Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Zertifizierung von KI-Systemen

Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN

WHITEPAPER

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

Inhalt

Zusammenfassung

Regulierung von KI-Systemen im Allgemeinen und Zertifizierung von KI-Systemen im Besonderen können entscheidend dazu beitragen, KI-Systeme in die Anwendung zu bringen und ihr volles Nutzenpotenzial auszuschöpfen. Die Zertifizierungsverfahren sollten dabei bestimmte Standards von KI-Systemen garantieren, gleichzeitig aber Überregulierung vermeiden und Innovation ermöglichen.

Das AG-übergreifende Whitepaper entstand unter Federführung der Arbeitsgruppen IT-Sicherheit, Privacy, Recht und Ethik sowie der Arbeitsgruppe Technologische Wegbereiter und Data Science der Plattform Lernende Systeme und knüpft an ein bereits veröffentlichtes Impulspapier zur Zertifizierung von KI-Systemen an (vgl. Heesen et al. 2020a). Die Expertinnen und Experten der Plattform Lernende Systeme sowie weitere Gastautoren adressieren im vorliegenden Papier offene Fragen, etwa dazu, wann KI-Systeme zertifiziert werden sollten, an welchen Kriterien sich diese Zertifizierung orientieren soll und wie eine effiziente Infrastruktur ausgestaltet sein sollte.

Zunächst wird im Whitepaper ein Überblick über den aktuellen Diskussionsstand gegeben sowie unterschiedliche Möglichkeiten zur Sicherstellung der Qualität von Produkten und Prozessen vorgestellt (Kapitel 1). Im Anschluss daran wird eine Übersicht über mögliche nationale sowie internationale Anknüpfungspunkte für gelungene Zertifizierungsinitiativen von KI-Systemen gegeben (Kapitel 2): Dazu gehören unter anderem politische Initiativen wie das Weißbuch zu Künstlicher Intelligenz der Europäischen Kommission, Good Practice-Beispiele aus der KI-Forschung und -Anwendung sowie weitere Initiativen zu technischen Lösungen, Standardisierung und zur Prüfung und Auditierung von KI-Systemen. Darauf aufbauend wird die Frage diskutiert, wie eine gelungene Zertifizierung von KI-Systemen ausgestaltet sein sollte (Kapitel 3). Hierbei müssen Anwendungsfälle unterschieden werden, in denen eine Zertifizierung von KI-Systemen notwendig ist und solche, in denen es keine Standardisierung braucht (Kapitel 3.1). Die Notwendigkeit einer Zertifizierung von KI-Systemen kann hier aus dem Ausmaß der Kritikalität in einem bestimmten Anwendungskontext abgeleitet werden. Dieses ist abhängig von der Einschätzung der Gefährdung von Menschenleben und anderen Rechtsgütern und dem Umfang der Handlungsoptionen von Menschen in bestimmten Anwendungskontexten. Anhand des Ausmaßes der Kritikalität kann auf den möglichen Regulierungsbedarf geschlossen werden. In einem nächsten Schritt werden verschiedene Arten von Zertifizierung sowie unterschiedliche Kriterien vorgestellt, die überprüft werden sollten – etwa zu Transparenz, Nachvollziehbarkeit, Sicherheit, Gerechtigkeit, Schutz der Privatheit und Persönlichkeit sowie zur Selbstbestimmtheit (Kapitel 3.2). Unterschiede – hinsichtlich Zielsetzung und Betrachtungsgegenstand – bestehen hierbei zwischen Produkt- und Prozesszertifizierung, weshalb manche Prüfkriterien besser im Rahmen einer Produktzertifizierung und andere besser innerhalb einer Prozesszertifizierung abgefragt bzw. umgesetzt werden können. Die anzulegenden Prüfkriterien lassen sich in Mindestkriterien, die immer erfüllt und abgeprüft werden müssen, sowie darüber hinausgehende freiwillige Kriterien unterteilen, die abgeprüft werden können und somit eine Art „Zertifizierung Plus“ ermöglichen.

KI-Systeme weisen eine besondere Dynamik auf, insbesondere weiterlernende Systeme entwickeln sich im laufenden Betrieb weiter, dies muss bei der Wahl des Zeitpunkts und des Detailgrads der Zertifizierung berücksichtigt werden (Kapitel 3.3). Die Zertifizierung sollte durchgeführt werden, bevor das Produkt oder die Dienstleistung in Verkehr gebracht wird, bei weiterlernenden Systemen sollte die Zertifizierung regelmäßig wiederholt werden. Detailgrad und Prüftiefe bei der Zertifizierung sollten sich ebenfalls am Kritikalitätslevel eines KI-Systems in seinem Anwendungsgebiet orientieren – je höher die Kritikalität im Anwendungskontext eingeschätzt wird, desto umfangreicher sollten der Detailgrad und die Prüftiefe der Zertifizierung ausfallen. Zudem wird eine Übersicht über die für die Umsetzung der Zertifizierung von KI-Systemen notwendige organisatorische und technische Infrastruktur gegeben (Kapitel 3.4). Damit die Konformitätsbewertung von KI-Systemen gelingt, sind etwa technische Voraussetzungen mit Blick auf Prüfwerkzeuge, Software und Testumgebungen zu erfüllen. Organisatorische Strukturen und Prozesse in Unternehmen sollten künftig zudem zu einem wichtigen, komplementären Baustein einer Zertifizierung von KI-Systemen werden. Die Kooperation zwischen Zertifizierungsstellen und Forschungsinstituten ist vor allem wichtig, um einer dynamischen Verfasstheit der Prüfstellen Rechnung zu tragen, die auf KI-Innovationen adäquat reagieren kann.

Basierend auf diesen Überlegungen skizzieren die Autorinnen und Autoren des Papiers konkrete Gestaltungsoptionen zur Etablierung einer gelungenen Zertifizierung von KI-Systemen, die verschiedene Akteursgruppen adressieren (Kapitel 4). Zuletzt wird ein Ausblick für die nächsten Schritte hin zu einem Zertifizierungsprozess von KI-Systemen gegeben: Zentral für die Beantwortung der Frage, in welchen Fällen eine Zertifizierung von KI-Systemen notwendig sein sollte, ist hierfür der kombinierte Ansatz aus empirischer Forschung zur Erarbeitung einer objektiven Basis auf der einen Seite und konzeptioneller und normativer Überlegungen auf der anderen Seite (Kapitel 5). Das Papier leistet damit einen Beitrag zur Diskussion, wie durch geeignete Zertifizierungsverfahren Nutzenpotenziale von KI-Systemen realisiert und potenzielle negative Effekte vermieden werden können.

1. Sicherstellung der Qualität von KI-Systemen – Status quo und Herausforderungen

Regulierung von KI-Systemen im Allgemeinen und Zertifizierung von KI-Systemen im Besonderen kann vor allem über den Aufbau von Vertrauen entscheidend dazu beitragen, KI-Systeme in die Anwendung zu bringen und ihr volles Nutzenpotenzial auszuschöpfen. Hierbei ist es entscheidend, einen Modus zu finden, der Überregulierung vermeidet und Innovationen ermöglicht (vgl. z. B. Heesen et al. 2020a). Die Suche nach diesem Ansatz ist Gegenstand einer breiten Debatte, die auch politisch geführt wird, wie beispielsweise an der Veröffentlichung des Weißbuchs der Europäischen Kommission oder der zugehörigen Stellungnahme der Bundesregierung zu erkennen ist. Auch die Plattform Lernende Systeme möchte dazu beitragen, eine gelingende Zertifizierung von KI-Systemen voranzutreiben. Einen ersten Schritt in diese Richtung stellt das bereits veröffentlichte Impulspapier zur Zertifizierung von KI-Systemen dar (vgl. Heesen et al. 2020a). Dieses schließt mit zahlreichen offenen Fragen, beispielsweise dazu, in welchen Fällen tatsächlich eine Zertifizierung notwendig ist, an welchen Kriterien sich diese orientieren soll und wie eine effiziente Infrastruktur ausgestaltet sein sollte. Die Autorinnen und Autoren wollen im Rahmen des vorliegenden Papiers an das bestehende Impulspapier anknüpfen und die dort aufgeworfenen offenen Fragen adressieren. Ziel ist es, eine Zertifizierung zu skizzieren, die Nutzenpotenziale von KI-Systemen realisieren und potenzielle negative Effekte vermeiden kann.

Da das vorliegende Whitepaper an das bereits erschienene Impulspapier anknüpft, werden zentrale Erkenntnisse aus dem Impulspapier nachfolgend kompakt dargelegt (vgl. Heesen et al. 2020a):

- **Nutzenpotenzial von KI heben**: Eine Zertifizierung von KI-Systemen wird als eine mögliche Schlüsselvoraussetzung gehandelt, um das Nutzenpotenzial von KI-Systemen ausschöpfen zu können und diese in mehr Bereichen einsetzen zu können. Dies funktioniert über mehrere Mechanismen: 1.) eine Zertifizierung von KI-Systemen ermöglicht die Einhaltung wichtiger gesellschaftlicher und ökonomischer Prinzipien; 2.) eine Zertifizierung von KI-Systemen kann bei Bürgerinnen und Bürgern Vertrauen schaffen und eine entlastende Entscheidungshilfe in Bezug auf Nutzungsoptionen geben; 3.) eine Zertifizierung kann zu besseren Produkten im Sinne europäischer Werte führen sowie 4.) eine Zertifizierung von KI-Systemen kann die nationale und internationale Marktdynamik beeinflussen.
- **Das richtige Maß halten**: Damit dies gelingen kann, ist es entscheidend, dass durch die Erfüllung bestimmter gesellschaftlicher, ökonomischer und regulatorischer Standards die Grundlage für einen europäischen Weg in der Entwicklung und Anwendung von KI-Systemen gelegt werden kann, in dessen Rahmen die Gestaltung von KI-Systemen zum Wohle der Menschen im Mittelpunkt steht und gleichzeitig Überregulierung

vermieden und innovative KI-Lösungen ermöglicht werden können. Damit Zertifizierungsverfahren keine nachteiligen Auswirkungen auf den Wirtschaftsstandort Deutschland haben, müssen internationale Abstimmungsprozesse und passgenaue Verfahren Teil der Regulierungsüberlegungen sein. Probleme sind hierbei beispielsweise Markteintrittshürden (v. a. für Start-ups oder KMU) oder ein Zurückfallen in der Forschung und Entwicklung von KI-Systemen, da ein Wettbewerber auf bestehende zertifizierungsfreie Technologien ausweichen könnte.

- **Weitere Herausforderungen und offene Fragen**: Weitere Herausforderungen bestehen hinsichtlich der Statik der Zertifikate im Gegensatz zur Dynamik der KI-Systeme und ihres Umfelds. Darüber hinaus gilt es das richtige Maß an Zertifizierung (allgemeine Zertifizierung vs. umfassende, kleinteilige Zertifizierung) zu ermitteln sowie Ansatzpunkte, Maßstäbe und Metriken richtig zu wählen. Im Vergleich mit aktuellen herkömmlichen IT-Systemen werden für die Entwicklung einer Zertifizierung von KI-Systemen neue Methoden und Technologien benötigt. Eine besondere Rolle spielt zudem der Einfluss des Anwendungskontextes eines KI-Systems auf die Kriterien der Regulierung. Bevor eine gelungene Zertifizierung von KI-Systemen etabliert werden kann, sind daher noch offene Fragen zu klären. Diese betreffen den Gegenstand der Zertifizierung, die Prüfkriterien, den Zeitpunkt und die Notwendigkeit der Zertifizierung, den Detailgrad der Zertifizierung sowie den Umgang mit weiterlernenden Systemen.

Vorliegendes Whitepaper knüpft an diese Überlegungen an und legt erste Antwortoptionen und Gestaltungsansätze für die aufgeworfenen Fragen dar.

Die Sicherstellung der Qualität von Technologien im Allgemeinen und von KI-Systemen im Besonderen kann auf Grundlage unterschiedlicher Herangehensweisen erfolgen, die sich auf rechtliche und ethische Vorgaben oder eben auf die standardisierten Verfahren der Zertifizierung beziehen. Zertifizierung lässt sich nicht komplett eigenständig verorten, da sie rechtlich vorgeschrieben sein kann – aber nicht muss – und unter Umständen auch ethische Vorgaben enthalten kann. Das folgende Papier bezieht sich auf Regulierung im Allgemeinen und Zertifizierung im Besonderen, das heißt, dass rechtliche und ethische Normen Teil oder Grundlage von Regulierungsprozessen sind. Die weiteren Möglichkeiten zur Sicherstellung der Qualität von Technologien im Allgemeinen und von KI-Systemen im Besonderen sind vielfältig (siehe Infobox). Eine Zertifizierung beschreibt nur eines dieser möglichen Verfahren: Sie ist die höchste der drei Stufen der Konformitätsbewertung (siehe Infobox) und beschreibt eine Überprüfung gegen national oder auch international anerkannte und gültige branchenabhängige Standards und Richtlinien. Eine Zertifizierung erlangt ihre Qualität unter anderem durch die Akkreditierung[1] der zertifizierenden Stelle durch die zuständige Akkreditierungsstelle; in Deutschland sind hierfür die Deutsche Akkreditierungsstelle DAkks GmbH sowie für bestimmte hoheitliche Aufgaben staatliche Stellen, wie beispielsweise das Bundesamt für Sicherheit in der Informationstechnik (BSI), zuständig.

1 Diese Akkreditierung wird zeitlich befristet vergeben und umfasst unter anderem eine Überprüfung der Prüfkataloge des Zertifizierers.

Konformitätsbewertungen
Konformität beschreibt die Übereinstimmung eines Produkts oder Systems mit vorher festgelegten Anforderungen (z. B. Standards oder auch selbst festgelegten Anforderungen). Eine Konformitätsbewertung ist die Analyse dieser Übereinstimmung mit den Anforderungen. Es existieren drei Stufen der Konformitätsbewertung: 1.) **First party**: Eigene Bewertung, 2.) **Second party**: Überprüfung durch den Abnehmer, 3.) **Third party**: Überprüfung durch unabhängige Dritte.

Möglichkeiten zur Sicherstellung der Qualität von Technologien im Allgemeinen und von KI-Systemen im Besonderen

Freiwillige Selbstverpflichtung
Unter einer freiwilligen Selbstverpflichtung wird im vorliegenden Kontext eine Erklärung von Organisationen (z. B. Wirtschaftsverbänden) oder Unternehmen verstanden, gewisse Regeln einzuhalten. Diese freiwillige Selbstverpflichtung ist rechtlich nicht bindend und umfasst keine staatliche Regulierung.

Gütesiegel
Ein Gütesiegel ist eine Produktkennzeichnung, dass ein Produkt gewisse Eigenschaften erfüllt (beispielsweise hinsichtlich Qualität, Sicherheit oder auch Umwelteigenschaften). Ein Gütesiegel kann prinzipiell von jedem konzipiert und ausgestellt werden – meist schließen sich branchenspezifisch verschiedene Hersteller oder Anbieter zusammen und legen die Kriterien für ein Gütesiegel fest.

(Freiwillige oder gesetzlich vorgeschriebene) Zertifizierung
Eine Zertifizierung ist eine meist zeitlich begrenzte Bestätigung, dass vorgegebene Standards, Normen oder Richtlinien eingehalten werden. Diese Bewertung wird von unabhängigen Dritten (Zertifizierungsstellen) durchgeführt und ist die höchste von drei Stufen der Konformitätsbewertung. Grundlage sind unterschiedliche national oder auch international anerkannte und gültige branchenabhängige Standards und Richtlinien. Es können sowohl Produkte und Dienstleistungen als auch Systeme, Prozesse und Personen zertifiziert werden. Meistens erfolgt eine Zertifizierung auf freiwilliger Basis, um die Qualität des zertifizierten Gegenstands nachzuweisen. Im Bereich KI-Systeme existieren aktuell (Stand November 2020) kaum gültige und anerkannte Standards und Normen, die konkret genug sind, um die Basis einer Zertifizierung bilden zu können.

Zertifizierung kann gesetzlich vorgeschrieben oder freiwillig erfolgen. Ein Beispiel für eine gesetzlich vorgeschriebene Zertifizierung ist die CE-Zertifizierung in der Medizin.

Zulassung
Die Zulassung beschreibt eine gesetzlich vorgeschriebene Prüfung gegen (europäische) Gesetze, in welchen die Anforderungen an dieses System oder Produkt konkret benannt sind. Die Zulassung wird von einer durch das jeweilige Bundesamt oder die jeweilige Bundesbehörde beauftragten Institution durchgeführt. Die Zulassung unterscheidet sich je nach Domäne – so heißt sie beispielsweise im Bereich der Automobilität Typzulassung und wird auf Veranlassung des Kraftfahrtbundesamts durch den TÜV vorgenommen.

Regulierung im Sinne von Ge- und Verboten
Der weitestgehende Eingriff ist eine staatliche Regulierung im Sinne von Ge- und Verboten. Diese können entweder gesamte Systeme oder Produkte oder bestimmte Anwendungskontexte dieser Produkte betreffen.

2. Bestehende Zertifizierungs-initiativen und -verfahren

Es existieren bereits einige Initiativen sowie Good Practice-Beispiele, auf denen die Diskussion um Zertifizierung von KI-Systemen aufbauen kann. Einige dieser Initiativen werden bereits im Impulspapier Zertifizierung von KI-Systemen der Plattform Lernende Systeme vorgestellt (vgl. Heesen et al. 2020a) und werden daher an dieser Stelle nur erwähnt, weitere, vor allem politische Initiativen und Good Practice-Beispiele werden ausführlicher dargestellt.

