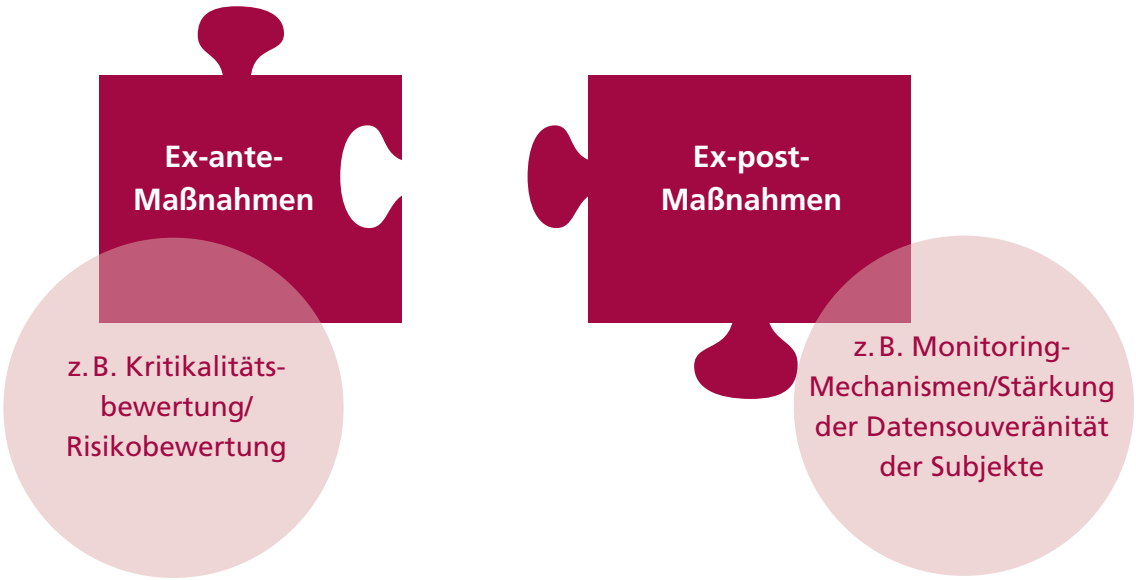
Regulierung, die die Korrektur oder Vermeidung unerwünschter Marktergebnisse zum Ziel hat, wird als Normsetzung und Ordnung im weiteren Sinne verstanden. Sie bezeichnet die Verhaltensbeeinflussung von Individuen, Institutionen oder Unternehmen. Sie hat zum Ziel, ein unerwünschtes Marktgeschehen und unerwünschte Nutzungsweisen von KI zu vermeiden oder zu korrigieren (siehe Wirtschaftslexikon Gabler 2020). Instrumente der Regulierung können beispielsweise Transparenzpflichten sowie Audits, Zulassungsverfahren, Zertifizierungen oder (in einigen Fällen) auch Verbote sein.

1 In dem genannten Kontinuum bestehen zahlreiche Abstufungen. Diese umfassen beispielsweise Selbstverpflichtungen, Zertifizierungen, aber auch unterschiedliche Arten von Zulassungen, wie zum Beispiel Zulassungen nach Prüfung (z. B. autonom fahrende U-Bahnen), zeitlich begrenzte Zulassungen (z. B. Aufzug im Hotel) sowie permanente Überwachung (z. B. selbstlernende Systeme). Der stärkste Eingriff ist das gesetzliche Verbot der Anwendung.

Ex-ante- und Ex-post-Maßnahmen

Die Ex-ante-Bewertung des Regulierungsbedarfs entspricht den staatlichen Fürsorgepflichten und dem Vorsorgeprinzip. Betrachtungen im Nachhinein würden von möglichen Schädigungen durch KI-Anwendungen Betroffene vor Herausforderungen stellen, die sie aufgrund fehlender Kompetenzen und Ressourcen nicht meistern könnten. Mithilfe dieser Ex-ante-Maßnahme können potenzielle Gefahren, die durch den Einsatz eines KI-Systems in einem Anwendungskontext entstehen können, bereits im Vorhinein eliminiert oder zumindest abgeschwächt werden. So können sicherere und vertrauenswürdigere Produkte entwickelt und genutzt werden (siehe Heesen, Müller-Quade, Wrobel et al. 2020). Die Bewertung des Regulierungsbedarfs im Vorhinein stellt allerdings nur eine notwendige, aber nicht hinreichende Bedingung für vertrauenswürdige KI-Systeme dar. Vor allem bei höheren Autonomiegraden sowie bei der Erwartung irreversibler Folgen oder gar Schäden, muss die Ex-ante-Bewertung von KI-Systemen um weitere Mechanismen ergänzt werden, die während oder nach der Anwendung (ex post) zum Einsatz kommen, aber bereits im Vorhinein festgelegt werden (siehe Abbildung 1). Die von der Europäischen Kommission vorgeschlagenen Monitoring-Mechanismen sollten um weitere Maßnahmen ergänzt werden, wie beispielsweise niedrigschwellige, zeitnahe und effektive Beschwerde- und Verbraucherschutzregime zur Stärkung der (Daten-)Souveränität der Bürgerinnen und Bürger (siehe Abbildung 1).

Abbildung 1: Bausteine für vertrauenswürdige KI-Systeme



Konzept der Kritikalität

Einen vielversprechenden Ansatz zur Ermittlung des Regulierungsbedarfs eines Systems vor dem Hintergrund seines Anwendungskontextes stellt das Konzept der Kritikalität dar. Auch die Europäische Kommission ermittelt anhand einer Risikobewertung den Regulierungsbedarf von KI-Systemen. Das Konzept weist bei der Anwendung auf KI-Systeme aber Schwächen auf, wie beispielsweise die starke Komplexitätsreduktion oder die mangelnde Vorherseh- und Absehbarkeit von Schäden. Wichtig ist in diesem Zusammenhang, dass die Regulierung von KI-Systemen immer vor dem Hintergrund des jeweiligen Anwendungskontextes erfolgt und dass KI-Regulierungsvorschläge unter Beachtung bestehender allgemeiner und domänenspezifischer Technikstandards reflektiert werden. Betrachtet werden KI-Systeme mit einem besonderen Fokus auf algorithmische Entscheidungssysteme (sog. ADM-Systeme/Algorithmic Decision Making-Systems) mit lernenden Komponenten.

Zielgruppe

Das vorliegende Papier möchte Anbietenden und Anwendenden von KI-Systemen sowie der Forschung und politischen Entscheidungsträgerinnen und -trägern weitere Aspekte mit an die Hand geben, die bei der Frage betrachtet werden sollten, welche Systeme vor dem Hintergrund welches Anwendungskontextes ex ante zu regulieren sind und wann dies nicht notwendig ist.

Aufbau des Papiers

Die Frage nach der Kritikalität von KI-Systemen wird von unterschiedlichen Disziplinen zum Teil sehr unterschiedlich betrachtet und auch beantwortet. Aufgrund dieser interdisziplinären Sichtweisen gliedert sich die Publikation in zwei Teile: Im ersten Teil (Kapitel 1–3) werden Ansatzpunkte zur Bewertung der Kritikalität und Operationalisierung des Konzepts vorgestellt, die von unterschiedlichen Disziplinen ähnlich betrachtet werden. Im zweiten Teil der Publikation (Kapitel 4) ergänzen und kommentieren Expertinnen und Experten der Plattform Lernende Systeme in kurzen Interviews die im ersten Teil vorgestellten Ansatzpunkte um weitere Faktoren aus ihrer jeweiligen Fachexpertise heraus. So sollen zum einen die Breite und zum anderen die Vielschichtigkeit von möglichen Antworten auf die Frage nach der Regulierung von KI-Systemen in Abhängigkeit von der Kritikalität abgebildet werden.

1.1 Untersuchungsgegenstand

In dem von der Europäischen Kommission vorgelegten Regulierungsvorschlag vom 21. April 2021 wird Künstliche Intelligenz sehr breit gefasst. So bezeichnet KI „eine Reihe von Technologien, die sich rasant entwickeln und zu vielfältigem Nutzen für Wirtschaft und Gesellschaft über das gesamte Spektrum industrieller und gesellschaftlicher Aktivitäten hinweg beitragen können“ (vgl. Europäische Kommission 2021b, S. 21). Laut der Verordnung erzeugt KI diesen Nutzen über die „Verbesserung der Vorhersage, Optimierung

der Abläufe, Ressourcenzuweisung und Personalisierung digitaler Lösungen, die Einzelpersonen und Organisationen zur Verfügung stehen" (ebenda). Die Autorinnen und Autoren dieses Papiers arbeiten dagegen mit einer engeren Definition von Künstlicher Intelligenz. Ihre hier vorgestellten Überlegungen beziehen sich ausschließlich auf KI-Systeme der neueren Generation, v. a. auf Lernende Systeme im Allgemeinen (siehe Infobox). Sie werden gleichzeitig aber auch für Algorithmic Decision Making-Systems (ADM-Systeme) mit lernenden Komponenten (siehe Infobox) ausgearbeitet.

Die meisten KI-Systeme und auch ADM-Systeme mit lernenden Komponenten sind deutlich komplexer als klassische IT-Systeme.[2] Diese Komplexität ist sowohl auf den Lernvorgang an sich als auch auf dem Umgang mit großen, häufig unübersichtlichen Datenmengen zurückzuführen. Sie kann dazu führen, dass selbstständige Weiterentwicklungen der Systeme möglich sind (weiterlernende Systeme) und Schäden unter anderem deshalb nur sehr schwer vorherzusehen und teilweise auch gar nicht abschätzbar sind (siehe auch Kapitel 2.1).

Lernende Systeme
Lernende Systeme sind Maschinen, Roboter und Softwaresysteme, die abstrakt beschriebene Aufgaben auf Basis von Daten, die ihnen als Lerngrundlage dienen, selbstständig erledigen, ohne dass jeder Schritt spezifisch vom Menschen programmiert wird. Um ihre Aufgabe zu lösen, setzen sie von Lernalgorithmen trainierte Modelle ein. Mithilfe des Lernalgorithmus können viele Systeme im laufenden Betrieb weiterlernen: Sie verbessern die vorab trainierten Modelle und erweitern ihre Wissensbasis. Lernende Systeme basieren auf Methoden der Künstlichen Intelligenz (KI), wobei das maschinelle Lernen eine Schlüsseltechnologie darstellt. Vor allem durch die Fortschritte im Deep Learning entwickelten sich Lernende Systeme in den letzten Jahren zum dynamischsten Bereich der KI-Forschung und -Anwendung (vgl. Plattform Lernende Systeme 2020b).

Algorithmic Decision Making-System (ADM-System) mit lernenden Komponenten
Ein Algorithmus ist eine genaue Berechnungsvorschrift, um eine Aufgabe zu lösen. Ein algorithmisches Entscheidungssystem (Algorithmic Decision Making-System) berechnet Entscheidungen auf Basis eines Algorithmus. ADM-Systeme können nur solche Entscheidungen treffen, für die ein Mensch die entsprechenden Rahmenbedingungen programmiert hat. Im Fall eines trainierten Modells legt der Mensch durch die Programmierung den Trainingsprozess fest, der eigentliche Entscheidungsalgorithmus entsteht durch das entsprechende Training. Unter einem ADM-System mit lernenden Komponenten wird ein System verstanden, das, mithilfe von Künstlicher Intelligenz (KI), Entscheidungen berechnet. Diese Systeme sind mit KI ausgestattete Software-Anwendungen und Algorithmen mit den für die Anwendung erforderlichen Daten, den für die Erhebung der Daten erforderlichen Sensoren, den Schnittstellen zu anderen Systemen sowie der dafür erforderlichen Hardware.

2 Zu Komplexitätsachsen von KI-Systemen siehe SafeTRANS e.V. 2019.

1.2 Begriff der Kritikalität und bestehende Konzepte

Allgemein ist Kritikalität (oder bei der Europäischen Kommission „Risiko") ein Maß für potenzielle Gefahren, die vom Einsatz eines KI-Systems in einem spezifischen Anwendungskontext ausgehen können. Aus der ex ante ermittelten Kritikalität lässt sich die Regulierungsbedürftigkeit (Normsetzung und Ordnung i. w. S.) ableiten: Ist die Kritikalität niedrig, ist wenig Regulierung nötig, während eine hohe Kritikalität mit einem höheren Regulierungsbedarf bis hin zum Verbot eines KI-Systems in bestimmten Anwendungskontexten einhergehen sollte. Bei einer Kritikalitätseinstufung werden neben technischer Komplexität vor allem der Einwirkungsgrad auf Individuen, Gemeinschaften und die Gesellschaft im Ganzen betrachtet. Entscheidend für die Kritikalitätseinstufung in einem Anwendungskontext sind zum einen das Ausmaß der möglichen Verletzungen von Rechtsgütern und Menschenleben, die durch den Einsatz eines KI-Systems in einem bestimmten Anwendungskontext entstehen können, sowie zum anderen der Umfang der Handlungsfreiheiten des Individuums bei der Auswahl und Nutzung der beiden Systemarten in einem bestimmten Anwendungskontext.

Der Regulierungsvorschlag zu KI der Europäischen Kommission unterscheidet vier Risikoklassen, an die jeweils unterschiedliche Anforderungen gestellt werden: unannehmbar, hoch, gering und minimal. Die Anforderungen gehen von einem Verbot über Vorgaben, die vor Markteinführung erfüllt sein müssen, und Transparenzpflichten bis hin zu keinerlei Vorgaben. Ein unannehmbares Risiko besteht bei KI-Systemen, die als Bedrohung für die Sicherheit und die Lebensgrundlage beziehungsweise die Rechte der Menschen gelten. Systeme, die ein unannehmbares Risiko darstellen, sollen verboten werden. Für den Einsatz von KI-Systemen, die unter ein hohes Risiko fallen, sollen spezielle Regulierungsanforderungen gelten.

