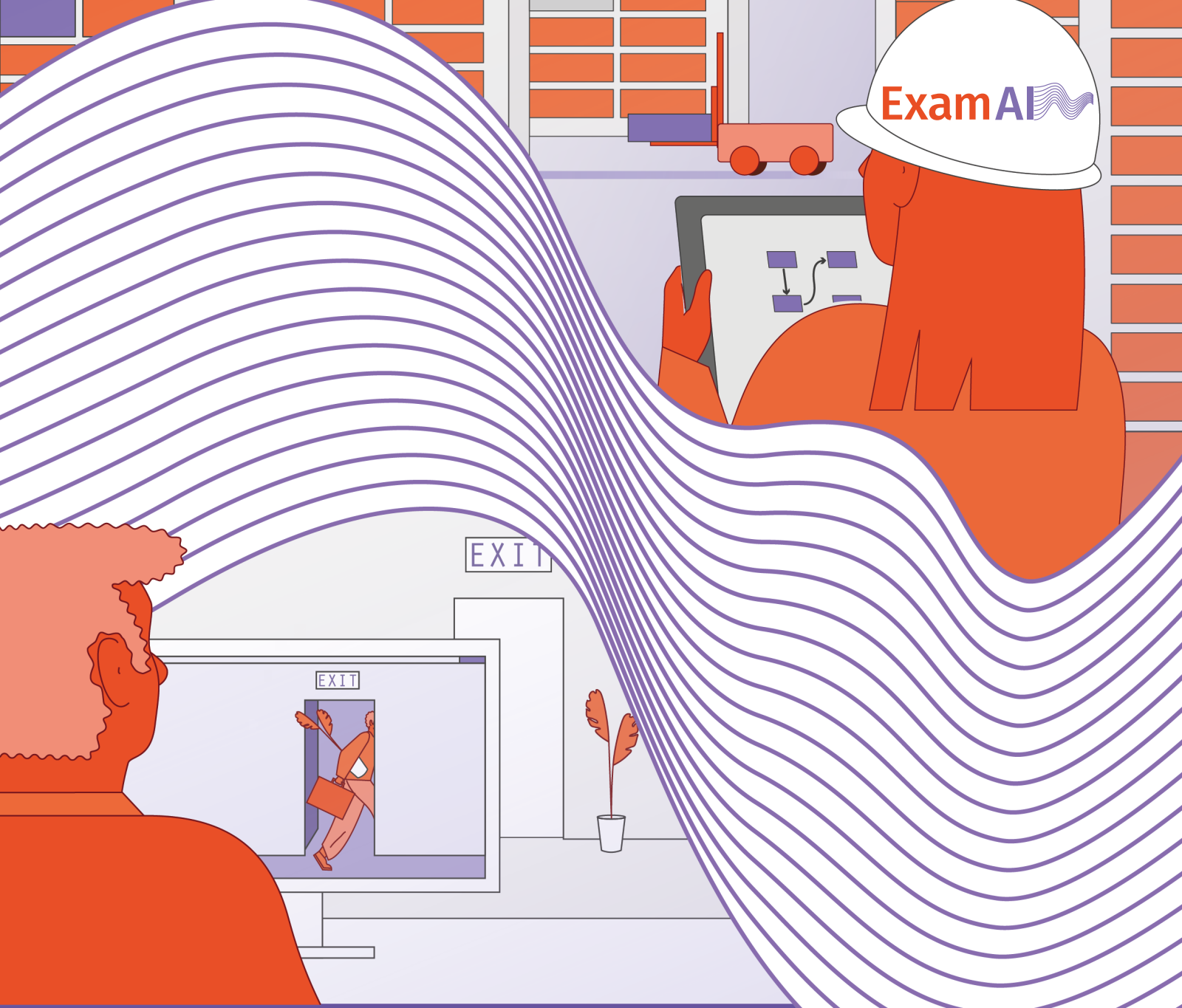

Abschlussbericht ExamAI – KI Testing und Auditing

Herausforderungen, Lösungsansätze und Handlungsempfehlungen für das Testen, Auditieren und Zertifizieren von KI

Abschlussbericht ExamAI – KI Testing und Auditing

Herausforderungen, Lösungsansätze und Handlungsempfehlungen für das Testen, Auditieren und Zertifizieren von KI

INHALTSVERZEICHNIS

1. Einleitung

Technologien, die Künstlicher Intelligenz (KI) zugerechnet werden, finden zunehmende Verbreitung: Das Personal- und Talentmanagement nutzt KI-basierte Vorschlagssysteme oder Persönlichkeitsbewertungen, intelligente Cobots werden in der Produktion eingesetzt oder fahrerlose Transportsysteme unterstützen die Logistik in Fabriken. Fragen nach der Zuverlässigkeit, Vertrauenswürdigkeit, Sicherheit und Fairness solcher Systeme sind auch Gegenstand der politischen Diskussion.

Der vorliegende Bericht stellt die zentralen Forschungsergebnisse des Konsortialprojekts „ExamAI - KI Testing & Auditing" (Laufzeit: März 2020 – November 2021) vor. Das Projekt wurde von der Gesellschaft für Informatik e.V. geleitet und bestand aus einem interdisziplinären Team aus (Sozio-)Informatiker*innen, SoftwareIngenieur*innen sowie Rechts- und Politikwissenschaftler*innen. Anhand der Anwendungsbereiche „Mensch-Maschine-Kooperation in der Industrieproduktion" und „KI-Systeme im Personal- und Talentmanagement sowie im Recruiting" ging das Team der Frage nach, wie sinnvolle Kontroll- und Testverfahren für KI-Systeme aussehen können.

Die Grundlage der Untersuchung bildeten insgesamt 11 Use Cases aus den Anwendungsbereichen Industrieproduktion und Personal- und Talentmanagement, die zu Beginn des Projekts identifiziert wurden (Arbeitspaket 1). Die Use Cases behandeln beispielsweise das Fehlverhalten von intelligenten Cobots (Kollaborative Roboter) und fahrerlosen Transportfahrzeugen oder befassen sich mit KI-basierten Background-Checks von Bewerber*innen und automatisierten Bewertungen von Lebensläufen. Neben dem Hauptfokus auf Möglichkeiten des Testens, der Auditierung und der Zertifizierung von KI-Systemen (Arbeitspaket 4 und 5) lagen die weiteren Schwerpunkte des Projekts auf dem aktuellen Stand der Normung und Standardisierung im Bereich KI (Arbeitspaket 2) und auf den rechtlichen Rahmenbedingungen die für den Einsatz von KI-Systemen gelten (Arbeitspaket 3). Zum Ende des Projekts wurde insbesondere der KI-Regulierungsvorschlags der EU-Kommission vom April 2021 hinsichtlich seiner rechtlichen Vorgaben für herstellende und anwendende Unternehmen untersucht. Zum Abschluss des Projekts wurden im Rahmen von zwei Expert*innenworkshops konkrete politische Handlungsempfehlungen erarbeitet und als Policy Paper veröffentlicht (Arbeitspaket 6).

Eine Übersicht über diese sowie die zahlreichen weiteren Publikationen aus dem Projekt finden Sie am Ende dieses Berichts.

Kapitel 2 gibt zunächst einen Überblick über die Kernaussagen, die in den einzelnen Arbeitspaketen des Projekts formuliert wurden. Es setzt die Kernaussagen zu den Forschungsfragen der Arbeitspakete in Beziehung. Kapitel 3 bis 8 widmen sich jeweils einem Arbeitspaket. Die Ergebnisse der Arbeitspakete werden zusammengefasst und die zugehörigen Veröffentlichungen in Beziehung gesetzt. Das Herzstück des Berichts ist Kapitel 7 mit den Ergebnissen des fünften Arbeitspaketes „Herausforderungen und Lösungsansätze für Testing, Auditing und Zertifizierung von KI“. In diesem Arbeitspaket wurden die Ergebnisse der vorherigen Arbeitspakete zusammengeführt, um die relevanten Themen für die Expert*innen-Workshops (Arbeitspaket 6) zu identifizieren und die abschließenden Handlungsempfehlungen vorzubereiten. Das Ziel des Kapitels ist es, Herausforderungen bezüglich der Realisierbarkeit technischer Lösungen aufzudecken und mögliche Lösungsansätze zu skizzieren. In diesem Zusammenhang wurden insbesondere auch technische und juristische Aspekte zusammengeführt. Kapitel 9 zieht ein Fazit, dass die wichtigsten Handlungsempfehlungen beinhaltet.

Um Missverständnisse aufgrund unterschiedlicher Terminologien in den verschiedenen Wissenschaftsdisziplinen zu vermeiden, wurden die im Glossar festgehaltenen Arbeitsdefinitionen gewählt.

2.
Kernaussagen aus den Arbeitspaketen

Abbildung 1 zeigt welche Kernaussagen sich aus den Ergebnissen der Arbeitspakete 1 bis 6 ableiten lassen. In den darauffolgenden Unterkapiteln werden die Kernaussagen und deren Beziehung zueinander erläutert.

Abbildung 1 – Darstellung von Kernaussagen aus den Arbeitspaketen 1 bis 5

2.1. Möglichkeiten und Grenzen von Anwendungsfällen

Im ersten Arbeitspaket wurden Anwendungsfälle mit KI-Einsatz in den Bereichen Produktionsautomatisierung sowie Personal- und Talentmanagement untersucht [1, 2]. Die wesentliche Hürde um das Potenzial von KI voll auszuschöpfen ist die Nutzung in kritischen Anwendungen.