2.1 Politische Initiativen

Sowohl auf der europäischen als auch der nationalen Ebene gibt es bereits politische Initiativen, die Anknüpfungspunkte für die Entwicklung von Konformitätsbewertungen von KI-Systemen bieten. Hierbei ist vor allem das Weißbuch zu Künstlicher Intelligenz der Europäischen Kommission zu nennen sowie die Stellungnahme zum Weißbuch der Bundesregierung.

Europäische Initiative: Das Weißbuch zu Künstlicher Intelligenz der Europäischen Kommission

Die EU-Kommission hat am 19.02.2020 das Weißbuch „Künstliche Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen“ veröffentlicht. Dieses enthält Ansätze für eine Konformitätsbewertung und freiwillige Kennzeichnung von KI-Systemen sowie für Bewertungskriterien und für Auflagen, die für KI-Systeme mit hohem Risiko gelten sollen. Das Weißbuch verweist mitunter auch auf die Ergebnisse der High Level Expert Group on Artificial Intelligence der EU (AI HLEG).

Konformitätsbewertung
Die EU-Kommission schlägt einen risikobasierten Ansatz vor, um die Notwendigkeit einer Konformitätsbewertung zu ermitteln, der zwischen hohem und nicht-hohem Risiko unterscheidet. Sie ist der Auffassung, dass eine Vorab-Konformitätsbewertung erforderlich wäre, um zu prüfen, ob KI-Anwendungen mit hohem Risiko bestimmte Anforderungen erfüllen. Sollte eine Konformitätsbewertung negativ ausfallen, müssten Mängel beseitigt werden. Die Dokumentation von KI-Anwendungen sollte Ex-post-Kontrollen durch Dritte (z. B. zuständige Behörden) ermöglichen. Zuständige Behörden sollten in die Lage versetzt werden, geltende Vorschriften durchzusetzen und sowohl Einzelfälle überprüfen als auch Auswirkungen auf die Gesellschaft bewerten zu können. Wo es möglich ist, sollte an bestehende Mechanismen und Verfahren angeknüpft werden. Die EU-Kommission betont, dass KMU unterstützt werden sollten, um an Konformitätsbewertungen teilzunehmen.

Bewertungskriterien
Bei der Definition der Kriterien orientiert sich die EU-Kommission an den Vorschlägen der AI HLEG. Diese sind: Vorrang menschlichen Handelns und menschlicher Aufsicht, technische Robustheit und Sicherheit, Privatsphäre und Datenqualitätsmanagement, Transparenz, Vielfalt, Nichtdiskriminierung und Fairness sowie gesellschaftliches und ökologisches Wohlergehen und Rechenschaftspflicht.

Freiwillige Kennzeichnung für KI-Anwendungen ohne hohes Risiko
Die EU-Kommission schlägt vor, dass Wirtschaftsakteure, die KI-Anwendungen ohne hohes Risiko anbieten, für diese Anwendungen auf freiwillige Kennzeichnungen (sogenannte „Gütesiegel", siehe Infobox) zurückgreifen können. Bei der Nutzung einer Kennzeichnung wären deren Anforderungen verbindlich. Durch eine Kombination aus Ex-ante- und Ex-post-Kontrollen müsste sichergestellt werden, dass die Anforderungen erfüllt werden.

Anforderungen für KI-Anwendungen mit hohem Risiko[2]
Für KI-Anwendungen mit einem hohen Risiko formuliert die Europäische Kommission spezifische Anforderungen. So sollte sichergestellt sein, dass die KI-Anwendungen auf der Grundlage adäquater Trainingsdaten trainiert wurden, sodass die Sicherheit von Anwendungen hinreichend gewährleistet ist, die Ergebnisse einer Nutzung der Anwendungen nicht zu Diskriminierung führen und der Datenschutz bzw. die Privatsphäre beachtet werden. Zudem könnte vorgesehen werden, dass relevante Aufzeichnungen zu Datensätzen und die Dokumentation zu Programmier- und Trainingsmethoden sowie gegebenenfalls die Datensätze selbst aufbewahrt werden. Weiterhin sollten Informationen bezüglich des Zwecks, der Fähigkeiten und Grenzen des Systems bereitgestellt werden. Für die Nutzenden sollte Klarheit darüber herrschen, ob sie mit einem KI-System interagieren. Auch Robustheit und Genauigkeit werden als Anforderungen angeführt. Die Kommission schlägt weiterhin Möglichkeiten vor, mit denen die menschliche Aufsicht eines KI-Systems hergestellt werden könnte. Sie hebt ebenfalls hervor, dass an Systeme für biometrische Fernidentifikation besondere Anforderungen gestellt werden sollten, da diese Grundrechte und den Datenschutz besonders berühren.[3]

Nationale Initiative: Die Stellungnahme der Bundesregierung zum Weißbuch der Europäischen Kommission

Die Bundesregierung hat im Rahmen einer Stellungnahme am 29.06.2020 auf das Weißbuch der Europäischen Kommission reagiert und darin unter anderem ihre Überlegungen zur Konformitätsbewertung und freiwilligen Kennzeichnung von KI-Systemen dargelegt. Dabei soll eine mögliche Überregulierung vermieden werden, um Innovationen zu ermöglichen und nicht zu hemmen.

2 Es handelt sich hierbei um Vorschläge für die Diskussion möglicher künftiger, rechtsverbindlicher Anforderungen.
3 Weitere Informationen: Weißbuch zu Künstlicher Intelligenz: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf; AI HLEG Ethisches Framework: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419; AI HLEG Assessment List: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60440

Konformitätsbewertung

Die Bundesregierung schlägt ebenfalls einen risikobasierten Ansatz vor, um die Notwendigkeit einer Konformitätsbewertung zu ermitteln. Dieser Ansatz sollte jedoch mehr als zwei Stufen umfassen und insbesondere Schadenspotenziale und Anwendungskontexte miteinbeziehen. Die Bundesregierung sieht eine objektive Konformitätsbewertung bei KI-Anwendungen mit hohem Risiko als notwendig an. Diese Bewertung sollte stattfinden, bevor eine Anwendung in den Verkehr kommt oder wenn sich diese wesentlich verändert. Eine wiederholte Prüfung bei weiterlernenden Systemen erachtet die Bundesregierung daher als sinnvoll. Für die Konformitätsbewertung sollte auf bestehende nationale Strukturen und Verfahren zurückgegriffen werden. Sofern es keine solche Behörden gibt, sollte es eine Pflicht zum Aufbau einer solchen Behörde oder zum Aufbau von Zuständigkeiten in bestehenden Behörden geben. KMU und der Nonprofit-Bereich sollten bei der Teilnahme an der Konformitätsbewertung unterstützt werden. Zudem könnte es eine Öffnungsklausel bzw. Ausnahmeregelungen geben, zum Beispiel für die Forschung oder in Krisensituationen.

Freiwillige Kennzeichnung für risikoarme KI-Anwendungen

Die Bundesregierung begrüßt eine freiwillige Kennzeichnung risikoarmer Anwendungen. An einem solchen Kennzeichnungssystem sollten Wirtschaftsakteure, öffentliche Organisationen, Behörden und Vereine gleichermaßen teilnehmen können. Die Kennzeichnung sollte jedoch befristet sein, sodass Teilnehmende das Gütesiegel ihres Produktes regelmäßig erneuern müssen. Es sollte zudem in europaweit anerkannten Stellen sowie im Binnenmarkt gegenseitig Anerkennung finden und von den Behörden der Mitgliedsstaaten kontrolliert werden. Ferner bedarf es wirkungsvoller, rechtlich durchsetzbarer Sanktionen, wenn Teilnehmende die Anforderungen nicht erfüllen bzw. das Gütesiegel missbräuchlich nutzen.

Anforderungen für KI-Anwendungen mit hohem Risiko

Aus Sicht der Bundesregierung eignen sich die von der EU-Kommission vorgeschlagenen Anforderungen als Maßstab einer Konformitätsbewertung (siehe Konformitätsbewertung siehe Infobox.).[4]

4 Weitere Informationen: Stellungnahme der Bundesregierung: https://www.ki-strategie-deutschland.de/files/downloads/Stellungnahme_BReg_Weissbuch_KI.pdf

2.2 Bereits etablierte Zertifizierungsverfahren

Es existieren bereits Zertifizierungsverfahren, die für die Diskussion um die Zertifizierung von KI-Systemen einen Orientierungspunkt darstellen können. Mit dem Zertifizierungsverfahren von KI-Systemen in Malta und einem Zertifizierungsverfahren im Bereich der IT-Sicherheit in Deutschland werden im Folgenden zwei Beispiele vorgestellt.

Kriterienkatalog und Umsetzung der Zertifizierung in Malta

Malta hat mit dem Digital Authority Act von 2018 eine Behörde zur Zertifizierung innovativer technologischer Anwendungen eingerichtet. Die Zertifizierung wird auf freiwilliger Basis durchgeführt. Die Kontrolle und Durchführung ist bei der Malta Digital Innovation Authority (MDIA) angesiedelt. In Kooperation mit lizenzierten Systemauditoren zertifiziert diese Behörde KI-Anwendungen.

Zertifizierungsprozess
Antragsteller können einen Antrag auf Zertifizierung bei der MDIA stellen. Basierend auf dem Innovative Technology Arrangements and Services (ITAS) Act führen zugelassene Systemprüfer gemeinsam mit der MDIA die Zertifizierung durch. Dabei führen sie einen mehrgliedrigen Erstzertifizierungsprozess durch:

- Dem Antrag bei der MDIA ist der sogenannte Blueprint beizufügen. Dieser gibt Folgendes an: welchen Zweck das System hat; welchen Nutzen Anwendende haben (Qualitäten des Produkts); welche Aspekte relevant für die Zertifizierung sind; welche Funktionen und Fähigkeiten das Produkt hat; wie sich das System verhält (beispielsweise bei unerwartetem Input oder Prozessen) und welche Grenzen das System hat.
- Die Systemprüfer (System Auditors) prüfen das System und erstellen ihren Bericht. Dieser beinhaltet ein Gutachten darüber, ob das System wie im Blueprint beschrieben arbeitet und ob das System alle durch den ITAS-Act vorgeschriebenen Regelungen erfüllt. Zudem muss das System eine Funktion haben, die es dem Menschen ermöglicht, im Fall eines technischen Fehlverhaltens einzugreifen.
- Auf Grundlage des Berichtes, des Blueprints und des Antrages überprüft die MDIA, ob eine Zertifizierung erfolgt.[5]

Eine Re-Zertifizierung wird periodisch über den gesamten Lebenszyklus des Systems durchgeführt. Hierbei werden die operationalen Funktionsweisen überprüft.

5 Weitere Informationen: Zertifizierungsprozess: https://mdia.gov.mt/wp-content/uploads/2019/08/ESA_Guidelines.pdf; Prüfkatalog für Systemprüfer: https://mdia.gov.mt/wp-content/uploads/2018/10/Systems-Audit-Control-Objectives-30Oct2018_Final.pdf; Maltas KI-Ethik-Guidelines: https://malta.ai/wp-content/uploads/2019/08/Malta_Towards_Ethical_and_Trustworthy_AI.pdf; Innovative Technology Arrangements and Services (ITAS) Act: https://legislation.mt/eli/bill/2018/43/eng/pdf

Bewertungskriterien

Die MDIA orientiert sich bei der Überprüfung eines Systems am KI-Ethik-Framework Maltas. Das Framework verfolgt dabei vier zentrale Ziele: 1.) einen menschenzentrierten Ansatz schaffen; 2.) die Beachtung und Befolgung aller anwendbaren Rechte und Normen, der Menschenrechte und demokratischer Werte; 3.) die Maximierung des Nutzens von KI bei gleichzeitiger Minimierung der Risiken und 4.) die Anpassung an künftige internationale Standards und Normen für KI-Ethik.

Neben diesen Zielen orientiert sich die MDIA zudem an den darauf aufbauenden vier KI-Prinzipien für eine vertrauenswürdige KI: 1.) Menschliche Autonomie, 2.) Verhinderung von Schaden, 3.) Fairness, 4.) Erklärbarkeit. Darüber hinaus betreffen weitere Kontrollpraktiken für KI beispielsweise die Schulung der Mitarbeitenden eines Unternehmens, die Unternehmensorganisation oder die Verwendung/Beschaffung der Anwendung. Auch diese Praktiken sollten laut des Ethik-Frameworks von der MDIA beachtet werden. Hinzu kommen außerdem die gesetzlich festgehaltenen Regeln des ITAS-Acts.

Fraunhofer Fokus CertLab: Prüfbegleitung bei Common Criteria-Zertifizierungsverfahren

Das Common Criteria Certification Lab des Fraunhofer Fokus (CertLab) begleitet seit 2010 Prüfstellen bei der Evaluierung von Software und Hardwareprodukten mit Fokus auf IT-Sicherheit. Grundlage der Evaluierung bilden die etablierten und anerkannten „Common Criteria for Information Technology Security Evaluation" (CC, ISO/IEC 15408). Das Bundesamt für Sicherheit in der Informationstechnik (BSI) besitzt die Verfahrenshoheit.

Zertifizierungsprozess

Der Antragsteller reicht einen Antrag beim BSI ein und beauftragt eine seitens des BSI anerkannte Prüfstelle mit der Evaluierung. Dafür sendet er die Produkt- und Herstellungsunterlagen an die Prüfstelle. Das CertLab wird wiederum durch das BSI beauftragt, die Prüfstelle bei der Evaluation zu begleiten. Diese erstellt einen Prüfbericht und übermittelt diesen zurück an den Antragsteller, das BSI und das CertLab, wobei das Prüfverfahren durch das CertLab und das gesamte Verfahren durch das BSI begleitet wird. Bei bestandener Untersuchung übermittelt das BSI dem Antragsteller das entsprechende Zertifikat. Das CC-Label und das Prüfverfahren sind bereits seit einigen Jahren etabliert und international anerkannt. Auch für die Bewertung der IT-Sicherheit von KI-Systemen stellen Common Criteria einen wichtigen Referenzpunkt dar.[6]

6 Weitere Informationen: Webseite des CertLab: https://www.fokus.fraunhofer.de/go/certlab; Webseite zu den Common Criteria: https://www.commoncriteriaportal.org/ccra/index.cfm

2.3 Weitere Initiativen

Eine Reihe weiterer Initiativen und Maßnahmen bieten Anknüpfungspunkte für die Bewertung von KI-Systemen und für die technische Umsetzung spezifischer Anforderungen an KI-Systeme. Einige solcher Initiativen wurden bereits im Impulspapier Zertifizierung von KI-Systemen vorgestellt (siehe Infobox). Ergänzend werden im Folgenden Initiativen zu technischen Lösungen, Standardisierung und zur Prüfung und Auditierung von KI-Systemen aufgegriffen (Liste ist nicht abschließend).

Initiativen, die im Impulspapier bereits vorgestellt werden
(vgl. Heesen et al. 2020a)

- AI Ethics Impact Group: Integrierter Ansatz zur wirksamen und normungsfähigen Verankerung von Ethik beim Einsatz von Künstlicher Intelligenz
  Die Gruppe hat ein Ethics Rating für KI-Systeme entwickelt, um KI-Ethik messbar zu machen.

- Zertifizierte KI: Zertifizierung zur Sicherstellung einer vertrauenswürdigen KI
  Ziel des Programms ist die standardisierungsreife Entwicklung von technischen Prüfkriterien, Prüfmethoden und Prüfwerkzeugen für KI-Anwendungen und die Etablierung der erforderlichen Prüfinfrastruktur im Bundesgebiet.

- Denkfabrik Digitale Arbeitsgesellschaft: KI-Observatorium
  Das Observatorium widmet sich dem Monitoring der Auswirkungen von KI auf Leben und Arbeit in Deutschland und versteht sich als „Kartograph" eines neuen technologischen Ökosystems.

- Nationale Initiative: Die KI-Normungsroadmap
  Die Roadmap wird eine Übersicht über bestehende Normen und Standards zu KI-Aspekten umfassen und Empfehlungen im Hinblick auf noch notwendige künftige Aktivitäten geben.