KI-Systeme mit hohem Risiko werden im Regulierungsvorschlag durch drei mögliche Aspekte charakterisiert:

1. Das KI-System wird als Safety-Komponente eines Produkts eingesetzt oder fällt bereits unter bestimmte EU-Harmonisierungsvorschriften (siehe Europäische Kommission 2021a, S. 44f., Artikel 6 und Europäische Kommission 2021c, S. 2f.) und das Produkt, dessen Sicherheitsbauteil das KI-System ist, oder das KI-System selbst muss gemäß Harmonisierungsrechtsvorschriften der Europäischen Union einer Konformitätsbewertung durch einen Dritten unterzogen werden (siehe ebenda).

2. Das KI-System ist ein System aus einer definierten Liste an kritischen KI-Systemen. Diese umfasst unter anderem KI-Systeme in folgenden Bereichen: Biometrische Fernidentifikation, kritische Infrastrukturen, Schul- oder Berufsausbildung, Beschäftigung, Personalmanagement und Zugang zu selbstständiger Tätigkeit, wichtige private und öffentliche Dienstleistungen, Strafverfolgung, Migration, Asyl und Grenzkontrollen sowie Rechtspflege und demokratische Prozesse.

3. Das KI-System wird in einem unter 2. aufgeführten Bereich eingesetzt und birgt ein Risiko der Verletzung von Gesundheit, Sicherheit oder Grundrechten, dessen Eintrittswahrscheinlichkeit und Schweregrad mindestens genauso hoch sind wie die der unter 2. aufgeführten Systeme. Bei der Risikoermittlung sollen unterschiedliche Einordnungskriterien[3] betrachtet werden (siehe Tabelle 1, Europäische Kommission 2021a, S. 45f.).

Tabelle 1: Einordnungskriterien zur Ermittlung des Risikos eines KI-Systems im Regulierungsvorschlag der Europäischen Union

Einordnungskriterien zur Ermittlung des Risikos eines KI-Systems

- die Anzahl der Nutzenden des KI-Systems
- das Ausmaß, in dem ein KI-System einen Schaden verursacht hat oder das Risiko eines Schadens besteht
- das potenzielle Ausmaß der Auswirkungen des Schadens
- die Wahrscheinlichkeit, dass das KI-System eine hohe Anzahl von Personen schädigt
- die Wahrscheinlichkeit, dass ein KI-System mehr als einen der speziell definierten Schäden verursacht
- Ausmaß der Abhängigkeit der potenziell von einem Ergebnis betroffenen Personen
- Ausmaß der Vulnerabilität der potenziell betroffenen Personen gegenüber der Nutzerin/dem Nutzer eines KI-Systems
- Ausmaß der Reversibilität des von einem KI-System erzeugten Ergebnisses
- Verfügbarkeit und Wirksamkeit von Rechtsmitteln (im Unionsrecht und im Recht der Mitgliedstaaten)
- Ausmaß, in dem die bestehenden Rechtsvorschriften der Union in der Lage sind, die vom KI-System ausgehenden Risiken zu verhindern/zu minimieren

Die von der Europäischen Kommission vorgeschlagenen Einordnungskriterien zur Ermittlung eines hohen Risikos von KI-Systemen basieren auf dem Konzept der Kritikalitätseinschätzung von KI-Systemen. Ähnliche Konzepte zur Kritikalitätseinstufung wurden bereits ausführlich diskutiert:

- Katharina Zweig und Tobias Krafft beziehen diesen Ansatz ausschließlich auf ADM-Systeme (unabhängig davon, ob diese auf Lernenden Systemen basieren) und geben einen Überblick darüber, welche Anforderungen hinsichtlich Transparenz und Nachvollziehbarkeit für welche Systeme gelten sollten (vgl. 2019). Hierfür erarbeiteten beide ein Konzept zur Einteilung von ADM-Systemen in unterschiedliche Regulierungsklassen. Die Einordnung erfolgt vor dem Einsatz des Systems anhand zweier Dimensionen: das Schadenspotenzial auf individuellem und gesamtgesellschaftlichem Level sowie die Möglichkeit einer Schadensbehebung oder einer Neubewertung im Falle einer Fehlbeurteilung. Anhand der Ausprägungen der jeweiligen Dimensionen

3 In einer Vorabversion des Regulierungsvorschlags wurden diese Kriterien noch nach Basiskriterien, die immer analysiert werden, sowie zusätzlichen Kriterien, die in Abhängigkeit des Einsatzzwecks analysiert werden, unterschieden. Dass diese Aufteilung entfallen ist und stets alle Einordnungskriterien abgefragt werden sollen, begrüßen die Autorinnen und Autoren.

erfolgt die Zuordnung zu einer von fünf unterschiedlichen Regulierungsklassen im Spektrum von „keine Regulierungsanforderungen" über „Erfüllung bestimmter Prüf- und Transparenzpflichten" bis hin zu „Verbot des jeweiligen ADM-Systems".

- Die Grundidee dieses Konzepts diente als Grundlage für das Rahmenwerk der Artificial Intelligence Ethics Impact Group, welche ein Ethics Rating entwickelt hat, um KI-Ethik messbar zu machen (2020). In der Technikfolgenabschätzung wird gelegentlich das Konzept der „Eingriffstiefe" verwendet, in dem ein ähnlicher Gedanke wie in vorliegendem Papier zu Kritikalitätsstufen verfolgt wird. Auch dort geht es darum, kritische von weniger kritischen Eingriffen – z. B. in Ökosysteme oder Organismen – zu unterscheiden.[4]

Auch in Strategien und Gutachten nationaler und internationaler Akteure findet sich bereits der Ansatz der Regulierung nach unterschiedlichen Risiken.

- Datenethikkommission: So baut beispielsweise die Datenethikkommission in ihrem Abschlussbericht auf dem Modell der Kritikalitätsstufen in Anlehnung an Zweig & Krafft auf (siehe 2019, S. 173).
- Bundesregierung: Die Bundesregierung spricht sich in ihrer Stellungnahme zum Weißbuch der Europäischen Kommission ebenfalls für einen risikobasierten Ansatz zur Regulierung von KI-Systemen aus (siehe 2020). Dieser Ansatz sollte mehr als zwei Stufen und insbesondere Schadenspotenziale und Anwendungskontexte umfassen (siehe 2020).
- UNESCO: Die UNESCO erarbeitet zurzeit Empfehlungen zur Ethik in der KI – auch hier wird ein differenzierter Regulierungsansatz aufgegriffen (siehe 2020).

Risikoabschätzungen sind ein bewährtes Konzept, das sich auch in bestehenden (domänenübergreifenden sowie domänenspezifischen) Normen und den dazugehörigen Prüfverfahren wiederfindet. Auch diese können einen Anknüpfungspunkt für Kritikalitätsüberlegungen darstellen. Da sie technologieunabhängig formuliert sind, gelten sie auch für KI-Systeme und müssen bei Überlegungen zur Kritikalität von KI-Systemen mitbedacht werden. Beispiele hierfür sind die IEC-Norm 61508 sowie daraus abgeleitet die ISO-Norm 26262 im Bereich Mobilität.

- IEC-Norm 61508: Diese Norm ist eine internationale Normserie zur Entwicklung von elektrischen, elektronischen und programmierbaren elektronischen Systemen, die eine Sicherheitsfunktion ausführen. Sie wird von der Internationalen Elektrotechnischen Kommission (IEC) herausgegeben. Ziel dieser Norm ist es, Verfahrensweisen festzulegen, mit denen Produkte hergestellt werden können, die keine unverhältnismäßigen oder unvertretbaren Gefahren für Anwendende und die Umwelt bergen. Die IEC 61508 ist eine Sicherheitsgrundnorm, welche die Basis für anwendungsbezogene Normen, wie beispielsweise die ISO-Norm 26262 für den Mobilitätssektor, bietet.[5]

4 Siehe hierfür beispielsweise von Gleich (1998).

5 Für folgende Sektoren wurden aus der Sicherheitsgrundnorm konkrete Normen abgeleitet: Prozessindustrie (IEC 61511), Kernkraftwerke (IEC 61513), Bahnanwendungen (EN 50128), Maschinen (IEC 62061) sowie Software-Entwicklung für landwirtschaftliche Traktoren und Maschinen (ISO 25119).

- ISO-Norm 26262 „Road vehicles – Functional Safety“: Diese Norm bezieht sich auf sicherheitsrelevante elektrische oder elektronische Systeme in Kraftfahrzeugen. Sie wird von der Internationalen Organisation für Standardisierung (ISO) herausgegeben. Ziel der ISO-Norm ist die Sicherstellung der funktionalen Sicherheit (d.h. korrekten Funktionsweise) eines Systems – auch in Wechselwirkungen mit anderen Systemen – im Kraftfahrzeug. Das Rahmenwerk der ISO-Norm 26262 wird bereits bei der Entwicklung von Systemen angewandt und umfasst unter anderem eine Folgenabschätzung sowie eine Risikobestimmung und -bewertung. Die Bewertung erfolgt über die Einteilung in eine von vier ASIL-Levels (Automotive Safety Integration Level, A–D). Aus der Einteilung in die ASIL-Kategorie ergeben sich dann je nach Stufe unterschiedliche Anforderungen, die im Entwicklungsprozess erfüllt werden müssen: Mit einem höheren ASIL-Level steigen die Anforderungen und umgekehrt.

2. Risiken und Schäden bei KI-Systemen

Die Risiken, die der Einsatz von KI-Systemen mit sich bringen kann, und die Schäden, die bei diesen Einsätzen entstehen können, sind vielfältig (vgl. Fraunhofer IAIS 2019, 2021). Gleichzeitig lassen sie sich in Bezug auf KI-Systeme nur schwer vorhersehen und abschätzen. Deshalb muss vor dem Einsatz dieser Systeme auch geklärt werden, wie im Falle von Schäden die Verantwortung unter den Akteuren aufgeteilt werden sollte. Das folgende Kapitel geht erst einen Schritt hinter den Regulierungsvorschlag für KI der Europäischen Kommission zurück und stellt unterschiedliche Akteure, Verwendungsintentionen und Anwendungskontexte vor und erklärt, warum es sich bei der Betrachtung der Kritikalität eher um eine Gefahrenvorsorge statt um eine konkrete Gefahrenabwehr handelt. Anschließend wird ein Modell zur Verantwortungsaufteilung im Fall von Schäden durch KI-Systeme vorgestellt.

2.1 Akteure, Verwendungsintentionen und Anwendungskontexte

Bei der Betrachtung der Kritikalität von KI-Systemen in bestimmten Anwendungskontexten müssen unterschiedliche Verantwortungs- und Betroffenheitsdimensionen betrachtet werden. Wir unterscheiden hierfür zwischen: Anbietenden (z. B. Hersteller einer KI-Software), Nutzenden (z. B. einer Ärztin in einem Krankenhaus) sowie Betroffenen (z. B. einem Patienten). In dieser Systematik kann ein Akteur – in Abhängigkeit vom Anwendungskontext – auch mehrere Rollen gleichzeitig einnehmen.

Bei der Verwendung von KI-Systemen werden, wie auch bei anderen technischen Systemen, vier verschiedene Verwendungsfolgen unterschieden: intendierte Folgen, intendierte verbotene Folgen (verbotene Einsatzzwecke), nicht-intendierte Folgen (Kollateralschäden) sowie nicht-intendierte verbotene Folgen (verbotene Kollateralschäden).[6] Für die Betrachtung der Kritikalität werden nur die intendierten Folgen, nicht-intendierten Folgen sowie nicht-intendierten verbotenen Folgen betrachtet. Intendierte verbotene Folgen sowie der Missbrauch der Systeme für illegale Zwecke oder auch für militärische Anwendungen[7] werden in dieser Betrachtung ausdrücklich nicht berücksichtigt. Neben diesen Verwendungsfolgen ist bei der Betrachtung der Kritikalität einer der beiden Systemarten auch die Art des Anwendungskontextes relevant. Hierbei kann zumindest zwischen wirtschaftlichen (Business-to-Consumer, B2C[8] und Business-to-Business, B2B)[9] und staatlichen (Government-to-Consumer, G2C[10] oder auch Government-to-Business, G2B)[11] Anwendungskontexten unterschieden werden.

6 Ein Beispiel wäre die Fahrt eines autonomen Fahrzeugs, das auf einen Zebrastreifen zufährt, über den eine Person geht. Die intendierte Folge wäre, dass das Auto abbremst, und die nicht-intendierte Folge, dass es so abrupt abbremst, dass die fahrende Person so stark in den Sicherheitsgurt gedrückt wird, dass sie oder er sich verletzt. Eine nicht-intendierte verbotene Folge würde vorliegen, wenn die Systeme des autonom fahrenden Autos die Person nicht erkennen und sie überfahren würde. Ein weiteres Beispiel wäre die Nutzung eines autonomen Rasenmähroboters: Die intendierte Folge wäre der gemähte Rasen, die nicht-intendierte Folge der Lärm des Rasenmähers und die nicht-intendierte verbotene Folge der Lärm des Rasenmähers in der Nacht.

7 Diese sind beabsichtigt und ordnungsgemäß, können aber durchaus äußerst kritisch sein.

8 Business-to-Consumer beschreibt die Beziehung zwischen Unternehmen und Konsumentinnen und Konsumenten.

9 Business-to-Business beschreibt die Beziehung von Unternehmen untereinander.

10 Government-to-Consumer beschreibt die Beziehung der staatlichen Institutionen zu Bürgerinnen und Bürgern.

11 Government-to-Business beschreibt die Beziehung der staatlichen Institutionen zu Unternehmen.

2.2 Gefahrenabwehr versus Vorsorge

Die Risikoforschung hat vielfach gezeigt, dass Risiken nicht nur aus technischen Eigenschaften, sondern auch aus der Wechselwirkung mit der gesellschaftlichen Umwelt und der eingesetzten Technik entstehen können. Dies gilt insbesondere in Anwendungskontexten, in denen beispielsweise unvorhergesehenes Verhalten von Menschen sozio-technische Risiken mit sich bringen kann. Um diese in den Blick nehmen zu können, ist es erforderlich, in der Bewertung der Kritikalität von KI-Systemen auch den gesellschaftlichen Kontext der technischen Systeme, einschließlich des Verhaltens von Nutzenden, Betreibern und anderer Akteure, mit einzubeziehen.