Im Bereich der Produktionsautomatisierung bezieht sich die Kritikalität in erster Linie auf physischen Personenschaden. Im Bereich des Personal- und Talentmanagement ist Diskriminierung die wesentliche Hürde. Die wesentlichen Qualitätseigenschaften eines Systems mit KI-Komponenten sind demzufolge Safety (Funktionale Sicherheit) im Bereich Produktionsautomatisierung bzw. Fairness im Bereich Personal- und Talentmanagement.

2.2. Bestandsaufnahme existierender technischer Standards

Die Bestandsaufnahme der Normen im zweiten Arbeitspaket [3] listet entsprechend relevante Normen für den Bereich der Software- und Maschinensicherheit, das Personalmanagement und das Testen von Software. Existierende Safety-Normen decken nicht die Verwendung von KI-Komponenten zur Realisierung von Sicherheitsfunktionen ab. Anforderungen für die Entwicklung traditioneller sicherheitskritischer Software sind für die Entwicklung einer KI-Komponente teilweise nicht erfüllbar, nicht sinnvoll und nicht ohne weiteres übertragbar auf vergleichbare Anforderungen. Der Stand der Technik für die Entwicklung einer safetykritischen KI-Komponente wird in aktuellen Safety-Normen, die mit der Maschinenrichtlinie harmonisiert sind und somit eine entscheidende Rolle bei der Markteinführung spielen, nicht abgebildet.

Bezüglich Fairness gibt es keine harmonisierten Normen, die für die Markteinführung ausschlaggebend wären. Es gibt nur eine Reihe von Normen, die sich generell mit ethischen Aspekten bezüglich KI und autonomen Systemen beschäftigen oder generell mit dem Testen von KI.

[1]
Adler, R., Heidrich, J., Jöckel, L., Kläs, M. (2020). Anwendungsszenarien: KI-Systeme in der Produktionsautomatisierung, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

[2]
Zweig, K., Hauer, M., Raudonat, F. (2020). Anwendungsszenarien: KI-Systeme im Personal- und Talentmanagement, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

[3]
Becker, N., Junginger, P., Martinez, L., Krupka, D. (2021). KI in der Arbeitswelt: Übersicht einschlägiger Normen und Standards, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V

2.3. Rechtliche Rahmenbedingungen des Einsatzes von KI-Systemen

Die rechtlichen Rahmenbedingungen werden maßgeblich durch das Einsatzgebiet des KI-Systems geprägt. Im Bereich der Produktionsautomatisierung werden insbesondere das Produktsicherheitsrecht und das Haftungsrecht relevant. Für den Einsatz von KI im Personalmanagement stellen sich die maßgeblichen Fragen hingegen im Bereich des Diskriminierungs- und Datenschutzrechts.

Für die Produktionsautomatisierung zeigt sich, dass das Produktsicherheitsrecht nach seinem gegenwärtigen Stand KI nicht explizit berücksichtigt und nur indirekt erfasst. Das Haftungsrecht kann KI zwar erfassen, jedoch stellen sich hier unter anderem im Einzelfall Fragen hinsichtlich der Sorgfaltspflichten bezüglich KI-Systemen sowie der Beweisführung. In beiden Fällen wird der „Stand der Technik“ erwähnt, zu deren Bestimmung technische bzw. harmonisierte Normen herangezogen werden können.

Bei dem Einsatz im Personalmanagement kann das Allgemeine Gleichbehandlungsgesetz (AGG) KI-Systeme erfassen, setzt aber voraus, dass für die Ungleichbehandlung ein verpöntes Merkmal kausal war. Hier zeigen sich wiederum Beweisschwierigkeiten. Zudem bedarf es der Kenntnis von dem Einsatz eines KI-System, um Betroffenenrechte durchzusetzen. Eine Testpflicht für KI-Systeme existiert de lege lata nicht. Weiterhin lassen sich auch Anforderungen an die Transparenz von KI-Systemen dem geltenden Recht nur schwer entnehmen. Entsprechende Regeln für Transparenz im Datenschutzrecht sind von ihrem Inhalt und ihrer Reichweite derzeit ungeklärt.

2.4. Möglichkeiten von Testing, Auditing und Zertifizierung von KI-Systemen

Das vierte Arbeitspaket untersuchte unterschiedliche Verfahren für das Testen, Auditing und Zertifizieren von KI unter Berücksichtigung der regulatorischen und normativen Rahmenbedingungen.

Im Hinblick auf Safety sollen diese Verfahren verwendet werden, um sicherzustellen, dass eine KI-Komponente keinen Personenschaden verursacht. Dies geht zwangsläufig mit der Frage „Wie sicher ist sicher genug?“ einher, denn Risiken lassen sich nicht vollständig eliminieren. Welche Testverfahren und andere Qualitätssicherungsmaßnahmen müssen wie kombiniert werden, damit die Risiken des sicherheitskritischen Einsatzes einer KI-Komponente akzeptabel sind? Es gibt viele Dokumente, die eine

Übersicht über die zahlreichen Verfahren geben, aber keinen Konsens darüber was ausreicht im Hinblick auf die Sicherheit von Maschinen beziehungsweise die Verwendung von KI in Sicherheitsfunktionen. Welche Auditing-Verfahren ein Maschinenhersteller anwenden kann, ist hingegen aktuell recht klar durch die Maschinenrichtlinie (und zukünftig durch KI-Verordnung und Maschinenverordnung) vorgegeben. Wenn die KI sicherheitskritisch ist, dann bleibt dem Hersteller in der Regel keine andere Wahl, als ein Audit durch Dritte durchführen zu lassen. Wenn die Verwendung von KI im sicherheitskritischen Kontext die Anwendung harmonisierter Normen verhindert und harmonisierte Normen für KI nicht existieren, lässt sich hiermit keine Konformitätsvermutung nach der Maschinenrichtlinie auslösen.

Im Hinblick auf Fairness soll das Testen, Auditing und die Zertifizierung von KI sicherstellen, dass eine KI-Komponente nicht unfair entscheidet. Analog zur Frage „Wie sicher ist sicher genug?“ stellt sich zwar auch die Frage „Wie fair ist fair genug?“, aber die primäre Frage ist zunächst „Was bedeutet fair?“ beziehungsweise „Was ist überhaupt das absolute Ziel?“. Wenn eine sicherheitskritische KI-Komponente einen Arbeiter in einem Gefahrenbereich detektieren soll, dann kann die Aufgabe recht gut spezifiziert werden und es ist eindeutig, wann die KI-Komponente fälschlicher Weise einen Arbeiter nicht detektiert. Wenn eine KI-Komponente die 10 besten Bewerber*innen aus einer langen Liste heraussuchen soll ohne dabei zu diskriminieren, dann ist der Fehlerfall nicht so klar. Für das Auditing von Produkten bezüglich Fairness gibt es keine Richtlinien, die genau festlegen welches Auditing-Verfahren anzuwenden ist. In dem Arbeitspaket wurde deswegen ein Konzept für das Auditing von KI im Personal und Talentmanagement erstellt [4]. Die Möglichkeiten der Zertifizierung von KI sind zwar vorhanden, weil es KI-Normen gibt, aber die Konformität zu diesen Normen ist im Hinblick auf Fairness nicht besonders aussagekräftig, da sie keine Hilfestellung geben wie man in einem bestimmten Anwendungsfall Fairness definiert.

Die Möglichkeiten für Betroffene dem Verdacht einer kausalen Benachteiligung nachzugehen, ohne dass Ihnen Zugriffe auf das System gewährt werden, die über eine normale Nutzung hinaus gehen (externe 3rd party Audits) wurden in diesem AP ebenfalls untersucht [5, 6].

2.5. Herausforderungen und Lösungsansätze

Die Herausforderung besteht darin gesetzlich und normativ festzulegen welche Maßnahmen ausreichen, um eine KI-Komponente in einem kritischen Kontext zu verwenden. Das gilt insbesondere für sicherheitskritische Anwendungen, da die Maßnahmen

[4]

Waltl, B.: Auditieren von KI-Systemen. KI-Audits für Personal- und Talentmanagement, ExamAI - KI Testing & Auditing, Gesellschaft für Informatik, in Veröffentlichung.

[5]

Krafft, T. D., Hauer, M. P., Zweig, K. A. (2020). Why Do We Need to Be Bots? What Prevents Society from Detecting Biases in Recommendation Systems In: Boratto L., Faralli S., Marras M., Stilo G. (eds) Bias and Social Aspects in Search and Recommendation. BIAS 2020. Communications in Computer and Information Science, vol 1245. Springer, Cham.