Initiative zu technischen Lösungen: IBM Toolbox 360 Fairness

Seit 2018 ist das von IBM entwickelte Open-Source Tool zur Implementierung von Fairness bei Machine Learning (ML)-Algorithmen verfügbar. Die Toolbox bietet einen Satz von Fairness-Metriken für Datensätze und Modelle (inklusive Erklärungen für diese Metriken). Zudem beinhaltet es Algorithmen zur Minderung von Verzerrungen in Datensätzen und Modellen. Durch ein interaktives Web-Erlebnis sollen Anwendende in die Konzepte und Möglichkeiten der Toolbox eingeführt werden. Mit weiteren Produkten wie AI Explainability 360 und AI Adversarial Robustness 360 werden weitere Kriterien für KI-Systeme implementierbar gemacht.[7]

Initiative zur Standardisierung: Institute of Electrical and Electronics Engineers (IEEE) – P7000 Serie

Das IEEE hat seit 2016 mit der P7000 Serie eine Reihe von Standardisierungsprozessen für technische, digitale und datenverarbeitende Systeme angestoßen. Darunter befinden sich etwa Standardisierungsprojekte zu Themen wie Transparenz, Data Privacy, Bias von Algorithmen, Wohlbefinden oder auch zur Berücksichtigung von ethischen Kriterien während der Entwicklung von Systemen generell (vgl. P7000 bis P7014). Die Projekte sind überwiegend noch nicht abgeschlossen – lediglich der IEEE 7010-2020 Standard „Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being" ist bereits erhältlich. Die IEEE hat ein Papier veröffentlicht, das sich mit ethischen Richtlinien für die Entwicklung autonomer und intelligenter Systeme befasst. Diese Ethik-Guidelines sind mit Unterstützung der P7000-Projektgruppen entstanden. Es werden unter anderem auch acht General Principles behandelt, die bei KI-Systemen berücksichtigt werden sollten. Dazu zählen vielfach genannte technische Anforderungen wie Transparenz und Accountability, aber auch Anforderungen an den Menschen, wie beispielsweise ein Bewusstsein für möglichen Fehlgebrauch (gerichtet an Entwickelnde) oder auch Kompetenz für den Umgang mit KI-Systemen (gerichtet an Nutzende).[8]

Initiative zu Prüfung und Auditierung von KI: ExamAI

Ein Verbundprojekt des KI-Observatoriums der Denkfabrik Digitale Arbeitsgesellschaft unter Federführung der Gesellschaft für Informatik sucht nach Lösungen, wie effektive Kontroll- und Testverfahren für KI-Systeme gestaltet sein müssen und wie die Bundesregierung solche Verfahren implementieren kann. Für das Projekt ist ein sechsstufiger Prozess vorgesehen. In einem ersten Schritt sollen Anwendungsfälle identifiziert und analysiert werden. Anschließend werden bestehende technische Standards untersucht. Als dritten Schritt will das Projekt rechtliche Anforderungen an KI-Systeme herausarbeiten.

7 Weitere Informationen: Informationen zur Toolbox https://aif360.mybluemix.net; Papier des Entwicklerteams, das sich differenziert mit dem Programm auseinandersetzt: https://arxiv.org/abs/1810.01943

8 Weitere Informationen: Übersicht über die IEEE-Projektgruppen: https://ethicsinaction.ieee.org/p7000/; IEEE Ethical Aligned Design: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

In einem vierten Schritt konzeptionieren die Mitglieder eigene Test-, Auditierungs- und Zertifizierungsverfahren. Die zuvor entworfenen Methoden sollen dann in einem fünften Schritt nochmals eingehend untersucht und geprüft werden, bevor schlussendlich konkrete Handlungsempfehlungen formuliert werden. Die Arbeit der Projektgruppe ist noch nicht abgeschlossen, sodass bisher noch keine Ergebnisse veröffentlicht wurden.[9]

Zertifizierte KI: Zertifizierung von Standard-KI-Anwendungen

In der Initiative werden Qualitätsstandards für die vertrauenswürdige Anwendung von KI erarbeitet sowie die Grundlagen und Anforderungen an einen Zertifizierungsprozess für die sachkundige Überprüfung von (technisch) zugesicherten Eigenschaften entwickelt. Im Rahmen von Pilotprojekten für branchen- und technologierelevante Use Cases von Unternehmen werden die konzipierten Kriterien und Methoden validiert. Der holistische Ansatz des Verbunds, welcher Expertise aus Informatik, Wirtschaftswissenschaften, Rechtswissenschaften und Philosophie vereint, zielt darauf, die gesellschaftliche Akzeptanz, Realisierbarkeit und Marktfähigkeit der zu konzipierenden Prüfmethodologie sicherzustellen. Beteiligte Institutionen sind das Fraunhofer IAIS, das Bundesamt für Sicherheit in der Informationstechnik (BSI), das Deutsche Institut für Normung e. V. (DIN), die Universität Bonn, die Universität zu Köln, die RWTH Aachen und die Kompetenzplattform KI.NRW sowie weitere Industriepartner. Die Projektergebnisse werden sowohl in einem öffentlichen Beteiligungsprozess diskutiert als auch in die Standardisierung eingebracht.[10]

9 Weitere Informationen: Für eine Übersicht zu den genauen Vorgängen der Prozesse: https://testing-ai.gi.de/ueber
10 Weitere Informationen: https://www.iais.fraunhofer.de/ki-zertifizierung

3. Wie kann die Zertifizierung von KI-Systemen gelingen?

Ausgangpunkt für eine Zertifizierung ist die Feststellung der Notwendigkeit eine Konformitätsprüfung durch eine unabhängige Drittstelle. Mit dem Konzept der Kritikalität wird in Kapitel 3.1 eine Orientierung angeboten, um eine solche Notwendigkeit festzustellen. Anschließend werden Systemeigenschaften und die entsprechend unterschiedlichen Gegenstandsbereiche der Zertifizierung in Kapitel 3.2 behandelt. Im Kapitel 3.3 wird schließlich auf die organisationale und technische Infrastruktur der Konformitätsbewertung von KI-Systemen eingegangen.

3.1 In welchen Fällen ist eine Zertifizierung von KI-Systemen notwendig?

Die Europäische Kommission sowie die Bundesregierung setzen auf einen risikobasierten Ansatz, um die Notwendigkeit einer Zertifizierung einzuschätzen (vgl. Bundesregierung 2020, Europäische Kommission 2020). Ein solcher Ansatz kann mit dem Konzept der Kritikalität näher spezifiziert werden. Der folgende Abschnitt fokussiert auf dieses Konzept, mit dessen Hilfe eine Einschätzung abgegeben werden kann, ob in bestimmten Fällen eine Zertifizierung notwendig sein soll und wer gegebenenfalls in welchem Ausmaß diese Form der Konformitätsprüfung durchführen soll. Das Konzept der Kritikalität bezieht sich im Rahmen dieses Whitepapers primär auf den Anwendungskontext von KI-Systemen und nicht lediglich auf die Eigenschaften von KI-Systemen selbst. Das bedeutet, dass die Kritikalität stets vor dem Hintergrund des Anwendungskontextes eingeschätzt werden sollte und nicht unabhängig von diesem auf der Basis systemeigener Charakteristika.

Wird eine Zertifizierung von KI-Systemen angestrebt, werden spezifische Konzepte nötig, um bestimmen zu können, wann eine Zertifizierung in welcher Art und Weise und in welchem Umfang erforderlich werden soll. Während ein Großteil der KI-Systeme unproblematisch sein dürfte – ein Beispiel wären lernende Algorithmen zur Identifizierung von Spam in der alltäglichen E-Mail-Kommunikation – gibt es Anwendungen, die einer genaueren Prüfung unterzogen werden sollten. Gerade deshalb ist es sinnvoll, Kriterien heranzuziehen, um Orientierung hinsichtlich einer Unterscheidung von kritischen und nicht-kritischen KI-Systemen zu bieten. Besonders hilfreich ist hierbei das Konzept der Kritikalitätsstufen. Einstufungen auf der Grundlage von Kritikalität wurden bereits im Bericht der Datenethikkommission als Ansatzpunkt hervorgehoben, um über die Notwendigkeit und das Ausmaß von Regulierung zu reflektieren (vgl. Datenethikkommission 2019, S. 177). Ausgehend von Zweig & Krafft (2019) entwickelt die Plattform Lernende Systeme die Überlegungen zur Kritikalität weiter. Die Plattform folgt dabei dem Ansatz eines abgestuften Regulierungsbedarfs entlang des Grades der Kritikalität.[11]

11 Die Plattform Lernende Systeme fokussiert auf Algorithmic Decision Making-Systemen (ADM-System) mit lernenden Komponenten im Besonderen und von KI-Systemen im Allgemeinen, weil dadurch die Konzepte im Weiteren vereinfacht erläutert werden können.

Die folgenden Ausführungen geben einen ersten Einblick in die Arbeit der Plattform Lernende Systeme zum Konzept der Kritikalität. Eingehendere Erläuterungen zum Kritikalitätskonzept finden sich bei Heesen et al. (i. E.). Innerhalb des Konzepts werden sowohl Anbietende und Nutzende als auch Betroffene betrachtet.

Einschätzung der Kritikalität

Zum einen ist bei der Einschätzung der Kritikalität der potenzielle physische und immaterielle Schaden zu berücksichtigen, der durch ein KI-System in einem spezifischen Anwendungskontext entstehen könnte. Der **Schaden** bezieht sich auf die Gefährdung von Menschenleben und weiterer Rechtsgüter wie etwa der Umwelt oder Persönlichkeitsrechte (Privatheit, Gerechtigkeit, Fairness). Hierbei ist sowohl die Eintrittswahrscheinlichkeit des Schadens mit Blick auf den Verbreitungsgrad eines KI-Systems einzubeziehen als auch der wahrscheinliche Schweregrad sowie die mögliche Persistenz und die Kontrollierbarkeit eines Schadens. Der Vernetzungsgrad des KI-Systems spielt für die Einschätzung der Kontrollierbarkeit eines Schadens eine besondere Rolle.[12] Mit zunehmendem Grad der Vernetzung eines KI-Systems mit weiteren Systemen (auch solchen ohne KI-Komponente) steigt die Kritikalität, wenn sich dadurch eine Bedrohungssituation oder ein Schadenspotenzial zunehmend ausbreiten und sich die entstandenen Schäden unter Umständen schwerer rückgängig machen lassen. Allerdings könnte ein KI-System durch eine Einbettung in andere Systeme auch als weniger kritisch gelten, etwa dann, wenn es hinsichtlich seiner Funktion oder Systemausgaben in die Absicherungsmaßnahmen anderer Systeme eingebunden ist. Hierbei ist also wiederum dem spezifischen Anwendungskontext des KI-Systems bei der Einschätzung der Kritikalität Rechnung zu tragen.

Zum anderen sollte eine Kritikalitätseinschätzung den Umfang der **Handlungsoptionen von Menschen** im Rahmen eines spezifischen Anwendungskontextes eines KI-Systems berücksichtigen. Hierbei ist zu betrachten, über welche Handlungsspielräume ein Individuum verfügt, das heißt unter anderem, inwiefern sich ein Individuum einem KI-System entziehen kann. Der Handlungsspielraum hängt davon ab, ob das Individuum das KI-System selbst konfigurieren kann, um etwa bestimmte Funktionen abzuschalten. Weiterhin gilt es zu reflektieren, inwiefern ein Individuum praktisch in der Lage ist, das System zu konfigurieren, etwa mit Blick auf die Verfügbarkeit notwendiger Informationen. Schließlich ist der Umfang der Handlungsoptionen ebenfalls von der Pluralität des Produkt- bzw. Dienstangebots abhängig, also davon, ob ein Individuum zum Beispiel die Möglichkeit hat, ein anderes Produkt mit anderen Eigenschaften zu nutzen. Die Handlungsoptionen von Individuen in spezifischen Anwendungskontexten werden von den jeweiligen Autonomiegraden der KI-Systeme berührt. Unter Autonomiegraden ist das Ausmaß zu verstehen, mit dem ein KI-System bzw. ein robotisches System selbstständig, also autonom in einem Anwendungskontext agiert. Kernelement der Einteilung von Autonomiegraden ist oft die Bestimmung von verschiedenen Stufen je nachdem, wie viel Kontrolle der Mensch über das KI-System und dessen „Verhalten“ ausüben kann bzw. je nachdem, wie stark der

12 Der Vernetzungsgrad spielt allgemein bei technischen Systemen eine Rolle, insofern sind die Ausführungen nicht spezifisch für KI. Der Vernetzungsgrad ist für die Einschätzung der Kritikalität von KI-Systemen jedoch elementar.

Anteil des Menschen in einer Mensch-Maschine-Interaktion ist.[13] Hierbei ist darauf hinzuweisen, dass auch die Gestaltung der Mensch-Maschine-Interaktion an sich die Handlungsoptionen eines Individuums beinträchtigen kann (relevant sind hier z. B. die Nachvollziehbarkeit der Dialogführung und Beschränkung der Handlungsauswahl). Mit zunehmenden Autonomiegraden werden potenzielle Handlungsmöglichkeiten von Individuen an das KI-System übertragen. Dadurch stellen sich vermehrt Fragen nach der Absicherung des KI-Systems, nach der Nachvollziehbarkeit und Erklärbarkeit seiner Funktionen sowie nach den Regeln, wie mit Fehlern des Systems umgegangen wird.

Schlussfolgerungen zur Einschätzung der Kritikalität

Aus den bisherigen Ausführungen kann zunächst eine Beziehung zwischen den vorgestellten Kriterien dargelegt werden, aus der auf das Ausmaß der Kritikalität und den möglichen Regulierungsbedarf geschlossen werden kann. Je höher das Ausmaß der möglichen Gefährdung von Menschenleben und weiteren Rechtsgütern (in Abhängigkeit vom Vernetzungsgrad des Systems) eingeschätzt wird und je geringer der Umfang der Handlungsoptionen des Individuums (in Abhängigkeit von der Autonomie des Systems), desto höher wird die Einschätzung der Kritikalität ausfallen und desto eher wird ein hoher Regulierungsbedarf begründbar (und vice versa). Aus dieser Überlegung heraus spannt sich ein Kontinuum zwischen niedriger und hoher Kritikalität auf, sodass je nach Einschätzung verschiedene Kritikalitätsgrade zwischen den beiden Extrempunkten unterscheidbar werden.

Es ist jedoch zu betonen, dass bei der Einschätzung der Kritikalität nicht in jedem Fall nach einem schlichten „je weniger bzw. je mehr, desto höher"-Verhältnis gegeneinander aufgewogen werden sollte. Vielmehr wird die Einschätzung meist eine Abwägung darstellen. So könnte beispielsweise eine KI-Anwendung in einem spezifischen Anwendungskontext eine Gefährdung für Menschenleben darstellen und zugleich den Individuen einen großen Umfang an Eingriffsmöglichkeiten bieten. Zwar könnte das Individuum durch dessen Eingriffsmöglichkeiten möglicherweise eine Gefährdung von Menschenleben abwenden, dass jedoch grundsätzlich Menschenleben gefährdet werden, wiegt mit Blick auf die Kritikalitätseinschätzung schwer. Aus dem Autonomiegrad selbst ergibt sich zudem nicht zwingend eine proportional höhere Kritikalität und damit ein höherer Regulierungsbedarf. Dies zeigt das Beispiel eines Staubsaugerroboters. Dessen hohe Autonomie in einem privaten Haushalt kann durchaus unproblematisch sein, obwohl die Nutzenden möglicherweise über keine Eingriffsmöglichkeit verfügen, wenn sie außer Haus sind. Trotzdem kann ein solcher Roboter eine höhere Kritikalität aufweisen, wenn dieser etwa eigenständig Daten im privaten Haushalt sammelt und an die Hersteller oder sogar Dritte weitergibt, denn dies berührt den Datenschutz und die Privatsphäre. Ferner kann ein gut abgesichertes,

13 Einteilungen von Autonomiegraden für robotische Systeme im weiteren Sinn sind gegenwärtig für viele Domänen schon vorhanden (vgl. Robotik: Beer et al., 2014: S. 87; autonomes Fahren: SAE International 2018; Chirurgie: Haidegger 2019; Industrieproduktion: Plattform Industrie 4.0, 2019: S. 12 ff.; Gamer et al. 2019). Autonomiegrade sollten jedoch nicht lediglich als diskrete Stufenmodelle gedacht werden, sondern auch im Sinne variabler Autonomiegrade (vgl. „sliding autonomy"; vgl. Beyerer et al. i. E.). Mit Blick auf die Zertifizierung sollten in solchen Fällen spezifische Anforderungen berücksichtigt werden. So sollte es beispielsweise bei robotischen Systemen möglich sein, dass in bestimmten rechtlich oder ethisch unklaren bzw. problematischen Fällen das System automatisch in niedrigere Autonomiestufen überführt wird oder dass bestimmte Anforderungen an die Mensch-Maschine-Interaktion gestellt werden, damit der Mensch die Steuerung problemlos übernehmen kann (z. B. beim autonomen Fahren).

stark autonomes robotisches System unter Umständen zwar wenig Eingriffsmöglichkeiten durch Individuen zulassen, dafür aber sicherer bzw. schadensbegrenzender sein als ein weniger autonomes System im gleichen Anwendungskontext. Für die Einschätzung der Kritikalität ist also eine eingehende Reflexion des spezifischen Falls eines KI-Systems im jeweiligen Anwendungskontext notwendig. Die vorgestellten Konzepte bieten hierfür eine notwendige Grundlage.

In welchen Fällen soll eine Zertifizierung erfolgen?

Grundsätzlich wirft eine abgestufte Perspektive auf den Regulierungsbedarf nach Kritikalität die Frage auf, wer über die Befugnis verfügen sollte, eine Einschätzung der Kritikalität vorzunehmen. Wenn der Staat für bestimmte Anwendungskontexte eine verpflichtende Zertifizierung oder Zulassung einführen möchte, obliegt es ihm, den konkreten Anwendungskontext zu regeln. In diesem Fall fällt dem Staat die Kritikalitätseinschätzung zu bzw. der Staat entscheidet darüber, ob und an wen (z. B. Hersteller oder unabhängige Institutionen) diese Einschätzung delegiert wird. In allen Fällen, in denen die Zertifizierung auf freiwilliger Basis initiiert wird, sollte die Kritikalitätseinschätzung den Unternehmen oder den Entwickelnden zufallen. Gegebenenfalls müssten zur Unterstützung einer solchen Einschätzung adäquate Abgrenzungskriterien für einzelne Kritikalitätsstufen spezifisch für die jeweilige Anwendung eines KI-Systems entwickelt werden. Dies erfordert eingehende wissenschaftliche Studien anhand einer größeren Anzahl an Fällen, die über den Rahmen dieses Papiers hinausgehen. Die Zuteilung der Befugnis hinsichtlich der Zertifizierung sollte im Weiteren über eine staatliche Akkreditierungsstelle ablaufen.

