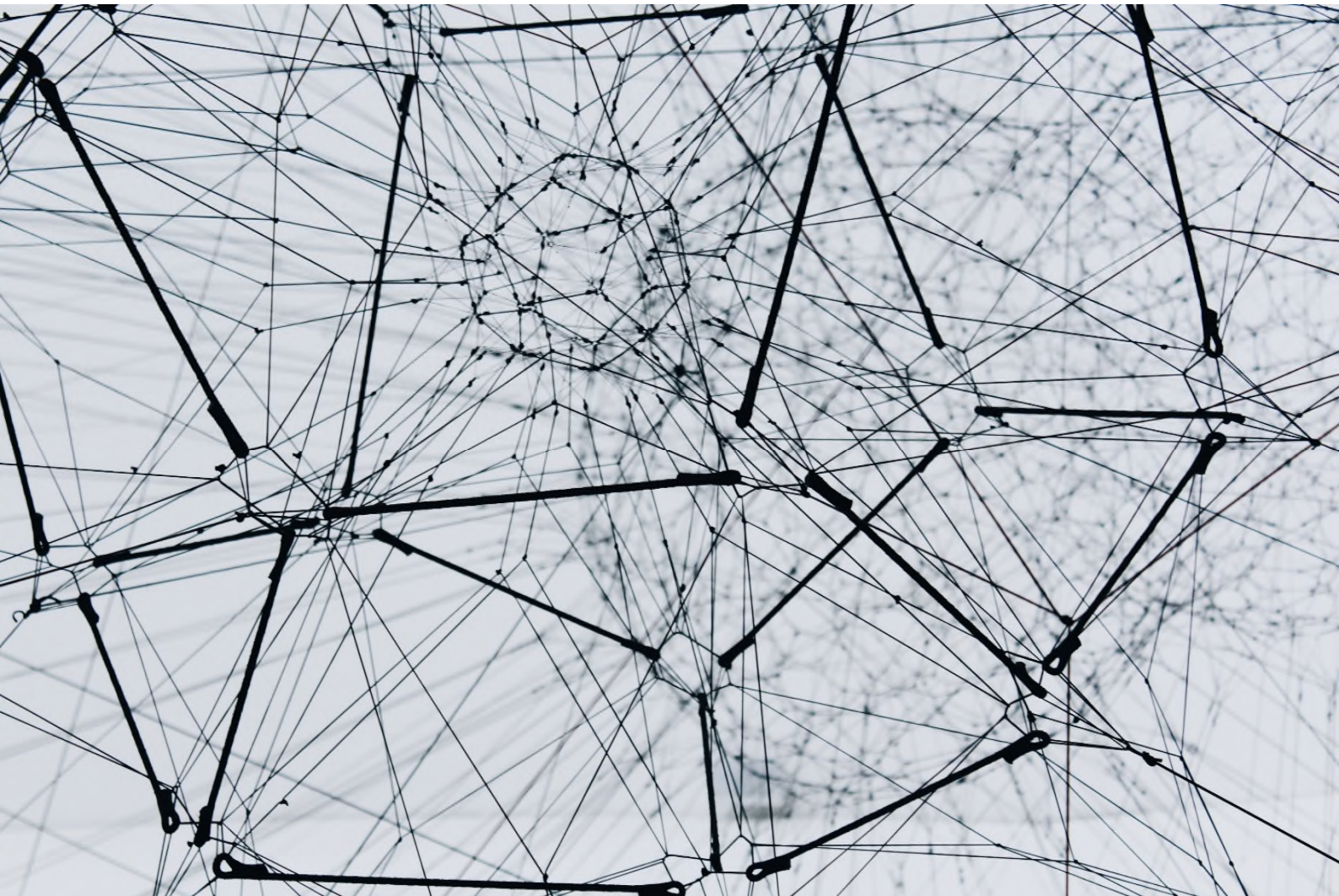
TÜV AI.LAB

Berlin
31. August 2021

Whitepaper des TÜV AI LAB

# Vorschlag für eine Risikoklassifizierung von KI-Systemen



## Zusammenfassung

**Mit diesem Whitepaper leistet das TÜV AI Lab einen Beitrag zur Risikoeinstufung von KI-Systemen. Wie sind Produkte und Anwendungen mit Künstlicher Intelligenz zu behandeln, die aus rechtlichen, ethischen oder gesellschaftlich erwünschten Gründen als riskant einzustufen sind? Dieser Frage hat sich die EU-Kommission in ihrem ersten Regulierungsentwurf angenommen. Vor allem geht es darum, wie KI-Systeme einzustufen sind, bei denen die KI-Komponente aufgrund ihrer Struktur zum sicherheitskritischen Bestandteil werden kann oder das Gesamtsystem aus rechtlichen, ethischen oder gesellschaftlich erwünschten Gründen als riskant einzustufen ist. Zusätzlich zu bereits automatisch als Hochrisiko-System eingestuften KI-Systemen sollten auch diese KI-Systeme Transparenzverpflichtungen unterliegen oder unabhängig geprüft werden.**

Eine risikobasierte Einstufung von Künstlicher Intelligenz (KI) ist bereits von verschiedenen gesellschaftspolitischen Akteuren vorgeschlagen worden.[1] Die Europäische Kommission hat zuletzt in ihrem Entwurf für einen KI-Rechtsrahmen vom 21. April 2021 solch einen risikobasierten Ansatz gewählt und kann damit als weltweit ersten KI-Regulierungsvorschlag Vorbildcharakter haben.[2] Dieser Regulierungsansatz wird deshalb nachfolgend auch beispielhaft herangezogen, um den hier beschriebenen Diskussionsbeitrag möglichst konkret werden zu lassen.

Der Regulierungsentwurf der EU-Kommission für KI-Systeme definiert vier Risikoklassen.[3] Der zur Zuordnung von KI-Systemen auf Risikoklassen verwendete Risikobegriff unterscheidet sich von den wohldefinierten und quantifizierbaren Risiken für Leib und Leben. So umfasst er sowohl weitere Risiken, wie solche für Vermögenswerte oder die Umwelt, als auch rechtliche, ethische und gesellschaftliche Risiken. Damit ergibt sich die Frage der Aufrechenbarkeit verschiedener Risiken, um sie in einer einzigen Risikoklasse zusammenfassen zu können. Weiterhin ist zu definieren, wie rechtliche, ethische und gesellschaftliche Risiken so zu fassen sind, dass sie sich einer Aufrechnung erschließen und wie mit möglichen systeminhärenten Zielkonflikten umgegangen wird.

Das TÜV AI Lab schlägt vor, zur Bestimmung von Risiken so weit wie möglich bestehende Regeln und Normen zu nutzen. Angelehnt an die Definition von Schutzzielen in der Informationssicherheit, schlagen wir für KI-Systeme Schutzziele vor, deren Erhalt bzw. Unversehrtheit zu gewährleisten sind.

[1] u.a. Tobias D. Krafft und Katharina A. Zweig, „Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse: Ein Regulierungsvorschlag aus sozioinformatischer Perspektive" (Berlin, 2019), https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf.

[2] „Für vertrauenswürdige Künstliche Intelligenz: EU-Kommission legt weltweit ersten Rechtsrahmen vor" (Brüssel, 2021), https://ec.europa.eu/germany/news/20210421-kuenstliche-intelligenz-eu_de.

[3] Unannehmbares, hohes, geringes und minimales Risiko. Zusammengefasst nachzulesen unter: https://ec.europa.eu/germany/news/20210421-kuenstliche-intelligenz-eu_de.

Diese sind:

- Leib und Leben;
- Sach- und Finanzeigentum;
- Umwelt;
- Grundrechte und
- ethische Prinzipien, hierbei auch - gesellschaftlich und sozial erwünschte Ergebnisse.

Für die ersten drei Schutzziele liegt bereits weitreichendes technisches Regelwerk vor, welches auf die spezifischen Herausforderungen der Risikobeurteilung von KI-Systemen angepasst werden muss. Für die anderen Schutzziele ist eine robuste und praktikable Herangehensweise noch zu erarbeiten.

Dort, wo sich aufgrund des Wirkungsprinzips der KI für deren Anwendung erhöhte Risiken ergeben, aber noch keine Regeln oder Normen existieren, sollen Analogieschlüsse aus bestehenden Regeln und Normen vorgenommen und kodifiziert werden.