[6]

Krafft, T. D., Hauer, M. P., Zweig, K. A. : Blackbox-testing and -auditing of unethical bias in ADM systems, in Veröffentlichung.

hier besonders effektiv sein müssen. Die Maßnahmen sollen ein Sicherheitsniveau bieten, das genauso hoch ist wie das Niveau, das für klassische Software gefordert wird. Dieses Niveau lässt sich aber nur schwer erreichen und nur schwer messen. Ein Lösungsansatz um mit dieser Problematik umzugehen liefern strukturierte Sicherheitsargumentationen (Assurance Cases). Die Sicherheitsargumentation basiert auf Testergebnissen und anderen Qualitätssicherungsmaßnahmen wie Sicherheitsanalysen. Sie erklärt, warum die Ergebnisse zeigen, dass das Gesamtrestrisiko akzeptabel ist und macht den Zusammenhang zu den gewählten Risikoakzeptanzkriterien transparent. Sicherheitsnormen könnten festlegen wie die Sicherheitsargumentation aussehen soll, wie sie auditiert werden soll, und wie sie genutzt werden soll, um gezielte Marktbeobachtungen durchzuführen.

Bei Fairnesskritischen Anwendungen von KI sind die Herausforderung anders als bei sicherheitskritischen Anwendungen. Es gibt aktuell keine Normen, die Maßnahmen für die Entwicklung und den Betrieb von Fairnesskritischer Software vorschreiben. Entsprechend besteht auch nicht der Anspruch, genauso effektive Maßnahmen für KI zu haben wie für klassische Software. Im Safety Engineering besteht dieser Anspruch. „Fairness Engineering“ ist hingegen eine sehr junge Engineering Disziplin ohne etablierten Namen. Eine Herausforderung in dieser Disziplin besteht darin den Begriff „Fairness“ im Hinblick auf den Umgang mit algorithmischen Entscheidungssystemen zu schärfen. Dies umfasst insbesondere Methoden um Akzeptanzkriterien für eine bestimmte fairnesskritische Anwendung festzulegen. Ein vielversprechender Lösungsansatz dafür ist die Akzeptanztestgetriebene Entwicklung (Acceptance Test Driven Development, kurz ATDD). Die Grundidee besteht darin, mit Anwendungsexpert*innen, KI-Expert*innen, Ethiker*innen, Rechtswissenschaftler*innen und anderen Stakeholdern so früh wie möglich Akzeptanzkriterien und Akzeptanztests festzulegen. Ein weiterer komplementärer Ansatz besteht darin, Assurance Cases zu nutzen um zu erklären warum Fairness hinreichend garantiert ist. Der Grundgedanke von Assurance Cases ist nicht auf Sicherheitsargumentationen beschränkt und lässt sich auf Fairnessargumentationen übertragen. Die Argumente und Gedanken der verschiedenen Stakeholder im ATDD Prozess werden im Assurance Case aufbereitet um Reviews und Audits zu unterstützen.

2.6. Handlungsempfehlungen

In diesem Arbeitspaket wurden die Lösungsansätze in Expert*innenworkshops vorgestellt, diskutiert und folgende politische Handlungsempfehlungen abgeleitet:

1. Eine wesentliche Handlungsempfehlung ist, Assurance Cases als zentrales Element für das Auditing und die Zertifizierung von KI zu etablieren und zumindest als Übergangslösung zu nutzen bis ausreichend Erfahrung bei der Anwendung von KI in kritischen Bereichen gesammelt wurde. Wenn genügend Erfahrung in speziellen Anwendungen vorhanden ist, dann können diese Erfahrungen als Grundlage für die Entwicklung anwendungsspezifischer Normen genutzt werden.
2. Um Erfahrungen in einem geschützten Umfeld zu sammeln müssen Experimentierfelder (regulatory sandboxes) aufgebaut und Fallstudien durchgeführt werden. Im Automotive-Bereich gibt es bereits viele Experimentierfelder in Deutschland und der Ansatz der Assurance Cases ist zentraler Bestandteil in Fallstudien und Forschungsprojekten wie KI-Absicherung oder Verifikations- und Validierungsmethoden. Die Bereiche Produktionsautomatisierung sowie Personal- und Talentmanagement sollten sich an dieser Vorgehensweise orientieren und die Erfahrungen auf ihren Anwendungsbereich übertragen.
3. Die Methoden- und Werkzeugentwicklung für die Qualitätssicherung von KI muss vorangetrieben und gefördert werden. In sicherheitskritischen Anwendungen ist die Qualifizierung von Werkzeugen unabdingbar und normativ geregelt. Aktuelle KI-Werkzeuge sind aber nicht qualifiziert. Noch wichtiger ist die Entwicklung und empirische Evaluierung von Methoden für die Qualitätssicherung von KI. Sinnvolle gesetzliche und normative Vorgaben für die Anwendung von Qualitätssicherungsmaßnahmen sind nur möglich, wenn es hinreichende empirische Nachweise für die Effektivität der Maßnahmen gibt.
4. Die Mitwirkung von Forschungseinrichtungen in Standardisierungs- und Normungsgremien muss unterstützt werden, um den Stand der Wissenschaft und Technik bezüglich der Qualitätssicherung von KI möglichst schnell in Normen zu verankern. Ansonsten besteht die Gefahr, dass Produkte oder Dienste auf den Markt kommen, deren Qualitätssicherungsniveau nicht so hoch ist und diese Produkte oder Dienste als de-facto Standard beziehungsweise als akzeptable Referenz angesehen werden.
5. Die Umsetzbarkeit der KI-Verordnung muss aus technischer und rechtswissenschaftlicher Perspektive untersucht werden. Die Untersuchung aus technischer Perspektive ist notwendig, um herauszufinden wie die regulatorischen Anforderungen umgesetzt und durch technische Normen adressiert werden sollten. Die Untersuchung aus rechtswissenschaftlicher Perspektive ist notwendig, um die Wechselwirkung mit bestehenden Gesetzen zu verstehen und mögliche Handlungsbedarfe frühzeitig zu erkennen.

3. Möglichkeiten und Grenzen von Anwendungsfällen

Im Folgenden werden die Möglichkeiten und Grenzen von KI-Anwendungen in den Bereichen Pro-duktionsautomatisierung und Personal- und Talentmanagement anhand von 11 konkreten Use Cases zusammengefasst und anschließend miteinander verglichen.

3.1. Anwendungsbereich: KI-Systeme in der Produktionsautomatisierung

Zunächst wurde eine Klassifizierung von Anwendungsbereichen durchgeführt, die hier [7] umfassender beschrieben ist. Gemäß der Klassifizierung kann KI angewendet werden um

1. Prozessplanung und -automatisierung zu optimieren
2. Wartungs- und Designvorschläge zu geben
3. Wissen bereitzustellen
4. Produkt-Qualitätssicherung zu automatisieren
5. Maschinen wie Cobots oder Fahrerlose Transportsysteme (FTS) zu verbessern

Die Klassifizierung nach diesen Bereichen wurde vorgenommen, um eine grobe Analyse möglicher Schäden und Risiken durchführen zu können, denn die Kritikalität des Einsatzes von KI ist die wesentliche Hürde, um das Potenzial von KI voll auszuschöpfen. Besonders hoch ist die Hürde bei Safetykritischen Anwendungen. KI in sicherheitsrelevanten Funktionen birgt aber in einigen Fällen großes Potenzial. Dies gilt insbesondere für Funktionen im Kontext von fahrerlosen Transportsystemen, autonomen mobilen Robotern und mobilen Cobots.
Entsprechend wurden vier Szenarien ausgewählt, in denen KI-basierte sicherheitsrelevante Funktionen versagen. Sie sind in [8] beschrieben:

1. Intelligenter Cobot montiert Klimakompressen falsch
2. Intelligenter Cobot verletzt Arbeiter*in am Auge
3. Diskriminierung bei der Routenplanung von Fahrerlosen Transportfahrzeugen (FTF) und Gabelstaplerfahrer*innen
4. Autonomes FTF fährt Arbeiterin an

[7] Adler, R., Heidrich, J., Jöckel, L., Kläs, M. (2020). Möglichkeiten und Grenzen von Anwendungen künstlicher Intelligenz in der Produktionsautomatisierung, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

[8] Adler, R., Heidrich, J., Jöckel, L., Kläs, M. (2020). Anwendungsszenarien: KI-Systeme in der Produktionsautomatisierung, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V

3.2. Anwendungsbereich: KI-Systeme im Personal- und Talentmanagement

Die im Projekt untersuchten Use Cases von KI im Personal- und Talentmanagement sind in [9] beschrieben. In diesem Bereich kann KI eingesetzt werden im Kontext von

1. automatisierten Vorschlägen auf Personalplattformen
2. Persönlichkeitsbewertungen per Lebenslauf (strukturierte Eingabe, Video, etc.)
3. Backgroundchecks
4. Chatbots zu HR-Fragen (CARL)
5. internem Jobprofilmatching
6. Vorhersagen der Jobkündigungsbereitschaft
7. automatischen Arbeitszeitzuweisung bei Gig-Workern

Für diesen Anwendungsbereich werden in Deutschland und Europa KI-Lösungen nur zurückhaltend eingesetzt. International lassen sich für alle Use Cases Beispiele finden, wodurch auch EU-Bürger*innen betroffen sind oder sein können. Teilweise gibt es bereits Einsätze im Deutschen und Europäischen Raum. Beispielhafte Hürden für die stärkere Verbreitung derartiger Systeme sind die unklare Rechtslage und die oft unzureichende Datengrundlage zum Trainieren einer KI.