Die Kritikalitätseinschätzung – Autonomie- und Vernetzungsgrade darin mitbedacht – kann genutzt werden, um eine flexible Antwort darauf zu finden, wann eine Zertifizierung von KI-Systemen notwendig sein sollte (siehe Tabelle 3). Eine solche abgestufte Vorgehensweise folgt dem grundlegenden Prinzip, dass bei zunehmender Kritikalität eines KI-Systems in einem spezifischen Anwendungskontext der Staat zunehmend als regulierende Instanz in Erscheinung treten könnte, wenn die Kriterien für entsprechende Kritikalitätslevel unzweifelhaft vorliegen. Eine solche Regulierung kann verschiedentlich zum Ausdruck kommen (nicht abschließende Liste).

Eine öffentliche Institution:

- spricht Verbote oder Einschränkungen aus;
- bestimmt die Kriterien, gegen die geprüft wird;
- bestimmt, ob eine Konformitätsbewertung durch Dritte erforderlich ist;
- gibt vor, ob eine Konformitätsbewertung oder eine Zulassung erforderlich ist.

Tabelle 3 bietet einen Überblick über Relationen zwischen Kritikalität, staatlicher Eingriffstiefe und verschiedene denkbare Formen der Prüfung von Kriterien und Regulierung (siehe Tabelle 3). Wir fokussieren im Folgenden auf die Zertifizierung, weitere Fälle in Tabelle 3 dienen der Illustration und der Verortung der Zertifizierung im Spektrum verschiedener Optionen. Lediglich bei einer Konformitätsprüfung durch eine unabhängige Drittstelle handelt es sich um eine Zertifizierung.

In vielen Fällen wird keine Zertifizierung notwendig sein, weil der Kritikalitätsgrad niedrig ist (beispielsweise bei Spamfiltern oder Plagiatserkennungssoftware). Hersteller können sich jedoch in solchen Fällen auf eigene Kriterien verpflichten (wie etwa Kodizes, Selbstverpflichtungen und Gütesiegel) oder mit anderen Unternehmen auf Kriterien einigen und die Einhaltung dieser Kriterien durch eine Drittstelle überprüfen lassen. Auch eine gegenseitige Überprüfung zwischen Unternehmen im Sinne einer Anwender-Anbieter-Überprüfung ist denkbar. Zudem könnten solche Selbstverpflichtungen oder auch Kodizes an unternehmensinterne Compliance-Prozesse gekoppelt werden, um die Einhaltung der Regeln zu überwachen. Konzepte für ethische Leitlinien und ihre Operationalisierung können hierbei berücksichtigt werden (vgl. AI Ethics Impact Group 2020). Für manche Produkte, wie zum Beispiel intelligente Assistenzsysteme, kann auch eine Orientierung an der freiwilligen Selbstkontrolle, wie sie im Medienbereich existiert, sinnvoll sein. Schließlich können Hersteller allerdings auch jederzeit freiwillig eine Zertifizierung anstoßen, auch wenn der Kritikalitätsgrad niedrig eingeschätzt wird.

Ist der Kritikalitätsgrad jedoch höher, sollte der Staat als regulierende Instanz auftreten. Dies kann flexibel erfolgen. So können Regelungen festgelegt werden, die bestimmen, ob in manchen Fällen bzw. Anwendungskontexten eine Selbstverpflichtung erforderlich wird oder ob eine Form von Ko-Regulierung durchgeführt werden soll, das heißt beispielsweise eine Verpflichtung der Hersteller, sich auf eigene Regeln zu einigen oder eine Anwender-Anbieter-Überprüfung durchzuführen. Reicht diese Selbstverpflichtung oder Ko-Regulierung aufgrund der Kritikalitätshöhe nicht aus, sollte eine Zertifizierung notwendig sein (Beispiel: Algorithmus zur Verteilung von Studienplätzen) oder bei noch höherer Kritikalität eine Zertifizierung durch staatliche Stellen (Beispiel: KI zur Ermittlung von Kreditwürdigkeit durch Akteure mit großer Marktmacht). Werden in Anbetracht der Kritikalitätshöhe tiefgreifendere Maßnahmen nötig, ist eine Zulassung erforderlich (siehe beispielsweise autonomes Fahren im Straßenverkehr). Bei sehr hoher Kritikalität könnten staatliche Verbote oder Einschränkungen hinsichtlich des Einsatzes von KI-Systemen ausgesprochen werden. Dies kann beispielsweise bei der biometrischen Fernidentifikation der Fall sein (vgl. Europäische Kommission 2020).

3.2 An welchen Gegenständen und Kriterien soll sich eine Zertifizierung von KI-Systemen orientieren?

Wurde die Notwendigkeit einer Zertifizierung festgestellt, so stellt sich die Frage, wie und anhand welcher Prüfkriterien KI-Systeme zertifiziert werden können. Die in diesem Abschnitt vorgestellten Vorschläge orientieren sich dabei ebenfalls stark an den Überlegungen zur Kritikalität von KI-Systemen in bestimmten Anwendungskontexten. Ziel ist es, zu skizzieren, wie eine sinnvolle und passende Zertifizierung aussehen kann – dafür ist es unerlässlich, dass die hierfür benötigten (finanziellen und zeitlichen) Ressourcen sich innerhalb eines angemessenen Rahmens bewegen. Unter den Überpunkten „Gegenstand der Zertifizierung“, „Zeitpunkt und Wiederholung der Zertifizierung“, „Kriterien der Zertifizierung“ sowie „Detailgrad und Prüftiefe“ werden erste mögliche Orientierungspunkte zur Zertifizierung von KI-Systemen vorgestellt.

Verhältnis zu Leitlinien, Standards und Regeln

Grundsätzlich sollte bei der Zertifizierung von KI-Systemen sowohl auf allgemeine als auch auf branchenspezifische Normen, Standards, Prüfverfahren und auch gesetzliche Vorschriften zurückgegriffen beziehungsweise an diese angeschlossen werden (ISO- und DIN-Normen, geltendes (EU-)Recht). Wo nötig, müssen diese unter Umständen auch neu interpretiert werden. Es sollte nicht zu einer Situation kommen, in der KI-spezifische Standards und Regularien mit weitergefassten Zertifizierungen und Regularien konkurrieren. Anknüpfungspunkte könnten beispielsweise folgende sein:

- gültige europäische Rechtsvorschriften beispielsweise in Bezug auf Grundrechte, Verbraucherschutz sowie Produktsicherheit und -haftung (in diesen werden allerdings KI-spezifische Risiken nicht ausreichend berücksichtigt),
- EU-Maschinenrichtlinie sowie ISO-Standards der Systemsicherheit,
- Kriterien der Medienselbstregulierung und
- bereits ausgearbeitete ethische Leitlinien für KI (AI Ethics Impact Group 2020; AI High Level Expert Group 2019; IAIS 2019; Heesen et al. 2020b).

Deshalb ist es mit Blick auf Zertifizierung und Regulierung lediglich notwendig, Lücken hinsichtlich KI-spezifischer Anforderungen zu schließen. Ob sich durch die technologische Entwicklung neue Lücken ergeben und Anpassungen beziehungsweise Ergänzungen zu den bestehenden Normen und Regeln notwendig sind, könnte durch ein staatlich eingesetztes Gremium, bestehend aus Expertinnen und Experten, in regelmäßigen Abständen überprüft werden. Entsprechend der Einschätzung des Gremiums könnten Prüfkataloge angepasst werden. Jenseits spezifischer Anforderungen an KI-Systeme ist es erforderlich, dass der Staat (hoch-)kritische Anwendungskontexte definiert. Es sollte festgelegt werden, wie diese Bereiche reguliert werden müssen und ob eine Zertifizierung notwendig werden soll.

Gegenstand der Zertifizierung (Was soll zertifiziert werden?)

Generell sollte kein Unterschied zwischen der Fokussierung einer Zertifizierung von KI-Systemen und der Fokussierung anderer Zertifizierungsverfahren gemacht werden. Lediglich in den Eigenschaften des untersuchten Systems und möglicherweise den Prüfkriterien sollten die Unterschiede liegen. Drei Fragen stellen sich hinsichtlich des Fokus einer KI-Zertifizierung im besonderen Maße: 1.) Was ist das Produkt (z. B. Auto)? 2.) Was ist das KI-System (z. B. Steuerungssystem)? 3. Was sollte zertifiziert werden? Wenn also ein Steuerungselement in einem Auto bisher immer unter eine Produktzertifizierung fiel, so soll auch ein KI-Steuerungselement in einem Auto durch eine Produktzertifizierung überprüft werden.

Grundsätzlich könnte das konkrete Produkt, der konkrete Prozess, aber auch das allgemeine Projektmanagement sowie allgemein die Entwickelnden oder Herstellenden zertifiziert werden. Die Autorinnen und Autoren empfehlen für den Bereich der KI-Systeme entweder eine **Produkt- oder eine Mischform aus Produkt- und Prozesszertifizierung**.

Eine **Produktzertifizierung** ist eine neutrale Überprüfung der Einhaltung zugesicherter Produkteigenschaften auf der physischen Produktebene. Es werden unter anderem folgende Fragen untersucht „Was macht das Produkt?“ und „Wie funktioniert das Produkt?“. Eine Produktzertifizierung sollte früh ansetzen (im Optimalfall bereits bei der Spezifikation des Produkts). Eine Eigenschaft der Produktzertifizierung ist, dass oft mehrere Verfahren kombiniert werden müssen, um die Anwendung abzubilden.

Eine Produktzertifizierung kann folgende Aspekte umfassen:

- Zertifizierung des Algorithmus[14] (z. B. Durchführen eines Realitätschecks, Prüfen der Unvoreingenommenheit im Sinne von Diskriminierungsfreiheit, Validierung des Modells [Cross-Validation], Optimierung der Zielfunktion[15] und der statistischen Verlustfunktion)
- Zertifizierung der Daten (z. B. Prüfen der Nachvollziehbarkeit, Vollständigkeit, Schutz vor unautorisiertem Zugriff, Sampling der Daten, Vermeidung von Verzerrungen und Prüfen der Repräsentativität der Daten)
- Zertifizierung der Spezifikation der Anwendung[16]
- Zertifizierung der Auswirkungen des Systems auf andere Technologien, auf Prozesse, auf Menschen, auf die Umwelt, auf die Gesellschaft (in Anlehnung an eine Technikfolgenabschätzung)
- Zertifizierung des Systemdesigns und der Entwicklungstools

Eine Alternative oder auch Ergänzung zu einer Produktzertifizierung kann eine **Prozesszertifizierung** sein. Sie untersucht die Qualität des Herstellungs- und Entwicklungs- sowie des Einführungsprozesses im Allgemeinen und der Implementation der KI-Lösung im Besonderen. Sie dient der Reflexion der zu prüfenden Prozesse und kann unter Umständen auch durch den Hersteller oder den Betreiber selbst vorgenommen werden. Dabei werden beispielsweise folgende Fragen gestellt: Wie wurde/ist die KI in das System integriert? Wurden/werden bestimmte Sicherheitsstandards eingehalten? Welche Prozessmodelle wurden zur Entwicklung angewendet (z. B. Wasserfall oder iterative und agile Modelle)? Wie wurden zukünftige Nutzende und andere Stakeholder des KI-Systems in Gestaltungs-, Entwicklungs- und Entscheidungsprozesse einbezogen? Welche Qualitäts- und Prüfkriterien leiten und beeinflussen die Gestaltungs- und Entscheidungsprozesse während der Entwicklung?

Bei der Entwicklung interaktiver technischer Systeme haben sich Vorgehensweisen der menschenzentrierten Gestaltung (human-centered Design, nach ISO 9241-210) in den letzten Jahren in der Industrie immer stärker etabliert. Eine Prozesszertifizierung kann hier überprüfen, ob menschenzentrierte Gestaltungs- und Entwicklungsprozesse durchgeführt worden sind und eingehalten werden. Auch bei der Implementation von KI-Systemen in

14 Der Algorithmus ist bewusst als Erstes genannt: Eine Möglichkeit wäre, zu Beginn bestimmte Verfahren oder auch Klassen von Algorithmen zu zertifizieren und erst wenn dort ein Rahmen gegeben ist, das gesamte Produkt zu zertifizieren.

15 Ein Beispiel hierfür ist der Facebook-Newsfeed: Die Zielfunktion optimiert auf eine hohe Interaktion. Das führt dazu, dass Posts mit negativen Emotionen als relevanter klassifiziert werden. Problematisch kann hierbei aber sein, dass die Zielfunktion teilweise erst im Betrieb gelernt werden muss. In der KI-Forschung ist darüber hinaus umstritten, inwieweit das System im Ziel „eingebettet“ sein darf. Statt einer Zielfunktion könnte sich ein System an einer Nutzenfunktion und absoluten Regeln orientieren.

16 Dies wird teilweise in der IT-Sicherheit ähnlich gehandhabt (vgl. hierzu Common Criteria). Einzelne KI-Komponenten sollten auch eine Spezifikation mitbringen, innerhalb der sie in einem größeren System eingesetzt werden.

Unternehmen kann eine Prozesszertifizierung Kriterien hinsichtlich der begleitenden Beteiligung und Gestaltung berücksichtigen (vgl. Stowasser & Suchy, i. E.).

In Ergänzung dazu funktioniert eine Prozesszertifizierung im Sinne eines Sicherheitsmanagements und ist vor allem bei einer nachträglichen Betrachtung bei einer möglichen Fehlfunktion des Systems wichtig. Wenn ein zertifiziertes Verfahren mit speziell auf KI zugeschnittenen Instrumenten angewandt wird, kann eine Prozesszertifizierung wichtige Implikationen für Fragen der Verantwortung und Haftung geben. Gleichzeitig können gut durchgeführte Prozesse auch möglichen Fehlfunktionen vorbeugen und so zu besseren Produkten führen.

Eine Prozesszertifizierung umfasst folgende Aspekte:
- Prozesse im Rahmen der Produktherstellung und -entwicklung
- Prozesse zur organisatorischen Einführung und Transformation im Zusammenhang mit KI-Anwendungen in Betrieben
- Prozesse zum Betrieb und zur kontinuierlichen Wartung und Weiterentwicklung/ Optimierung von KI-Systemen
- Prozesse im Zusammenhang mit der Nutzung von KI-Systemen

Neben dem Produkt und dem Prozess könnte auch das allgemeine **Projektmanagement** sowie allgemein die **Entwickelnden oder Hersteller** zertifiziert werden. Die Zertifizierung von Entwickelnden wird allerdings kritisch gesehen. Sie könnte möglicherweise über gesonderte, speziell dafür zu entwickelnde Lehrgänge für Software-Entwickelnde umgesetzt werden, an deren Ende ein Zertifikat steht, welches bescheinigt, dass die Risiken bekannt sind sowie welche Abläufe in der Entwicklung wie abzulaufen haben. Dies schließt auch an die ISO 9001-Norm an, welche vorschreibt, dass Produzenten einer Ware sicherstellen müssen, dass ihr Personal hinreichend qualifiziert ist. Es ist jedoch umstritten, ob der finanzielle Aufwand mit dem erreichbaren Ergebnis in einem sinnvollen Verhältnis steht. Gleiches gilt für die Zertifizierung von Herstellern, diese ist aus Praktikabilitätsgründen oft nicht möglich. Folglich sollten Hersteller gegebenenfalls bewertet, aber nicht zertifiziert werden.

Prüfkriterien der Zertifizierung

Die Einhaltung des definierten rechtlichen Rahmens ist die Basisannahme für die Entwicklung und Anwendung von KI-Systemen. Eine Zertifizierung baut auf rechtlichen Kriterien auf und kann auch weitergehende Kriterien abprüfen. Alle zu überprüfenden Kriterien beziehen sich auf KI in sozio-technischen Systemen.[17] Deshalb werden sowohl Kriterien überprüft, die sich unmittelbar auf die Sachtechnik, als auch Kriterien, die sich anwendungsbezogen aus der Mensch-Maschine-Interaktion (MMI) ableiten. Da die verschiedenen Aspekte miteinander verbunden sind und aufeinander einwirken, ist eine strikte Binnendifferenzierung nicht möglich. Die zu überprüfenden Kriterien lassen sich in zwei Kategorien hinsichtlich ihrer Verbindlichkeit im Rahmen einer Zertifizierung unterteilen.

17 Ein soziotechnisches System beschreibt ein System, in welchem Beziehungen und Wechselwirkungen zwischen einem sozialen und einem technischen System bestehen und damit selbst wieder einen eigenen Handlungskontext generieren, in dem der Einfluss technischer und menschlicher Akteure kaum noch zu unterscheiden ist (Mensch-Technik-Ensembles).

Die Basis bilden **Mindestkriterien**, die geprüft werden müssen. Diese Mindestkriterien sollten bei jedem System in seinem Anwendungskontext überprüft werden. Abbildung 1 gibt einen Überblick über diese Mindestkriterien.