Menschliche Einfluss- oder Vermeidungsmöglichkeiten bei Entscheidungen eines KI-Systems verstehen wir, anders als beispielsweise von DIN/DKE vorgeschlagen, nicht ausschließlich als „Einschränkung der Handlungsfreiheit des Individuums".[4] Wir schlagen vor, diese in zwei unterschiedliche Aspekte aufzuteilen:

**„Controllability"**

Der menschliche Eingriff auf ein KI-System, um den Eintritt eines Risikos zu beeinflussen, was auch konform zur allgemein akzeptierten Vorgehensweise bei der Bestimmung von „Safety Integration Levels" (SIL) oder „Performance Levels" (PL) ist.

**„Human Control"**

Dies erfasst die Möglichkeiten des von der Entscheidung eines KI-Systems betroffenen Menschen. Betrachtet wird, ob sich dieser der Entscheidung entziehen oder in Frage stellen kann bzw. die Möglichkeit besteht, Folgen rückabwickeln zu können.

Zur Risikoabschätzung bei nicht-materiellen Risiken regen wir die Verwendung des bei materiellen Risiken bewährten und praktizierten, semiquantitativen Ansatzes[5] an. Bei der Festlegung und Normierung der Risikodimensionen besteht jedoch umfassender Klärungsbedarf. Das TÜV AI Lab und die TÜV-Unternehmen unterstützen hier gerne, verorten sich aber eher in der Rolle des Verfahrens- und Prozessbegleiters als des Entwicklers von Verfahren zur Einschätzung ethischer Anforderungen und Normen. Der TÜV-Verband wiederum unterstützt den Einsatz von KI-basierten Systemen zum Nutzen der Gesellschaft und bringt sich mit seinen Expert:innen in eine konstruktive und faktenbasierte Diskussion von Forschung, Gesellschaft und Politik ein.

[4] DIN e. V. und DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE, „Ethik und Künstliche Intelligenz: Was können technische Normen und Standards leisten?", 4. August 2020, 75, https://www.din.de/resource/blob/754724/00dcbccc21399e13872b2b6120369e74/whitepaper-ki-ethikaspekte-data.pdf.

[5] Vgl. „Safety Integrity Levels", welche ebenfalls qualitative Sub-Kriterien verwenden und diese aggregieren. Eine Übersicht ist nachzulesen unter https://en.wikipedia.org/wiki/Safety_integrity_level.

# 1. Einleitung

Systeme künstlicher Intelligenz (KI) sind bereits heute Bestandteil unseres täglichen Lebens, deren Bedeutung in Zukunft aller Wahrscheinlichkeit nach weiter zunehmen wird. *Machine*- und *Deep Learning*-Algorithmen sind dabei Kerntechnologien, welche die Grundlage für immer mehr Anwendungen darstellen. Anders als bei klassischer Software ist dabei nicht immer einfach nachvollziehbar, wie die Algorithmen zu den ausgegebenen Ergebnissen kommen. Damit die Nutzung von KI-Systemen allgemein akzeptiert wird, ist Vertrauen sowohl in die von der KI produzierten Ergebnisse als auch die Art und Weise, wie diese Ergebnisse erzielt wurden, notwendig. Vertrauen ist damit eine Funktion der Erfüllung von Kriterien funktionaler Sicherheit, aber beispielsweise auch der informationellen Selbstbestimmung, der Einhaltung ethischer Standards oder um gesellschaftlich erwünschte Ziele zu erreichen.

Eine Risikobeurteilung von KI wird nicht nur politisch gefordert, sondern ist im Interesse von Verbraucher:innen sowie Herstellern. So ergibt eine Unternehmensbefragung des TÜV-Verbands vom Oktober 2020, dass 90 Prozent der befragten Unternehmen gesetzliche Regelungen fordern, um Haftungsfragen zu klären.[6] 87 Prozent wünschen sich, dass KI-Anwendungen in Abhängigkeit von ihrem Risiko reguliert werden sollten und 84 Prozent, dass Produkte und Anwendungen mit KI für die Nutzer:innen klar gekennzeichnet werden. Gleichzeitig fordern sowohl etablierte Unternehmen als auch Startups, dass eine KI-Regulierung nicht innovationsverhindernd sein darf und keine parallellaufenden Regelwerke geschaffen werden sollen. Stattdessen sollen gesetzliche Regelungen im Idealfall Innovationen ermöglichen, da das damit geschaffene Vertrauen in KI-Anwendungen erst zur Nutzung führt. Letztendlich gibt es Stimmen, die argumentieren, dass eine proaktive Regulierung Entwicklungen zu vermeiden hilft, die nicht mit dem europäischen Wertekanon vereinbar sind, wodurch wettbewerbliche Differenzierung für KI „Made in Europe" ermöglicht wird.

Der Vorschlag der EU-Kommission zur Schaffung des weltweit ersten rechtlichen Rahmens für sichere und vertrauenswürdige KI geht in diese Richtung. Aus Sicht des TÜV-Verbands ist der vorliegende Regelungsentwurf jedoch nicht hinreichend ambitioniert und bleibt hinter dem eigenen Anspruch der EU-Kommission zurück, ein „Ökosystem für Vertrauen" zu schaffen. Vorschläge für Nachbesserungen hat der TÜV-Verband daher im August 2021 formuliert.[7] Davon unabhängig bedarf die Risikoklassifizierung einer detaillierteren Operationalisierung. Derzeit werden die Risikoklassen primär anhand von Beispielen (z.B. Social Scoring, Gesichtserkennung) definiert.[8] Notwendig ist jedoch eine allgemein anwendbare und

[6] „Künstliche Intelligenz in Unternehmen: TÜV Studienbericht 2020" (Berlin, 2020), https://www.tuev-verband.de/pressemitteilungen/ki-studie.

[7] Die Stellungnahme des TÜV-Verbands zum Vorschlag der EU-Kommission für ein Gesetz über Künstliche Intelligenz ist abrufbar unter https://www.tuev-verband.de/stellungnahmen/stellungnahme-zum-gesetzesvorschlag-der-eu-kommission-fuer-kuenstliche-intelligenz.

[8] Europäische Kommission, „Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union: COM

vorab bekannte Systematik zur Definition von Risikoklassen sowie zur Zuordnung von KI-Anwendungen in diese Risikoklassen.

Bei der Entwicklung einer solchen Systematik ist zu berücksichtigen, dass der Begriff „Risikoklasse" oft unterschiedlich verstanden wird. Eine Risikoklassifizierung von technischen Systemen erfolgt über Standards, wie sie beispielsweise in der funktionalen Sicherheit verwendet werden, sowie auf Basis einer systematischen, semiquantitativen[9] Risikoanalyse und Risikobeurteilung. Abhängig von den Ergebnissen dieser Analyse wird das betrachtete technische System in je nach Industriedomäne definierte, vier bis fünf „Sicherheitsanforderungsstufen" [10] eingeteilt, mit denen die Sicherheitsanforderungen an das System beschrieben werden.

Der Gesetzesentwurf der EU-Kommission sieht vier Risikoklassen vor, in die KI-Systeme je nach ihrem Einsatzzweck eingeteilt werden. Jedoch umfassen die zugrunde gelegten qualitativen Kriterien sowohl materielle Risiken – vordringlich für Leben und Gesundheit – als auch ethisch-gesellschaftlich motivierte Risiken. Es werden aber keine Mechanismen vorgeschlagen, wie unterschiedliche Risiken kategorisiert, gemessen und vergleichbar gemacht werden können.

Dieses Whitepaper entwickelt erste Vorschläge zur Operationalisierung dieser notwendigen Kategorisierung, Messung und Herstellung von Vergleichbarkeit, auch um eine faktenbasierte Diskussion von Zielkonflikten zwischen verschiedenen Risiken zuzulassen. Anspruch ist dabei nicht, eine Gesamtlösung zu entwickeln, sondern Ansätze für ein konsistentes Gesamtsystem und Vorschläge für eine mögliche Vertiefung aufzuzeigen.