Anders als bei einer klassischen Risikoeinschätzung, bei der in sicherheitsrelevanten Domänen oft Hazard-Analysen (siehe Infobox) angewandt werden, wird im vorliegenden Papier der Fokus auf den Umgang mit Risiken in einem schwer operationalisierbaren Handlungsfeld gelegt.

Hazard-Analyse
Unter einer Hazard-Analyse wird eine Gefahrenanalyse verstanden. Sie ist meist der erste Schritt bei einer Risikobewertung und hat das Ziel, potenziell bedrohliche Situationen – englisch: Hazards – in der Nutzung des Systems und systembedingte Ursachen für deren Auftreten zu ermitteln. Diese Situationen können sich in Kombination mit weiteren Faktoren zu einem Unfall weiterentwickeln; in einem zweiten Schritt wird mithilfe einer Hazard-Analyse der Schweregrad von möglichen Schäden klassifiziert.

Es existieren zwei Arten von Hazard-Analysen: In einer qualitativen Hazard-Analyse (z. B. der HAZOP-Analyse) werden kausale Wirkketten im System identifiziert, die zu einer potenziell bedrohlichen Situation führen können. Quantitative Hazard-Analysen setzen voraus, dass für die kausalen Ursachen Auftretenswahrscheinlichkeiten bekannt sind, sodass auch die Auftretenswahrscheinlichkeit von Bedrohungssituationen berechnet werden kann.

Diese Entscheidung ist darin begründet, dass sich Hazard-Analysen größtenteils auf Situationen erstrecken, in denen die Systemarchitektur bekannt ist und belastbare Wahrscheinlichkeitsaussagen im Vorhinein vorliegen. Diese basieren im Idealfall auf empirischer Erfahrung mit entsprechenden Systemen oder validierten Modellaussagen. Vielfach ist dieses Verfahren bei neuen Technologien nicht oder nur begrenzt anwendbar, da entsprechende Schadenspotenziale nur sehr schwer vorherseh- oder abschätzbar sind. Diese werden oft eher vermutet und befürchtet, als dass sie bereits auf Basis belastbaren Wissens bekannt sind. Bei KI-Systemen liegt diese Situation aufgrund der Komplexität der Systeme noch einmal in verschärfter Form vor. Dies betrifft vor allem vermutete individuelle wie gesellschaftliche Risiken, die beispielsweise mögliche Verletzungen von Normen oder Verfassungsprinzipien beinhalten und sich wegen der unsicheren Prädiktionsbasis weder hinreichend quantifizieren noch belastbar qualitativ erfassen lassen. In diesen Situationen geht

es zum gegenwärtigen Zeitpunkt der Forschung daher auch um Vorsorgemaßnahmen wie beispielsweise die Schärfung des Problembewusstseins und weitere Forschung zur Verbesserung des Wissensstandes und der Risikoabschätzung.

2.3 Verantwortung, Verantwortungsketten und Haftung

Regulierung oder das Haftungsrecht allein können nicht sicherstellen, dass beim Einsatz von Diensten, die auf KI-Systemen basieren, stets alle von möglichen Risiken und Schäden Betroffenen vollkommen geschützt werden. Trotz des Ziels einer optimalen Schadensvermeidung kann jedoch nicht ausgeschlossen werden, dass bestimmte Personen(-gruppen) beim Einsatz von KI-Systemen – wie auch beim Einsatz anderer IT-Systeme – Nachteile bis hin zur ungerechtfertigten Diskriminierung erleiden können. Deshalb sollten betroffene vulnerable Personen(-gruppen) sowohl bei der Ermittlung der Kritikalität als auch bei der Festlegung von Ex-post-Maßnahmen im Vorhinein hinsichtlich beider Systemarten Beachtung finden. So ist es unerlässlich, bereits vor dem Einsatz des Systems die Datensouveränität (vgl. Hummel et al. 2021) der Datensubjekte zu stärken. Dabei geht es darum, dass betroffene Datensubjekte im Datenverarbeitungsprozess effektive Kontrollier- und Überprüfungsmöglichkeiten erhalten. Dies geschieht durch Verbraucherschutzregime, die beispielsweise niedrigschwellige, zeitnahe und vor allem effektive Beschwerdemöglichkeiten beinhalten müssen.

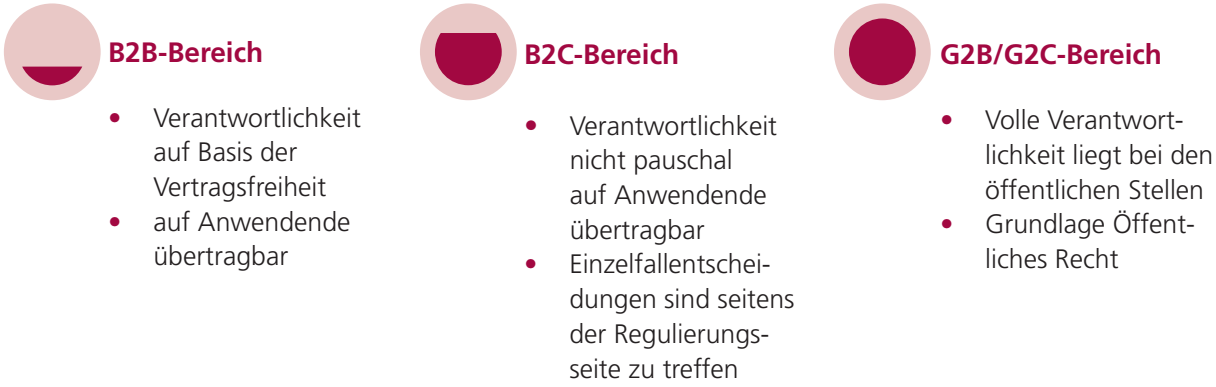
Neben der Regulierung anhand der Kritikalität im Vorhinein und dem Schaffen von Verbraucherschutzregimen für mögliche Schadensfälle bildet auch die Aufteilung der Verantwortung des Risikos über Haftungsregelungen einen wichtigen Baustein hin zu vertrauenswürdigen KI-Systemen. Ziel der (Unternehmens-)Haftung ist es, das technische Risiko nach Gesichtspunkten der Fairness und Funktionalität unter den unterschiedlichen Akteuren bestmöglich zu teilen. Dabei geht es zentral um die Frage, wer ein Risiko besser einschätzen, besser tragen und gegebenenfalls beseitigen kann.[12] Überlegungen zur Verantwortung und auch Haftung bei beiden Systemarten dürfen jedoch nicht nur national gedacht werden, sondern sind in Europa auch unter dem Gesichtspunkt des einheitlichen europäischen Marktes zu betrachten, damit keine nicht-tarifären Handelshemmnisse entstehen. Es ist damit für jede Regulierungsdiskussion von KI essenziell, die laufende Diskussion auf dem europäischen Binnenmarkt zu beobachten und gegebenenfalls zu beeinflussen. Dies gilt besonders für die laufende Diskussion zur Erneuerung der europäischen Produkthaftungsrichtlinie (product liability directive, 85/374/EEC), da diese zukünftig auch digitale Produkte mit einschließen soll.

Der Regulierungsvorschlag für KI-Systeme der Europäischen Kommission verlangt, dass ein eindeutig identifizierbarer Wirtschaftsakteur die rechtliche Verantwortung für das Endprodukt als Ganzes tragen muss (vgl. 2021, S. 31).

12 Für eine Diskussion zur Unternehmenshaftung siehe beispielsweise H. Koziol (2014), K. Wantzen (2007) oder G. Brüggemeier (2006).

Die Autorinnen und Autoren empfehlen folgende Präzisierung: Die Verantwortungsketten[13] bei KI-Systemen sind nach dem Adressaten und gegebenenfalls nach der Kritikalität des jeweiligen Anwendungsgebiets zu unterscheiden: So ist es ein fundamentaler Unterschied, ob Entscheidungen der Systeme primär Einfluss auf die Integrität von Gesundheit und Leben haben oder Einfluss auf Sachwerte nehmen. So sollten bei Haftungsfragen im B2B-Bereich grundsätzlich die Bedingungen des Vertragsrechts gelten (außer in gesellschaftlich kritischen Bereichen), während im B2C-Bereich Entscheidungen in Abhängigkeit von der Kritikalität getroffen werden sollten. Für die Konsumierenden ändert sich somit nichts im Vergleich zur Verantwortungskette bei herkömmlicher Hardware – obwohl ihnen die rechtliche Sachlage bei der Anwendung oft nicht bewusst ist (z. B. Telefon-Hotline-Service auf Basis eines Chatbots). Für den G2B- und G2C-Bereich gelten aufgrund der besonderen Beziehung der öffentlichen Hand zur Bevölkerung (inkl. Unternehmen und Privatpersonen) die höchsten Anforderungen an die Betreiber beider Systemarten (siehe Abbildung 2).

Abbildung 2: Verantwortungsketten nach Bereich



Verantwortungsketten im…

- B2B-Bereich: Für den B2B-Bereich bedeutet das konkret, dass die Verantwortlichkeit auf Basis der Vertragsfreiheit grundsätzlich auf die Anwenderin oder den Anwender übertragen werden kann. Ausnahmen müssen je nach Gesamtbewertung für gesellschaftlich kritische Bereiche geschaffen werden, bei denen
  - entweder im rechtlichen Sinne diskriminierungsfreie Verfügbarkeit gelten soll, wie zum Beispiel bei Personalentscheidungen, oder
  - die Anwendung Einfluss auf die persönliche und körperliche Integrität hat, wie zum Beispiel Systeme im medizinischen Kontext, oder
  - der Einsatz in als kritische Infrastrukturen angesehenen Bereichen, wie beispielsweise Energie- und Wasserversorgung, Transportinfrastruktur, und den wesentlichen darauf aufbauenden Dienstleistungen – beispielsweise im Gesundheitswesen etc. – erfolgt.

13 Die Verantwortungsketten beschreiben eine systemübergreifende Aufteilung von Verantwortlichkeiten. Hierbei kann weder eine eindeutige/alleinige Verantwortungszuweisung erfolgen, noch liegt eine eindeutig gemeinschaftliche Verantwortung vor.

- B2C-Bereich: Im B2C-Bereich sind Verantwortungszuschreibungen von Einzelfallentscheidungen über den Kritikalitätsstatus durch die Regulierungsseite abhängig. Diese stehen etwa in Abhängigkeit zu der Frage, ob die oder der einzelne Anwendende sich der Entscheidung eines KI-Systems im Zweifel entziehen kann beziehungsweise sich ihrer bewusst ist. Darüber hinaus ist eine Betrachtung der Eingriffstiefe des Systems in Bezug auf die Persönlichkeitsebene zu berücksichtigen. Konkret auf ein Beispiel übertragen heißt das: Konsumierende müssten ihrem intelligenten Kühlschrank explizit gestatten, auf Basis angelernter Verbrauchsmuster selbsttätig rechtzeitig Nachschubbestellungen bestimmter Lebensmittel durchzuführen – idealerweise in einem zu bestimmenden Rahmen (Menge oder Preis). Alternativ könnte die Konsumentin oder der Konsument auch vor die Wahl gestellt werden, jede einzelne Kühlschrankbestelung freizugeben oder abzulehnen. Ein weiteres Beispiel ist der Einsatz von ADM-Systemen mit lernenden Komponenten im Aktienhandel. Hier werden sowohl der Kauf als auch der Verkauf automatisiert durchgeführt und als Entscheidungsparameter nur Parameter eingetragen beziehungsweise verändert. Wie bereits ausgeführt, zeigt dieses Beispiel, wie auch in jeder anderen B2C-Orientierung, dass der Transparenz der gewählten Automatisierungsmodelle und der Frage, ob sich eine einzelne Person der Entscheidungsinstanz entziehen kann beziehungsweise ausweichen kann, eine Schlüsselrolle zukommen. Gleichzeitig ist immer abzuwägen, ob die Nutzenden in den jeweiligen technischen Handlungssituationen überhaupt in der Lage sind, zu einer wohlinformierten Entscheidung zu kommen. Bei einer hohen Kritikalität sollte erwogen werden, die Vertragsfreiheit insofern einzuschränken und die Verantwortlichkeit beim Hersteller zu belassen, da die Vertragsbeziehung eine sachgerechte Verantwortungsbeziehung aller Voraussicht nach nicht garantieren kann. Dies ist beispielsweise der Fall, wenn die Entscheidungen des KI-Systems natürliche Personen betreffen und zu einer psychischen oder physischen Beeinträchtigung führen könnten.[14] Bei grober Fahrlässigkeit oder kriminellem Verhalten ist jedoch nach wie vor die Anwenderin oder der Anwender in der Verantwortlichkeit (Bedienungsfehler sind zum Beispiel über Log-Dateien auslesbar).
- G2B- und G2C-Bereich: Im G2B- und G2C-Bereich müssen sich öffentliche Stellen jederzeit und gegebenenfalls flankierend automatisiert für ihre KI- und auch ADM-gestützten Verwaltungsvorgänge erklären können. Sie tragen die volle Verantwortung für entstehende ungerechtfertigt diskriminierende oder schädliche Konsequenzen im Rahmen des Öffentlichen Rechts. Diese grundlegende Anforderung ergibt sich aus dem Umstand, dass die öffentliche Hand in einer umfassenden Rechtfertigungsposition gegenüber der Bevölkerung steht – sowohl im allgemeinen Sinne vor allem im Bereich G2B, aber auch im Speziellen mit einer Auswirkung auf einzelne Bürgerinnen und Bürger (G2C).