Abbildung 1: Mindestkriterien, die im Rahmen einer Zertifizierung überprüft werden müssen

Mindestkriterien

- Transparenz, Nachvollziehbarkeit, Nachprüfbarkeit und Verantwortlichkeit
- Funktionale Sicherheit/Safety/inkl. Produktsicherheit und Zuverlässigkeit
- Vermeidung von nicht-intendierten Folgewirkungen (auf andere Systeme, Menschen und die Umwelt)
- Gerechtigkeit im Sinne von Gleichheit und Diskriminierungsfreiheit
- Schutz der Privatheit und der Persönlichkeit
- Selbstbestimmung inkl. Transparenz über den Einsatz des KI-Systems und die Rolle des Menschen im Entscheidungsprozess

Anmerkung: Siehe Tabelle 1 und Tabelle 2 sowie den Anhang für eine detailliertere Darstellung.

In Ergänzung dazu können **darüber hinausgehende Kriterien** abgeprüft werden. Diese ergänzenden Kriterien stellen eine Art „Zertifizierung plus" oder eine „Zertifizierung nach dem Goldstandard" dar. Sie sind von großer Bedeutung für eine positive und wertorientierte Entwicklung von KI und gehen über die Mindestanforderungen, die der Verhinderung von evidenten und unmittelbaren Gefährdungen dienen, hinaus. Abbildung 2 stellt die freiwilligen, darüber hinausgehenden Kriterien vor. Dazu zählen insbesondere ökologische Nachhaltigkeit, offene Schnittstellen und Systemoperabilität sowie Nutzerfreundlichkeit (Usability).

Abbildung 2: Darüber hinausgehende Kriterien, die im Rahmen einer Zertifizierung überprüft werden können

Darüber hinausgehende Kriterien

- Offene Schnittstellen und Systemoperabilität
- Menschenzentrierung und Nutzerfreundlichkeit (Usability) inkl. Partizipation, Schutz des Einzelnen, sinnvolle Arbeitsteilung und förderliche Arbeitsbedingungen
- Nachhaltigkeit
- Kennzeichnung und Begrenzung der Systemfunktionalität

Anmerkung: Siehe Tabelle 1 und Tabelle 2 sowie den Anhang für eine detailliertere Darstellung.

Wie im vorherigen Abschnitt dargelegt, unterscheiden sich Produkt- und Prozesszertifizierung sowohl hinsichtlich ihres Ziels als auch hinsichtlich des Betrachtungsgegenstandes. Folglich können manche Prüfkriterien besser im Rahmen einer Produktzertifizierung abgefragt werden, während andere wiederum besser innerhalb von einer Prozesszertifizierung umgesetzt werden können. Tabelle 1 und 2 geben einen Überblick, anhand welcher Fragen die vorgestellten Prüfkriterien im Rahmen einer Produkt- und einer Prozesszertifizierung abgefragt werden können. Die Fragen sind auf Grundlage verschiedener Vorarbeiten, wie beispielsweise dem Leitfaden Ethik-Briefing der Plattform Lernende Systeme (vgl. Heesen et al. 2020b), den Kriterien für die Mensch-Maschine-Interaktion bei KI der Plattform Lernende Systeme (vgl. Huchler et al. 2020), den Empfehlungen der AI High Level Expert Group (vgl. 2019), einem Whitepaper zu Diskriminierung und KI-Systemen der Plattform Lernende Systeme (vgl. Beck et al. 2019) und der ISO-Norm 9241-210 sowie eigener Ergänzungen formuliert worden.

Tabelle 1: Kriterien für die Produktzertifizierung

Mindestkriterien
Transparenz, Nachvollziehbarkeit, Nachprüfbarkeit und Verantwortlichkeit
✓ Ist das System transparent?
✓ Ist eine klare Zuordnung von Verantwortlichkeiten möglich?
✓ Sind die Entscheidungen des Systems nachvollzieh- und nachprüfbar (Datengrundlage und Entscheidungsweg)? Ist das System als erklärbare KI gestaltet (explainable AI)?
Funktionale Sicherheit/Safety/inkl. Produktsicherheit und Zuverlässigkeit
✓ Ist die Zuverlässigkeit/die Genauigkeit des Algorithmus (Validitätswert) transparent kenntlich gemacht?
✓ Funktioniert das Produkt wie beabsichtigt? Ist das System so robust, dass es zu keinem Zeitpunkt ein unannehmbares Sicherheitsrisiko darstellt? Beinhaltet das System systeminterne Kontrollmaßnahmen, die eine Fehlfunktion verhindern?
✓ Ist das Produkt „sicher" (im Sinne von IT-Sicherheit)?
✓ Sind in dem System Verfahren zum sicheren Abbrechen eines KI-gesteuerten Vorgangs implementiert?
Vermeidung von nicht-intendierten Folgewirkungen auf andere Systeme, Menschen und die Umwelt
✓ Werden mögliche (unbeabsichtigte) negative Folgewirkungen auf andere Systeme, Menschen (z. B. in Bezug auf Gesundheit, Arbeit) und die Umwelt verhindert bzw. möglichst minimiert?
✓ Beinhaltet das System Mechanismen zur Vermeidung längerfristiger negativer Folgewirkungen (gesellschaftliche Legitimität)?[18]
Gerechtigkeit i. S. v. Gleichheit und Diskriminierungsfreiheit
✓ Handelt das System gerecht und diskriminierungsfrei?
✓ Beinhaltet das System keine als diskriminierend bewerteten Merkmale und Klassifizierungen in den Inputdaten?
✓ Werden alle Sorgfaltspflichten in Hinsicht auf mögliche statistische Diskriminierungen erfüllt? Ist die Vorhersage „well calibrated"?[19]
Schutz der Privatheit und der Persönlichkeit: Informationelle Selbstbestimmung, Datenschutz, Datenqualität und Datensicherheit
✓ Werden Daten möglichst sparsam und zweckgebunden erhoben und verarbeitet?
✓ Wird das KI-System möglichst mit anonymisierten oder pseudonymisierten Datensätzen trainiert?
✓ Wird für die Erhebung und Verarbeitung eine Einverständniserklärung eingeholt?
✓ Herrscht Transparenz über die Art der Erhebung, Auswertung und Verwendung von Daten?
Selbstbestimmung inkl. Transparenz über den Einsatz des KI-Systems und die Rolle des Menschen im Entscheidungsprozess
✓ Ist dafür gesorgt, dass die oder der Nutzende weiß, dass sie oder er von einem KI-System betroffen ist/ein KI-System nutzt?
✓ Herrscht Transparenz über die Art der Erhebung, Auswertung und Verwendung von Daten?
✓ Herrscht Transparenz über die Handlungsmöglichkeiten (bin ich passiv oder habe ich eine Handlungsoption?)?

18 Ein Beispiel hierfür wären Policy-Enforcement-Strukturen bei robotischen Systemen (siehe Plattform Lernende Systeme 2019; Beyerer et al. i. E.).

19 „Well calibrated" bedeutet: Wenn ein Algorithmus vorhersagt, dass eine bestimmte Eigenschaft mit einer bestimmten Wahrscheinlichkeit auf eine Gruppe zutrifft, sollte ein der Wahrscheinlichkeit entsprechender Anteil der Gruppe auch diese Eigenschaft haben (siehe Beck et al. 2019).

Darüber hinausgehende Kriterien
Offene Schnittstellen und Systemoperabilität
✓ Weist das System offene Schnittstellen auf?
✓ Ist das System interoperabel konzipiert?
Menschenzentrierung und Nutzerfreundlichkeit (Usability), inkl. Partizipation, Schutz des Einzelnen, sinnvolle Arbeitsteilung und förderliche Arbeitsbedingungen
✓ Bekommt der Mensch alle relevanten Informationen, um das System übernehmen zu können?
✓ Vermeidet das System Risiken für die physische und psychische Gesundheit der Menschen?
✓ Entlastet das KI-basierte Assistenzsystem nachhaltig und unterstützt die Anwendenden?
✓ Ist das System lernförderlich und lässt Handlungsräume zu?
✓ Kann sich das System flexibel und situationsspezifisch an den Bedarfen und Bedürfnissen der Arbeitspraxis der Nutzenden ausrichten?
✓ Ermöglicht das System wechselseitiges Lernen von der Maschine zum Menschen und umgekehrt?
Nachhaltigkeit
✓ Trägt das System zur Erreichung der UN-Nachhaltigkeitsziele bei?
✓ Sind das System und seine Teilkomponenten ressourcenschonend konzipiert?
Kennzeichnung und Begrenzung der Systemfunktionalität
✓ Sind die Funktionalitätsgrenzen des KI-Systems gekennzeichnet?
✓ Kann das System angewiesen werden, sich in bestimmten Situationen menschliche Unterstützung hinzuzuholen (beispielsweise in rechtlich oder ethisch schwierigen Situationen)?

Tabelle 2: Kriterien für die Prozesszertifizierung

Mindestkriterien	
Transparenz, Nachvollziehbarkeit, Nachprüfbarkeit und Verantwortlichkeit	
✓	Ist der Entwicklungs- und Anwendungsprozess transparent?
✓	Gibt es ein Verfahren, um die Nachvollziehbarkeit sicherzustellen?
✓	Ist der Entwicklungsvorgang des KI-Systems lückenlos begleitet und überwacht worden?
Funktionale Sicherheit/Safety/inkl. Produktsicherheit und Zuverlässigkeit	
✓	Wurden im Prozess Maßnahmen zur Erreichung der funktionalen Sicherheit unternommen?
Vermeidung von nicht-intendierten Folgewirkungen auf andere Systeme, Menschen und die Umwelt	
✓	Beinhaltet der Entwicklungsprozess systematisch eine Abschätzung, welche (unbeabsichtigten) negativen Folgewirkungen auf andere Systeme, Menschen (z. B. Gesundheit, Arbeit) und die Umwelt auftreten können und wurde dies bei der Entwicklung antizipiert?
✓	Beinhaltet der Entwicklungsprozess Mechanismen zur Technikfolgenbewertung und zur Vermeidung längerfristiger negativer Folgewirkungen (gesellschaftliche Legitimität)?
Gerechtigkeit i. S. v. Gleichheit und Diskriminierungsfreiheit	
✓	Existieren unternehmensinterne Prozesse und Kontrollmechanismen, um Diskriminierungen durch KI-Systeme zu vermeiden?
Schutz der Privatheit und der Persönlichkeit: Informationelle Selbstbestimmung, Datenschutz, Datenqualität und Datensicherheit	
✓	Existieren unternehmensinterne Prozesse und Kontrollmechanismen, um den Schutz der Privatheit und der Persönlichkeit sicherzustellen?
Selbstbestimmung inkl. Transparenz über den Einsatz des KI-Systems und die Rolle des Menschen im Entscheidungsprozess	
✓	Existieren unternehmensinterne Prozesse und Kontrollmechanismen, um das nötige Bewusstsein für die Wahrung der Selbstbestimmung sicherzustellen?
Darüber hinausgehende Kriterien	
Offene Schnittstellen und Systemoperabilität	
✓	Wurde standardmäßig im Entwicklungsprozess eine interne Review in Bezug auf offene Schnittstellen und eine interoperable Konzeption des Systems gemacht?
Menschenzentrierung und Nutzerfreundlichkeit (Usability), inkl. Partizipation, Schutz des Einzelnen, sinnvolle Arbeitsteilung und förderliche Arbeitsbedingungen	
✓	Werden Nutzende in allen Phasen der Gestaltung und Entwicklung aktiv einbezogen?
✓	Wird das Systemdesign durch Evaluationen mit Nutzenden vorangetrieben und verfeinert?
✓	Ist der Design-/Entwicklungsprozess iterativ?
✓	Umfasst das Designteam multidisziplinäre Kompetenzen und Perspektiven?
✓	Wurden in der Entwicklung die Erfahrungen von Testpersonen berücksichtigt?
✓	Existieren Prozesse, um eine gelungene Mensch-Maschine sicherzustellen?
Nachhaltigkeit	
✓	Existieren unternehmensinterne Prozesse und Kontrollmechanismen, um sicherzustellen, dass das Produkt zur Erreichung der UN-Nachhaltigkeitsziele beiträgt?
Kennzeichnung und Begrenzung der Systemfunktionalität	
✓	Existieren unternehmensinterne Prozesse, um zu überprüfen, ob die Entscheidungen des KI-Systems in bestimmten Fällen begrenzt werden können?

3.3 Wann und wie detailliert sollte eine Zertifizierung von KI-Systemen erfolgen?

Zeitpunkt und Wiederholung der Zertifizierung

Um bereits bei der Produktentwicklung die von einer Zertifizierung vorgeschriebenen Kriterien einhalten zu können, könnte eine ergänzende Vorab-Zertifizierung bereits bei der Spezifikation ansetzen. Die erste Produktzertifizierung soll sowohl bei ausgelernten als auch bei weiterlernenden KI-Systemen analog zu anderen technischen Systemen zum Auslieferungszeitpunkt erfolgen. Um die Vorabzertifizierung und die Zertifizierung bei Inbetriebnahme leichter aufeinander abzustimmen, sollen zwei Möglichkeiten zur Verfügung stehen: 1.) Sind Teilkomponenten eines KI-Systems im Rahmen der Vorabzertifizierung bereits zertifiziert, soll lediglich eine Zertifizierung mit Blick auf das Zusammenspiel der Komponenten durchgeführt werden. 2.) Im Rahmen der Vorabzertifizierung soll der Prozess im Sinne einer entwicklungsbegleitenden Prüfung mitzertifiziert werden – an Produkte, die eine bestimmte Prozesszertifizierung durchlaufen haben, sollen bei der folgenden Zertifizierung niedrigere Hürden angelegt werden.

Wie bereits ausgeführt, weisen KI-Systeme eine besondere Dynamik auf und insbesondere weiterlernende Systeme entwickeln sich im Betrieb weiter (siehe Kapitel 1). Deshalb soll bei weiterlernenden Systemen die Zertifizierung bei Inbetriebnahme wiederholt werden – dies trifft vorrangig auf die Anwendung in besonders kritischen Bereichen zu. Da sich weiterlernende Systeme auch nach Inbetriebnahme weiterentwickeln können, sollen diese anschließend turnusmäßig rezertifiziert werden (ähnlich einer Hauptuntersuchung eines Autos, die regelmäßig wiederholt werden muss). Die Abstände zwischen den Rezertifizierungen sollen verlängert werden können, wenn keine Ungereimtheiten in der Funktionsweise des Systems auftreten (vgl. Müller-Quade et al. 2020) oder wenn nachgewiesen werden kann, dass das System durch den laufenden Lernprozess nicht unsicherer werden kann.

Neben weiterlernenden Systemen gibt es auch KI-Systeme mit strukturellen Updates – hier ist der Zeitpunkt, zu dem eine Wiederholung der Zertifizierung durchgeführt werden soll, klarer zu ermitteln. Muss ein System zurückgerufen werden, soll eine Rezertifizierung ebenfalls verpflichtend sein (vgl. Müller-Quade et al. 2020). Eine Rezertifizierung könnte mithilfe eines Freeze-Verfahrens durchgeführt werden. Hierbei wird eine alte Version des Systems „konserviert" und gegen eine aktuelle Version abgeglichen. So kann Erkenntnis darüber erlangt werden, ob und wie sich das System verändert hat.

Bei Updates oder Weiterentwicklung eines (Teil-)Systems soll vermieden werden, dass das gesamte System in seiner Anwendung neu zertifiziert werden muss. Hier soll von der Aufgabe aus gedacht werden, die das KI-System lösen können muss. Deshalb kommt eine Teilzertifizierung oder ein modulares System in Frage. So wäre bei einem Austausch bestimmter Komponenten denkbar, dass nur diese Komponenten neu zertifiziert werden (z. B. eine Innovation bei einem Spurhalteassistenten, um das Auto in der Kurve zu halten,

oder bei einer Innovation beim Computer-Vision-System des Fahrzeugs zur Fußgängererkennung). Diese Vorgehensweise erfordert gemeinsame Überlegungen mit den Herstellern hinsichtlich des Aufbaus von Systemen und der Identifikation von Modulen, die einzeln erneuert und gegebenenfalls auch einzeln überprüft werden können. Gleichzeitig ist es wichtig, im Blick zu halten, wie sich der Austausch von Komponenten auf das Gesamtsystem auswirken könnte, da unter Umständen nicht-intendierte Folgen im Zusammenspiel mit anderen Komponenten auftreten könnten.

Eine Prozesszertifizierung wäre eine Möglichkeit, KI-Systeme unabhängig von möglichen späteren Veränderungen zu zertifizieren, da der Prozess von den späteren Veränderungen des Systems unabhängig ist.

Detailgrad und Prüftiefe der Zertifizierung

Auch der Detailgrad und die Prüftiefe sollen sich an dem Kritikalitätslevel eines KI-Systems in einem Anwendungsgebiet orientieren. Je höher die Kritikalität im Anwendungskontext eingeschätzt wird, desto umfangreicher soll der Detailgrad und die Prüftiefe der Zertifizierung ausfallen. Bei der Einschätzung des Umfangs der Zertifizierung soll jedoch berücksichtigt werden, ob es sich um eine Neuentwicklung handelt oder ob ein bestehendes System rezertifiziert wird. Während bei einer Neuentwicklung umfassender getestet werden soll, ist dies bei einer Rezertifizierung nicht im gleichen Maße durchzuführen.