(2021) 206" (2021), Art. 6, https://eur-lex.europa.eu/legal-content/DE/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

[9] Unter semiquantitativ ist zu verstehen, dass die Kriterien nicht ausschließlich objektiv messbar sind, sondern auch weitere Kriterien herangezogen werden

[10] SIL = Safety Integrity Level nach IEC 61508 bei Beurteilung elektrischer/elektronischer/programmierbarer elektronischer (E/E/PE)-Systeme, ASIL nach ISO 26262 im Automobilbereich, PL = performance level nach ISO 13849 in der Maschinensicherheit, usw.

## 2. Grundsätzliches

Der Verordnungsentwurf der EU-Kommission gibt vom Standpunkt der Risikobewertung her verschiedene Kategorien von KI-Anwendungen vor:

**Kategorie 1: KI-Systeme, welche als Sicherheitskomponente von Produkten vorgesehen sind, die vor Inverkehrbringung einer Konformitätsbewertung durch einen unabhängigen Dritten unterliegen.[11]**

a. Im jeweiligen Gebrauchskontext bewertet wird also die KI, welche sich als Systembestandteil auf die Sicherheit des Gesamtsystems auswirkt.[12] Damit besteht eine Widerspruchsfreiheit mit bereits geltenden Gebrauchskontext-spezifischen Regulierungen. Auch lassen sich Risikoklassen eines KI-Systems[13] und des darin verwendeten Algorithmus über Sicherheitsanforderungen aus bereits bestehenden Standards und Normen ableiten. Das heißt im Umkehrschluss, dass der gleiche Algorithmus, auch wenn er in der gleichen Hardware- und Software-Umgebung läuft, in einem anderen Anwendungsfall in eine andere Risikoklasse eingruppiert werden kann und eine „Typzulassung" eines KI-Systems nur mit eindeutig definierter Verbindung zu seiner Verwendung möglich sein wird. Um keinen zu hohen Aufwand über Einzelzulassungen von Algorithmen betreiben zu müssen, erscheint die Entwicklung von klar abgegrenzten, gebrauchskontext-spezifischen Gruppen angeraten.

b. In manchen der heute verwendeten Systematiken zur Risikobeurteilung werden Fähigkeiten eines Menschen als Kriterium benutzt. So stützt sich der Risikofaktor „Controllability" bei der im Automobilbereich angewendeten ISO 26262 auf Fähigkeiten von menschlichen Fahrer:innen. Damit müsste einem KI-System, welches Teile oder Aspekte des menschlichen „Systemelements" im Automobilbereich ersetzen sollen, grundlegende Fähigkeiten eines Menschen als systemimmanente Eigenschaften („innate property") unterstellt werden bzw. diese im Gebrauchskontext beschrieben werden. Hier ist noch Entwicklungsarbeit zu leisten, um die aus der Norm ergebenden Anforderungen an ein KI-System entsprechend zu übersetzen, eindeutig zu definieren und im Kontext der spezifischen Anwendung mit Akzeptanzkriterien zu versehen.

[11] „KI-Systeme, die als Sicherheitskomponenten von Produkten, die einer Vorab-Konformitätsbewertung durch Dritte unterliegen, verwendet werden sollen", Europäische Kommission, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union: COM (2021) 206, 15.

[12] Beispielsweise bei der Verkehrsschilderkennung im Auto nach ISO 26262 oder analog bei der Maschinensicherheit nach ISO 12100 und 13849.

[13] Ein KI-System kann, muss aber nicht den Algorithmus, die Sensorik, eine Schnittstelle zur KI-Anwendung oder eine Gesamtanwendung wie z.B. ein Fahrzeug umfassen.

**Kategorie 2: KI als „Stand-Alone" Lösungen, welche hauptsächlich auf fundamentale Rechte wirken.[14]**

c. Hier wirken die Ergebnisse einer KI entweder direkt auf die Betroffenen, zum Beispiel bei der Gesichtserkennung, beim Kreditrisiko-Scoring sowie bei der Bewertung von Lebensläufen, oder es wird unterstellt, dass keine signifikanten Sicherheitsrisiken vorliegen. Damit sind auch keine Konformitätsbewertungen durch unabhängige Dritte vorgesehen.

Damit ergibt sich unserer Einschätzung nach eine Definitionslücke für KI-Systeme, welche ein sicherheits- oder funktionsbestimmendes Element eines Systems sind, für die z.Zt. keine Konformitätsbewertungen durch unabhängige Dritte vorgesehen sind. In Anlehnung und Erweiterung der Argumentation in 2.b. können KI-Elemente eine im Vergleich zum System mit menschlichem Operator zusätzliche Risikokomponente darstellen, die mit herkömmlichen Ansätzen zur Risikobeurteilung schwer oder nicht zu identifizieren und zu quantifizieren sind. So zum Beispiel durch Data Drift[15] der zum Training eines Algorithmus verwendeten Datenbasis oder durch Edge Cases[16] ausgelöste Fehlinterpretationen. Menschen können solche Ausnahmen aufgrund ihrer Kontextintelligenz kompensieren, KI-Systeme jedoch nicht. Zudem können KI-Systeme grundlegend anderen Wirkmechanismen unterliegen – wie zum Beispiel bei der Verhinderung von Bränden über frühzeitige optische Branderkennung und zielgerichtete Brandbekämpfung mit geringen Wassermengen als Alternative zu Sprinkleranlagen. Damit ist es möglich, dass diese KI-Anwendungen noch nicht über bereits bekannte Ansätze zur Gefährdungsbeurteilung, wie beispielsweise in der Norm IEC 61508 und ISO 31000 dargelegt, erfasst werden, aber aufgrund ihres Risikoprofils in Risikokategorie 1 einzuordnen sind. Hier sind die entsprechenden Grundlagen zur Einordnung noch zu erarbeiten.

[14] „[S]onstige eigenständige KI-Systeme, die ausdrücklich in Anhang III genannt werden und sich vor allem auf die Grundrechte auswirken", Europäische Kommission, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union: COM (2021) 206, 15.

[15] Definition „Data Drift": Das Modell schneidet bei unbekannten Datenregionen schlechter ab. Eine Ursache kann die Veränderung der Datenverteilungen sein, siehe Elena Samuylova, „Machine Learning in Production: Why You Should Care About Data and Concept Drift", 7. Dezember 2020, https://towardsdatascience.com/machine-learning-in-production-why-you-should-care-about-data-and-concept-drift-d96d0bc907fb.

[16] Definition „Edge Case": Ein Problem oder eine Situation, in der ein extremer, aber immer noch möglicher Parameter („operating parameter") auftritt. Solche nicht-triviale. Grenzfälle können in der Entwurfsphase möglicherweise übersehen werden, siehe Michael Sayre, „The significance of 'edge cases' and the cost of imperfection as it pertains to AI adoption", 25. April 2019, https://medium.com/@livewithai/the-significance-of-edge-cases-and-the-cost-of-imperfection-as-it-pertains-to-ai-adoption-dc1cebeef72c.

## 3. Aufgabenstellung und mögliches Vorgehen

Offen ist, nach welchen Kriterien überprüft werden kann, ob KI-Systeme in der o.a. Kategorie 2 einzustufen sind, oder sie einem nicht kategorisierten KI-System mit hohem Risiko und ggf. damit verbundenen Prüfpflichten jenseits der geforderten Transparenz einzuordnen sind. Es lässt sich argumentieren, dass es durchaus Anforderungen beispielsweise an die Robustheit eines KI-Systems gibt, die vor dem Hintergrund des Risikos des KI-Systems im jeweiligen Kontext prüfpflichtig sind. Es ist ansonsten inkonsistent bei Systemen der Kategorie 1, bei denen bereits über von KI unabhängigen Normen eine Prüfung vor der Inverkehrbringung durch einen unabhängigen Dritten vorgesehen ist, dies bei Systemen mit gleichem Risiko mit KI-Normen zu unterlassen. Dabei kann man nicht davon ausgehen, dass man sich auf bereits bestehende Normen und Standards stützen und direkt daraus Kriterien für die Risikobewertung eines KI-Systems ableiten kann. Praktikable Ansätze zur Risikobewertung müssen daher genauso neu entwickelt werden wie Ansätze, um Prüfpflichten jenseits bestehender Konformitätsbewertungsverfahren abzuleiten.