14 An dieser Stelle ist darauf hinzuweisen, dass die gewählten Beispiele nicht zwingend in jedem Fall und auf alle Arten der verwendeten Systeme angewendet werden können. Sie dienen vor allem der Verdeutlichung der unterschiedlichen zu diskutierenden Fragestellungen.

3. Kritikalitätsbewertung von KI-Systemen vor dem Hintergrund ihrer Anwendungskontexte

Die Autorinnen und Autoren dieses Papiers weisen darauf hin, dass das Konzept der Kritikalitätsbewertung von KI-Systemen Schwachstellen aufweist. Gründe hierfür sind die starke Komplexitätsreduktion, die bei einer Einordnung vorgenommen wird, sowie die eingangs dargestellte kaum vorhandene Vorherseh- und Abschätzbarkeit von Schäden, die beim Einsatz von KI-Systemen vorliegen können (siehe Kapitel 1.1 und 2.1). Dennoch ist das Konzept eine vielversprechende Lösung und bildet die Basis des Regulierungsvorschlags für KI der Europäischen Kommission. Deshalb möchten die Autorinnen und Autoren auf Aspekte hinweisen, die ihrer Einschätzung nach bei der Ermittlung der Kritikalität eines KI-Systems in seinem jeweiligen Anwendungskontext beachtet werden müssen, um für diese Klassifizierungsform ein praktikables, wenn auch nicht endgültiges Ergebnis zu erzielen. Hierfür wird in folgendem Kapitel zunächst auf Kriterien zur Ermittlung der Kritikalität eingegangen und diese näher erläutert ehe anschließend Aspekte thematisiert werden, die bei einer Operationalisierung des Konzepts beachtet werden sollten.

3.1 Dimensionen zur Bewertung

Die Europäische Kommission schlägt in ihrem Regulierungsvorschlag sowohl eine Liste mit Domänen, deren Systeme immer als hochriskant eingeschätzt werden sollten, als auch eine Bewertung anhand unterschiedlicher Kriterien vor (siehe Tabelle 1). Die Autorinnen und Autoren des Whitepapers plädieren jedoch für eine Einzelfallbewertung von KI-Systemen vor dem Hintergrund des jeweiligen Anwendungskontextes anhand bestimmter Kriterien. Die im EU-Regulierungsvorschlag vorgestellten Kriterien sind nach Meinung der Autorinnen und Autoren noch durch weitere Kriterien zu ergänzen. Folgendes Unterkapitel stellt mögliche weitere Kriterien und deren Präzisierung vor.

Die Autorinnen und Autoren schlagen vor, die Kriterien zur Bewertung der Kritikalität eines KI-Systems in einem bestimmten Anwendungskontext in zwei Dimensionen einzuteilen: 1. Möglichkeit und Ausmaß der Verletzung von Menschenleben und weiteren hohen Rechtsgütern und 2. Umfang der Handlungsfreiheiten des Individuums (siehe Abbildung 3). Je höher das Ausmaß der möglichen Verletzung von Menschenleben und weiteren hohen Rechtsgütern und je geringer der Umfang der Handlungsmöglichkeiten des Individuums, desto höher ist die Kritikalität – und daraus abgeleitet der Regulierungsbedarf – und vice versa. Wichtig ist darüber hinaus, dass diese Dimensionen stets abgefragt werden sollten.

Abbildung 3: Kritikalität von KI-Systemen vor dem Hintergrund ihres jeweiligen Anwendungskontextes

Höhere Kritikalität

- Möglichkeit und Ausmaß der Verletzung von Menschenleben und weiteren hohen Rechtsgütern: **groß**
- Umfang der Handlungsmöglichkeiten der Individuen: **gering**

Niedrigere Kritikalität

- Möglichkeit und Ausmaß der Verletzung von Menschenleben und weiteren hohen Rechtsgütern: **gering**
- Umfang der Handlungsmöglichkeiten der Individuen: **groß**

Möglichkeit und Ausmaß der Verletzung von Menschenleben und weiteren Rechtsgütern

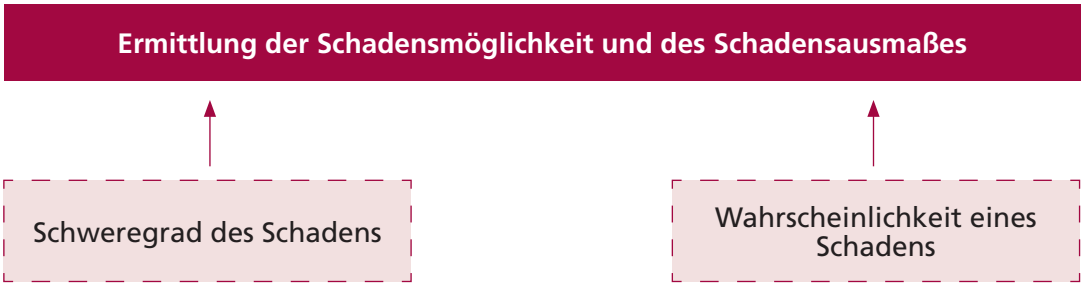
Entscheidend für die Bestimmung der Kritikalität sind nach Meinung der Europäischen Kommission Kriterien, die laut Einschätzung der Autorinnen und Autoren auch als „die Möglichkeit und das Ausmaß der möglichen Verletzung von Menschenleben und weiteren Rechtsgütern“[15] bezeichnet werden können. Sie können dann im nächsten Schritt differenziert hergeleitet und untergliedert werden. Diese Kriterien können allerdings nicht immer quantifiziert werden, da verschiedene gesellschaftliche Gruppen häufig unterschiedlich betroffen sind.[16]

Die Möglichkeit und das Ausmaß der Verletzung von Menschenleben und weiteren hohen Rechtsgütern könnten nach Meinung der Autorinnen und Autoren durch zwei Aspekte genauer beschrieben werden: den Schweregrad des Schadens sowie die Wahrscheinlichkeit eines Schadens (siehe Abbildung 4).

15 Beispiele für weitere Rechtsgüter sind Diskriminierungsfreiheit, Persönlichkeitsschutz und Teilhabechancen.

16 Die Ermittlung des Ausmaßes der möglichen Verletzung von Rechtsgütern und Menschenleben lehnt sich an eine klassische quantitative Hazard-Analyse an. Aus den oben dargestellten Gründen (siehe Kapitel 2.2, v. a. oft fehlende hinreichend belastbare Informationen über Wahrscheinlichkeit, Anzahl und Ausmaß möglicher Schäden) wird das Konzept der Hazard-Analyse jedoch nur als Orientierung genutzt.

Abbildung 4: Ermittlung der Möglichkeit und des Ausmaßes der Verletzung von Menschenleben und weiteren Rechtsgütern



Schweregrad des Schadens

In einer Vorabversion des Regulierungsvorschlags der Europäischen Kommission beschreiben zwei Einordnungskriterien den Schweregrad des Schadens: „das Ausmaß, in dem ein KI-System einen Schaden verursacht hat oder das Risiko eines Schadens besteht" und „das potenzielle Ausmaß der Auswirkungen des Schadens" (siehe auch Tabelle 1). Auch im finalen Vorschlag orientieren sich die Kriterien an diesen Fragen. Wie diese Kriterien konkret operationalisiert werden können, ist zum gegenwärtigen Zeitpunkt noch zu klären.

Mögliche Schäden beim Einsatz von KI-Systemen sind vielfältig und können unterschiedliche Akteure betreffen (siehe auch Kapitel 2.1). Obwohl der Schweregrad des Schadens in einigen Bereichen bereits domänenspezifisch definiert wird,[17] ist in vielen anderen Bereichen noch ein gesellschaftlicher Diskurs über mögliche Schäden und deren Bewertung zu führen. Deshalb kann an dieser Stelle nur eine Beschreibung möglicher Schäden erfolgen.

KI-Systeme können sowohl einzelnen Individuen (individueller Schaden) als auch Gruppen wie der gesamten Gesellschaft (gesamtgesellschaftlicher Schaden) Schaden zufügen. Zu beachten ist, dass der gesamtgesellschaftliche Schaden aufgrund unterschiedlicher Verstärkereffekte größer sein kann als die bloße Summe der einzelnen individuellen Schäden.[18] Sowohl die individuellen als auch die gesamtgesellschaftlichen Schäden können entweder in Form von materiellen oder in Form von immateriellen Schäden auftreten. Materielle Schäden beschreiben im juristischen Sinne (hohe) Sachschäden oder Verletzungen der körperlichen Unversehrtheit bis hin zur Gefährdung von Menschenleben. Beispiele für materielle Schäden sind Schäden an materiellen Gütern, Schäden an kritischen Infrastrukturen und ökologische Schäden sowie Umweltkatastrophen. Unter immateriellen Schäden versteht man das Ergebnis der Verletzung von Persönlichkeitsrechten, also beispielsweise die Verletzung von Privatheit oder die Verletzung von sozialen Normen wie Gerechtigkeit

17 Siehe z. B. Association for the Advancement of Automotive Medicine (2020).

18 Außer Acht bleiben muss an dieser Stelle, ob es unter Umständen gerechtfertigt ist oder in Kauf genommen werden kann, dass bestimmte Individuen oder Gruppen benachteiligt werden, während der gesamtgesellschaftliche Nutzen maximiert wird und dem Gemeinwohl zugutekommt.

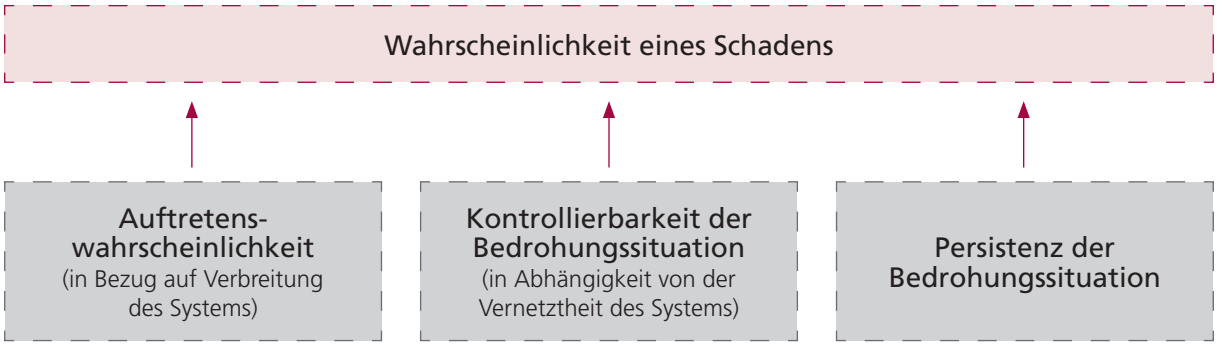
und Fairness, oder die Verletzung normativ-regulativer Vorgaben sowie die Gefährdung der Stabilität der Demokratie. Im Fall von KI-Systemen werden unter immateriellen Schäden zum Beispiel die Benachteiligung von einzelnen Individuen oder ganzen Gruppen anhand bestimmter, als diskriminierend zu bewertender Merkmale, die Verletzung der Privatsphäre, Rechtsgutverletzungen (z. B. informationelle Selbstbestimmung) sowie die ungerechten Behandlungen durch eine der beiden Systemarten verstanden. Es ist wichtig und notwendig, dass bei der Schadensbewertung sowohl immaterielle als auch materielle Schäden berücksichtigt werden: Beide Kategorien stehen in einem komplexen Verhältnis zueinander. Um zur Gewichtung von materiellen und immateriellen Schäden als breit geteilte Kriterien zu gelangen, sollte ein möglichst konsensfähiges Konzept unter Einbeziehung aller relevanten Akteure gefunden werden.

Wahrscheinlichkeit eines Schadens

Im Regulierungsvorschlag der Europäischen Kommission bezieht sich ein Kriterium auf die Wahrscheinlichkeit eines Schadens: die „Anzahl der Nutzenden" (siehe Tabelle 1). Die Wahrscheinlichkeit eines Schadens hängt nach Meinung der Autorinnen und Autoren insgesamt von drei Faktoren ab: Auftretenswahrscheinlichkeit, Dauer der Bedrohungssituation und Kontrollierbarkeit. Diese Erweiterung ist folglich präziser und könnte bei der konkreten Umsetzung des EU-Regulierungsvorschlags zu Rate gezogen werden.

Der erste Faktor bezieht sich direkt auf die Nutzung des Systems und die beiden anderen Faktoren beziehen sich auf das Auftreten und den Umgang mit einer Bedrohungssituation (siehe Abbildung 5).

Abbildung 5: Ermittlung der Wahrscheinlichkeit eines Schadens



1. Hier ist zuerst die Auftretenswahrscheinlichkeit von Schäden zu betrachten. Da die Wahrscheinlichkeit und die genaue Anzahl möglicher Schäden bei KI-Systemen meist nicht bekannt sind (siehe Kapitel 1.1 und 2.2), ist – wie auch im Regulierungsvorschlag vorgesehen – eine Hilfskonstruktion zielführend: Die Anzahl der Schäden korreliert mit

der Anzahl der Nutzenden und der Nutzungshäufigkeit dieser. Die Verbreitung eines Systems könnte also mit ausschlaggebend für die Höhe der Wahrscheinlichkeit sein, sodass mit der Nutzung des Systems zusammenhängende Schäden auftreten.[19]

2. Ein weiterer Faktor ist die Dauer, für die eine Bedrohungssituation, aus welcher ein Schaden entstehen kann, anhält (Persistenz). Eine Bedrohungssituation kann oft beendet werden, indem der Einsatz des Systems beendet wird. Je länger die Persistenz einer Bedrohungssituation ist, desto höher ist die Wahrscheinlichkeit für einen Schaden.