Die Minimalanforderungen bei KI-Systemen in Anwendungsgebieten mit einer eher niedrigeren Kritikalität lassen sich folgendermaßen definieren: Es sollen grundsätzlich sowohl einzelne Komponenten als auch das Gesamtsystem beziehungsweise Produkt zertifiziert werden, weil das Zusammenspiel der Komponenten zu nicht-intendierten Effekten führen kann, beziehungsweise Emergenzphänomene auftreten können. Hierzu gehört auch, dass einzelne „sichere" Komponenten nur in den dafür vorgesehenen Spezifikationen eingesetzt werden sollen. Bei der Betrachtung des Gesamtsystems muss mindestens das Ergebnis als solches geprüft werden. Ebenso müssen auch bei geringen Anforderungen Dokumentations- und Prüfpflichten vorgeschrieben und eingehalten werden (hinsichtlich Sicherheit, Datenqualität, Usability, Systemarchitektur, vgl. ISO 9000).

Bei KI-Systemen in kritikalitätsarmen Anwendungskontexten kann eine Dokumentation ausreichend sein, während bei höherer Kritikalität und Autonomie eigenständige Tests durchgeführt werden müssen. Bei einer Rezertifizierung von beispielsweise weiterlernenden Systemen können sowohl die Prüftiefe als auch der Detailgrad geringer ausfallen. Hier könnten möglicherweise auch Benchmarks, die von Herstellern bereitgestellt werden, ausreichen (auch Unit-Tests sind eine Option, also die Überprüfung von Komponenten hinsichtlich ihrer korrekten Funktionalität).

Wie sollte auf technologische Weiterentwicklungen reagiert werden?

Die bestehenden Zertifizierungssysteme sind häufig eher träge und schwerfällig. Dies kann unter Umständen in einigen Fällen dazu führen, dass bereits bei konventionellen IT-Systemen Weiterentwicklungen teilweise ausbleiben, weil die damit verbundenen Zertifizierungen zu teuer und aufwendig sind. Hinzu kommt, dass es zu lange dauert, bis Standards verbindlich werden. Diese Probleme gelten in verschärfter Weise für moderne KI-Systeme. Da die technologische Entwicklung im KI-Bereich einer großen Dynamik unterliegt (siehe Kapitel 1), spielen Weiterentwicklungen noch einmal eine größere Rolle als bei konventionellen Systemen. Zudem treten Veränderungen nicht nur bei Updates und Systemänderungen auf. Sie entstehen bei weiterlernenden KI-Systemen auch durch eben diese Lernprozesse des Systems.

Ziel ist deshalb die Etablierung einer Agilitätskomponente im Standardisierungswesen für KI-Systeme. Das bedeutet, dass eine qualitativ hochwertige und sinnvolle Zertifizierung in diesem Bereich ein offener Prozess sein soll, der auf technologische Weiterentwicklungen reagieren kann. Ein gutes Zertifikat muss auch bei Veränderungen gelten, also seine Gültigkeit auch unabhängig von der Datensituation und dem technologischen Fortschritt bewahren. Diese Agilitätskomponente soll eine angemessene Abwägung zwischen drei Anliegen darstellen: 1.) KI marktfähig zu halten, 2.) notwendige Rezertifizierungen durchzuführen sowie 3.) den richtigen Detailgrad zu wählen.

Ein agiles Zertifizierungswesen und Standardisierungswesen im Allgemeinen und für KI-Systeme im Besonderen zeichnet sich durch folgende Aspekte aus:

- Es existieren Feedback-Mechanismen, die mit Weiterentwicklung umgehen können – dazu gehört auch, dass die angelegten Kriterien regelmäßig an den technologischen Fortschritt angepasst werden müssen.
- Die Zertifizierungsstellen sind dynamisch verfasst. So können sie auf neue technologische Entwicklungen oder Fehlnutzungen reagieren.
- Die Prüfverfahren und -prozesse werden so dokumentiert, dass über die Zeit Erfahrungswerte gesammelt und neue Tendenzen und Entwicklungen frühzeitig identifiziert werden können.

Anwendungsbeispiel: Zertifizierung von Trainingsdaten für ein KI-System
Eine Zertifizierung von KI-Systemen sollte das gesamte System in die Prüfung einbeziehen. Dies kann, wie in Kapitel 3.2 bereits angesprochen, im Rahmen einer Produktzertifizierung auch eine Zertifizierung von Trainingsdaten umfassen. Dies ist naheliegend, da die Trainingsdaten gemeinsam mit dem Algorithmus die Grundlage für das spätere Produkt bilden.[20]

Wie die Zertifizierung abläuft, ist abhängig von der Kritikalität des KI-Systems, das mit den Trainingsdaten trainiert wird, in seinem späteren Einsatzgebiet. Liegt eine niedrige Kritikalität vor, kann die Zertifizierung der Trainingsdaten freiwillig durch die Hersteller angestoßen werden. Die freiwillige Zertifizierung wird von einer unabhängigen akkreditierten Prüfstelle durchgeführt. Weist das KI-System, das mit den Trainingsdaten trainiert wird, in einem bestimmten Anwendungskontext eine einschlägige Kritikalität auf, ist eine Zertifizierung angezeigt. Diese vorgeschriebene Zertifizierung wird dann von einer staatlich legitimierten dritten Stelle oder einer staatlichen Behörde durchgeführt.

Die Schwellen dieses Bereiches sollten von offizieller Seite definiert werden. Die Einordnung der Kritikalität sollte im Rahmen der Produktentwicklung erfolgen. Die gegebenenfalls notwendige Zertifizierung der Trainingsdaten sollte im Idealfall bereits im Rahmen der Produktentwicklung durchgeführt werden, spätestens aber, bevor das Produkt in Verkehr gebracht wird. Eine Zertifizierung der Trainingsdaten würde anhand der Daten, mit denen das KI-System trainiert wird, und gegebenenfalls anhand der Dokumentation der Datensätze erfolgen, die etwa auch die Herkunft der Trainingsdatensätze transparent darlegt. Diese Dokumentation sollte von den Unternehmen vorgehalten werden, da sie zur Überprüfung der Datensätze oder auch im Rahmen einer späteren Rezertifizierung (beispielsweise über ein Freeze-Verfahren) verwendet werden kann.

Von den Prüfkriterien können anhand der Trainingsdaten folgende Mindestanforderungen überprüft werden: 1.) Transparenz, Nachvollziehbarkeit und Nachprüfbarkeit, 2.) Gerechtigkeit im Sinne von Gleichheit und Diskriminierungsfreiheit und 3.) Schutz der Privatheit und der Persönlichkeit: Informationelle Selbstbestimmung, Datenschutz, Datenqualität und Datensicherheit. Hierbei sind konkret folgende Fragen zu stellen:

- Ist die Datengrundlage nachvollziehbar?
- Ist die Datengrundlage valide?
- Beinhaltet das System keine als diskriminierend bewerteten Daten oder Klassifizierungen in den Inputdaten?
- Werden Daten möglichst sparsam und zweckgebunden erhoben und verarbeitet?
- Wird das KI-System möglichst mit anonymisierten oder pseudonymisierten Datensätzen trainiert?

20 Die Trainingsdaten können jedoch nicht als alleiniger Ansatzpunkt für eine Zertifizierung von KI-Systemen genutzt werden. Zum einen, weil auch auf der Basis eines Datensatzes mit hoher Qualität defizitäre KI-Systeme erstellt werden können und zum anderen, weil durch eine geeignete Datenaufbereitung auch mit einem Datensatz minderer Qualität akzeptable Systeme umgesetzt werden können. Dennoch können bestimmte Eigenschaften der Datensätze zur Unterstützung der Entwickler bescheinigt werden.

3.4 Wie sollte die Infrastruktur der Konformitätsbewertung von KI-Systemen aussehen?

Wie im Kapitel 3.1 bereits erläutert, kann die Notwendigkeit einer Zertifizierung über eine Einschätzung der Kritikalität bestimmt werden. Eine Zertifizierung sollte nur dann notwendig sein, wenn eine höhere Kritikalität des KI-Systems in einem spezifischen Anwendungskontext festgestellt wurde. Eine freiwillige Zertifizierung kann allerdings fortwährend durch die Hersteller angestoßen werden. Für die Umsetzung einer Zertifizierung sind allerdings nicht nur die Einschätzung der Kritikalität sowie ein adäquater Prüfkatalog nötig, sondern eine gewisse organisatorische und technische Infrastruktur. Hinsichtlich dieser Infrastruktur ist es sinnvoll, an bereits bestehende Strukturen in den jeweiligen Branchen anzuschließen. KI-Systeme stellen die Prüfinfrastruktur vor Herausforderungen aufgrund der Dynamik der Systeme und der KI-Forschung selbst, aber auch weil in einigen Fällen nicht nur bestimmte Systemeigenschaften geprüft werden sollten, sondern auch das Systemverhalten in spezifischen Testszenarien. Um eine nachhaltige Zertifizierung von hoher Güte umzusetzen, ist daher auch die Kooperation unterschiedlicher Akteure sinnvoll. Im Folgenden wird dargelegt, auf welche bestehenden Strukturen für die Konformitätsbewertung von KI-Systemen zurückgegriffen werden könnte, ob ein Ausbau bzw. Ergänzungen notwendig sind und welche Optionen sinnvoll sein könnten.

Allgemeine Einschätzung zur organisatorischen und technischen Infrastruktur

In mancherlei Hinsicht wird ein Ausbau bestehender Strukturen notwendig sein oder es kann sinnvoll sein, neue Einrichtungen zu gründen. Verkäufer und Vermarkter, aber auch etablierte Prüfstellen und Behörden können derzeit eine Konformitätsbewertung nicht leisten, weil es größtenteils an Standards fehlt, gegen die eine solche Bewertung durchgeführt werden könnte. Für die Durchführung von Konformitätsbewertungen wird zudem eine hohe Expertise sowie ein ausgeprägtes Wissen hinsichtlich des gegenwärtigen Forschungsstandes notwendig sein. Diese Expertise ist mitunter an den bestehenden Stellen noch nicht ausreichend vorhanden und sollte entsprechend ausgebaut werden. Alternativ könnten auch neue Prüfstellen aufgebaut werden, deren fester Bestandteil auch andere Partner sind, wie etwa Forschungseinrichtungen. Darüber hinaus fehlt es häufig an Prüfwerkzeugen und unterstützender Software, um die Prüfung spezifischer Kriterien einer Konformitätsbewertung umsetzen zu können. Es ist weiterhin davon auszugehen, dass verschiedene Formen der Konformitätsbewertungen und Instanzen der Konformitätsprüfung durch unabhängige Dritte notwendig werden. Dies kann auch von Branche zu Branche variieren, da beispielsweise in international ausgerichteten Branchen internationale Einrichtungen eine größere Rolle spielen (z. B. sind in maritimen Branchen die Versicherer und Klassifikationsgesellschaften gewichtige Akteure). Es sollten jedoch klare Kriterien dafür entwickelt werden, wer zertifizieren darf und welche Maßstäbe für eine Zertifizierung herangezogen werden. Zertifizierungsstellen sollten daher durch die Deutsche Akkreditierungsstelle (DAkkS) akkreditiert sein sowie für bestimmte hoheitliche Bereiche auch andere Stellen, wie etwa das BSI bei der IT-Sicherheit.

Technische Infrastruktur

Damit die Konformitätsbewertung von KI-Systemen gelingt, sind die technischen Voraussetzungen mit Blick auf Prüfwerkzeuge, Software und Testumgebungen zu erfüllen. So sind etwa für die Validierung spezifischer Eigenschaften von KI-Systemen entsprechende Prüfwerkzeuge notwendig. Ein Beispiel macht dies besonders deutlich: Soll etwa der Programmcode einer KI-Anwendung manuell überprüft werden, ist dies bei einigen Terabyte Daten Code unrealistisch. Hierfür werden passende Prüfwerkzeuge und unterstützende Software benötigt. Da bei einer Zertifizierung unter Umständen nicht nur (technische) Eigenschaften des KI-Systems bewertet werden, sondern auch das Verhalten des KI-Systems in spezifischen Testszenarien, werden für einige Anwendungen virtuelle oder auch physische Testumgebungen erforderlich sein. Auf diese Weise kann das Abschneiden des KI-Systems in einem Testszenario bewertet werden. Dies wird vor allem bei autonomen Systemen und insbesondere bei robotischen Systemen eine wichtige Rolle spielen, um testen zu können, ob sich das System wie vorgesehen verhält. Sowohl die Prüfwerkzeuge als auch die Testumgebungen müssen standardisiert sein. Entsprechende Standardisierungs- und Normungsprozesse sollten daher angestoßen werden, um die jeweiligen Anforderungen zu klären. Zudem kann eine gewisse Flexibilität bei den Prüfwerkzeugen und Testumgebungen erforderlich werden, um der schnellen technischen Entwicklung Rechnung zu tragen.

Freiwillige Kennzeichnung

Nicht immer sind gesetzliche Vorgaben oder eine umfassende Zertifizierung sinnvoll oder notwendig, gerade bei niedriger Kritikalität (siehe Tabelle 3). Gütesiegel auf freiwilliger Basis oder freiwillige Selbstverpflichtungen können ein sinnvolles Instrument sein, um die Einhaltung bestimmter Kriterien zu signalisieren und somit gegebenenfalls einen Wettbewerbsvorteil zu erlangen. Sie können Orientierung bieten und Bewusstsein für bestimmte Aspekte von KI-Systemen bei Nutzenden und Anwendenden schaffen. Aus Sicht der Verbrauchenden wäre dies wünschenswert. Es sollte allerdings nicht zu viele unterschiedliche Gütesiegel geben, da diese nur zu Unübersichtlichkeit führen würden und so an Aussagekraft verlieren. Gegenwärtig gibt es beispielsweise bereits das Gütesiegel des KI-Bundesverbandes.[21] Die Mitglieder des Verbandes verpflichten sich, bestimmte Werte und Normen zu berücksichtigen. Darunter befinden sich ethische Prinzipien wie Transparenz, Sicherheit, Datenschutz und Unvoreingenommenheit. Der Kriterienkatalog ist zwar nicht umfassend und differenziert genug, aber dennoch ein erster Ansatzpunkt. Hinsichtlich der ethischen Prinzipien wäre zum Beispiel eine Orientierung an bestehenden Leitlinien zu ethischen Kriterien und ihrer Operationalisierung sinnvoll (vgl. AI Impact Group 2019). Gütesiegel sollten weiterhin europaweit gültig sein und im Binnenmarkt gegenseitige Anerkennung finden (vgl. Bundesregierung 2020). Ferner bedarf es wirkungsvoller, rechtlich durchsetzbarer Sanktionen, wenn Nutzende eines Siegels die Anforderungen nicht

21 Für weitere Informationen zum Gütesiegel siehe KI-Bundesverband: https://ki-verband.de/wp-content/uploads/2019/02/KIBV_Guetesiegel.pdf

erfüllen bzw. das Gütesiegel missbräuchlich nutzen (ebenda). Eine Alternative wäre ein Siegel für vertrauenswürdige KI ähnlich dem Trust E-Commerce Label, das durch ein privatwirtschaftlich organisiertes Institut auf der Basis rechtlicher Kriterien vergeben wird. Die Unternehmen könnten sich jedoch auch auf Kriterien einigen und die Überprüfung an eine unabhängige Drittstelle delegieren.

Unternehmensprozesse als Gegenstand der Zertifizierung

Prozesse in Unternehmen sollten künftig ein wichtiger, wenn auch nicht alleiniger Baustein einer Zertifizierung von KI-Systemen sein. Schon heute gibt es freiwillige KI-Kodizes in Unternehmen wie Bosch, SAP oder der Telekom, die innerhalb von Unternehmen einen Rahmen für einen ethisch reflektierten Umgang mit KI bilden.[22] Es werden jedoch Standards und Normen für die Zertifizierung der Herstellungs- und Entwicklungs- und Anwendungsprozesse mit Blick auf KI in den Unternehmen notwendig werden. Dies trifft sowohl für eine freiwillige Zertifizierung zu als auch für eine obligatorische Zertifizierung bei der Entwicklung von KI-Anwendungen, deren Kritikalität für ihre Anwendungskontexte hoch eingeschätzt wird. Teil dieser Unternehmensprozesse könnten auch Compliance-Systeme sein mit einer oder einem Beauftragten, die oder der die Einhaltung von Regeln mit Blick auf die Entwicklung von KI-Systemen überwacht – hier kann die DSGVO bezüglich des Datenschutzbeauftragten einen Orientierungspunkt darstellen. Zuständig für die Zertifizierung solcher Unternehmensprozesse könnten bestehende Prüfstellen werden, wie etwa der TÜV oder auch Wirtschaftsprüfer.