Mögliche Vorgehen hierzu sind:

- Katalog bzw. „living document", so wie von der EU-Kommission vorgeschlagen;
- „Analogieschlüsse" aus vergleichbaren, bereits über Normen und Standards, ggf. auch über Regulierung, abgebildeten KI- und Nicht-KI Anwendungen;
- Frameworks, welche einfach bestimmbare, nachvollziehbare Kriterien zur Zuweisung von KI-Systemen in Risikoklassen bereitstellen.[17] Solche Frameworks dürfen dabei nicht im Widerspruch zu den beiden vorab genannten Vorgehen stehen, sondern müssen diese unterstützen, präzisieren und die Zuordnung erleichtern.

[17] Analogie zum Blitzschutz: Auch hier werden aus technischen Kriterien einfach identifizierbare „Gebäudeklassen" abgeleitet, welche mit jeweils spezifischen Prüfpflichten verbunden werden.

## 4. Schutzziele

Angelehnt an die Definition von Schutzzielen in der Informationssicherheit, schlagen wir für KI-Systeme fünf Ziele vor, deren Erhalt bzw. Unversehrtheit es zu sichern gilt und die sich in Unterziele gliedern können:

i. Leib und Leben;

ii. Sach- und Finanzeigentum;

iii. Umwelt;

iv. Grundrechte, sofern nicht in i.-iii. geregelt. Diese umfassen z.B. Freiheitsrechte, das Recht auf informationelle Selbstbestimmung, Diskriminierungsfreiheit;

v. ethische Prinzipien, hier auch - gesellschaftlich und sozial erwünschte Ergebnisse.[18] [19]

Zu prüfen ist nun, inwiefern bereits bekannte und geübte semiquantitative Ansätze zur Risikobeurteilung zur Bestimmung der domänen- und anwendungsbezogenen Schutzziele herangezogen werden können:

d. Schutzziele i.- iii. können mit bestehenden, semiquantitativen Ansätzen zur Gefährdungsbeurteilung (SIL und ähnlichen Systematiken) beschrieben und mit den darunterliegenden Normen (z.B. IEC 61508) gut gegriffen und präzisiert werden.

Diese stellen sicher, dass das Risikopotenzial eines Systems im jeweiligen Gebrauchskontext präzise beschrieben wird und so die Grundlage dafür geschaffen wird, dass eine sicherheitstechnische Einrichtung (Kombination aus Teilsystemen, die eine oder mehrere Sicherheitsfunktionen ausführen) mit der notwendigen Zuverlässigkeit richtig arbeitet.

Bei Schutzziel iv. und den damit verbundenen Unterzielen gehen wir davon aus, dass ein semiquantitativer Ansatz zur Risikobeurteilung ebenfalls praktikabel ist. Es muss aber im Detail geprüft werden, inwiefern dies konform zu bestehenden gesetzlichen Regelungen ist, die normative Anforderungen ohne Abstufungen stellen. Beispielsweise sieht die Datenschutzgrundverordnung (DSGVO) eine binäre Betrachtung vor, wonach die gesetzlichen Anforderungen entweder erfüllt oder nicht erfüllt werden. Gleichwohl enthält die DSGVO – neben den Vorgaben zum Datenschutz -Vorgaben für Systeme, die auf Basis personenbezogener Daten automatisierte Entscheidungen treffen.[20] Die

[18] Beispiel in einem anderen Kontext: Unisex-Tarife bei der Krankenversicherung.

[19] Eine Herausforderung ist, dass die Schutzzieldefinitionen nicht disjunkt sind und ein Kriterium in verschiedenen Schutzzielen vorkommen kann.

[20] Krafft und Zweig, „Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse: Ein Regulierungsvorschlag aus

DSGVO kann daher als Grundlage zur Ableitung von Kriterien dienen, muss jedoch spezifisch konkretisiert werden, um der Fragestellung des Datenschutzes bei KI-Systemen gerecht zu werden. So kann man beispielsweise aus dem Output eines KI-Systems in Grenzen auf deren Input schließen. Damit stellt sich die Frage, inwiefern man das Gebot zur Datensparsamkeit in „Privacy by Design"-Kriterien übersetzen und deren Erfüllung prüfen kann.

Bei Schutzziel v. besteht noch Klärungsbedarf. Es gibt zwar Ansätze für Standards zur Abbildung dieses Schutzziels über Anforderungen an das Design von KI-Systemen, so wie zum Beispiel in IEEE p7000[21] beschrieben. Diese adressiert jedoch die Abbildung und Nachverfolgbarkeit von ethischen Werten anhand eines Betriebskonzepts und die Berücksichtigung von Wertversprechen und -disposition beim System-Design. Nicht dort adressiert wird, welche ethischen Werte abgebildet und verfolgt werden sowie welche Akzeptanzkriterien dafür anzulegen sind. Es muss überlegt werden, ob ein semiquantitativer Ansatz zur Risikobeurteilung, wenn auch mit entsprechenden angepassten Kriterien, angewendet werden kann, bzw. alternative Systeme nicht besser geeignet sind. Hier besteht Arbeitsbedarf, vor allem bei der Umsetzung in praktisch anwendbaren Ansätzen zur Quantifizierung der Erfüllung von ethischen Kriterien durch KI-Systeme.

## 5. Risikoklassen

Entsprechend des jeweiligen Erfüllungsgrades der fünf oben definierten Schutzziele können die Risikoklassen entwickelt werden. Eine trennscharfe und einfach umzusetzende Systematik ist dabei unumgänglich. Das von der EU-Kommission vorgeschlagene Risikoklassenmodell sieht bei höheren Risikoklassen verpflichtende Anforderungen vor, welche geprüft werden sollen und somit mit entsprechendem Aufwand verbunden ist. Freiwillige Zertifizierungen im Sinne eines Qualitäts-Labels bleiben unabhängig von Risikoklassen möglich. Risikoklassen entsprechend des Regulierungsentwurfs der EU-Kommission und damit verbundene Prüfanforderungen sind:

i. Minimales Risiko – Einhaltung des allgemein gültigen Rechts, d.h. ohne Beachtung zusätzlicher, die KI als solche betreffenden, rechtlichen Verpflichtungen. Es besteht keine Pflicht zur Prüfung durch unabhängige Dritte, wobei eine freiwillige Qualitätsaussage beispielsweise über ein noch zu definierendes „KI-Qualitäts-Gütesiegel"[22] unbenommen bleibt und es dem Lizenzgeber solcher Gütesiegel überlassen ist, die Feststellung der Qualitätsaussage an gewisse Verfahren und Regeln zu binden.

sozioinformatischer Perspektive", 14.

[21] „IEEE 7000-2021: IEEE Approved Draft Model Process for Addressing Ethical Concerns During System Design" (Piscataway, NJ, 2021), https://standards.ieee.org/standard/7000-2021.html.

[22] Zur Zeit in Entwicklung durch das „Digital Trust Forum", 2021, https://www.digitaltrustforum.org.

ii. Geringes Risiko – den KI-Systemen in dieser Klasse werden besondere Transparenzverpflichtungen auferlegt. Deren Einhaltung soll vor der Inverkehrbringung vom Entwickler bzw. Betreiber durch Selbsterklärung rechtsverbindlich nachgewiesen werden. Dies kann Qualitätsmanagement[23], damit auch Befähigungs- und Prozessqualitätsnachweise erfordern. Weiterhin besteht eine Offenlegungspflicht gegenüber Aufsichtsorganen. So können von der Marktaufsicht stichprobenhafte Untersucherungen durch beauftragte Institute angeordnet und so die Einhaltung der Transparenzverpflichtungen nachgehalten werden.