3.3. Gemeinsamkeiten und Unterschiede der Anwendungsbereiche

Je kritischer ein Anwendungsfall ist, desto größer ist die Hürde den Anwendungsfall mithilfe von KI zu realisieren. Dies gilt grundsätzlich für beide Anwendungsbereiche, vorausgesetzt, dass die KI-Komponente auch wirklich auf dem kritischen Pfad liegt und somit direkt zur Verletzung eines wesentlichen Rechtsguts beitragen kann.

Unterschiede ergeben sich aus den Rechtsgütern, die verletzt werden können. Im Bereich der Produktionsautomatisierung sind die wesentlichen Hürden aufgrund potenzieller Sachschäden und Personenschäden. Im Bereich Personalmanagement geht es primär um Diskriminierung, Datenschutz und Nachvollziehbarkeit für die Betroffenen (Recht auf Erklärung). Dies lässt sich aber nicht pauscha-lisieren und es sollte immer der konkrete Anwendungsfall im Hinblick auf Risiken untersucht werden.

[9]

Zweig, K., Hauer, M., Raudonat, F. (2020). Anwendungsszenarien: KI-Systeme im Personal- und Talentmanagement, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

4. Bestandsaufnahme existierender technischer Standards

Die Ergebnisse der Bestandsaufnahme existierender technischer Standards und Normen ist in [10] dokumentiert.

[10]

Becker, N., Junginger, P., Martinez, L., Krupka, D. (2021). KI in der Arbeitswelt: Übersicht einschlägiger Normen und Standards, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

Für den Bereich der Produktionsautomatisierung sind in erster Linie Normen relevant, die mit der Maschinenrichtlinie harmonisiert sind. Die Anwendung von harmonisierten Normen kann die Konformitätsvermutung der Maschinenrichtlinie auslösen und die Erfüllung rechtlicher Voraussetzungen hinsichtlich der Einhaltung grundlegender Sicherheits- und Gesundheitsschutzanforderungen vereinfachen. Die Einhaltung dieser Anforderungen ist eine Voraussetzung für die CE-Kennzeichnung von Maschinen und deren Vertrieb in Europa.

Die harmonisierten Sicherheitsnormen adressieren auch sicherheitskritische Software in Maschinen. Sie legen fest, welche Maßnahmen in Abhängigkeit von der Sicherheitskritikalität der Software gewählt werden sollten. Die Maßnahmen wurden aber im Hinblick auf klassische Software geschrieben und sind größtenteils unpassend für KI beziehungsweise datengetriebene Modelle.

Es gibt eine Reihe von Normen bezüglich KI und autonomen Systemen. Die Anwendung dieser Normen ist aber nicht ausreichend, um die Konformitätsvermutung zur Maschinenrichtlinie auszulösen. Sie machen auch keine Vorgaben in Abhängigkeit von der Sicherheitskritikalität der KI.

Für den Anwendungsbereich „Steuerung von Arbeitskarrieren“ gibt es deutlich weniger relevante Normen. Die Qualitätseigenschaft „Fairness“ wird durch Normen kaum reguliert. Bezüglich „Safety“ gibt es die Maschinenrichtlinie mit über 500 harmonisierten Normen. Eine Richtlinie für „Fairness“ und Normen für „fairnesskritische“ Software gibt es nicht. Es entstehen Ethik-Normen für autonome und intelligente Systeme (z.B. die IEEE P7000), aber diese Normen adressieren nicht speziell „Fairness“ oder “Schutz vor Diskriminierung” und sind nicht mit Richtlinien harmonisiert. Außerdem adressieren sie autonome Systeme und nicht KI. Ob ein autonomes System eine Aufgabe mithilfe von KI erledigt oder nicht, ist für viele ethische Fragestellungen irrelevant. Es geht häufig zunächst darum, erstmal so gut wie möglich zu klären was es bedeutet die Aufgabe moralisch gut zu erledigen.

5. Rechtliche Rahmenbedingungen des Einsatzes von KI-Systemen

Im Hinblick auf die rechtlichen Rahmenbedingungen zeichnet sich für die beiden untersuchten Anwendungsszenarien jeweils ein unterschiedliches Bild, maßgeblich bedingt durch die unterschiedlichen KI-spezifischen Risiken des Einsatzes intelligenter Systeme. Für beide Anwendungsszenarien wurde der bestehende Rechtsrahmen mit Blick auf die Besonderheiten von KI-Systemen untersucht und jeweils geprüft, auf welche Eigenschaften entsprechende Systeme getestet bzw. auditiert werden müssen.

5.1. KI-Systeme in der Produktionsautomatisierung

Das maßgebliche Risiko des Einsatzes von KI in der Produktionsautomatisierung – etwa bei Cobots oder fahrerlosen Transportsystemen (FTS) – besteht in der möglichen Schädigung von Personen und Sachen [11]. Bei der rechtlichen Untersuchung, wie diesen Risiken angemessen zu begegnen ist, ist zwischen den Anforderungen des Produktsicherheitsrechts und den Anforderungen des Haftungsrechts zu unterscheiden. Dabei stellt das Produktsicherheitsrecht Anforderungen an das Bereitstellen eines Produktes auf dem Markt, um dessen Sicherheit ex ante zu gewährleisten, während das Haftungsrecht für die Kompensation im Schadensfall einschlägig ist und allein durch die Haftungsandrohung Anreize für Hersteller setzt, möglichst sichere Produkte in Verkehr zu bringen.

[11] Hilpisch, S. T., Kreutzer, A., Sesing, A. (2021). KI-Systeme in der Produktionsautomatisierung – Rechtsfragen im Überblick, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V., in Veröffentlichung.

Im Bereich des Produktsicherheitsrechts zeigen sich Defizite hinsichtlich einer ausdrücklichen Berücksichtigung von KI-Systemen. Der rechtliche Rahmen aus der Maschinenrichtlinie 2006/42/EG und dem Produktsicherheitsgesetz sowie der 9. Produktsicherheitsverordnung enthält keine spezifischen KI-Regelungen. KI-Systeme werden jedoch von generellen Regelungen erfasst, woraus sich zumindest einige allgemeine Anforderungen an KI-Systeme herleiten lassen.

Besondere Bedeutung kommt Testing und Auditing von KI-Systemen im Bereich der Konformitätsbewertung zu, wo die Einhaltung der produktsicherheitsrechtlichen Vorschriften überprüft werden soll. Hier bedarf es des Testings und Auditings, um eine

entsprechende Bewertung durchzuführen, die Erfüllung der Vorschriften nachzuweisen und ein Produkt rechtmäßig auf den Markt bringen zu können. Sollte in diesem Kontext von einer Zertifizierung gesprochen werden, ist es jedoch notwendig, dass Dritte an dieser Bewertung beteiligt werden.

Allerdings könnte das Produktsicherheitsrecht in Zukunft reformiert werden, da es mit einer neuen Maschinenverordnung [12] und einer Verordnung für ein KI-Gesetz [13] Reformbestrebungen gibt, um KI-Systeme mit ihren Charakteristiken zu erfassen. So sollen unter anderem in den Produktanforderungen auch Security- und Transparenzaspekte sowie Kontrollmöglichkeiten bei KI-Systemen berücksichtigt werden. Hierzu gehört auch eine Risikoeinteilung von KI-Systemen, die bei KI-Sicherheitsfunktionen eine Konformitätsbewertung mit Drittbeteiligung verlangt. Offen ist hierbei jedoch, wie eine klare Trennung zwischen einer KI als Sicherheitsfunktion und einer KI für die Nominalfunktion eines Systems erfolgen soll.

[12]

Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über Maschinenprodukte vom 21.4.2021, COM(2021) 202 final.

[13]

Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union vom 21.04.2021, COM(2021) 206 final.

Für das Haftungsrecht zeigten sich unterschiedliche Herausforderungen hinsichtlich der Verantwortung für Schäden, die durch ein KI-System hervorgerufen werden. De lege lata ist das Haftungsrecht bereits in der Lage, grundlegend KI-Systeme zu erfassen, wobei sich im Einzelnen verschiedene Fragestellungen ergeben, die weiterer Forschung bedürfen.

Zunächst sind bei dem Einsatz eines KI-Systems verschiedene Parteien beteiligt, was die Anzahl möglicher Haftungsadressaten und damit die Komplexität der Schadenskompensation erhöht. Eine Zurechnung des Verhaltens bzw. eines Verschuldens von KI-Systemen scheidet dabei nach geltender Rechtslage aus.

Für eine potenzielle Haftung muss somit weiterhin an das Verhalten des Haftungsadressaten angeknüpft werden. Hierbei wird mittelbar über das Instrument der Sorgfaltspflichten möglicher Haftungsadressaten bestimmt, über welche Eigenschaften ein KI-System verfügen muss. Dabei ist teilweise ungeklärt, welche Anforderungen das Haftungsrecht an die potenziellen Haftungsadressaten im Hinblick auf den Umgang mit KI-Systemen im Einzelnen stellt. Die Erfüllung möglicher Sorgfaltspflichten könnte materiell daran anknüpfen, nur getestete bzw. auditierte Systeme überhaupt zum Einsatz zu bringen. Daneben kann Testing und Auditing im Bereich der Beweisführung innerhalb des Haftungsrechts von essenzieller Bedeutung sein. So kann Testing und Auditing beispielsweise ein Verschulden beweisen oder dieses widerlegen, wobei bislang weitgehend ungeklärt ist, welche konkreten Anforderungen an diesen Nachweis gestellt werden.