3. Zuletzt ist noch die Kontrollierbarkeit einer Bedrohungssituation (in Abhängigkeit der Vernetztheit eines Systems) zu betrachten. Sie ist ein Maß für die Möglichkeit, damit die oder der Nutzende die Bedrohungssituation beenden oder so weit begrenzen kann, dass kein (weiterer/größerer) Schaden auftritt. Dies hängt unter anderem von der Vernetztheit des Systems ab. Diese gibt Auskunft darüber, ob – und wenn ja, wie stark und schnell – sich eine Bedrohungssituation oder auch ein Schaden in einem System auf weitere Systeme auswirken kann. Wenn über eine Vernetztheit Fehleinschätzungen an weitere Systeme übertragen werden, treffen diese möglicherweise wiederum ihre Entscheidungen auf Grundlage falscher Informationen, sodass diese Bedrohungssituationen in allen miteinander vernetzten Systemen auftreten können (vorausgesetzt, es werden keine Maßnahmen zur Erkennung und Isolation der falschen Informationen getroffen). Betrachtet man den Grad der Vernetztheit, bilden in der Außenbeziehung vollständig isolierte Systeme das eine Ende des Kontinuums und in Innen- wie Außenbeziehung vollständig vernetzte Systeme das andere Ende des Kontinuums. Die Ausbreitung der Bedrohungssituation steigt mit dem Grad der Vernetztheit, während die Kontrollierbarkeit der Bedrohungssituation mit dem Grad der Vernetztheit sinkt. Dies gilt vor allem dann, wenn die Kontrollmittel schwach sind.

Als Fazit lässt sich festhalten: Neben der Auftretenswahrscheinlichkeit in Bezug auf die Verbreitung des Systems sollten nach Auffassung der Autorinnen und Autoren auch die Persistenz der Bedrohungssituation sowie der Grad der Kontrollierbarkeit der Bedrohungssituation in die Wahrscheinlichkeitsbewertung für einen Schaden – und somit in die Kriterien des Konzepts der Europäischen Kommission – mit einfließen.

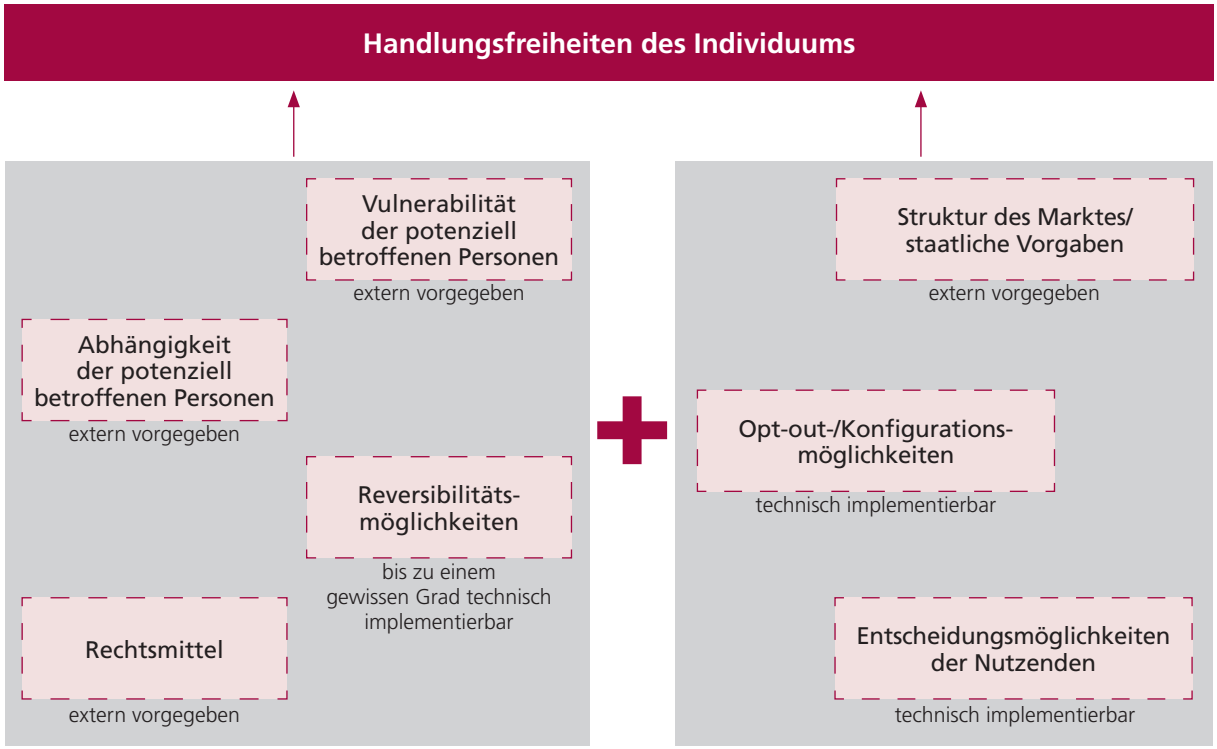
Umfang der Handlungsfreiheiten des Individuums

Von Relevanz ist darüber hinaus, bis zu welchem Grad ein Individuum einem KI-System in einem bestimmten Anwendungsgebiet vorwiegend passiv ausgesetzt ist und/oder über welchen eigenen Handlungsspielraum es verfügt. Im EU-Regulierungsvorschlag werden hierzu ebenfalls Kriterien abgefragt. Diese sind: die Abhängigkeit der potenziell von einem Ergebnis betroffenen Personen, die Vulnerabilität der potenziell von einem Ergebnis

19 Dennoch kann auch ein punktuell eingesetztes System zu einem hohen Schweregrad des Schadens führen – beispielsweise ein Unfall in einem Atomkraftwerk (siehe S. 14, Schweregrad des Schadens). Die Wahrscheinlichkeit für ein Auftreten dieses Schadens, die hier betrachtet wird, ist in Anbetracht der Summe aller Anwendungen aber geringer.

betroffenen Personen, die Reversibilität des Ergebnisses sowie die Verfügbarkeit und Wirksamkeit von Rechtsmitteln (im Unionsrecht und im Recht der Mitgliedstaaten). Diese Kriterien sollten durch weitere Kriterien ergänzt werden. Diese sind: Opt-out-/Konfigurationsmöglichkeiten der Nutzenden, Entscheidungsmöglichkeiten der Nutzenden sowie die Struktur des Marktes beziehungsweise die Pluralität des Dienstangebots. Der Handlungsspielraum der Individuen sollte folglich durch sieben Aspekte detailliert charakterisiert werden (siehe Abbildung 6). Von diesen Aspekten können einige technisch in die Anwendung implementiert werden, während die Struktur des Marktes ein extern gegebener Aspekt ist und Reversibilitätsmöglichkeiten bis zu einem gewissen Grad technisch implementiert werden können (siehe Abbildung 6).

Abbildung 6: Ermittlung des Umfangs der Handlungsfreiheiten des Individuums



Die von den Autorinnen und Autoren vorgeschlagenen Kriterien lauten:

1. *Opt-out-/Konfigurationsmöglichkeiten* der Nutzenden: Entscheidend ist, ob die von den Auswirkungen des Systems Betroffenen Einfluss auf die Funktionsweise und den Einflussbereich des Systems nehmen können.[20] Dies ist abhängig von dem Grad der Autonomie des speziellen Systems sowie der Autonomie des Gesamtsystems (in welchem das KI-System nur eine Teilkomponente ist). Ein Beispiel für diesen Aspekt ist, dass Nutzende beispielsweise der Verwendung ihrer Daten für eine speziell auf sie

20 Eine Opt-in-Lösung als Voreinstellung bedeutet, dass Betroffene der Erhebung und Verarbeitung ihrer Daten ausdrücklich zustimmen müssen. Dies wird von der Datenschutzgrundverordnung vor allem für personenbezogene Daten vorgeschrieben. Da es hier über personenbezogene Daten hinaus um allgemeine Systemfunktionalitäten und Daten geht, wird hier von einer Opt-out-Möglichkeit gesprochen.

zugeschnittene Suchmaschinenoptimierung oder der Verwendung der Daten nur für bestimmte Websites widersprechen können. Entscheidend ist darüber hinaus mit Blick auf die Autonomie des Systems, ob die vorgeschlagenen Ergebnisse/nächsten Schritte angenommen werden müssen/autonom ausgeführt werden, oder ob diese noch bestätigt werden müssen oder auch widerrufen werden können. Sind diese Möglichkeiten in einem bestimmten Anwendungsgebiet nicht oder nur teilweise gegeben, könnte das unter sonst gleichen Umständen (ceteris paribus) zu einem höheren Regulierungsbedarf des Systems in diesem Anwendungsgebiet führen.[21]

2. *Entscheidungsmöglichkeiten* der Nutzenden: Es ist wichtig, dass die Nutzenden nicht nur theoretisch, sondern auch praktisch in der Lage sind, das System zu konfigurieren oder sich dem System zu entziehen. Um diese Einstellungen bewusst vornehmen zu können, müssen die Individuen selbstbestimmt in ihrem eigenen Handeln sein. Fehlende Informationen oder auch bewusste oder unbewusste Beeinflussungen können hier mögliche Barrieren bilden. Sind die Nutzenden beispielsweise aufgrund eines Informationsdefizits oder einer zu hohen Komplexität nicht in der Lage, eine informierte und bewusste Entscheidung zu treffen, sollte dies ebenfalls zu einem höheren Regulierungsbedarf des KI-Systems in diesem Anwendungsgebiet führen. Eine transparente Kennzeichnung und nachvollziehbare Information der Nutzenden über den Einsatz und die Funktionsweise einer der beiden Systemarten sollten hingegen unter sonst gleichen Umständen (ceteris paribus) zu einem niedrigeren Regulierungsbedarf des Systems in diesem Anwendungsgebiet führen.

3. *Reversibilitätsmöglichkeiten* der Nutzenden: Der Umfang der Handlungsfreiheiten des Individuums hängt entscheidend davon ab, ob – und wenn ja, inwieweit – Nutzende Entscheidungen des Systems korrigieren oder rückgängig machen können (Reversibilität). Hierzu gehört auch, ob Nutzende mögliche Schäden einfach beheben oder eine Neubewertung vornehmen können. Dies ist stark abhängig vom konkreten Einsatzkontext des jeweiligen Systems und weist zwei unterschiedliche Dimensionen auf: die systeminterne Reversibilität und die Reversibilität in der Umwelt des Systems. Wird ein KI-System beispielsweise in einem Auto eingesetzt, um Menschenleben zu schützen, und wird aufgrund einer Fehlfunktion des Systems ein Mensch verletzt oder verliert sein Leben, ist keine Reversibilitätsmöglichkeit in Bezug auf die Rettung des Menschenlebens gegeben. Ex post kann jedoch das System verbessert werden, sodass solche Unfälle nicht mehr passieren. Eine falsche Online-Bestellung, die für eine Kundin oder einen Kunden durch ein KI-System getätigt wurde, ist sowohl systemintern (Umprogrammierung) als auch extern (Rücksendung) leicht rückgängig zu machen. Sind die Reversibilitätsmöglichkeiten gering, so ist mehr Regulierung des Systems in diesem Anwendungskontext notwendig und vice versa.

21 Im Hochfrequenzhandel beispielsweise trifft und führt das ADM-System mit lernenden Komponenten Entscheidungen so schnell aus, dass Nutzende nicht eingreifen können.

4. *Struktur des Marktes bzw. Pluralität des Dienstangebots:* Die Handlungsmöglichkeiten des einzelnen Individuums werden darüber hinaus auch von der Marktstruktur beziehungsweise der Abhängigkeit von einem KI-System in einem bestimmten Anwendungsgebiet beeinflusst. Entscheidend sind dabei die Angebotsstruktur in dem bestimmten Anwendungsgebiet sowie die vorgegebenen Regularien. Hat ein Anbieter in dem untersuchten Anwendungsgebiet ein Monopol auf das System oder ist die Nutzung dessen beispielsweise aufgrund staatlicher Vorgaben verpflichtend, so sind kaum Handlungsmöglichkeiten des Individuums gegeben. Ist in dem untersuchten Anwendungsgebiet jedoch ein dem System angemessener freier Wettbewerb und auch eine Wahlmöglichkeit der Nutzenden zu beobachten, so ist unter sonst gleichen Umständen (ceteris paribus) der Regulierungsbedarf des Anwendungsgebiets geringer.

Der Umfang der Handlungsfreiheiten des Individuums steigt folglich mit Möglichkeiten des Individuums, sich dem Einfluss des Systems zu entziehen oder dieses zu konfigurieren sowie echte Entscheidungen zu treffen. Zu einem größeren Umfang an Handlungsfreiheiten führen auch Möglichkeiten der Nutzenden, die Entscheidungen des eingesetzten Systems korrigieren oder rückgängig machen zu können, sowie eine plurale Marktstruktur, wobei letzterer Punkt in einem äußeren Verhältnis zur Kritikalität des Systems vor dem Hintergrund seines Anwendungskontextes steht. Der Umfang der Handlungsmöglichkeiten des Individuums ist niedrig bei wenig Entziehungs-, Konfigurations-, Entscheidungs- und Reversibilitätsmöglichkeiten der Nutzenden sowie einer monopolartigen Struktur.

Insgesamt lässt sich Folgendes festhalten: Ein hohes Ausmaß der möglichen Verletzung von Menschenleben und weiteren hohen Rechtsgütern sollte nach Meinung der Autorinnen und Autoren in Kombination mit einem niedrigen Umfang der Handlungsfreiheiten des Individuums zu einer hohen Kritikalität führen und vice versa.