iii. Hohes Risiko – aufgrund des hohen Schadenspotentials sowie des Grades der Einschränkung der Handlungsfähigkeit des Individuums, beispielsweise durch Abhängigkeit bzw. Unumkehrbarkeit von Entscheidungen, ist für KI-Systeme dieser Risikoklasse vor der Inverkehrbringung ausnahmslos eine Konformitätsbestätigung, entweder durch Herstellerselbsterklärung oder durch einen unabhängigen, ggf. beliehenen Dritten zu erbringen. Danach sind unter anderem Nachvollziehbarkeit und Überprüfbarkeit zu belegen. Diese muss über den gesamten Lebenszyklus der KI aufrecht erhalten werden.[24] Damit ist im jeweiligen Anwendungsfall zu prüfen, ob sich die Natur der KI oder deren Leistungscharakteristika während des Betriebs maßgeblich verändern oder ob die KI an durch den Stand der Technik bestimmte steigende Qualitäts- oder Konformitätsanforderungen angepasst werden muss. In diesen Fällen ist, wie heute schon bei Automobilen oder Aufzügen praktiziert, eine wiederholende Konformitätsbestätigung notwendig. Deren Frequenz[25] sowie die Art und Weise[26] sind domänen- und anwendungsspezifisch festzulegen, die Regularien dazu zu entwickeln. Da sich Softwarekomponenten eines Systems oft schneller entwickeln als die zugrundeliegende Hardware, ist ein kürzerer Prüfzyklus für KI-Systeme, als wir es heute beispielsweise bei Fahrzeugen oder Industrieanlagen beobachten, angebracht.

iv. Unannehmbares Risiko: Eine sehr geringe Zahl besonders schädlicher KI-Anwendungen, die gegen die Werte der EU bzw. deren Grundwerte verstoßen, wird verboten. Das betrifft z. B. die Bewertung des (sozialen) Verhaltens durch Behörden (Social Scoring), die Ausnutzung der Schutzbedürftigkeit von Kindern, den Einsatz von Techniken zur unterschwelligen Beeinflussung – mit eng gefassten Ausnahmen[27]. Hier ist keine Konformitäts- oder sonstige Prüfung vonnöten,

[23] Z.B. ISO 9000-Reihe, einschlägige Zertifikate wie C4, C5.

[24] Vgl. Europäische Kommission, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union: COM (2021) 206, Teil 15 (1).

[25] Periodische Häufigkeit bzw. bei Eintritt bestimmter Ereignisse.

[26] Beispielsweise online vs. offline; Modell/Simulation vs. im Betrieb; „Back-to-Back-Test", etc.

[27] Europäische Kommission, „Neue Vorschriften für künstliche Intelligenz: Fragen und Antworten" (Brüssel, 2021), https://ec.europa.eu/commission/presscorner/detail/de/QANDA_21_1683#1.

da Anwendungen in dieser Risikoklasse nicht in Verkehr gebracht oder benutzt werden dürfen. Nichtdestotrotz ist klar zu definieren, welche KI-Anwendungen in diese Klasse fallen und ob es ggf. Ausnahmen gibt, bei denen dann eine notwendige und vorgeschriebene Prüfung besondere Kriterien zu erfüllen hat.

Generell ist zu berücksichtigen, dass Subsysteme eines KI-Systems inklusive der damit verbundenen Sensoren und Aktoren, beispielsweise über „over the air"-Updates von Lieferanten, sich im Betrieb ändern und damit auf das Risikoprofil des Gesamtsystems Einfluss nehmen können.

## 6. Risikomatrix-Ansatz zur Bestimmung von Risikoklassen

Ein Großteil[28] der heute in Europa diskutieren Ansätze zur Bestimmung der Risikoklasse von KI-Systemen fußt auf einem von T. D. Krafft und K. A. Zweig[29] im Jahre 2019 vorgeschlagen Ansatz. Dieser bedient sich einer zweidimensionalen, sogenannten „Risiko-Matrix" mit den beiden Hauptachsen „Schadenspotential" und „Einschränkungen der Handlungsfreiheiten des Individuums" (Abb. 1).

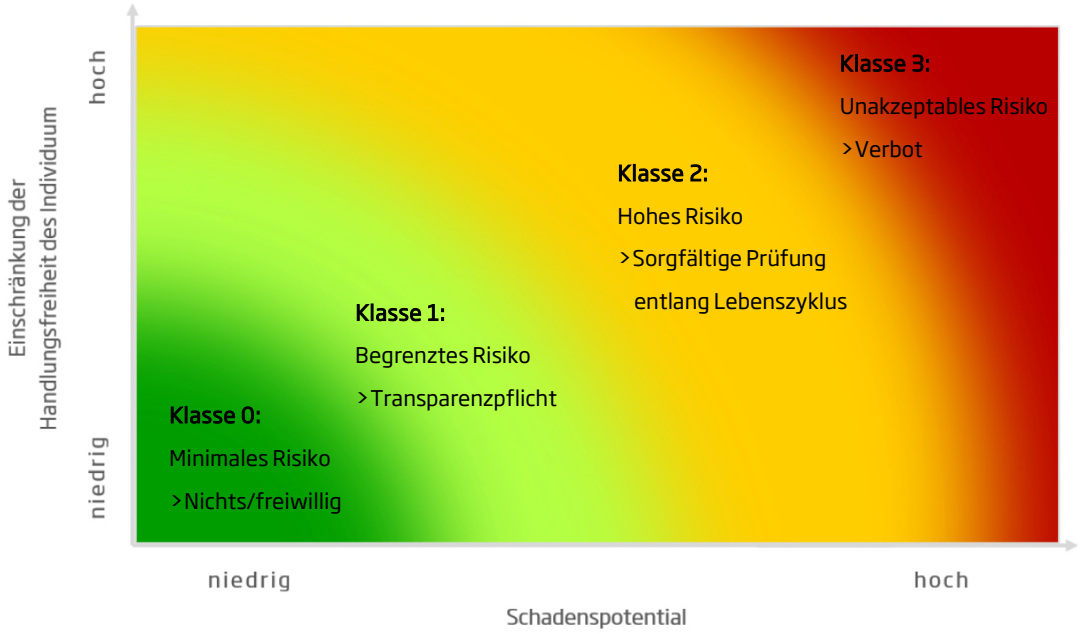



Abbildung 1: Risikomatrix nach Kritikalitätsmodell der „Normungsroadmap KI" und „Risikomatrix" von T. D. Krafft und K. A. Zweig.[30]

Auf der Achse **„Schadenspotential"** wird die Einordnung eines KI-Systems über eine aggregierte Betrachtung des Risikos für die in Abschnitt 4 definierten Schutzziele vorgenommen. Wie dort ausgeführt, sind Schutzziele 4.i bis 4.iii in Mehrheit durch bestehende Regeln und Normen beschrieben, die Prozess zur Risikoermittlung klar und allgemein akzeptiert. Die semiquantitative Bewertung mit Risikoklassen, wie beispielsweise bei der „Safety Integrity Level"-Systematik, ist Industrie- wie Prüfunternehmen bekannt und gängige betriebliche Praxis. Zudem existieren Transformationstabellen zum Vergleich von verschiedenen Systemen u.a. für verschiedene

[28] Z.B. AI Ethics Impact Group, „From Principles to Practice: An interdisciplinary framework to operationalise AI ethic" (Bertelsmann Stiftung, 1. April 2020), https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf.

[29] Krafft und Zweig, „Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse: Ein Regulierungsvorschlag aus sozioinformatischer Perspektive".

[30] Krafft und Zweig, „Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse: Ein Regulierungsvorschlag aus sozioinformatischer Perspektive", 29.

Industrien (Tabelle 1).

| Domain | Domain-Specific Safety Levels | | | | | |
|---|---|---|---|---|---|---|
| Automotive (ISO 26262) | QM | ASIL-A | ASIL-B | ASIL-C | ASIL-D | - |
| General (IEC 61508) | - | SIL-1 | SIL-2 | | SIL-3 | SIL-4 |
| Railway (CENELEC 50126/128/129) | - | SIL-1 | SIL-2 | | SIL-3 | SIL-4 |
| Space (ED-12/DO-178/DO-254) | Category E | Category D | Category C | | Category B | Category A |
| Aviation: airborne (ED-12/DO-178/DO-254) | DAL-E | DAL-D | DAL-C | | DAL-B | DAL-A |
| Aviation: ground (ED-109/DO-278) | AL6 | AL5 | AL4 | AL3 | AL2 | AL1 |
| Medical (IEC 62304) | Class A | Class B | | | Class C | - |
| Household (IEC 60730) | Class A | Class B | | | Class C | - |
| Machinery (ISO 13849) | PL a | PL b | PL c | PL d | PL e | - |

Tabelle 1: Bereichsübergreifende Abbildung von ASIL. Entsprechung von „Safety Integrity Level" in verschiedenen Industrien.[31]

Die angeführten semiquantitativen Verfahren sind auf oberster Ebene in der Regel gleich aufgebaut:

> Risiko = Schadensausmaß bzw. Schwere der Verletzung * Eintrittswahrscheinlichkeit * Exposition (Häufigkeit & Dauer des Aufenthalts) * Möglichkeit zur Vermeidung/Begrenzung des Schadens

In manchen Systematiken werden Eintrittswahrscheinlichkeit und Exposition zusammengefasst, was jedoch keinen Unterschied in der damit gemachten Aussage macht.