Sowohl im Produktsicherheitsrecht als auch im Haftungsrecht wird der „Stand der Technik“ relevant. Zur näheren Bestimmung des „Stand der Technik“ können harmonisierte bzw. technische Normen herangezogen werden. Hier existiert somit aus rechtlicher Sicht Bedarf an diesen Normen.

Es zeigt sich somit bei dem Einsatz von KI-Systemen in der Industrieproduktion, dass Testing und Auditing eine wichtige Rolle im Bereich der Beweisführung einnehmen können. So erlauben sie es einerseits, Produkte zu bewerten und in den Verkehr zu bringen und andererseits, einen Anspruch im Schadensfall erfolgreich geltend zu machen.

5.2. KI-Systeme im Personal- und Talentmanagement

Bei dem Einsatz von KI-Systemen im Personalmanagement liegt das wesentliche Risiko in einer möglichen Diskriminierung aufgrund verzerrter Daten („bias in the data“) und in Persönlichkeitsrechtsverletzungen durch ein KI-System [14]. Hierzu kann festgestellt werden, dass *de lege lata* die Vorschriften des Allgemeinen Gleichbehandlungsgesetzes (AGG) auch im Falle des Einsatzes von KI-Systemen uneingeschränkt Anwendung finden. Des Weiteren ist der Einsatz von KI-Systemen für das Datenschutz- und Diskriminierungsrecht bezüglich Maßstab und Rechtsdurchsetzung problematisch.

[14]

Siehe zu dem Bereich Personal- und Talentmanagement Hoffman, R., Sesing, A., Borges, G. (2021). KI-Systeme im Personal- und Talentmanagement – Rechtsfragen im Überblick, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V., in Veröffentlichung.

Fraglich ist dabei, ob der „klassische“ Diskriminierungsbegriff auf KI angewendet werden kann, wenn das KI-System den Grund für seine Entscheidung nicht preisgibt bzw. die Entscheidung nicht erklärbar ist. Darüber hinaus zeigte sich der Bedarf an Testing und Auditing von KI-Systemen an Fragen der Beweisführung. So mag zwar das Opfer einer durch KI verursachten Diskriminierung Kenntnis vom Einsatz des KI-Systems erlangen (was für sich bereits zweifelhaft erscheint), aber es muss zumindest Indizien dafür vorbringen, dass es zu einer kausalen Diskriminierung durch das System gekommen ist. Ein zentrales Problem der Anwendung des AGG besteht darin, dass das AGG ausschließlich Ungleichbehandlungen erfasst, für die ein sog. „verpöntes“ Merkmal (z.B. Geschlecht, Herkunft oder sexuelle Orientierung) (mit-)ursächlich, also kausal ist. Da die Kausalität von Einzelmerkmalen jedoch bei datengetriebenen KI-Systemen nach bisherigem Stand weder sicher nachgewiesen noch sicher ausgeschlossen werden kann, da entsprechende Systeme an Korrelationen anknüpfen, kann dies erhebliche Beweisproblematiken erzeugen und die Rechtsdurchsetzung behindern. Hiergegen könnten eine Testpflicht (auch für „eingriffsvorbereitende Systeme“) für KI-Systeme

zur Bewältigung dieser Herausforderungen eingesetzt werden [15]. *De lege lata* ist eine explizite Testpflicht für KI-Systeme nicht ausdrücklich geregelt. Darüber hinaus erscheint es unklar, ob und woher sich eine solche Testpflicht überhaupt aus dem geltenden Recht herleiten ließe. Die dafür denkbaren Anhaltspunkte (bspw. § 241 Abs. 2 Bürgerliches Gesetzbuch (BGB), §§ 823 ff. BGB, sowie § 12 AGG) können die KI-spezifischen Sorgfaltsmaßstäbe derzeit nicht sachgerecht abbilden.

[15]

Sesing, A./Tschech, A., Vermeidung von Diskriminierung durch KI – Rechtliche Ankerpunkte und Ausblick auf die KI-Regulierung der EU, in: Arbeitspapier: Diskriminierende KI? Risiken algorithmischer Entscheidungen in der Personalauswahl, Gesellschaft für Informatik e.V. (GI)

Zudem zeigen sich Probleme hinsichtlich rechtlich geforderter Transparenz bei KI-Assistenzsystemen. Entsprechende Pflichten lassen sich dem Recht kaum entnehmen. Wichtigster Anknüpfungspunkt ist hier Art. 22 der Datenschutz-Grundverordnung (DS-GVO), wobei dessen genauer Regelungsinhalt und der Anwendungsbereich noch einer Klärung bedürfen. Nach derzeit herrschender Lesart bezieht sich diese Regelung ausschließlich auf den Einsatz von KI-Systemen, die vollständig autonom und automatisiert Entscheidungen mit Wirkung für Individuen treffen; demgegenüber sollen bloß vorbereitende Systeme („human in the loop“) nicht adressiert sein. Dessen ungeachtet fragt sich, wie weit etwaige Transparenz- und Auskunftspflichten reichen; dies hängt nicht zuletzt mit der Unsicherheit beim Begriff der „involvierten Logik“ (Art. 13 Abs. 2 lit. f), 14 Abs. 2 lit. g), 15 Abs. 1 lit. h) DS-GVO) und der Abwägung mit gleichermaßen gewichtigen Interessen des Schutzes von Geschäftsgeheimnissen zusammen.

Im Ergebnis zeigte sich bei dem Einsatz von KI im Personalmanagement, dass im Rahmen des Diskriminierungsrechts KI Systeme grundlegend erfasst sind, wobei sich hier Beweisprobleme ergeben. Daneben finden sich Regeln zur Transparenz von KI-Systemen im Wesentlichen im Datenschutzrecht, aber auch hier ist die genaue Reichweite der geforderten Transparenz nicht zweifelsfrei geklärt.

Durch die geplante KI-Verordnung soll eine risikobasierte Klassifizierung von KI-Systemen eingeführt werden, wobei KI-Anwendungen im Bereich des Personalmanagements überwiegend als Hochrisiko-KI eingestuft werden. Somit würden besondere Anforderungen – unter anderem hinsichtlich Testing und Transparenz – an KI im Personalmanagement gestellt werden. Hierbei besteht jedenfalls potenziell das Risiko, dass die umfassende Erfassung aller KI-Systeme durch den Entwurf zu einer Überregulierung führt.

Zusammenfassend lässt sich festhalten, dass der Einsatz von KI-Systemen in der Industrieproduktion und im Personalmanagement in unterschiedlichen Rechtsgebieten eine Vielzahl an Fragestellungen aufwirft. Vieles ist hiervon noch ungeklärt und wird sich durch die derzeitigen europäischen Gesetzgebungsbestrebungen in der Zukunft nochmals verändern.

6. Möglichkeiten von Testing, Auditing und Zertifizierung von KI-Systemen

6.1. Möglichkeiten, KI zu testen

In diesem Arbeitspaket wurde eine Übersicht zu Blackbox-Testverfahren und -konzepten erstellt [16, 17] und eine Terminologie erarbeitet, um den Austausch zwischen Testwissenschaftler*innen und Datenwissenschaftler*innen zu vereinfachen [18].

Diese und andere Arbeiten zeigen, dass es sehr viele Möglichkeiten gibt, um zu testen, ob eine KI-Komponente sich so verhält wie sie soll. Die große Herausforderung besteht darin festzulegen wann welche Möglichkeit wie genutzt werden soll. Es mangelt an Guidelines, Standards und Normen, die diese Frage beantworten.

Eine besondere Herausforderung bezüglich „Safety“ besteht dabei darin, dass man sehr geringe Fehlerraten nachweisen muss. Traditionell wird dieser Nachweis nur für zufällige Hardwarefehler geführt aber nicht für Softwarefehler. Im Hinblick auf datengetriebene Modelle werden aber zunehmend auch probabilistische Nachweise diskutiert. Es ist allerdings kaum möglich, mit hinreichender statistischer Signifikanz Fehlerraten von bis zu 10-9 Fehler pro Stunde für eine KI-Komponente nachzuweisen.

Im Hinblick auf Fairness liegt das Problem vielmehr bei der Definition der Zielsetzung. Wenn es keinen Konsens darüber gibt, was eine „faire“ Entscheidung ist, dann kann man grundsätzlich nicht überprüfen, ob eine KI-Komponente „fair“ entscheidet. Auf diese Herausforderungen wird detaillierter in Kapitel 7 eingegangen.

6.2. Möglichkeiten des Auditings von KI

Eine Veröffentlichung [19] aus dem Projekt widmete sich dem Auditieren von KI-Systemen mit besonderem Schwerpunkt auf KI-Audits im Personal- und Talentmanagement. Dabei wurde gemäß [20] zwischen folgenden Möglichkeiten unterschieden:

[16]

Krafft, T. D., Hauer, M. P., Zweig, K. A. (2020). Why Do We Need to Be Bots? What Prevents Society from Detecting Biases in Recommendation Systems In: Boratto L., Faralli S., Marras M., Stilo G. (eds) Bias and Social Aspects in Search and Recommendation. BIAS 2020. Communications in Computer and Information Science, vol 1245. Springer, Cham.

[17]

Krafft, T. D., Hauer, M. P., Zweig, K. A. : Blackbox-testing and -auditing of unethical bias in ADM systems, in Veröffentlichung.