Zertifizierung durch zivilgesellschaftliche und öffentliche Einrichtungen sowie technische Prüfstellen

Bei höherer Kritikalität sollte eine staatlich vorgeschriebene Zertifizierung notwendig sein (siehe Tabelle 3). Die Zertifizierung sollte durch akkreditierte Drittstellen durchgeführt werden (Vereine, Stiftungen, GmbHs etc.). Bestehende Prüfstellen könnten eine solche Zertifizierung übernehmen, müssten jedoch entsprechend ausgebaut werden, um dies im Fall von KI-Systemen leisten zu können. So kann etwa die Prüfung von technischen Eigenschaften wie Robustheit, Verlässlichkeit oder auch IT-Sicherheit, die mittels objektiver Messverfahren geprüft werden, von solchen Stellen durchgeführt werden. Im Falle der funktionalen Sicherheit könnte beispielsweise der TÜV zuständig werden und bei IT-Sicherheit das BSI und die entsprechenden bestehenden und anerkannten Prüfstellen. Im Gegensatz zu den technischen Eigenschaften von KI-Systemen ist für die Bewertung des Anwendungskontextes von KI-Systemen ein interdisziplinärer Ansatz sinnvoll. In einem Dialogprozess mit relevanten Stakeholdern können besondere Anforderungen an KI-Systeme für spezifische Anwendungskontexte sowie ihre Umsetzung geklärt werden. Bei hoher Kritikalität sollte eine verpflichtende Zertifizierung durch eine öffentliche Einrichtung erfolgen. Mit dem Bundesamt

22 Für weitere Informationen zu den Kodizes siehe Bosch https://www.bosch.com/de/stories/ethische-leitlinien-fuer-kuenstliche-intelligenz/; SAP https://www.sap.com/products/intelligent-technologies.html?pdf-asset=940c6047-1c7d-0010-87a3-c30de2ffd8ff&page=1 ; Telekom https://www.telekom.com/de/konzern/digitale-verantwortung/details/ki-leitlinien-der-telekom-523904

für Sicherheit in der Informationstechnik und den Datenschutzbehörden existieren bereits Behörden mit hoher Kompetenz, die für eine Prüfung hinsichtlich IT-Sicherheit und Datenschutz herangezogen werden können.

Die Rolle von Forschungsinstituten bei der Konformitätsbewertung

Die Kooperation zwischen Zertifizierungsstellen und Forschungsinstituten ist in verschiedener Hinsicht bedeutend, damit die Zertifizierung von KI-Systemen gelingen kann. Sie ist in erster Linie ein wichtiger Baustein, um einer dynamischen Verfasstheit der Prüfstellen Rechnung zu tragen, die auf KI-Innovationen adäquat reagieren kann. Zudem können technische Lösungen unter Umständen einen hohen Grad an Komplexität erreichen. Die Kooperation mit Forschungsinstituten kann den Prüfstellen einerseits den momentanen Stand der Forschung aufzeigen und andererseits dazu beitragen, dass Prüfmethoden mit der aktuellen Entwicklung Schritt halten. So können öffentliche und private Forschungseinrichtungen die Entwicklung von Prüfwerkzeugen vorantreiben. Die Institute können weiterhin gegebenenfalls schnell Expertise zu neuen KI-Entwicklungen zuliefern, um eine bestmögliche Absicherung von KI-Systemen zu erreichen. Handelt es sich jedoch um technisch sehr anspruchsvolle Fälle, sollten Forschungsinstitute in den Zertifizierungsprozess eingebunden werden bzw. Kooperationen mit den Instituten angestrebt werden. Wenn für KI-Anwendungen Testumgebungen notwendig werden, bietet sich ebenfalls eine solche Kooperation an, da etwa für die Prüfung von robotischen Systemen schon heute geeignete Testumgebungen an Forschungseinrichtungen existieren. Letztlich sind solche Kooperationen auch sinnvoll, um im Rahmen von Lehrgängen künftige Prüferinnen und Prüfer im Umgang mit solchen Prüfwerkzeugen oder in der Nutzung von Testumgebungen zu schulen. Perspektivisch sollten die Curricula solcher Lehrgänge standardisiert sein.

Tabelle 3: Verhältnis von Kritikalität zu staatlicher Regulierung

Kritikalität	Zuständigkeit	Staatliche Eingriffstiefe	Formen staatlicher Regulierung (Beispiele)	Zertifizierung
niedrig	• Kein Prüfer notwendig • Hersteller • (akkreditierte) Drittstellen (Vereine, Stiftungen, GmbHs etc.)	Keine / freiwillig durch Anbieter	• Keine Form der Prüfung • Verpflichtung auf eigene Kriterien (Herstellererklärung, Selbstverpflichtung, Gütesiegel) • Anwender-Anbieter-Überprüfung	
			• Unternehmen einigen sich auf Kriterien und delegieren Überprüfung an Drittstelle • Konformitätsbewertung gegen Standards und Normen durch Dritte	
	Hersteller	Staatlich vorgeschrieben	• Herstellerselbsterklärung • Selbstverpflichtung auf eigene Kriterien • Anwender-Anbieter-Überprüfung	
	Akkreditierte Drittstellen (Vereine, Stiftungen, GmbHs etc.)	Staatlich vorgeschrieben	Unternehmen einigen sich auf Kriterien und delegieren Überprüfung an Drittstelle	
	Akkreditierte Drittstellen (Vereine, Stiftungen, GmbHs etc.)	Staatlich vorgeschrieben	Konformitätsbewertung gegen Standards und Normen durch staatlich berufene Dritte	
	Staatliche Stellen	Staatlich vorgeschrieben	Konformitätsbewertung gegen Standards und Normen durch staatliche Behörden	
	Staatliche Stellen	Staatlich vorgeschriebene Zulassungsverfahren	Zulassung, Prüfung gegen Gesetze durch staatlich berufene Dritte (vgl. Typengenehmigung bei Autos, bestimmte Medizinprodukte)	
hoch	Staat	Staatliche Regulierung i. S. v. Verboten oder Einschränkungen	• Einschränkung des KI-Einsatzes in bestimmten Domänen • Verbote von Entwicklung und Einsatz bestimmter KI-Systeme (gänzlich oder für bestimmte Anwendungskontexte)	

Anmerkung: Die Tabelle ist eine Systematisierung der Regulierungsformen und trifft keine Aussage über die Quantität der Anwendungen, die in den jeweiligen Bereich fallen.

4. Mögliche Gestaltungsoptionen

Zur Etablierung einer gelungenen Zertifizierung von KI-Systemen stehen folgende mögliche Ansatzpunkte zur Verfügung:

Die politischen Entscheidungsträgerinnen und Entscheidungsträger könnten…

- sich auf nationaler und internationaler Ebene für die Schaffung klarer Rahmenbedingungen für die Zertifizierung von KI-Systemen einsetzen. Dies umfasst sowohl die Standards als auch die sachkundige und neutrale Überprüfung durch Dritte.

- Initiativen für den Aufbau einer Infrastruktur für die Zertifizierung von KI-Systemen sowie die hierfür erforderliche Durchführung von Use Cases und Pilotprüfungen sowie Testgelände und Testumgebungen unterstützen und finanziell fördern.

- sich für die Etablierung eines Zertifizierungssystems für KI-Systeme einsetzen, das sich am Anwendungskontext orientiert und somit Überregulierung vermeidet und gleichzeitig Innovationen ermöglicht. Dazu gehört, mit Augenmaß eine Regulierung für hoch-kritische KI-Systeme zu erarbeiten und zugleich die Entwicklung von KI-Anwendungen in unkritischen Bereichen zu unterstützen und diese Freiräume zu erhalten.

- den ersten Schritt hin zu einem Zertifizierungssystem für KI-Systeme gehen und gemeinsam mit Vertreterinnen und Vertretern aus Wissenschaft, Wirtschaft und der Zivilgesellschaft Schwellwerte für die unterschiedlichen Kritikalitäten definieren und so die Grundlage für eine detailliertere Festlegung auf bestimmte Kriterien legen. Ein Fokus sollte hierbei auf die Bereiche gelegt werden, in denen autonome Systeme schützen oder retten können.

- dafür Sorge tragen, dass in der schulischen Bildung die Wissensvermittlung über die Funktionsweise und Auswirkungen von KI-Systemen gestärkt wird und Schülerinnen und Schüler einen souveränen Umgang mit KI-Systemen erlernen und somit auch die Aussagekraft von Gütesiegeln oder Zertifizierungen besser einordnen können.

- Zertifizierungsprozesse mit Bürgerbeteiligungsverfahren und Stakeholderkonsultationen begleiten, um einerseits die Vielfalt der Interessenlagen einzubeziehen und andererseits die Anwendung von KI-Systemen durch einen gesellschaftlichen Konsens abzusichern.

Die Forschung könnte…

- die Details der Zertifizierungsverfahren in interdisziplinären Forschungsverbünden eingehender erforschen, um dazu beizutragen, Prüfwerkzeuge zur Evaluation von KI-Systemen zu entwickeln und allgemeine Kriterien wie „Transparenz" für Wirtschaft, Nutzende und Technikentwicklung operationalisierbar zu machen. Auf dieser Basis kann die Forschung Politik, Unternehmen und Zivilgesellschaft noch eingehender zu den Chancen, Risiken und Konsequenzen der einzelnen Technologien und Anwendungsbereiche beraten.

- interdisziplinär technologische Lösungen und Methoden entwickeln, um sicherzustellen, dass KI-Systeme vertrauenswürdig sind.

- mit Unternehmen zusammen vertrauenswürdige KI-Methoden entwickeln (erklärbare, explainable AI, XAI).

- ihre modernsten Infrastrukturen zur Verfügung stellen, damit diese Anknüpfungspunkte für erste Zertifizierungsvorhaben bilden können.

- erforschen, wo traditionelle Signalverarbeitungsmethoden aufhören und wo KI anfängt, um eine genauere Gesetzgebung zu ermöglichen.

- sich an der Entwicklung eines Konzepts zur Ausbildung von KI-Prüfingenieuren beteiligen.

Die Unternehmen könnten…

- die Bildung von Vertrauen in KI-Systeme unterstützen, indem sie freiwillig ethische und technische Standards ausarbeiten und offenlegen und sich verstärkter dem Einsatz von erklärbarer KI widmen. Dies stellt eine Basis für die Debatte um die Zertifizierung und Regulierung von KI-Systemen dar.

- sich an der Schaffung entsprechender Standards beteiligen und entsprechende Bedarfe identifizieren.

- sich jeweils branchenspezifisch darüber austauschen, welche Aspekte von KI-Systemen in ihrem Anwendungskontext als kritisch zu betrachten sind und wie von Seiten der Hersteller Best Practices für solche Fälle etabliert werden könnten. Als ein Orientierungspunkt für einen solchen Austausch können bestehende Konzepte und Gestaltungsrichtlinien dienen. Ferner könnte dieser Austausch eine Basis dafür darstellen dass Unternehmen in vertrauenswürdige KI investieren und entsprechende Geschäftsmodelle entwickeln.

- ihre modernsten Infrastrukturen zur Verfügung stellen, damit diese Anknüpfungspunkte für erste Zertifizierungsvorhaben bilden können.

- Beschäftigten im Rahmen der innerbetrieblichen Weiterbildung spezielle Schulungen anbieten, die einen souveränen Umgang mit KI-Systemen zum Ziel haben.

Die Zivilgesellschaft könnte…

- Bereiche identifizieren, für die aus Sicht von Verbrauchenden und Bürgerinnen und Bürgern eine Regulierung erforderlich ist. Genauso könnten Bereiche identifiziert werden, für die keine Regulierung notwendig ist und die sich für eine zivilgesellschaftliche Konformitätsprüfung etwa über ein Gütesiegel anbieten könnten.

- auf der Basis bestehender und künftig entwickelter Kriterien und Gestaltungsrichtlinien sowie der rechtlichen Rahmenbedingungen die Rolle eines „Watchdogs" einnehmen und so auf die Einhaltung der Kriterien, Richtlinien und Regeln drängen, um auf diese Weise den Einsatz von KI mitzugestalten.

Darüber hinaus bedürfen einige Aspekte eines **gesellschaftlichen Diskurses** unter Einbeziehung aller relevanten Stakeholder aus Wirtschaft, Wissenschaft und Zivilgesellschaft. Diese betreffen:

- die Definition von Kritikalitätsstufen. Dies erfordert eine Diskussion zu tragbaren Risiken und zur gerechten Verteilung der Vorteile, die aus KI-Anwendungen hervorgehen. Ferner ist eine Aufklärung über die Funktions- und Wirkungsweise und Einsatzmöglichkeiten von KI notwendig, um zu einer realistischen Einschätzung ihrer Potenziale zu kommen.

- die Notwendigkeit der Zertifizierung von KI-Systemen: Dies gilt vor allem für die Frage, inwiefern weitere Normen und Standards über die bereits existierenden Sicherheits- und Transparenzstandards von technischen (industriellen) Systemen hinaus notwendig sind.

- die Art, wie wir zukünftig mit KI leben, lernen und arbeiten wollen: Ziel ist eine Entwicklung von KI-Systemen, die so umgesetzt sind, dass sie menschliche Kompetenzen erweitern und nicht beschneiden. Ein solcher breiter Diskurs von KI stellt die Basis dar, auf der künftig über Bewertungs- und Prüfkriterien für KI-Systeme diskutiert werden könnte.

5. Fazit und Ausblick

Wie zu Beginn dargestellt, sind in Bezug auf die Ausgestaltung einer Zertifizierung von KI-Systemen, die Überregulierung vermeidet und Innovationen ermöglicht, noch zahlreiche Fragen offen. In vorliegendem Whitepaper wurden einige dieser offenen Fragen adressiert und erste Gestaltungsoptionen dafür vorgestellt. So ist festzuhalten, dass die Notwendigkeit einer Zertifizierung von KI-Systemen sich aus dem Maß der Kritikalität in einem bestimmten Anwendungskontext ableitet. Diese muss immer in Abhängigkeit vom konkreten Einsatzgebiet der KI ermittelt werden. Die Schwellenwerte, ab welcher Kritikalität eine Zertifizierung oder auch eine verpflichtende Zulassung notwendig werden, müssen regulatorisch festgelegt werden. Dies bedarf einer gesellschaftlichen Debatte, die der Komplexität der Bestimmung solcher Schwellenwerte in unterschiedlichen Anwendungskontexten Rechnung trägt. Der Gegenstand einer Zertifizierung kann entweder das Produkt oder der Prozess oder beides gemeinsam sein. Beide Verfahren haben Vor- und Nachteile und legen unterschiedliche Schwerpunkte. Die anzulegenden Prüfkriterien lassen sich in Mindestkriterien, die immer erfüllt und abgeprüft werden müssen, sowie darüber hinausgehende freiwillige Kriterien, die abgeprüft werden können, unterteilen. Die Zertifizierung sollte durchgeführt werden, bevor das Produkt oder die Dienstleistung in Verkehr gebracht wird. Bei weiterlernenden Systemen sollte die Zertifizierung regelmäßig wiederholt werden. Der Detailgrad und die Prüftiefe orientieren sich ebenfalls an der Kritikalität. Damit die skizzierte Zertifizierung von KI-Systemen umgesetzt werden kann, wird eine Anpassung und gegebenenfalls auch Erweiterung der bestehenden Infrastruktur benötigt.

All die aufgezählten Punkte sind Ansatzpunkte, wie eine gelungene Zertifizierung von KI-Systemen gestaltet werden könnte. Entscheidend ist aktuell die Beantwortung der Frage, in welchen Fällen eine Zertifizierung von KI-Systemen notwendig sein sollte. Zur Beantwortung dieser Frage bedarf es eines kombinierten Ansatzes aus empirischer Forschung zur Erarbeitung einer objektiven Basis auf der einen Seite und konzeptioneller und normativer Überlegungen auf der anderen Seite. Durch die Verzahnung beider Herangehensweisen können wichtige Informationen gefunden und Schwellenwerte definiert werden. Mithilfe dieser Schwellenwerte können dann die nächsten Schritte für ein Zertifizierungssystem von KI-Systemen eingeleitet werden. Ziel ist es, Überregulierung zu vermeiden, Innovationen zu ermöglichen und KI-Systeme in die Anwendung zu bringen.

Literatur

AI Ethics Impact Group (2020): From Principles to Practice. An interdisciplinary framework to operationalise AIethics. https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf (abgerufen am 23.09.2020).

AI High Level Expert Group (HILEG) (2019): Ethics Guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (abgerufen am 23.09.2020).

Beck, S. et al. (2019): Künstliche Intelligenz und Diskriminierung Herausforderungen und Lösungsansätze. Whitepaper aus der Plattform Lernende Systeme, München. https://www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-loesungsansaetze.html?file=files/Downloads/Publikationen/AG3_Whitepaper_250619.pdf (abgerufen am 23.09.2020).

Beer, J.M., Fisk, A. D. & Rogers, W.A. (2014): Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. https://dl.acm.org/doi/pdf/10.5898/JHRI.3.2.Beer (abgerufen am 23.09.2020).

Beyerer et al. (i. E.): Kompetent im Einsatz: Variable Autonomie Lernender Systeme in lebensfeindlichen Umgebungen. Whitepaper aus der Plattform Lernende Systeme, München.

Bundesregierung (2018): Strategie Künstliche Intelligenz der Bundesregierung. https://www.bmbf.de/files/Nationale_KI-Strategie.pdf (abgerufen am 23.09.2020).

Bundesregierung (2020): Stellungnahme zum Weißbuch KI der EU-Kommission. https://www.ki-strategie-deutschland.de/files/downloads/Stellungnahme_BReg_Weissbuch_KI.pdf (abgerufen am 23.09.2020).

Datenethikkommission (2019): Gutachten der Datenethikkommission. https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf;jsessionid=1B58624B275151B999917CB93C661013.1_cid364?__blob=publicationFile&v=6 (abgerufen am 23.09.2020).