Damit beziehen sie alle die Möglichkeit zur Vermeidung bzw. Begrenzung des Schadens („Controllability") mit ein. Controllability erfasst die Beherrschbarkeit des Fehlers vor dem Eintritt und betrachtet damit auch die Befähigung der menschlichen Komponente eines Systems zu Fehlervermeidung. Beispielsweise wird im Automobilbereich für Safety-Level C0 die sichere Beherrschung der Situation durch alle Fahrer:innen als Forderung gestellt, bei Level C3 – schwierige Beherrschbarkeit – beherrschen weniger als 90 Prozent der Fahrer:innen die Situation. Aus der Sicht der Prüfunternehmen ist der Einbezug der Controllability zur Bestimmung des Schadenspotentials sinnvoll, da so der effektive Gesamtschaden bei Fehlurteilen des KI-Systems erfasst werden kann.

[31] Abbildung nach „Automotive Safety Integrity Level", in Wikipedia, 1. August 2021, https://en.wikipedia.org/wiki/Automotive_Safety_Integrity_Level.

Hier zeigt sich ein Unterschied zur von DIN/DKE und VDE vorgeschlagenen Systematik.[32] Dort wird „Controllability" als „Human Control" zumindest als Begriff auf der Achse „Einschränkung der Handlungsfreiheit des Individuums" verortet. Daher bedarf es einer Klärung, inwiefern dies eine reale Abweichung zur beispielsweise von DIN/DKE oder VDE vorgeschlagenen Risikomatrix ist oder ob hier komplementäre Kriterien vorliegen. So könnte beispielsweise „Controllability" in der Systematik der *Safety Integrity Levels* die Möglichkeit widerspiegeln, dass ein Mensch direkt in die Entscheidung einer KI eingreift – beispielsweise als Steuereingriff des Fahrers eines autonomen Fahrzeugs und so den Schadenseintritt verhindert. Human Control hingegen könnte die Möglichkeit beschreiben, dass ein Mensch die Entscheidungen der KI, nachdem sie getroffen worden ist, nachvollziehen, bewerten und in Frage stellen kann.

Auf der Achse **„Einschränkung der Handlungsfreiheit des Individuums"** wird abgebildet, inwiefern Betroffene der Entscheidung einer KI ausgeliefert sind. Die Einschränkung der Handlungsfreiheit muss sowohl auf individueller Ebene (z.B. personalisierte Medizin, Kredite) als auch auf Organisations- oder gesellschaftlicher Ebene (z.B. Chatbots, Social Scoring) bewertet werden.

Diese Dimension ist nicht über die in Abschnitt 4 vorgeschlagenen Schutzziele beschreibbar, da sie nicht nur den spezifischen Anwendungsfall des KI-Systems, sondern zusätzlich die Charakteristika und Möglichkeiten der von der Entscheidung einer KI betroffenen Menschen (Nutzer:innen und Umfeld) berücksichtigen muss. Das Risiko eines KI-Systems wird damit zusätzlich zum Schadenspotential (vgl. Abschnitt 6a) auch von den Möglichkeiten zur Umgehung der Entscheidung der KI, von Einspruchs- und Zweitmeinungsmöglichkeiten sowie Prozessen zur Korrektur oder Rückabwicklung der Folgen der Entscheidung der KI bestimmt.

„Human Control" – in der Deutung der Autoren dieses Papiers, die Möglichkeit für den Menschen, die Entscheidung einer KI anzunehmen oder zu verweigern bzw. auf ein von Menschen bestimmtes System auszuweichen. Dies weicht ab von der von DIN & DKE verwendeten Definition.

„Switchability" – Möglichkeiten, eine spezifischen KI und deren Entscheidungen zu umgehen, beispielsweise durch Nutzung eines anderen Anbieters/Dienstleisters. Eine Einschätzung muss sowohl den Markt (Vorhandensein eines Monopols/Oligopols) als auch den Kontext oder rechtliche Überlegungen (z.B.: die Polizei benutzt nur ein KI-System, für Übersetzung von Texten vor Gericht ist nur eine KI zugelassen) einbeziehen.

„Redress" – Möglichkeit, bereits getroffene Entscheidungen einer KI zu re-evaluieren und ggf. deren Folgen aufzuheben. Hier stellen sich auch Fragen der Rechts- und Prozesssicherheit.

[32] DIN e. V. und DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE, „Ethik und Künstliche Intelligenz: Was können technische Normen und Standards leisten?"

## 7. Herausforderung Aggregation

Beim in Abschnitt 6 beschriebenen Ansatz zur Bestimmung der Risikoklassen von KI-Systemen müssen eine Vielzahl verschiedene Größen miteinander verknüpft werden, um eine Positionierung auf einer der beiden Matrix-Achsen zuzulassen: das Risiko für die fünf Schutzziele auf der x-Achse sowie die Risiken für die Selbstbestimmung auf der y-Achse der Matrix. Die Art und Weise dieser Verknüpfung und der damit verbundenen Aggregation von Risiko-Einschätzungen ist idealerweise eindeutig und einfach nachvollziehbar festzulegen.

Auf der Achse „Einschränkung der Handlungsfreiheit", also der y-Achse der Matrix, ist eine Reduktion der drei in Abschnitt 6b.i und 6b.ii vorgestellten Unterkriterien auf eine Dimension gut darstellbar. Die Unterkriterien haben keine inhärenten Zielkonflikte. Viele Anbieter und ein geregelter Prozess zur Rückabwicklung einer auf der Basis einer KI-Empfehlung getroffenen Entscheidung stehen nicht im Widerspruch zueinander, sondern zahlen zur Risikominimierung aufeinander ein. Eine einfache Aggregation der Unterkriterien aus Abschnitt 6b erscheint, abgesehen von den Schwierigkeiten diese Unterkriterien „mess- und damit vergleichbar" zu machen, grundsätzlich möglich. Noch zu diskutieren ist hingegen, inwiefern ein niedriger Wert in einem Unterkriterium durch einen hohen Wert bei einem anderen Unterkriterium kompensiert werden kann, es nicht zu unterschreitende Mindestanforderungen gibt oder ob das Unterziel mit dem höchsten Risikowert den Gesamtwert festlegt. Diese Überlegungen werden bestimmen, ob eine additive, multiplikative, regelbasierte oder eine Mischform zur Verknüpfung bei der Aggregation zur Anwendung kommt.-

Auf der Achse „Schadenspotential", also der x-Achse, sind die Unterkriterien aus 4.i bis 4.v – Leib und Leben, Sach- und Finanzeigentum, Umwelt, weitere Grundrechte und ethische Prinzipien – nicht mehr ohne inhärenten Zielkonflikt. So kann zum Beispiel eine bessere Leistung eines Algorithmus bei der Fehlererkennung auf Kosten der Diskriminierungsfreiheit in den zugrundeliegenden Daten gehen. Um vom Risiko für die vorgeschlagenen fünf Schutzziele auf einen Wert für das „Schadenspotential" zu kommen, muss eine Bewertung und ein Vergleich der Erfüllung der verschiedenen Schutzziele stattfinden, um eine Aggregation zu ermöglichen. Das ist bei den Schutzzielen 4i. bis 4.iii materieller Risiken machbar und wird bereits heute praktiziert.[33] Auch hier stellt sich die bereits in 7a angeführte Frage, inwiefern ein niedriger Wert bei einem Schutzziel durch einen hohen Wert bei einem anderen Schutzziel kompensiert werden kann.