[18]

Jöckel, L., Bauer, T., Kläs, M., Hauer, M. P., Groß, J. (2021). Towards a Common Testing Terminology for Software Engineering and Data Science Experts, 22nd International Conference on Product-Focused Software Process Improvement (Profes 2021), Turin, Italy, 2021.

- 1st party Audit oder Self-assessment: Das Audit wird durch die Organisation durchgeführt, die auch für die Entwicklung verantwortlich ist.
- 2nd party Audit (internes Audit): Das Audit wird durch einen Kunden, Lieferanten, oder einer anderen Partei, die in einer Beziehung zur Organisation steht, verantwortet.
- 3rd party Audit (internes oder externes Audit): Das Audit wird durch eine unabhängige Organisation durchgeführt. Hat diese Organisation Einsicht, bzw. Zugriff auf interne Informationen und Prozesse handelt es sich dabei um ein internes Audit. Das ist zum Beispiel bei akkreditierten Zertifizierungsverfahren der Fall. Hat die Organisation keine Einsichten oder Zugriffe, handelt es sich um ein externes Audit. In diesem Fall spricht man auch von einem Black Box Audit.

In der Europäischen Union bildet der „New Approach“ zusammen mit dem „New Legislative Framework (NLF)“ den Rahmen für das Auditing von Produkten. Dieser Rahmen bestimmt die Regeln für die Konformitätsbewertung, bei der geprüft wird, ob ein Produkt den EU-Vorschriften und -Normen entspricht. Außerdem legt der Rahmen fest, welche Organisationen überhaupt prüfen dürfen – und wer die Prüfenden überprüft.
Für den Einsatz von KI in der Produktionsautomatisierung ist die Konformitätsbewertung zur Maschinenrichtlinie von besonderer Bedeutung. Cobots, fahrerlose Transportsysteme und viele andere Systeme fallen in den Anwendungsbereich der Maschinenrichtlinie.

[19]

Waltl, B.: Auditieren von KI-Systemen. KI-Audits für Personal- und Talentmanagement, ExamAI - KI Testing & Auditing, Gesellschaft für Informatik e.V., in Veröffentlichung.

[20]

ISO/IEC 17000:2020(en) Conformity assessment — Vocabulary and general principles

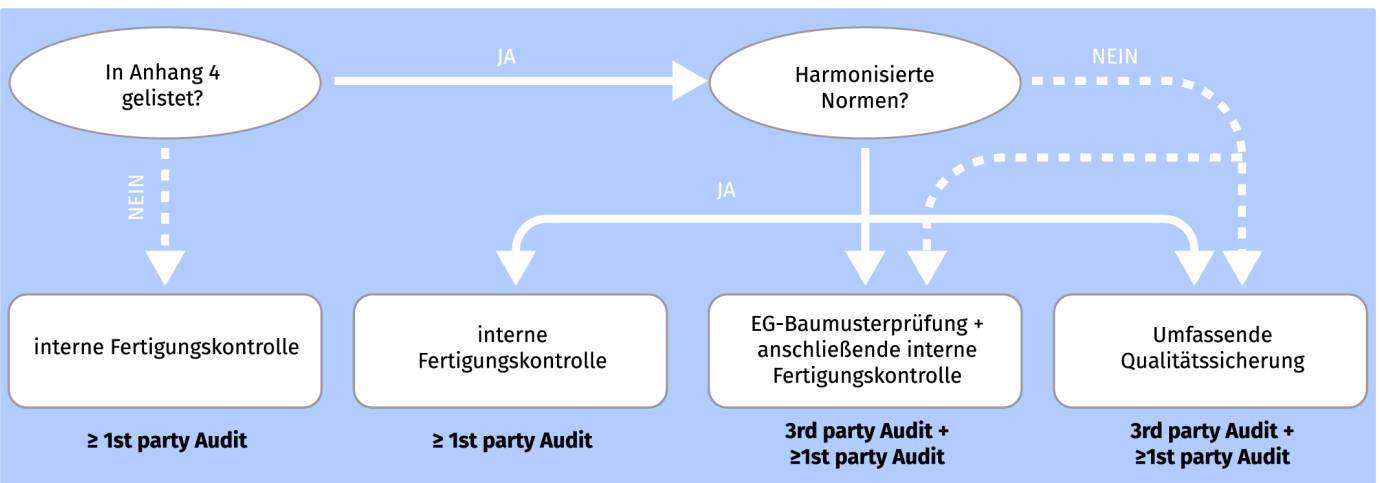

Abbildung 2 – Mögliche Konformitätsbewertungsverfahren für Maschinen

Das Auditing bezüglich der Konformität zur Maschinenrichtlinie ist in Abbildung 2 dargestellt. Welches Auditing-Verfahren ein Hersteller wählen darf hängt davon ab,

ob das Produkt in Anhang 4 gelistet ist und ob es harmonisierte Normen gibt. Anhang 4 beinhaltet eine Auflistung „besonders gefährlicher Maschinen“. Wenn das Produkt nicht zu diesen besonders gefährlichen Maschinen zählt, dann darf der Hersteller eigenständig das Konformitätsbewertungsverfahren durchführen (1st party Audit). Die Maschinenrichtlinie verlangt, dass alle erforderlichen Maßnahmen ergriffen werden, damit durch den Herstellungsprozess gewährleistet ist, dass die hergestellten Maschinen konform sind. Die notwendigen Maßnahmen werden als „interne Fertigungskontrollen“ bezeichnet. Die interne Fertigungskontrolle kann der Hersteller auch wählen, wenn das Produkt in Anhang 4 gelistet ist, aber es harmonisierte Normen gibt. Wenn es keine harmonisierten Normen gibt, dann muss er eine „EG-Baumusterprüfung mit anschließender Fertigungskontrolle“ wählen oder eine „umfassende Qualitätssicherung“ wählen. Beide Varianten gehen mit einem 3rd party Audit einher.

Wird ein KI-System nun bei einer bereits gefährlichen Maschine aus Anhang 4 verwendet, kann der Hersteller nur eine interne Fertigungskontrolle als Bewertungsverfahren wählen, wenn er harmonisierte Normen anwendet, die alle einschlägigen grundlegenden Sicherheits- und Gesundheitsschutzanforderungen abdecken. Ansonsten hat der Hersteller eine EG-Baumusterprüfung mit anschließender Fertigungskontrolle oder eine umfassende Qualitätssicherung durchzuführen. Wird das KI-System nun in einem sicherheitskritischen Teil der Maschine verwendet, sind die derzeitigen harmonisierten Normen nicht mehr anwendbar, da sie datengetriebene Modelle nicht berücksichtigen oder explizit ausschließen. In diesem Fall muss daher bei der Verwendung eines KI-Systems stets auf eine dritte Stelle zur Bewertung zurückgegriffen werden. Sollte das KI-System nicht im sicherheitskritischen Teil angewandt werden, lassen sich die derzeitigen harmonisierten Normen anwenden. Der Hersteller kann (unter der Voraussetzung, dass alle grundlegenden Anforderungen durch harmonisierte Normen abgedeckt sind) somit auch auf die interne Fertigungskontrolle zurückgreifen.

Während der Laufzeit des Projektes hat die EU-Kommission einen Vorschlag für eine neue Maschinenverordnung veröffentlicht. In dem Projekt wurden die Änderungen herausgearbeitet und aus technischer sowie rechtlicher Perspektive betrachtet. Ein wesentlicher Diskussionspunkt war dabei die Bestimmung des anzuwenden Konformitätsbewertungsverfahrens. Die Bestimmung welches Konformitätsbewertungsverfahren anzuwenden ist, ist nach dem neuen Vorschlag unabhängig von der Anwendung harmonisierter Normen. Generell wird dadurch übersichtlicher welches Verfahren anzuwenden ist. Der Hersteller ist nun nicht mehr dem Risiko ausgesetzt, dass die Normen nicht vollumfänglich die Anforderungen abdecken und er so ein falsches Bewertungsverfahren durchführt. Bei einer Hochrisiko-Maschine muss zwingend eine dritte

Partei beteiligt werden, selbst wenn die entsprechende Maschine vollumfänglich nach harmonisierten Normen konstruiert wurde. Eine solche Prüfung kann sich positiv auf die Sicherheit auswirken, geht aber zugleich mit einer Mehrbelastung des Herstellers und der Bewertungsstellen einher. Es stellt sich somit die Frage, ob die Mehrbelastung in einem gesunden Verhältnis zum Nutzen steht. Hersteller haben ein hohes Eigeninteresse daran, dass ihre Produkte sicher sind. Die Aufwände um diesen eigenen Anspruch zu genügen sollten nicht reduziert werden nur weil neue Aufwände für verpflichtende Audits benötigt werden.

Bezüglich Fairness und Erklärbarkeit gibt es bisher keine festgelegten Auditverfahren. Eine Veröffentlichung über KI-Audits, die sich auf diese Aspekte fokussiert wurde vorbereitet, konnte jedoch noch nicht abgeschlossen werden.