Europäische Kommission (2020): White Paper on Artificial Intelligence – A European approach towards excellence and trust. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (abgerufen am 23.09.2020).

Gamer, T., Kloepper, B. & Hoernicke, M. (2019): The way toward autonomy in industry – taxonomy, process framework, enablers, and implications. In: IECON 2019 – 45th Annual Conference of the IEEE Industrial Electronics Society, pp. 565-570. IEEE.

Haidegger, T. (2019): Autonomy for Surgical Robots: Concepts and Paradigms. https://www.semanticscholar.org/paper/Autonomy-for-Surgical-Robots%3A-Concepts-and-Haidegger/0af00e43658fe5dacaec577a943619351b8ad230/figure/7 (abgerufen am 23.09.2020).

Hessen, J. et al. (i. E.): Kritikalitätsstufen und ihre Nutzung in unterschiedlichen Anwendungsbieten von KI-Systemen. Whitepaper aus der Plattform Lernende Systeme, München.

Heesen, J. et al. (Hrsg.) (2020a): Zertifizierung von KI-Systemen – Impulspapier aus der Plattform Lernende Systeme. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Impulspapier_290420.pdf (abgerufen am 23.09.2020).

Heesen, J. et al. (Hrsg.) (2020b): Ethik-Briefing. Leitfaden für eine verantwortungsvolle Entwicklung und Anwendung von KI-Systemen. Whitepaper aus der Plattform Lernende Systeme, München. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_EB_200831.pdf (abgerufen am 01.10.2020).

Huchler, N. et al. (Hrsg.) (2020): Kriterien für die Mensch-Maschine-Interaktion bei KI. Ansätze für die menschengerechte Gestaltung in der Arbeitswelt. Whitepaper aus der Plattform Lernende Systeme, München. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG2_Whitepaper2_220620.pdf (abgerufen am 23.09.2020).

IAIS (Hrsg.) (2019): Vertrauenswürdiger Einsatz von künstlicher Intelligenz. https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf (abgerufen am 23.09.2020).

Müller-Quade, J. et al. (Hrsg.) (2020): Sichere KI-Systeme für die Medizin. Datenmanagement und IT-Sicherheit in der Krebsbehandlung der Zukunft. Whitepaper aus der Plattform Lernende Systeme, München. https://www.acatech.de/publikation/sichere-ki-systeme-fuer-die-medizin/download-pdf?lang=de (abgerufen am 23.09.2020).

Plattform Industrie 4.0 (2019): Technologieszenario „Künstliche Intelligenz in der Industrie 4.0". https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/KI-industrie-40.pdf?__blob=publicationFile&v=10 (abgerufen am 23.09.2020).

Plattform Lernende Systeme (2019): Lernende Systeme in lebensfeindlichen Umgebungen. Potenziale, Herausforderungen und Gestaltungsoptionen. Bericht der Arbeitsgruppe Lebensfeindliche Umgebungen, München. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG-7_Bericht_web_final.pdf (abgerufen am 23.09.2020).

Stowasser, S. & Suchy, O. et al. (Hrsg.) (2020): Einführung von KI-Systemen in Unternehmen. Gestaltungsansätze für das Change-Management. Whitepaper aus der Plattform Lernende Systeme, München. www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG2_Whitepaper_Change_Management.pdf (abgerufen am 17.11.2020).

SAE International (2018): SAE International Releases Updated Visual Chart for Its "Levels of Driving Automation" Standard for Self-Driving Vehicles. https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles (abgerufen am 23.09.2020).

Zweig, K. & Krafft, T. (2019): Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse. Ein Regulierungsvorschlag aus sozioinformatischer Perspektive. https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf (abgerufen am 23.09.2020).

Über dieses Whitepaper

Vorliegendes Whitepaper wurde auf der Basis von Experteninterviews mit Mitgliedern und Vertreterinnen und Vertretern der in der Plattform Lernende Systeme beteiligten Forschungseinrichtungen und Unternehmen sowie Gastautorinnen und -autoren erstellt. Federführend waren PD Dr. Jessica Heesen und Prof. Dr. Jörn Müller-Quade für die Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik sowie Prof. Dr. Stefan Wrobel für die Arbeitsgruppe Technologische Wegbereiter und Data Science. Beteiligt waren darüber hinaus Mitglieder aller Arbeitsgruppen der Plattform Lernende Systeme: das heißt Mitglieder der Arbeitsgruppen Arbeit/Qualifikation, Mensch-Maschine-Interaktion, Mobilität und intelligente Verkehrssysteme, Geschäftsmodellinnovationen, Gesundheit, Medizintechnik, Pflege und der Arbeitsgruppe Lebensfeindliche Umgebungen.

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. Jürgen Beyerer, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB
Dr. Gunnar Brink, ROSEN Technology and Research Center GmbH
Dr. Wolfgang Faisst, ValueWorks GmbH
Dr. Martin Hoffmann, ABB AG
Dr. Norbert Huchler, Institut für sozialwissenschaftliche Information und Forschung e.V.
Dr. Elsa Kirchner, Deutsches Forschungszentrum für Künstliche Intelligenz
Prof. Dr. Tobias Matzner, Universität Paderborn
Dr. Matthias Peissner, Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO
Dr. Christoph Peylo, Robert Bosch GmbH
Thomas Schauf, Deutsche Telekom AG
Dr. Sirko Straube, Deutsches Forschungszentrum für Künstliche Intelligenz
Oliver Suchy, Deutscher Gewerkschaftsbund
Dr. Susann Wolfgram, SAP

Autorinnen und Autoren mit Gaststatus
Prof. Dr. Ute Schmidt, Universität Bamberg
Prof. Dr. Dr. Frauke Rostalski, Universität Köln
Dr. Maximilian Poretschkin, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Dr. Pascal Birnstill, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung

Redaktion
Stephanie Dachsberger, Geschäftsstelle der Plattform Lernende Systeme
Maximilian Hösl, Geschäftsstelle der Plattform Lernende Systeme
Dr. Ursula Ohliger, 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.

Anhang: Prüfkriterien für die Zertifizierung von KI-Systemen

Mindestkriterien[23]

Transparenz, Nachvollziehbarkeit, Nachprüfbarkeit und Verantwortlichkeit

Transparenz muss sowohl darüber bestehen, dass ein KI-System eingesetzt wird, als auch über die Art der Erhebung, Auswertung und Verwendung von Daten sowie die Funktionsweise des Systems. Welche Anforderungen die Transparenz an den Einsatz von KI konkret richtet, ist abhängig vom jeweiligen Betroffenen und dem Einsatzgebiet beziehungsweise der Aufgabe des KI-Systems.[24] So muss beispielsweise für den für die Überwachung des Systems Zuständigen Transparenz auf technischer Ebene bestehen.

Funktionale Sicherheit/Safety/inkl. Produktsicherheit und Zuverlässigkeit

Das eingesetzte KI-System muss die notwendige technische Robustheit gewährleisten, das heißt, so zuverlässig sein, dass es zu keinem Zeitpunkt ein unannehmbares Risiko darstellt.

Vermeidung von nicht-intendierten Folgewirkungen auf andere Systeme, Menschen und die Umwelt[25]

KI-Systeme müssen so entwickelt werden, dass Schaden für Einzelpersonen, Gemeinschaften, andere Systeme oder beispielsweise auch die Umwelt vermieden oder in gerechtfertigter Abwägung eines kleinen Schadens gegen einen größeren Nutzen minimiert wird. Besonders zu beachten ist die Vermeidung von Schäden an Gesundheit und Leben durch KI-Systeme.

Gerechtigkeit i. S. v. Gleichheit und Diskriminierungsfreiheit

Gerechtigkeit ist das Prinzip eines individuellen oder gesellschaftlichen Handelns, das jedem gleichermaßen sein Recht gewährt und dabei eine kontextbezogene Beurteilung von Ausgleich und Gleichbehandlung im Blick hat. Gerechtigkeit kann unter anderem durch die weiterführenden Prinzipien Gleichheit und Diskriminierungsfreiheit beschrieben werden:

- Das Prinzip der **Gleichheit** umfasst die Gleichheit vor dem Gesetz, die Uneinschränkbarkeit der Menschenwürde sowie Diskriminierungsfreiheit und Chancengerechtigkeit.[26]
- **Diskriminierung** (von lat. *discriminare*, unterscheiden) bedeutet in normativer Perspektive Ungleichbehandlung von Gleichem oder Gleichbehandlung von Ungleichem ohne sachlichen Grund. Dies ist besonders im Hinblick auf diskriminierungsfreie Lernprozesse und Datengrundlagen relevant

Schutz der Privatheit und der Persönlichkeit: Informationelle Selbstbestimmung, Datenschutz, Datenqualität und Datensicherheit

Privatheit meint die Abgrenzung eines Bereiches von der Öffentlichkeit, der geschützt ist vor äußeren Eingriffen. In diesem Raum kann der Mensch seine Persönlichkeit und Identität frei entfalten und entwickeln. Der Schutz der Privatheit umfasst den Schutz der Privatsphäre, Anonymität als Privatheit im öffentlichen Raum, das Recht auf informationelle Selbstbestimmung sowie die Integrität der Persönlichkeit. Schutz der Privatheit und der Persönlichkeit kann unter anderem durch die weiterführenden Prinzipien informationelle Selbstbestimmung sowie die Integrität der persönlichen Identität und Privacy by Design beschrieben werden.

- **Informationelle Selbstbestimmung** beschreibt das Recht, grundsätzlich selbst über die Preisgabe und Verwendung von personenbezogenen Daten zu bestimmen.
- Die **Integrität der persönlichen Identität** umfasst das Recht auf Selbstdarstellung als Recht am eigenen Wort und Bild sowie den Schutz vor verfälschenden Darstellungen. Darunter fallen Beleidigungen und Verleumdungen ebenso wie die Erstellung von (falsch) prognostizierenden Persönlichkeitsprofilen (etwa in Bezug auf Straffälligkeit).
- Die Umsetzung des Schutzes der Privatheit kann nicht allein durch gesetzliche Regelungen (wie etwa Vorschriften zur Datenminimierung oder Zweckbindung) normiert werden, sondern muss durch das Design der Technologie realisiert werden (**Privacy by Design**).

23 Falls nicht anders angegeben, sind diese Definitionen dem Whitepaper „Ethik-Briefing" (siehe Heesen et al. 2020b) entnommen.

24 Die grundlegenden Prüfkriterien sollten wenn möglich immer erfüllt werden. In Fällen, in denen Gründe für eine Interessenbeeinträchtigung durch den Einsatz des KI-Systems vorliegen, muss dies begründet und im Sinne von Abwägungsentscheidungen beschlossen werden. Beispiele sind etwa der Einsatz eines KI-Systems für die Gefahrenabwehr oder die Wahrung von Geschäftsgeheimnissen. In beiden Fällen kann es sinnvoll sein, zwar den Einsatz, aber nicht die Funktionsweise des Systems offenzulegen.

25 Unintendierte Folgewirkungen können auch in Bezug auf Menschen auftreten, wenn sich die intendierten Folgewirkungen des Systems nicht auf Menschen beziehen.

26 Siehe hierzu Allgemeine Erklärung über die Menschenrechte Art.1 und Art.7, Art. 20 – 23 EU-GRC und Art. 3 GG.

Selbstbestimmung inkl. Transparenz über den Einsatz des KI-Systems und die Rolle des Menschen im Entscheidungsprozess

Selbstbestimmung ist das grundlegende Prinzip demokratischer Gesellschaften und garantiert jedem die freie Entfaltung seiner Persönlichkeit, soweit die Rechte anderer nicht verletzt werden. Die Selbstbestimmung des Einzelnen muss im Kontakt mit anderen immer abgewogen werden gegen die Prinzipien der Gleichheit und Gerechtigkeit.

- Ein wesentliches Element, um auch im Umgang mit KI-Systemen Selbstbestimmung zu erhalten, ist **Transparenz**. Nur wenn die oder der Endnutzende weiß, dass sie oder er mit einem KI-System interagiert (und gegebenenfalls auch, wie dieses funktioniert), kann sie oder er selbstbestimmt und eigenverantwortlich Entscheidungen fällen (AI High Level Expert Group 2019).
- Wichtig ist auch **Transparenz über die Rolle des Menschen im Entscheidungsprozess**. Hier werden drei verschiedene Rollen unterschieden: 1. Human-in-the-loop: Mensch muss Entscheidungen der Systeme bestätigen oder kann zwischen verschiedenen Entscheidungsoptionen wählen, 2. Human-on-the-loop: Mensch hat eine Beobachterrolle inkl. Vetorecht, 3. Human-out-of-the-loop: Das System trifft die Entscheidungen gänzlich alleine.[27]

Darüber hinausgehende Kriterien

Offene Schnittstellen und Systemoperabilität

Offene Schnittstellen ermöglichen es, Softwaredienste externer Anbietender in eigene Applikationen zu **integrieren**. Damit steigt die Verwendbarkeit der Software für verschiedene Anbietende und Nutzende und das mögliche Angebot an Diensten. Systemoperabilität meint die **Kompatibilität** von zwei oder mehr Systemen. Ist ein erfolgreiches KI-System nicht auf Interoperabilität mit anderen Anwendungen ausgelegt, kann sie diese langfristig vom Markt verdrängen. Offene Schnittstellen und Systemoperabilität beugen monopolartigen Strukturen vor und tragen so zu selbstbestimmten Wahlmöglichkeiten bei. Durch die Auswahlmöglichkeit verschiedener Dienste wird Selbstbestimmung gestärkt.

Menschenzentrierung und Nutzerfreundlichkeit (Usability), inkl. Partizipation, Schutz des Einzelnen, sinnvolle Arbeitsteilung und förderliche Arbeitsbedingungen

Nutzerfreundlichkeit (Usability) beschreibt das Ausmaß, in dem ein System durch einen Nutzenden in einem bestimmten Anwendungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen (siehe DIN EN ISO 9241-11).

Für eine gelungene Gestaltung der Mensch-Maschine-Interaktion ist wichtig, dass die Technologie an die Vorteile und Potenziale menschlichen Denkens anknüpft und die wechselseitige Ergänzung – nicht Ersatz oder Konflikt – in den Mittelpunkt der Interaktion gestellt wird. Eine gelungene Mensch-Maschine-Interaktion umfasst unter anderem folgende Cluster aus Kriterien:[28]

- Partizipation bedeutet Teilhabe, Beteiligung und Mitwirkung. In vorliegendem Kontext bedeutet Partizipation, dass der Nutzende aktiv in alle Phasen des Entwicklungs- und Anwendungsprozesses einbezogen wird und an der Ausgestaltung des KI-Systems teilhaben kann.
- Ein wesentliches Element ist der Schutz des Einzelnen. Dieser Punkt umfasst neben bereits genannten Aspekten die Aspekte Sicherheits- und Gesundheitsschutz und verantwortungsbewusste Leistungserfassung.
- Ein weiterer wichtiger Aspekt ist sinnvolle Arbeitsteilung. Dies bedeutet Angemessenheit, Entlastung und Unterstützung, Handlungsträgerschaft und Situationskontrolle, Adaptivität, Fehlertoleranz und Individualisierbarkeit.
- Hinzu kommt die Forderung nach förderlichen Arbeitsbedingungen. Hierzu zählen Handlungsräume und reichhaltige Arbeit, Lern- und Erfahrungsförderlichkeit sowie Kommunikation, Kooperation und soziale Einbindung.

27 Siehe Plattform Lernende Systeme 2019.

28 Grundlage ist das Whitepaper zu Kriterien für die Mensch-Maschine-Interaktion bei KI (siehe Huchler et al. 2020). Da manche dieser Kriterien bereits eingeführt worden sind, werden an dieser Stelle nur noch die zusätzlichen Kriterien vorgestellt.

Nachhaltigkeit
Intergenerationelle Gerechtigkeit beschreibt die Verpflichtung gegenüber **künftigen Generationen**, für einen Erhalt der Lebensumstände zu sorgen. KI-Systeme sollen zur Entwicklung einer nachhaltigen Gesellschaft beitragen und nachhaltig entwickelt und angewandt werden. Nachhaltigkeit hat dabei eine ökonomische, ökologische und soziale Dimension (Datenethikkommission 2019). Die UN haben hierfür 17 Nachhaltigkeitsziele[29] definiert. Diese wurden 2015 verabschiedet.
Kennzeichnung und Begrenzung der Systemfunktionalität
Die Kennzeichnung und Begrenzung der Systemfunktionalität sieht vor, dass – soweit möglich – das System auf vorhergehende Anweisung in bestimmten Situationen menschliche Unterstützung einholt (beispielsweise in rechtlich oder ethisch schwierigen Situationen). Bei Systemen mit höheren Autonomiegraden ist das nicht möglich – hier müsste das System in die Lage versetzt werden, den potenziellen Nutzen möglicher Aktionen zu bestimmen. Wenn keine Aktion mit einem positiven Nutzen gefunden wird, so muss das System Unterstützung einholen beziehungsweise sich in einen sicheren Zustand begeben.

29 Weitere Informationen zu den Nachhaltigkeitszielen der UN sind hier zu finden: https://www.un.org/sustainabledevelopment/sustainable-development-goals/

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 2020

Bildnachweis
Tierney/Adobe Stock/Titel

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. (Hrsg.): Zertifizierung von KI-Systemen – Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme. Whitepaper aus der Plattform Lernende Systeme, München 2020.