Eine Aufrechnung von Schutzzielen 4.iv und 4.v (Grundrechte und ethische Prinzipien) gegeneinander und gegenüber den Schutzzielen 4.i-4.iii (Leib und Leben, Sach- und Finanzeigentum, Umwelt) materieller Risiken ist aber nicht mehr ohne weiteres möglich, da vollständig verschiedene Wertedimensionen miteinander verglichen werden müssen. Hier ist eine Abwägung von Rechtsgütern

[33] Beispielsweise die Zahlung von Schmerzensgeld, bei der die Verletzung des Schutzziels „Leben und Gesundheit" monetär bewertet wird, oder die finanzielle Bewertung von Umweltschäden.

und eine Diskussion moralisch-ethischer Ebene unumgänglich. Sowohl um die notwendigen *Trade-Offs* festzulegen, aber auch zu bestimmen, wie man diese in der Praxis nachhält. Die Diskussion dieser *Trade-Offs* muss unserer Erfahrung nach angelehnt an Anwendungsfällen stattfinden und kann nicht von Unternehmen oder dem TIC-Sektor geleistet werden: Es ist ein gesellschaftlicher Diskurs im Rechts- und Werterahmens im jeweiligen Wertekanon und eine politische Willensbildung notwendig. Fragen, die hierbei zu adressieren sind, umfassen unter anderem bei:

Schutzziel iv. (gesetzlich unterlegte Grundrechte):

- Informationelle Selbstbestimmung: Sensitivität der gespeicherten und verarbeiteten Daten; wie „schützenswert" sind diese und wo besteht ein berechtigtes öffentliches Interesse zur Verwendung? Wie viele und welche höchstpersönliche Informationen bin ich bereit für eine höhere Ergebnisqualität von KI-Systemen preiszugeben, beispielsweise bei medizinischen Diagnosen? Eine Krankengeschichte kann qualitativ anders bewertet werden als Position und Geschwindigkeit eines Automobils, obwohl beide Informationen personenbezogene Daten sind.

- Einhaltung der in der Charta der in den Grundrechten der Europäischen Union definierten Werte, v.a. Diskriminierungsfreiheit (Alter, Geschlecht, Hautfarbe, Religion, Nationalität, etc.). Welche anderen Rechte sind im Rechtekanon unter- oder übergeordnet? Die Frage stellt sich beispielsweise bei *Predictive Policing*, wenn Kriminalitätsdaten dazu genutzt werden, um künftiges kriminelles Verhalten vorherzusehen und zu unterbinden. Dies kann zur Diskriminierung von benachteiligten Bevölkerungsgruppen führen, aber im Idealfall auch tatsächlich die Kriminalitätsrate senken.[34]

Schutzziel v. (ethische Prinzipien, gesellschaftlich erwünschte Ergebnisse):

- Beeinflussung der öffentlichen Meinung und des politischen Prozesses: KI unterstützte psychometrische Verfahren erlauben eine bessere Modellierung von menschlichen Neigungen und Erwartungen als zum Beispiel Meinungsumfragen. Wie schon in „Everybody Lies"[35] aus verschiedensten Blickwinkeln beleuchtet, offenbart die Auswertung der *Google*-Suchhistorie oder des *Facebook*-Profils mehr über einen Menschen, als er oder sie bei einer direkten Befragung zugeben würde, manchmal sich sogar einer Person selber bewusst ist.

[34] Katherine Aguirre, Emile Badran, und Robert Muggah, „Case study Crime prediction for more agile policing in cities Rio de Janeiro Brazil", 30. September 2019, https://www.itu.int/myitu/-/media/Publications/2019-Publications/TSB-2019/Case-study--Crime-prediction-for-more-agile-policing-in-cities--Rio-de-Janeiro-Brazil.pdf; Berfin Karakurt, „Predictive Policing und die Gefahr algorithmischer Diskriminierung", 18. April 2019, http://grundundmenschenrechtsblog.de/predictive-policing-und-die-gefahr-algorithmische-diskriminierung/.

[35] Seth Stephens-Davidowitz, *Everybody lies: Big data, new data, and what the internet can tell us about who we really are*, First edition (New York, NY: Dey St and HarperCollins Publishers, 2017).

Dieses Wissen kann bzw. wurde schon über Mikrotargeting zur Beeinflussung des politischen Prozesses ausgenutzt.[36] Gleichzeitig können das Social Media- oder Suchverhalten bei der Behandlung von psychischen Störungen oder der Identifikation von verfassungswidrigen Aktivitäten helfen. Die zu beantwortende Frage besteht also darin, ob eine Regelung des KI-Algorithmus oder des Datenzugangs der bessere Weg ist, um die gewünschten Ergebnisse zu erzielen.

- Beeinflussung von Gerichtsverfahren und Wahrheitsfindung: Deep Fakes, egal ob als Audio, Bild oder Video, sind nicht nur eine Gefahr für die informationelle Selbstbestimmung von Menschen. Sofern die Ergebnisse ausreichend echt erscheinen, können Sachverhalte verzerrt werden, bis hin zur Diskreditierung persönlicher oder politischer Gegner oder zu Verfälschung von Indizien. Ein einfaches Verbot des momentan bevorzugten Werkzeugs zur Herstellung solcher Fälschungen, der *generative adversarial networks*, wäre allerdings nicht zielführend. Denn dieses kann auch für nützliche Zwecke genutzt werden, beispielsweise in der Cybersecurity, der Tumorentdeckung oder schlicht für den Entwurf animierter Modelle. Beantwortet werden muss also die Frage, wie bei der Verwendung des gleichen Algorithmus in verschiedenen Anwendungsszenarien die Risikoklasse gefasst werden kann, ohne dass erwünschte Anwendungen ausgeschlossen werden.
- Regelung des Zugangs zu knappen Ressourcen (Heilfürsorge, Rechtsbeistand, Kredit): Oft ist der Sinn und Zweck eines KI-Systems durch Korrelation von Kriterien Gruppen zu bilden, also zu „differenzieren" – auch außerhalb der vom Menschen nachvollziehbaren Kausalität. Hier stellt sich die Frage, welche Anwendungen besonders kritisch sind und wie Diskriminierung wirkungsvoll und nachprüfbar vermieden werden kann.
- Wie lässt sich das Potential zur Gefährdung besonders schutzbedürftiger Gruppen, wie beispielsweise von Kindern, in ein messbares Schutzziel überführen?

Eine ergebnisgetriebene Klärung, verbunden mit der Umsetzung in technische Regelwerke und juristisch belastbare Handlungsanweisungen, ist dringend notwendig. Nur dadurch kann Unternehmen Rechtsunsicherheit genommen werden, die nach Umsetzung des EU-Kommissionsvorschlags drohen würde. Fragen, welche in der Praxis schon an das TÜV AI Lab herangetragen wurden, waren beispielsweise:

- Wie kann die Geschäftsführung einschätzen, dass die von der Personalabteilung genutzte HR-Software keine Hochrisikoanwendung ist? Wie ist nachvollziehbar, dass die Anwendung die für Hochrisikoanwendung nachzuweisenden Transparenz- und Diskriminierungsfreiheitspflichten erfüllt?

[36] Wong und Julia Carrie, „The Cambridge Analytica scandal changed the world – but it didn't change Facebook" (London, 2019), https://www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook.

- Wie kann ein:e Datenschutzbeauftragte:r eines Krankenhauses oder größeren Radiologiepraxis feststellen, ob ein KI-gestütztes Werkzeug zur Tumorerkennung diskriminierungsfreie Trainingsdaten verwendet? Wie kann der Datenschutz entsprechend der DSGVO oder künftiger KI-Gesetze sichergestellt werden, wenn die Daten von Patient:innen in der Cloud verarbeitet werden und dem KI-Anbieter die Möglichkeit zur Verbesserung seines Algorithmus mittels dieser Daten erlaubt wurden?

## 8. Anwendung des vorgestellten Ansatzes am Beispiel einer biometrischen Zugangskontrolle

Das folgende Beispiel verdeutlicht, inwieweit unser Ansatz in der Praxis angewandt werden kann:

1. Ein Gesichtserkennungssystem ist im Empfangsbereich eines Unternehmensstandortes und damit in einem öffentlich zugänglichen Raum installiert. Ein Kamera-unterstütztes KI-System gleicht erkannte Gesichter mit einer biometrischen Datenbank bekannter Personen ab, um Mitarbeitenden Zutritt zum Gebäude zu ermöglichen sowie Zeiterfassung zur Arbeitszeitbestimmung vorzunehmen. Ein von einem Menschen besetzter Empfang ist nicht vorhanden - nicht erkannte Mitarbeiter:innen oder Gäste müssen per Videoanruf um Einlass nachsuchen.