6.3. Zertifizierung von KI

Zertifizierung bezieht sich auf die Bestätigung, dass ein Produkt, Prozess, ein System, eine Person oder eine Stelle festgelegte Anforderungen erfüllt [21]. Grundlage für Zertifizierung sind somit festgelegte Anforderungen. Die KI-Verordnung legt einige Anforderungen fest aber diese Anforderungen sind viel zu grob im Hinblick auf zahlreiche technische Designentscheidungen. Die Anforderungen müssen durch entsprechende Standards und Normen konkretisiert werden. Diese Standards und Normen gibt es allerdings noch nicht und der europäische Regulierungsvorschlag ist noch nicht rechtskräftig. Somit gibt es aktuell keine verbindlichen, KI-spezifischen Anforderungen.

[21] ISO/IEC 17000:2020(en) Conformity assessment — Vocabulary and general principles

Der Ergebnisbericht [22] von AP 2 listet allerdings bereits einige Dokumente aus der Standardisierung, die Anforderungen bezüglich KI formulieren. Die Erfüllung dieser Anforderungen ist aber nicht aussagekräftig bezüglich Sicherheit und Fairness. Die Frage „Welche Maßnahmen sind ausreichend um gefährliches Verhalten einer KI-Komponente zu adressieren?“ wird nicht beantwortet. Die Frage „Welche Maßnahmen sind ausreichend um unfaires Verhalten einer KI-Komponente zu adressieren?“ wird nicht ebenfalls nicht beantwortet.

[22] Becker, N., Junginger, P., Martinez, L., Krupka, D. (2021). KI in der Arbeitswelt: Übersicht einschlägiger Normen und Standards, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

7. Herausforderungen und Lösungsansätze

Die zentrale Herausforderung im Hinblick auf die funktionale Sicherheit besteht darin, festzulegen, welche Sicherheitsmaßnahmen ausreichen, um eine KI-Komponente in einem sicherheitskritischen Kontext zu verwendet.

Im Hinblick auf Fairness gibt es zunächst die Herausforderung, den Begriff „Fairness“ zu operationalisieren oder Methoden an die Hand zu geben, um ihn bezüglich einer konkreten Anwendung zu schärfen. Im Zusammenhang damit steht die Herausforderung, Kriterien zu definieren, die messbar machen, ob das intendierte Maß an Fairness erreicht wurde. Im Folgenden wird auf Herausforderungen und Lösungsansätze bezüglich Sicherheit und Fairness eingegangen. Anschließend werden Gemeinsamkeiten und Unterschiede diskutiert.

7.1. Herausforderungen sicherheitskritischer KI

Für sicherheitskritische Software in Maschinen gibt es bereits etablierte Normen und Regeln die festlegen, welche Maßnahmen ausreichen. KI ist Software; zumindest gemäß der Definition des europäische Regulierungsvorschlags. Demnach sollten die existierenden Anforderungen für sicherheitskritische Software auch für sicherheitskritische KI gelten. Diese Anforderungen sind aber nicht sinnvoll und/oder nicht erfüllbar für KI. Neue Anforderungen für KI abzuleiten, die ein vergleichbares Sicherheitsniveau bieten, ist aus verschiedenen Gründen eine große Herausforderung. Um diese Herausforderung zu verstehen, wird im Folgenden zunächst die generelle Vorgehensweise bei der sicherheitsgerichteten Entwicklung einer Maschine beschrieben.

Aktuelle Vorgehensweise – „Maschinensicherheit“ (DIN EN ISO 12100)

Die folgende Vorgehensweise ist in der DIN EN ISO 12100 beschrieben.

Zunächst werden die Grenzen der Maschine inklusive der bestimmungsgemäßen Verwendung festgelegt. Eine Maschine kann dabei aus mehreren miteinander verbundenen Teilen oder Baugruppen bestehen und gilt somit auch für Maschinenanlagen. Die Art der Verbindung ist nicht weiter definiert aber umfasst grundsätzlich auch drahtlose Verbindungen. Für fahrerlose Transportfahrzeuge (FTF) bedeutet dies beispielsweise, dass das ganze fahrerlose Transportsystem mit allen Transportfahrzeugen als eine Maschine angesehen werden kann. Auch Kameras an den Decken könnten Teil dieser Maschine sein, um Kollisionen besser zu vermeiden. Wenn in einer Fabrik alle Arbeitsabläufe dynamisch miteinander gekoppelt werden und dazu alle Systeme miteinander vernetzt sind, dann wäre es sogar denkbar, dass eine ganze Fabrik als eine Maschine angesehen wird. Es stellt sich allerdings die Frage, ob der bisherige Ansatz der Maschinensicherheit gemäß DIN EN ISO 12100 noch ausreichend ist für so „große“ Maschinen und die Bezeichnung Maschine noch treffend ist. Das Industrielle Internet der Dinge (IoT) erschwert die Festlegung der Grenzen der Maschine. Diese Herausforderung kommt zwar primär durch die Vernetzung der Systeme, wird aber trotzdem bei den Lösungsvorschlägen mit betrachtet, da die Vernetzung viele KI-Anwendungen erst ermöglicht und die Möglichkeiten durch KI auch ein Treiber für die Vernetzung sind.

Im nächsten Schritt werden Gefährdungen beziehungsweise „potenzielle Schadensquellen“ identifiziert. Dazu zählen mechanische Gefährdungen, elektrische Gefährdungen, thermische Gefährdungen, Gefährdung durch Lärm, Gefährdung durch Schwingung, Gefährdung durch Strahlung, Gefährdung durch Materialen und Substanzen, und weitere, die in der DIN EN ISO 12100 gelistet sind. KI (und Software generell) ist keine Gefährdung und wird auch nicht zu neuen Gefährdungen führen. Dieser Schritt ist somit weitgehend unabhängig von KI.

Anschließend wird für jede Gefährdung das Risiko abgeschätzt und bewertet. Wenn das Risiko akzeptabel ist, dann sind keine weiteren Schritte notwendig. Wenn es inakzeptabel ist, dann muss ein Sicherheitskonzept entwickelt werden oder das bereits vorhandene Sicherheitskonzept verbessert werden. Die Abschätzung und Bewertung der Risiken berücksichtigt alle Schutzmaßnahmen im Sicherheitskonzept. Dies gilt auch für softwarebasierte Funktionen wie eine Kollisionsvermeidung. Die Risikobewertung in diesem Schritt berücksichtigt dabei die fehlerfreie Funktion. Risiken aufgrund von Fehlfunktionen werden dediziert in Normen der funktionalen Sicherheit wie der IEC 61508 behandelt.

Die Sicherheitskonzeptentwicklung folgt einem 3-Stufenmodell. Im Folgenden werden die 3 Stufen und der Bezug zu KI erläutert.

Die erste Stufe betrifft die inhärent sichere Konstruktion. Hier wird versucht Gefährdungen zu vermeiden oder Risiken zu vermindern durch eine geeignete Auswahl von Konstruktionsmerkmalen der Maschine selbst und/oder Wechselwirkungen zwischen den gefährdeten Personen und der Maschine. Bezüglich der Risiken aufgrund von Kollisionen könnte beispielsweise das Risiko durch eine Polsterung und eine Reduktion des Fahrzeuggewichtes erzielt werden. Auch eine Limitierung der Antriebskräfte durch einen kleinen Motor würde zum inhärent sicheren Design zählen. Softwarebasierte und somit auch KI-basierte Ansätze gehören zur zweiten Stufe des 3-Stufenmodells.
In dieser zweiten Stufe werden technische Schutzmaßnahmen und ergänzende Schutzmaßnahmen festgelegt. Die technischen Schutzmaßnahmen können auf Software basieren. Typischerweise sind es sehr primitive Sicherheitsmechanismen wie beispielsweise eine Abschaltfunktion basierend auf einer Lichtschranke oder einem manuellen Notaus-Schalter. Es können aber auch ausgeklügelte Funktionen sein, die auf KI basieren. Beispielsweise eine Kollisionsvermeidung von einem mobilen Roboter, die Personen erkennt und ausgeklügelte Ausweichmanöver umsetzt. Neuronale Netze könnten hier verwendet werden, um Objekte zu klassifizieren und Personen zu erkennen. Die Kollisionsvermeidung könnte das Risiko möglicher Ausweichmanöver im laufenden Betrieb bestimmen und bewerten, um ein passendes Ausweichmanöver auszuwählen. In diesem Kontext könnten Bayessche Netze eingesetzt werden, um die Wahrscheinlichkeit einer Kollision zu bestimmen. Ein anderes Beispiel ist eine Funktion, die konstruktiv versucht gefährliche Situationen zu vermeiden, indem Aufgaben so zugewiesen werden, dass sich Arbeiter*innen und mobile Roboter möglichst selten begegnen. Dazu müssen die Arbeiter*innen und ihre Position erkannt werden. Auch für diese Personenerkennung ist der Einsatz von neuronalen Netzen naheliegend. Die DIN EN ISO 12100 berücksichtigt aber nicht welche Art von Software im Spiel ist. Sie verweist beim Thema Software auf die IEC 61508 und fordert im Wesentlichen nur, dass die „Wahrscheinlichkeit von zufälligen Ausfällen der Hardware und von systematischen Ausfällen, die die Leistung der sicherheitsbezogenen Steuerfunktion(en) beeinträchtigen können, ausreichend gering ist“.