3.2 Von den Dimensionen zur Kritikalität

Zwischen den vorgestellten Dimensionen und der Höhe der Kritikalität/des Risikos besteht eine Beziehung, woraus auf mögliche Maßnahmen zur Sicherstellung der Qualität, Sicherheit und Vertrauenswürdigkeit von Systemen geschlossen werden kann. Aufgrund der in Kapitel 2.2 vorgestellten Einschränkung, dass sich bei KI-Systemen gesellschaftliche Schadenspotenziale im Vorhinein weder hinreichend quantifizierend ermitteln noch qualitativ belastbar erfassen lassen, ist eine Operationalisierung des Konzepts im Sinne von konkreten Kategorien, Stufen oder Gruppen derzeit nur sehr begrenzt möglich. Dies ist aktuell noch offen und bedarf der Forschung sowie eines gesellschaftlichen Diskurses unter Einbeziehung von Vertreterinnen und Vertretern aus Gesellschaft, Wissenschaft, Wirtschaft und Politik.

Zum gegenwärtigen Zeitpunkt schlagen die Autorinnen und Autoren vor, den Regulierungsvorschlag wie in Abbildung 7 dargestellt zu konkretisieren:

Abbildung 7: Anpassungsvorschläge zum Regulierungsvorschlag der Europäischen Kommission

EU-Regulierung	Anpassungsvorschläge
Möglichkeit und Ausmaß der Verletzung von Menschenleben und weiteren Rechtsgütern	
Schweregrad des Schadens • Ausmaß, in dem ein KI-System einen Schaden verursacht hat oder das Risiko eines Schadens besteht • Ausmaß der Auswirkungen des Schadens • Wahrscheinlichkeit, dass das KI-System eine hohe Anzahl von Personen schädigt • Wahrscheinlichkeit, dass ein KI-System mehr als einen der speziell definierten Schäden verursacht	**Problem:** schwierige Quantifizierbarkeit → Betrachtung gesamtgesellschaftlicher und individueller Schäden → Betrachtung materieller und immaterieller Schäden
Wahrscheinlichkeit eines Schadens • Anzahl der Nutzenden des KI-Systems	+ • Anzahl der Nutzungen des KI-Systems durch die Nutzenden • Persistenz der Bedrohungssituation • Kontrollierbarkeit der Bedrohungssituation (Grad der Vernetztheit)
Umfang der Handlungsfreiheiten der Individuen	
• Ausmaß der Abhängigkeit der potenziell von einem Ergebnis betroffenen Personen • Ausmaß der Vulnerabilität der potenziell betroffenen Personen gegenüber dem/der Nutzenden eines KI-Systems • Ausmaß der Reversibilität des von einem KI-System erzeugten Ergebnisses • Verfügbarkeit und Wirksamkeit von Rechtsmitteln (im Unionsrecht und im Recht der Mitgliedstaaten) • Ausmaß, in dem die bestehenden Rechtsvorschriften der Union in der Lage sind, die vom KI-System ausgehenden Risiken zu verhindern/zu minimieren	+ • Opt-out-/Konfigurationsmöglichkeiten der Nutzenden • Entscheidungsmöglichkeiten der Nutzenden • Marktstruktur/Pluralität des Dienstangebots

Darüber hinaus geben die Autorinnen und Autoren folgende Ansatzpunkte zur Einordnung von KI-Systemen vor dem Hintergrund ihres jeweiligen Anwendungskontextes:

- Die Kritikalitätseinordnung von KI-Systemen sollte sich nicht nur auf die Eigenschaften der Systeme selbst fokussieren, sondern immer vor dem Hintergrund des Anwendungskontextes vorgenommen werden. So kann die Kritikalität eines KI-Systems vor dem Hintergrund eines bestimmten Anwendungskontextes als niedrig und gleichzeitig vor dem Hintergrund eines anderen Anwendungskontextes als hoch eingeschätzt werden.
- Die Einordnung sollte zunächst immer anhand der einzelnen Teilkomponenten sektor- oder domänenspezifisch vorgenommen werden. Hierbei soll jedem Risikofaktor eine Stufe zugewiesen werden.

- Werden zwei oder mehr KI-Systeme miteinander kombiniert, so muss diese Kombination als neues System betrachtet werden. Dieses neue System muss vor dem Hintergrund seines Anwendungskontextes noch einmal gesondert bewertet werden.
- Die Einordnung der Kritikalität eines Systems vor dem Hintergrund seines Anwendungskontextes soll anhand einer Beachtung der Ausprägungen aller oben erwähnten Dimensionen geschehen.
- Eine erste Einschätzung der Kritikalität eines KI-Systems sollte idealtypisch vor dem Inverkehrbringen des Systems im konkreten Kontext geschehen. Diese Ex-ante-Maßnahme soll durch effektive Ex-post-Maßnahmen, wie Verbraucherschutz- oder Beschwerderegime, ergänzt werden. Nur diese sichern die Datensouveränität des Datensubjekts über die Einwilligung am Beginn des Datenverarbeitungsprozesses hinaus.

Wer über die Befugnis verfügen sollte, eine Einschätzung der Kritikalität vorzunehmen, sollte abhängig von der Anwendungsdomäne sein. Wenn staatliche Stellen für bestimmte Anwendungskontexte eine verpflichtende Regulierung einführen wollen, muss die Einschätzung der Kritikalität des KI-Systems vor dem Hintergrund seines Anwendungskontextes entweder von den staatlichen Stellen selbst vorgenommen werden, oder die staatlichen Stellen müssen darüber entscheiden, ob und an wen (z. B. Herstellende oder unabhängige Institutionen) diese Einschätzung delegiert wird. Sieht die öffentliche Hand keine Regulierung vor, sollte die Kritikalitätseinschätzung den Unternehmen oder den Entwickelnden zufallen (siehe Heesen, Müller-Quade, Wrobel et al. 2020, S. 20).

4. Aktuelle Fragen und Antworten zum Thema Kritikalität

Die Frage nach der Kritikalität von KI-Systemen in ihrem Anwendungskontext ist sehr facettenreich und vielschichtig. Unterschiedliche Disziplinen betrachten und beurteilen das Thema aus verschiedenen Perspektiven. Im Folgenden kommen einzelne Expertinnen und Experten der Plattform Lernende Systeme im Interviewstil zu Wort, um ausgewählte Aspekte zu vertiefen oder noch weiter zu erörtern, aber auch um unterschiedliche Sichtweisen aufzuzeigen.

Prof. Dr. Peter Dabrock

Peter Dabrock
Schwierigkeit der Operationalisierung des Kritikalitätskonzepts

Das Konzept der Kritikalitätsstufen zu operationalisieren, ist nicht trivial (siehe Kapitel 3.2). Die Problematik liegt vor allem darin, dass wichtige, dem Konzept zugrundeliegende Kategorien schwer voneinander abzugrenzen sind und teilweise sogar ineinander übergehen. Diese Problematik sowie mögliche Implikationen erläutert Peter Dabrock, Professor an der Universität Erlangen-Nürnberg.

Warum ist das Konzept der Kritikalität schwer zu operationalisieren?

„Gegenüber dem bisherigen Kritikalitätsdiskurs liest man in diesem Paper, dass Kritikalität nicht in ein statisches Modell gepackt werden darf, stattdessen domänenspezifisch und nur notwendig, aber nicht hinreichend für vertrauenswürdige KI sein soll. Wenn aber sowohl das Lernen der Systeme als auch der jeweilige Kontext die Kritikalität bestimmen, gibt es zwei Optionen:
Erstens: Die Regulierung, will sie mit der hohen Dynamik zwischen System und Kontexten Schritt halten, muss extrem engmaschig sein, was notwendigerweise zu Innovationshemmnissen führt.
Zweitens: Man setzt ex post entscheidend auf haftungs- und verbraucherschutzrechtliche Standards, die das Kriterium der Datensouveränität umsetzen. Dann stellt sich die Frage, ob über eine erste notwendige Heuristik für die technische Robustheit und die Angemessenheit des Einsatzes von Lernenden Systemen (LS) und Algorithmic Decision Making-Systemen (ADM) in einer Domäne hinaus der Ansatz der Kritikalität überhaupt dazu geeignet ist, die Vertrauenswürdigkeit von KI-Systemen zu befördern oder gar zu operationalisieren.“

Welche Probleme, gerade mit Blick auf Regulierung, entstehen, wenn wichtige Kategorien nicht hinreichend voneinander abgegrenzt werden?

„Gern wird behauptet, dass Kritikalität eines sozio-technisch verstandenen KI- oder ADM-Systems ex ante und ex post, dynamisch und kontextabhängig eruiert werden könne. Wie unter erstens erwähnt, ist aber bestenfalls eine statische, nur technische, domänenspezifische Ex-ante-Bewertung möglich. Die Gefahr, das Kritikalitätskonzept für den gesamten Prozess der rechtlichen und ethischen Bewertung über die Vertrauenswürdigkeit von LS zu erheben und sie darüber gegebenenfalls noch zertifizieren zu wollen, besteht darin, eine Sicherheit vorzugaukeln, die ohne flankierende, nicht-technische Maßnahmen nicht gewährleistet werden kann. Regulatorisch würde Kritikalität mehr versprechen, als sie halten kann. Leidtragende wären vor allem betroffene Datensubjekte, wenn sie nicht weitgehender geschützt würden.“

Wie detailliert sollten die Kategorien benannt sein, um das Konzept praktikabel zu gestalten?

„Wie detailliert die Kategorien sein sollen, um das Konzept praktikabel zu gestalten, lässt sich a priori nicht beantworten. Wird der Ansatz ausschließlich ex ante für die Bewertung technischer Robustheit verwendet, ist jede Detailgenauigkeit zu begrüßen. Für regulatorische und ethische Ex-post-Einordnung eines Systems muss entscheidend sein, dass ein betroffenes Datensubjekt schnell und effektiv – und sei es via Datentreuhänder – die Nutzung ihrer oder seiner personenbezogenen Daten überprüfen und kontrollieren kann, Standards des Haftungsrechts greifen und somit keine Verantwortungsdiffusion befördert wird. All das geht über Kritikalitätseinordnungen hinaus.“

Prof. Dr. Werner Damm

Werner Damm

Hazard-Analysen als Instrument zur Ermittlung von Risiken

In sicherheitsrelevanten Domänen werden zur Ermittlung eines Risikos im Sinne einer Gefahrenabwehr oft Hazard-Analysen angewandt (siehe Kapitel 2.2). Was darunter zu verstehen ist und wie sie funktionieren, erläutert Werner Damm, Direktor des Forschungszentrums Human Cyber Physical Systems an der Carl von Ossietzky Universität Oldenburg.

Wofür werden Hazard-Analysen aktuell genutzt?

„In sicherheitsrelevanten Anwendungsdomänen sind Methoden der Hazard-Analyse zwingender Bestandteil des den Systementwurf begleitenden Sicherheitsprozesses und vom Hersteller des Systems durchzuführen. Diese Aussage gilt unbeschadet der Art der Realisierung des Systems, also auch dann, wenn die Systemrealisierung auf Lernenden Systemen (LS) oder ADM-Systemen basiert.“

Wie wird in einer klassischen Hazard-Analyse ein potenzielles Risiko ermittelt?

„Eine Preliminary Hazard Analysis (PHA) wird während der Systemkonzeption eingesetzt, in der lediglich eine Dekomposition des Systems in Teilsysteme vorausgesetzt wird. In der sogenannten Hazard- and Operability-Analyse (HAZOP) untersucht ein Review-Team anhand von Guide-Words für jedes Teilsystem und für jeden Informationsfluss, welche Auswirkungen Ausfälle auf die Sicherheit des Gesamtsystems haben. Diese Methode ist ausgezeichnet für KI-basierte Implementierungen anwendbar. Entscheidend ist, dass das Risiko, dass eine Fehlfunktion zu einem Unfall führt, für jede Funktion bekannt ist. Dies gilt also unbeschadet der Frage, ob Lernende Systeme oder ADM-Systeme für deren Implementierung verwendet werden. Damit kann für jede angenommene Fehlfunktion und für alle möglichen daraus resultierenden Hazards ein nach Eintrittswahrscheinlichkeit gewichteter Schweregrad daraus resultierender Unfälle ermittelt werden.“

An welche Grenzen stoßen Hazard-Analysen aktuell bei der Ermittlung der Kritikalität von KI-Systemen?

„In späteren Phasen wird die PHA durch Sicherheitsanalysen ersetzt, die detailliertere Kenntnisse über die Realisierung des Systems voraussetzen, wie etwa Fault Tree Analysis (FTA). Bei qualitativen Varianten der FTA werden während der PHA erkannte Hazards des Systems „Top-down” über den Aufbau des Systems auf Ausfälle von Basiskomponenten reduziert. Somit können die Kombinationen von Ausfällen von Basiskomponenten berechnet werden, die gleichzeitig auftreten müssen, damit der analysierte Hazard des Gesamtsystems auftritt. Insbesondere können so Single Point of Failures erkannt werden, die natürlich durch konstruktive Maßnahmen eliminiert werden müssen. Während qualitative FTA somit auch für LS einsetzbar ist, gilt dies nicht für quantitative Varianten der FTA, da es keine sicheren Verfahren zur Ermittlung der Auftretenswahrscheinlichkeiten von Fehlklassifikationen oder Fehlentscheidungen bei LS gibt.“

Prof. Dr. Armin Grunwald

Armin Grunwald
Gefahrenabwehr versus Vorsorge

Die Autorinnen und Autoren dieses Whitepapers plädieren dafür, das Konzept der Kritikalität eher als Vorsorgemaßnahme im Sinne einer Technikfolgenabschätzung denn als klassische Risikoanalyse zu verstehen (siehe Kapitel 2.2). Gründe und Auswirkungen auf die Bewertung von KI-Systemen erläutert Armin Grunwald, Leiter des Instituts für Technikfolgenabschätzung und Systemanalyse (ITAS) am Karlsruher Institut für Technologie und Leiter des Büros für Technikfolgenabschätzung beim Deutschen Bundestag.