Das Risiko für die verschiedenen Schutzziele ist:

- Leben und Gesundheit: gering. Es muss allerdings sichergestellt werden, dass in Notfällen (z.B. Brandfall, medizinischer Notfall) die Zugangskontrollen händisch außer Kraft gesetzt werden können.
- Sach- und Finanzeigentum: mittel bis hoch. Bei zu vielen „false negatives“ wird Mitarbeitenden der Zugriff verwehrt und bei Fehlerkennung werden falsche Arbeitszeiten gebucht. Der Einlass unbefugter Personen kann jedoch hohe Folgeschäden durch Diebstahl, Spionage oder Sabotage haben. Das Risikopotential ist also auch von der spezifischen Gebäudenutzung abhängig.
- Umwelt: sehr gering.
- Grundrechte: hoch. Die KI könnte Bewegungsmuster von Mitarbeitenden erkennen und daraus höchstpersönliche Daten ableiten. So könnte aus einem langsamen oder unsicheren Gang eine medizinische Diagnose oder dem Gesichtsausdruck die psychische Gesundheit abgeleitet werden. Zudem ist das Potential für ein Bias bei biometrischen Daten hoch. Eine Diskriminierung, beispielsweise durch Nichterkennung nicht-weißer Menschen oder Menschen mit religiöser Kopfbedeckung, ist zu vermeiden.
- Ethik: gering.

In diesem Anwendungsfall ergibt sich kein Zielkonflikt zwischen Schutzzielen i. bis iii. (Leib und Leben,

Sach- und Finanzeigentum, Umwelt) und den Schutzzielen iv. und v. (Grundrechte und ethische Prinzipien). Eine geringere Fehlerhäufigkeit des Algorithmus sowie das Training mit einem möglichst diversen Datensatz hilft sowohl Erkennungsfehler zu reduzieren als auch einen Bias zu vermeiden. Zudem besteht die Möglichkeit, über einen für alle berechtigte:n Mitarbeiter:innen und Kund:innen problemlos verfügbaren Videoanruf Einlass zu bekommen - es besteht also „human control". Anforderungen an Datenschutz und Cybersecurity sind unabhängig von der Verwendung einer KI. Allerdings eröffnet die Möglichkeit *adverserial attacks* auf die Biometriesoftware und ggf. die Trainingsdatenbank auszuführen, eine neue Angriffsmöglichkeit und muss entsprechend berücksichtigt werden.

## 9. Zukünftiger Beitrag des TÜV-Verbands und des TÜV AI Labs zur Diskussion

Die TÜV-Unternehmen besitzen die Expertise, praktische Erfahrungen bei der Definition der Anforderungen der in Abschnitt 2 genannten KI-Systemen der Kategorie 1 einzubringen. Hierbei geht es darum, die notwendige „Übersetzungsleistung" zu erbringen, welche eine Konformitätsbewertung von KI-Systemen erlaubt und welche sicherheitsrelevante Komponenten von Produkten sind, die von unabhängigen Dritten geprüft werden. Dies umfasst sowohl die Entwicklung von Anforderungen an Prüfverfahren, wie auch die Festlegung von Akzeptanzkriterien für KI-Systeme im Kontext von bereits existierenden domänenspezifischen Standards. Weiterhin geht es darum, die vorhandenen technischen Regelwerke zu konkretisieren und wo notwendig, neue Regelwerke zu schaffen. Ein Ansatz dazu ist der „Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz" des Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme IAIS, der praxistaugliche Prüfverfahren für Entwickler:innen und Prüfer:innen vorschlägt.[37]

Als weiteres Arbeitsgebiet fokussieren sich die TÜV-Unternehmen und das TÜV AI Lab auf die Weiterentwicklung von Ansätzen zur Messbarmachung und Aggregation des Schadenspotentials, also den materiellen Risiken auf der x-Achse der Risikomatrix. Zu nennen sind hier vor allem:

- Die Zuordnung der existierenden „Safety Integration Levels" und analoger Äquivalente der materiellen Schutzziele 4.i bis 4.iii (Leib und Leben, Sach- und Finanzeigentum, Umwelt) zu Werten für das „Schadenspotential", deren Verortung in bestehenden Normen und Standards und die Weiterentwicklung von Normen und Standards, wie im KI-Regulierungsvorschlag der EU-Kommission vorgesehen.
- Die Erarbeitung eines Vorschlages zur Aggregation der Schutzziele 4.i bis 4.iii.

[37] Maximilian Poretschkin u. a., „Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz: KI-Prüfkatalog" (Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, 7. Juli 2021), https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf.

Gerne unterstützen wir bei der Erarbeitung, Detaillierung und Messbarmachung des Schutzziels 4.iv (Grundrechte), zum Beispiel indem wir *Best Practices* aus der Risikobewertung oder Cybersecurity-Dienstleistungen der TÜV-Unternehmen teilen. Auch bei der Diskussion zur Quantifizierung ethischer und gesellschaftlich erwünschter Schutzziele nach 4.v können wir einen Beitrag leisten, Allerdings ist die Rolle der TÜV-Unternehmen nicht diejenige, ethische Fragen zu definieren und gewünschte Antworten zu entwickeln. Dies ist eine Aufgabe für den politischen und gesellschaftlichen Dialog.

Wir sehen eine Rolle für den TÜV-Verband mit den TÜV-Unternehmen bei der Entwicklung von Verfahren zur Bewertung von ethischen Fragen für KI-Anwendungen. Als unabhängige Dritte können sie KI-Anwendungen praxisnah und mit umfangreicher Erfahrung aus der Einführung und Verbesserung von Regelwerken zu neuen Technologien bewerten und somit wichtige technische Aspekte zu deren Einflussmöglichkeiten einschätzen. Einen weiteren Beitrag können wir leisten, wenn wir mitgestalten, wie der Prozess, ethische Fragen auditierungsfest und rechtssicher zu stellen und zu beantworten, aussehen soll und wie deren Beantwortung Teil einer Marktzulassung werden kann.

Insgesamt sollte die Beurteilung der Einflussmöglichkeit einer KI-Anwendung gemeinsam durch Politik, Gesellschaft, Wissenschaft, Hersteller, Verbraucher:innen und anderen Stakeholdern erfolgen – nur dann erreichen wir das Ziel, das Vertrauen in KI-Systeme zu stärken und unsere hohen Standards als Voraussetzung dafür zu verankern, dass KI dem Menschen dient und eine breite gesellschaftliche Akzeptanz erreicht.

## Autoren

Sascha Dörfel (TÜV Thüringen)

Marc Fliehe (TÜV-Verband)

Christian Kolf (TÜV AI Lab und TÜViT)

Lars Komrowski (TÜV Hessen)

David Schimanko (TÜV Nord)

Dr. Dirk Schlesinger (TÜV AI Lab und TÜV Süd)

Sebastian Steinbach (TÜV-Verband)

Dr. Robert Walter (TÜV AI Lab und TÜV Rheinland)

## Über das TÜV AI Lab:

Im TÜV AI Lab arbeiten KI-Expert:innen aus den TÜV-Organisationen an praktischen Prüfszenarien. Arbeitsfelder sind zum Beispiel, wie sicher automatisierte Fahrzeuge mit KI-Systemen Personen, Verkehrszeichen oder bestimmte Hindernisse erkennen und darauf reagieren. Dabei spielen sowohl Performanz als auch Aspekte der Cybersecurity und der Robustheit von KI eine wichtige Rolle. Das TÜV AI Lab entwickelt Kriterien, wie die Eignung von Trainingsdaten für bestimmte KI-Anwendungen beurteilt werden kann.

Herausgeber
Verband der TÜV e. V.
Friedrichstraße 136
10117 Berlin

Tel.: +49 30 760095-400
Fax: +49 30 760095-401
E-Mail: berlin@vdtuev.de
www.tuev-verband.de