Was ist eine Technikfolgenabschätzung – in Abgrenzung zu einer Risikoanalyse?

„Technikfolgenabschätzung (TA) nutzt immer wieder die Risikoanalyse. Sie ist jedoch allgemeiner und setzt weder voraus, dass klar ist, dass etwas ein Risiko ist, noch, dass quantitative Daten für eine Risikoanalyse verfügbar sind. Insbesondere bei visionären Technologien ist in der TA oft eher die Frage zentral, warum Technik von manchen als Risiko und von anderen als Chance wahrgenommen wird. Risiken sind nicht einfach Risiken an sich, sondern immer Risiken für bestimmte Gruppen, Akteure, Regionen, Branchen oder andere Objekte. TA versucht hier, Verständigungen vor konkreten Risikoanalysen zwischen verschiedenen Perspektiven auf das jeweilige Feld zu erzielen.“

Wie unterscheidet sich die Technikfolgenabschätzung bei KI-Systemen von einer Gefahrenabwehr bei konventionellen IT-Systemen?

„Für konventionelle IT-Systeme ist belastbares Wissen über Gefahren und Risiken (mögliche Gefahren) zumindest in Teilen verfügbar, etwa über die Eintrittswahrscheinlichkeit und die Ausprägung möglicher Gefahren. KI hingegen ist häufig in visionäre Erzählungen eingebettet, in denen dieses Wissen nicht besteht. Weiterhin verändern sich KI-Systeme durch maschinelles Lernen, während konventionelle IT ihre Eigenschaften nicht selbst verändern kann. TA für KI-Systeme muss daher auch mögliche Risiken nicht-intendierten Lernens betrachten und über einprogrammierte Leitplanken des Lernens nachdenken.“

Warum können klassische Instrumente zur Risikoermittlung aktuell nicht auf KI-Systeme angewandt werden?

„Teilweise können Verfahren der Risikoermittlung durchaus auch auf KI-Systeme angewendet werden, vor allem, wenn deren Software-Eigenschaften, aber auch ihre Anwendungen bereits gut bekannt sind. Häufig ist dies bei erst noch zu entwickelnden KI-basierten Systemen jedoch kaum der Fall, bei eher visionären oder spekulativen KI-Systemen in der Regel gar nicht. Dann kann es nur darum gehen, frühzeitig mögliche Schäden in Form von plausiblen Erzählungen oder Szenarien aufzuzeigen, um Sensibilität für die weitere Entwicklung zu wecken und den Boden für spätere konkrete Risikoermittlungen zu bereiten.“

PD Dr. Jessica Heesen

Jessica Heesen

Orientierungsfunktion von ethischen Werten im technischen Handeln

Die Frage nach der Kritikalität von KI-Systemen ist eng mit der Frage verbunden, welche ethischen Ansprüche in Hinsicht auf die Verteilung von Risiken und die Zuschreibung von Verantwortung gestellt werden können. Wie die Orientierungsfunktion von ethischen Werten im technischen Handeln aussieht und welche Implikationen dies hat, erklärt Jessica Heesen, Leiterin des Schwerpunkts Medienethik und Informationstechnik im Ethikzentrum der Universität Tübingen.

In welchem Zusammenhang stehen ethische Werte und das Konzept der Kritikalität von KI-Systemen?

„Das Konzept der Kritikalität beruht auf der Einschätzung der Risiken von KI-Systemen in bestimmten Handlungskontexten. In beiden Feldern sind ethische Fragestellungen grundlegend und für die Beurteilung von Gefährdungslagen leitend. Bereits ganz grundsätzlich beruht die Bereitschaft zu einer Technikbewertung auf ethischen Anforderungen wie Verantwortungsübernahme, dem Vorsorgeprinzip und dem Nichtschädigungsgrundsatz. Bei der konkreten Beurteilung der Kritikalität geht es dann um Grundwerte wie Diskriminierungsfreiheit, Selbstbestimmung und Persönlichkeitsschutz.“

Müssen ethische Werte bei jeder Entwicklung und Anwendung von KI-Systemen zwingend beachtet werden?

„Die Beachtung ethischer Werte ist kein Zwang im Sinne einer von außen vorgegebenen Handlungsanweisung, sondern ihre leitende Funktion ergibt sich aus der Wertebasis der demokratischen Rechtsstaaten. Wenn Entwicklungsabteilungen und Unternehmen ihre Verantwortung nicht wahrnehmen, wenn der Staat beim Schutz seiner Bürgerinnen und Bürger vor Diskriminierung oder unannehmbaren Risiken versagt, dann erodiert auch das Vertrauen in Politik, Staat und den Markt. In diesem Sinne ist für ein gutes gesellschaftliches Zusammenleben die Beachtung ethischer Werte unabdingbar, also zwingend erforderlich.“

Wie kann das auf pragmatische Art und Weise erfolgen?

„Wenn es um die Bewertung der Kritikalität geht, sind normative Anforderungen an KI-Systeme explizit zu benennen und Wertkonflikte müssen identifiziert und diskutiert werden. Bei der konkreten Anwendung von KI-Systemen ist es entscheidend, allgemeine ethische Prinzipien wie zum Beispiel. den Nichtschädigungsgrundsatz für die Praxis umzusetzen. Welche Risiken sind für wen akzeptabel? Kann der Nutzen aus einer Anwendung mit einem möglichen Schaden verrechnet werden? Für solche und ähnliche Fragen leistet die Ethik als Wissenschaft vom guten und richtigen Handeln einen unerlässlichen Beitrag.“

Dr. Detlef Houdeau

Detlef Houdeau

Mögliche Eingriffe und Sofortmaßnahmen bei Fehlfunktionen

Trotz einer Kritikalitätseinschätzung und einer damit verbundenen Regulierung von KI-Systemen in ihren jeweiligen Anwendungskontexten lassen sich Fehlfunktionen nie ganz ausschließen. Wie mit Fehlfunktionen von KI-Systemen umgegangen werden sollte, erläutert Detlef Houdeau, Senior Director bei Infineon Technologies AG.

Welche Eingriffsoptionen und Sofortmaßnahmen sollten generell bei Fehlfunktionen der KI-Systeme zur Verfügung stehen?

„Opt-out muss immer gegeben sein! Je geringer der menschliche Entscheidungsraum und je höher die Gefährdungslage, und damit die Kritikalität eines KI-Systems, eingeschätzt wird, desto mehr Handlungsoptionen sollten dem Individuum zur Verfügung stehen. Beispiele hierfür sind, den Anbieter zu wechseln oder alternative Konfigurationen zu wählen bis hin zum Opt-out. Wenn eine erhöhte Kritikalität vorliegt, muss die jeweilige Maßnahme sofort wirksam sein, um mögliche Gefährdungen für Menschen und Rechtsgüter abzuwenden.“

Sollten diese Interventionsmöglichkeiten abhängig von der Kritikalität des KI-Systems in seinem Anwendungskontext sein?

„Die Interventionsmöglichkeit „KI-Not-Aus“ sollte immer bestehen, unabhängig von der Kritikalität! Die Gefährdungslage, bestehend aus Schutzbedarf, Bedrohung und Risiko, ist ein wichtiger Parameter für die Interventionsoptionen. Um den dynamischen Aspekt einer Gefährdungslage abzubilden, könnten Überwachungssysteme erforderlich werden. Besteht ein Anbietermonopol und ein hoher möglicher Gesamtschaden bei Fehlurteilen, ist von einem KI-System mit automatischer Entscheidungsfindung sogar abzuraten.“

Wer sollte in der Lage sein, diese Fehlfunktionen festzustellen?

„Sowohl der Hersteller von KI-Systemen als auch die oder der Anwendende selbst können Fehlfunktionen feststellen und sollten diese zeitnah melden bzw. anzeigen. Die Ursachen für Fehlfunktionen können in fehlenden und/oder fehlerhaften Anforderungen, in fehlerhaften Implementierungen, aber auch in falschen Trainingsdaten liegen. Die Ermittlung der Fehlfunktion ist dabei genauso wichtig wie ihre Ursachenfindung. Da die oder der Anwendende durch den täglichen Gebrauch eher auf Fehlerreaktionen der KI treffen wird, sollte die Rückmeldung an den Hersteller technisch möglich sein. Im Idealzustand würde diese Meldung automatisiert erfolgen.“

Thomas Schauf

Thomas Schauf

Anknüpfungspunkte an die Verordnung zur Bestimmung Kritischer Infrastrukturen

Das BSI-Gesetz sowie die Verordnung zur Bestimmung Kritischer Infrastrukturen (KRITIS-VO) nach dem Gesetz über das Bundesamt für Sicherheit in der Informationstechnik (BSI-Gesetz – BSIG) sehen vor, dass IT-Systeme, die für die Funktionsfähigkeit von kritischen Infrastrukturen maßgeblich sind, von den Betreibern eigens abzusichern und Vorfälle dem Bundesamt für Sicherheit in der Informationstechnik (BSI) zu melden sind. Welche Anknüpfungspunkte und Lücken die Verordnung für KI-Systeme bietet, erläutert Thomas Schauf, Senior Expert Public and Regulatory Affairs bei der Deutschen Telekom AG.

Was deckt die bestehende Verordnung zur Bestimmung Kritischer Infrastrukturen in Bezug auf KI-Systeme ab?

„Das IT-Sicherheitsgesetz und damit auch die KRITIS-VO sind Technologie-agnostisch auf die Sicherheit informationstechnischer Systeme und den Schutz kritischer Infrastrukturen ausgerichtet. Die KRITIS-VO definiert, welche (funktionalen) Teile der kritischen Infrastrukturen (also beispielsweise im Sektor Verkehr die Leitzentrale der Eisenbahn) und ab welcher Größe (meist auf Basis eines Mengenbezugs, beispielsweise 22 Millionen m³ gewonnenes Trinkwasser pro Jahr) Infrastrukturbetreiber von der Regulierung erfasst sind. Das heißt, KI-Systeme sind zwar nicht spezifisch reguliert, aber KI-Nutzungen im Bereich von KRITIS-Kernkomponenten sind erfasst, auch wenn keine spezifischen technischen Standards definiert sind. Wenn beispielsweise KI-Systeme in der Kraftwerksleittechnik oder in der Steuerungstechnik von Trinkwasserversorgungsanlagen zum Einsatz kommen, dann sind diese Funktionalitäten von der Gesetzgebung erfasst. Das heißt, dass diese Systeme besonders zu schützen sind. Aussagen über das „Wie" werden aber nicht getroffen. Dabei eröffnet der Gesetzgeber den betroffenen Branchen grundsätzlich die Möglichkeit, spezifische Sicherheitsstandards zu entwickeln und vom BSI anerkennen zu lassen. Diese branchenspezifischen Standards können dann auch Anforderungen an eingesetzte KI-Systeme enthalten."

Welche Anknüpfungspunkte bietet sie für die Ermittlung der Kritikalität von KI-Systemen?

„Das Schutzziel der Gesetzgebung erstreckt sich primär auf die Aufrechterhaltung der öffentlichen Sicherheit und Ordnung. Damit wirken die klassischen Ziele der IT-Sicherheit (Vertraulichkeit, Integrität und Verfügbarkeit) vor allem in Richtung des Gemeinwesens. Hier hat der Staat auch eine Verantwortung im Sinne seiner Fürsorgepflicht. Die zweite besondere Ebene, die aber mit der IT-Sicherheitsgesetzgebung in Summe nicht abgedeckt wird, ist die spezifische Schutzdimension Mensch und Umgebung, bei der der Staat auch eine besondere Sorgfaltspflicht innehat. Für die Kritikalität von KI-Systemen wird dadurch deutlich, dass eine Einsatzkontextbewertung von KI-Systemen dringend angeraten ist, da die Kritikalität des Systems nicht zwingend aus sich selbst entsteht."

Sind die Vorgaben der Verordnung zur Bestimmung Kritischer Infrastrukturen im Hinblick auf KI-Systeme ausreichend?

„Nein, da eben nur kritische Sektoren betrachtet werden. Diese sind Energie (Elektrizität, Gas, Mineralöl), Wasser (Öffentliche Wasserversorgung, öffentliche Abwasserbeseitigung), Ernährung (Ernährungswirtschaft, Lebensmittelhandel), Informationstechnik und Telekommunikation, Gesundheit (Medizinische Versorgung, Arzneimittel und Impfstoffe, Labore), Finanz- und Versicherungswesen (Kreditinstitute, Börsen, Versicherungen, Finanzdienstleister), Transport und Verkehr (Luftfahrt, Seeschifffahrt, Binnenschifffahrt, Schienenverkehr, Straßenverkehr, Logistik) sowie nach der Novellierung des IT-Sicherheitsgesetzes auch der Bereich der Siedlungsabfallwirtschaft. Fragen der Produkt- und Betriebssicherheit inklusive Haftungsfragen jenseits KRITIS bleiben hier im Wesentlichen unbeantwortet. Gleiches gilt für die Frage der Kritikalität aus dem System selbst.“

Prof. Dr. Stefan Wrobel

Dr. Maximilian Poretschkin

Stefan Wrobel & Maximilian Poretschkin
Zusammenhang von Kritikalität und Regulierung

Dem Konzept der Kritikalitätsstufen kommt auch im Whitepaper der Plattform Lernende Systeme zur Zertifizierung von KI-Systemen eine Schlüsselrolle zu. Über den Zusammenhang zwischen der Kritikalität eines KI-Systems in einem Anwendungskontext und dessen Regulierungsbedürftigkeit sowie über die Hürden für die Regulierung sprechen Stefan Wrobel, Leiter des Fraunhofer-Instituts für Intelligente Analyse und Informationssysteme (IAIS), und Maximilian Poretschkin, Senior Data Scientist und Projektleiter KI-Zertifizierung am Fraunhofer IAIS.

Wie hängt das Prinzip der Kritikalitätsstufen mit entsprechenden Prüfungen von KI-Systemen und gegebenenfalls der Ausstellung von Zertifikaten zusammen?

„Die Kritikalitätsstufen sind ein Indikator für die Risiken, welche mit dem Einsatz eines KI-Systems verbunden sind. Für hohe Kritikalitätsstufen ist es also wichtig, dass potenzielle KI-Risiken hinreichend gut mitigiert sind, was etwa durch eine unabhängige Prüfung und die Verleihung eines Zertifikats erfolgen kann. Gleichzeitig können Zertifikate für Systeme mit geringer Kritikalität bescheinigen, dass zugesicherte Eigenschaften dieser Systeme erfüllt werden. Dies kann für Anbieter dieser Systeme einen Wettbewerbsvorteil darstellen und Transparenz für Nutzende dieser Systeme schaffen.“

Was ist entscheidend, damit das Konzept der Kritikalitätsstufen angewandt werden kann?

„Hierzu ist es wichtig, Risiken, welche mit dem Einsatz von KI-Systemen verbunden sein können, systematisch zu bewerten. Weiterhin sollten branchenspezifische Regulierungen wie beispielsweise im Bereich Medizin und Medizinprodukte und technologische Anforderungen mitberücksichtigt werden. Hierbei ist es wichtig, dass unkritische Systeme nicht überreguliert werden und KI-Prüfverfahren marktfähig sind.“

Wie kann die Prüfung von KI-Systemen praktisch umgesetzt werden?

„Wie auch für andere Software- und IT-Systeme ist eine Prüfung von KI-Systemen ein Zusammenspiel aus der Validierung von zugesicherten Produkteigenschaften und der Auditierung von Prozessen (vgl. Fraunhofer IAIS 2021). In einem ersten Schritt der Prüfung muss zunächst der Prüfgegenstand inklusive der Spezifikation des Einsatzkontextes bestimmt werden. Als Nächstes sind die Ziele der Prüfung festzulegen – das heißt, welche Qualitätsstandards sollen geprüft werden und welche Anforderungen an das KI-System ergeben sich aus dem konkreten Einsatzkontext? Schließlich erfolgt die eigentliche Prüfung, bei der Entwicklungsdokumentationen begutachtet und wichtige Systemeigenschaften mithilfe von KI-Prüfwerkzeugen getestet werden.“

5. Fazit

Das vorliegende Papier zeigt, dass das von der Europäischen Kommission gewählte Konzept der Risiko- oder auch Kritikalitätsbewertung zur Sicherstellung der Qualität von KI-Systemen eine gute Orientierungsfunktion zur Bewertung und Regulierung hat. Gleichzeitig besteht aber noch weiterer Forschungsbedarf zur Überwindung von Schwachstellen. Diese liegen vor allem in der starken Komplexitätsreduktion sowie der mangelnden quantitativen Ermittlung und qualitativen Erfassung von Schäden, die durch KI-Systeme verursacht werden. Dennoch ist es aktuell eine vielversprechende zur Verfügung stehende Lösung – wenn eine passende Operationalisierung gefunden wird. Diese muss einerseits die Qualität von KI-Systemen sicherstellen, um das Vertrauen der Bürgerinnen und Bürger zu gewinnen, und andererseits gleichzeitig aber Überregulierung vermeiden und Innovationen ermöglichen.

Hierfür ist es wichtig, dass die Bewertung der Kritikalität eines KI-Systems immer vor dem Hintergrund des Anwendungskontextes erfolgt. Für alle Teilkomponenten müssen stets eigene Bewertungen vorgenommen werden, die dann miteinander kombiniert werden. Entscheidend für die Höhe der Kritikalität sollten zum einen die Möglichkeit und das Ausmaß der Verletzung von Menschenleben und weiteren Rechtsgütern sein sowie zum anderen der Umfang der Handlungsfreiheiten des Individuums. Hierfür sind die zusätzlich vorgestellten Kriterien ebenfalls heranzuziehen. Diese Aspekte sollten stets analysiert und evaluiert werden.

Wichtig ist, dass die Kritikalitätsbetrachtung bereits im Vorhinein durch weitere Maßnahmen ergänzt wird, die im Nachhinein greifen. Diese Maßnahmen im Nachhinein (ex post) können im Hinblick auf Transparenz und Nachvollziehbarkeit sowie Haftung und Schadensersatz nur dann erfolgversprechend sein, wenn der Grundstein hierfür im Vorhinein gelegt wurde. Ein Beispiel sind effektive, niedrigschwellige und zeitnahe Beschwerde- und Verbraucherschutzregime.

Bis die Kritikalitätsbetrachtung umfassend angewandt werden kann, müssen die Instrumente der Technikfolgenabschätzung im Hinblick auf KI-Systeme verfeinert und geschärft werden. Ziel ist die Erreichung objektiver Ergebnisse sowie die Reflexion auf Standards – und dies ohne dabei rein klassifizierend oder quantifizierend vorzugehen.

Literatur

Artificial Intelligence Ethics Impact Group (2020): From Principles to Practice. An interdisciplinary framework to operationalise AI ethics.

Bundesregierung (2020): Stellungnahme der Bundesregierung der Bundesrepublik Deutschland zum Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen. https://www.ki-strategie-deutschland.de/files/downloads/Stellungnahme_BReg_Weissbuch_KI.pdf (abgerufen am 15.10.2021).

Brüggemeier, G. (2006): Haftungsrecht. Struktur, Prinzipien, Schutzbereich. Wiesbaden.

Datenethikkommission (DEK) (2019): Gutachten der Datenethikkommission der Bundesregierung. https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf;jsessionid=1A60B9612A-0459BD9245B91BC042CB8E.2_cid364?__blob=publicationFile&v=4 (abgerufen am 15.10.2021).

Europäische Kommission (2020): Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf (abgerufen am 15.10.2021).

Europäische Kommission (2021a): 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. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence (abgerufen am 15.10.2021).

Europäische Kommission (2021b): 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. Deutsche Übersetzung. Abrufbar unter: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence (abgerufen am 15.10.2021).

Europäische Kommission (2021c): Annexes to the Proposal for a Regulation on a European approach for Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence (abgerufen am 15.10.2021).

Fraunhofer IAIS (Hrsg.) (2019): Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf (abgerufen am 15.10.2021).

Fraunhofer IAIS (Hrsg.) (2021): Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz – KI-Prüfkatalog. https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (abgerufen am 15.10.2021).

Heesen, J. et al. (Hrsg.) (2020): Zertifizierung von KI-Systemen – Impulspapier aus der Plattform Lernende Systeme, München.

Heesen, J., Müller-Quade, J., Wrobel, S. et al. (Hrsg.) (2020): Zertifizierung von KI-Systemen – Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme. Whitepaper aus der Plattform Lernende Systeme, München 2020.

Hummel, P. et al. (2021): Data sovereignty. A review, in: Big Data & Society 8 (2021), S. 1–17. https://doi.org/10.1177%2F2053951720982012

Koziol, H. (Hrsg.) (2014): Grundfragen des Schadensersatzrechts aus rechtsvergleichender Perspektive. Wien.

Plattform Lernende Systeme (Hrsg.) (2019a): Lernende Systeme im Gesundheitswesen – Bericht der Arbeitsgruppe Gesundheit, Medizintechnik, Pflege, München.

Plattform Lernende Systeme (Hrsg.) (2019b): Auf dem Weg zu einem intelligenten Mobilitätsraum – Bericht der Arbeitsgruppe Mobilität und intelligente Verkehrssysteme, München.

Plattform Lernende Systeme (Hrsg.) (2020a): Zukunftsfähigkeit mit KI sichern – Ansätze für mehr Resilienz und digitale Souveränität – Positionspapier, München.

Plattform Lernende Systeme (2020b): Glossar.
Abrufbar unter: https://www.plattform-lernende-systeme.de/glossar.html

SafeTRANS e. V. (2019): Safety, Security, and Certifiability of Future Man-Machine Systems. https://www.safetrans-de.org/de/Uploads/AK_2018_RLE_CPS/SafeTRANS_RM_SSC_FMMS_Roadmap_V2.pdf?m=1611136486 (abgerufen am 15.10.2021).

UNESCO (2020): Outcome Document. First version of a draft text of a recommendation on the ethics of artificial intelligence.

von Gleich, A. (Hrsg.) (1998): Bionik – Ökologische Technik nach dem Vorbild der Natur? Stuttgart.

Wantzen, K. (2007): Unternehmenshaftung und Enterprise Liability. Tübingen.

Zweig, K. & Krafft, T. (2019): Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse. Ein Regulierungsvorschlag aus sozioinformatischer Perspektive.

Über dieses Whitepaper

Die Autorinnen und Autoren des Whitepapers sind Mitglieder der Arbeitsgruppen IT-Sicherheit, Privacy, Recht und Ethik sowie Technologische Wegbereiter und Data Science der Plattform Lernende Systeme.

Als eine von insgesamt sieben Arbeitsgruppen thematisiert die Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik Fragen zur Sicherheit (Security), Zuverlässigkeit (Safety) und zum Umgang mit Privatheit (Privacy) bei der Entwicklung und Anwendung von Lernenden Systemen. Sie analysiert zudem damit verbundene rechtliche sowie ethische Anforderungen und steht in engem Austausch mit allen weiteren Arbeitsgruppen der Plattform Lernende Systeme. Die Arbeitsgruppe Technologische Wegbereiter und Data Science befasst sich mit den technologischen Grundlagen und Enablern von Künstlicher Intelligenz (KI). Dabei geht es beispielsweise um die Anforderungen an die Forschung, die Ausbildung von KI-Fachleuten oder den Transfer von Forschungsergebnissen in erfolgreiche Anwendungen.

Autorinnen und Autoren
PD Dr. Jessica Heesen, Universität Tübingen (Projektleitung)
Prof. Dr. Jörn Müller-Quade, Karlsruher Institut für Technologie (Projektleitung)
Prof. Dr. Stefan Wrobel, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS (Projektleitung)

Prof. Dr. Peter Dabrock, Friedrich-Alexander-Universität Erlangen-Nürnberg
Prof. Dr. Michael Decker, Karlsruher Institut für Technologie (KIT)
Prof. Dr. Werner Damm, Universität Oldenburg
Prof. Dr. Armin Grunwald, Karlsruher Institut für Technologie (KIT)
Prof. Dr. Klaus Heine, Universität Rotterdam, Jean Monnet Centre of Excellence on Digital Governance
Dr. Detlef Houdeau, Infineon Technologies AG
Prof. Dr. Tobias Matzner, Universität Paderborn
Peter Rost, secunet Security Networks AG
Thomas Schauf, Deutsche Telekom AG
Prof. Dr. Katharina Zweig, Technische Universität Kaiserslautern

Autorinnen und Autoren mit Gaststatus
Tobias Krafft, Technische Universität Kaiserslautern
Dr. Maximilian Poretschkin, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS

Redaktion
Stephanie Dachsberger, Geschäftsstelle der Plattform Lernende Systeme
Christine Wirth, Geschäftsstelle der Plattform Lernende Systeme

Über die Plattform Lernende Systeme

Lernende Systeme im Sinne der Gesellschaft zu gestalten – mit diesem Anspruch wurde die Plattform Lernende Systeme im Jahr 2017 vom Bundesministerium für Bildung und Forschung (BMBF) auf Anregung des Fachforums Autonome Systeme des Hightech-Forums und acatech – Deutsche Akademie der Technikwissenschaften initiiert. Die Plattform bündelt die vorhandene Expertise im Bereich Künstliche Intelligenz und unterstützt den weiteren Weg Deutschlands zu einem international führenden Technologieanbieter. Die rund 200 Mitglieder der Plattform sind in Arbeitsgruppen und einem Lenkungskreis organisiert. Sie zeigen den persönlichen, gesellschaftlichen und wirtschaftlichen Nutzen von Lernenden Systemen auf und benennen Herausforderungen und Gestaltungsoptionen.

Impressum

Herausgeber
Lernende Systeme –
Die Plattform für Künstliche Intelligenz
Geschäftsstelle | c/o acatech
Karolinenplatz 4 | 80333 München
www.plattform-lernende-systeme.de

Gestaltung und Produktion
PRpetuum GmbH, München

Stand
November 2021

Bildnachweis
DedMityay/shutterstock/Titel
Dt. Ethikrat, R. Zensen/S. 28
OFFIS _ Institut für Informatik/S. 29
Anne Behrendt/S. 31
Thilo Schoch/S. 32
Infineon Technologies AG 2018/S. 33
Thomas Schauf, privat/S. 34
Fraunhofer IAIS/S. 35

Bei Fragen oder Anmerkungen zu dieser Publikation kontaktieren Sie bitte Johannes Winter (Leiter der Geschäftsstelle):
kontakt@plattform-lernende-systeme.de

Folgen Sie uns auf Twitter: @LernendeSysteme

Empfohlene Zitierweise
Jessica Heesen, Jörn Müller-Quade, Stefan Wrobel et al.: Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten – Ein notwendiger, aber nicht hinreichender Baustein für Vertrauenswürdigkeit. Whitepaper aus der Plattform Lernende Systeme, München. DOI: https://doi.org/10.48669/pls_2021-3

Dieses Werk ist urheberrechtlich geschützt.
Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, der Entnahme von Abbildungen, der Wiedergabe auf fotomechanischem oder ähnlichem Wege und der Speicherung in Datenverarbeitungsanlagen, bleiben – auch bei nur auszugsweiser Verwendung – vorbehalten.