In der dritten Stufe des 3-Stufenmodells werden Benutzerinformation hinsichtlich des Restrisikos gegeben. Diese Benutzerinformationen dürfen dabei kein Ersatz für die korrekte Anwendung der inhärent sicheren Konstruktion, der technischen Schutzmaßnahmen oder der ergänzenden Schutzmaßnahmen sein. Auf besondere Risiken aufgrund von KI in technischen Schutzmaßnahmen hinzuweisen, ist somit nur ausreichend, wenn diese Risiken nicht durch inhärent sichere Konstruktion eliminiert oder durch klassische Schutzmaßnahmen ohne KI minimiert werden können. Im Hinblick auf die zunehmende Digitalisierung von Dokumenten und Informationen könnten die

Benutzerinformation auch mit mithilfe von Software gegeben werden. Auch in diesem Kontext ist der Einsatz von KI vorstellbar.

Die generelle Vorgehensweise in der DIN EN ISO 12100 muss im Hinblick auf KI nicht angepasst werden, da die Norm nicht auf das Thema Software eingeht und stattdessen auf eine andere Norm verweist. Lediglich bei der Anwendung des 3-Stufenmodells sollte darauf geachtet werden, dass man technischen Schutzmaßnahmen zwar so einfach wie möglich hält aber nicht kategorisch KI-basierte Schutzmaßnahmen ausschließt. Insbesondere Funktionen, die im fehlerfreien Fall Risiken reduzieren und im Fehlerfall kein zusätzliches Risiko bergen sollte man nicht ausschließen, nur weil absehbar ist, dass man diese Funktion nur mithilfe von KI implementieren kann.

Aktuelle Vorgehensweise – „Funktionale Sicherheit“ (IEC 615098)

Die wesentliche Herausforderung betrifft die Vorgehensweise bei der funktionalen Sicherheit. Die funktionale Sicherheit ist der Teil der Gesamtsicherheit, der von der korrekten Funktion des elektrisch/elektronisch/programmierbar elektronischen sicherheitsbezogenen Systems und anderer risikomindernder Maßnahmen abhängt. Diese generelle Vorgehensweise bei der funktionalen Sicherheit berücksichtigt KI nicht und ist auch nicht ohne weiteres anpassbar. Vereinfacht besteht die Vorgehensweise aus den folgenden 2 Schritten:

1. Bestimmung des Sicherheitsintegritätslevels
2. Auswahl der Maßnahmen gemäß des Sicherheitsintegritätslevels

Der erste Schritt ist unabhängig von der Art der Software und somit auch für KI möglich. Der zweite Schritt betrifft Maßnahmen gegen systematische Fehler und zufällige Hardwarefehler, wobei Softwarefehler generell als systematische Fehler angesehen werden. Für die verschiedenen Entwicklungsschritte bei der Softwareentwicklung wird festgelegt, welche Maßnahmen in Abhängigkeit von dem Integritätslevel getroffen werden müssen. Dies wird typischerweise mithilfe von Tabellen dargestellt. Wie in Tabelle 1 dargestellt wird in Abhängigkeit von dem Sicherheitsintegritätslevel (SIL) eine bestimmte Maßnahme stark empfohlen (HR für Highly Recommended), empfohlen (R für Recommended), keine Empfehlung gegeben (--) oder nicht empfohlen (NR – Not Recommended).

	Technique / Measure	SIL 4	SIL 4	SIL 4	SIL 4
1	technique_1a	—	Recommended	Highly Recommended	Highly Recommended
2	technique_1b	Recommended	Recommended	Highly Recommended	Highly Recommended

Tabelle 1 – Allgemeine Darstellung wie Maßnahmen in Abhängigkeit vom SIL empfohlen werden

Die Tabellen gibt es für die Entwicklungsschritte von klassischer Software aber nicht für KI. Es gibt beispielsweise keinen Entwicklungsschritt “Trainingsdaten sammeln”. Es können zwar einige Maßnahmen berücksichtigt werden, aber viele Maßnahmen passen nicht. Dies wird insbesondere bei Maßnahmen für das Softwaredesign und die Implementierung deutlich. Eine Tabelle in der IEC 61508 bezieht sich beispielsweise auf die Auswahl von geeigneten Programmiersprachen. Dabei geht es um Sprachen, mit der die Lösung für das Problem explizit programmiert und nicht aus Daten gelernt wird. Dies ist nur eins von vielen möglichen Beispielen an denen deutlich wird, dass die Maßnahmen nicht alle anwendbar oder ungeeignet sind für KI-basierte Sicherheitsfunktionen. Insgesamt sind aktuelle Sicherheitsstands ungenügend für die Gewährleitung der funktionalen Sicherheit einer KI-basierten Sicherheitsfunktion.
Ein Ansatz um den Mangel an geeigneten Maßnahmen zu adressieren, besteht darin neue Maßnahmen für die verschiedenen Entwicklungsschritte bei der Entwicklung von KI definieren. Diese neuen Maßnahmen bezüglich KI sollten, dann aber zur funktionalen Sicherheit in gleichem Maße beitragen wie die bisherigen Maßnahmen. Die Effektivität der Maßnahmen ist aber weder gut vergleichbar noch gut messbar. Eine Möglichkeit die Effektivität der Maßnahmen abzuschätzen besteht darin, Rückschlüsse aus Feldbeobachtungen zu ziehen und systematisch die Ursache von Unfällen zu analysieren. Diese empirische Vorgehensweise eignet sich aber nur sehr bedingt um die Effektivität von einzelnen Maßnahmen zu messen. Deswegen führt dieser Ansatz in eine Sackgasse.

Anstatt die tatsächliche Effektivität von den aktuellen Maßnahmen zu messen, könnte man sich auch an probabilistischen Zielvorgaben für zufällige Fehler orientieren. Die IEC 61508 definiert für Sicherheitsfunktionen in Abhängigkeit vom Integritätslevel zu erreichende Zielfehlerraten und Zielfehlerwahrscheinlichkeiten. Aktuell werden diese probabilistischen Zielwerte nur im Hinblick auf zufällige Hardwarefehler berücksichtigt. Für klassische Software werden diese Werte nicht berücksichtigt, da Softwarefehler grundsätzlich als systematische Fehler angesehen werden und somit nicht statistisch beschrieben werden. Im Hinblick auf datengetriebene Modelle zeichnet sich allerdings ein Paradigmenwechsel ab. Die Anwendungsregel VDE-AR-E 2842-61 führt eine neue

Fehlerart neben systematischen und zufälligen Fehlern ein. Diese Fehler beziehen sich auf die Unsicherheit mit der eine KI-Komponente eine korrekte Ausgabe liefert. Die Unsicherheit ist eine probabilistische Größe und bietet die Möglichkeit quantitativ zu zeigen, dass die Zielfehlerraten und Zielwahrscheinlichkeiten erreicht werden. Die Unsicherheit der Ausgabe einer KI-Komponente kann während des Betriebs bestimmt werden und entsprechende Sicherheitsmechanismen könnten das Systemverhalten so anpassen, dass die funktionale Sicherheit trotz dieser Unsicherheit gewährleistet ist [23]. Eine hohe Unsicherheit ist somit grundsätzlich unproblematisch für die Sicherheit. Der kritische Fehler ist das unterschätzen der Unsicherheit. Dies wirft die Frage auf mit welcher Wahrscheinlichkeit die Unsicherheit unterschätzt werden darf, da es immer eine gewisse Wahrscheinlichkeit gibt, wenn die Unsicherheit nicht 100% ist. Diese Frage lässt sich nicht so einfach beantworten, da es für die Zielfehlerwahrscheinlichkeit keine weitere Wahrscheinlichkeit gibt die beschreibt mit welcher Wahrscheinlichkeit die Zielfehlerwahrscheinlichkeit verfehlt werden darf. Dieses konkrete Beispiel und die vorherige Diskussion veranschaulichen die Herausforderung festzulegen, welche Sicherheitsmaßnahmen im Hinblick auf KI ausreichen.

[23]

Kläs, M., Adler, R., Sorokos, I., Jöckel, L., Reich, J., „Handling Uncertainties of Data-Driven Models in Compliance with Safety Constraints for Autonomous Behaviour," Proceedings of European Dependable Computing Conference (EDCC 2021), Munich, Germany, IEEE, 2021.

Welche Sicherheitsmaßnahmen ausreichen lässt sich nicht ohne weiteres verallgemeinern und in einfache Regeln gießen. Dies betrifft auch das Regelwerk, das mithilfe der Integritätslevel definiert wird. Wie an folgenden Beispielen veranschaulicht, ist dieses Regelwerk in einigen Fällen unpassend.

Sobald eine Funktion im ersten Schritt ein Integritätslevel zugewiesen bekommt, sind die Hürden KI für die Implementierung zu verwenden so hoch, dass Sie möglicherweise gar nicht realisiert wird, wenn die Funktion nur mit KI umgesetzt werden kann. Betrachtet man eine Funktion, die im fehlerfreien Fall Risiken reduzieren kann und im Fehlerfall kein zusätzliches Risiko darstellt, dann ist dieser Effekt bedenklich, da die Gesamtsicherheit auf einem unnötig niedrigen Niveau bleibt, obwohl sie mithilfe von KI angehoben werden könnte.

Ein anderer Fall sind Funktionen um sicherheitskritische Aufgaben zu erledigen, die zuvor Menschen erledigt haben. Diese Funktionen sind häufig ebenfalls nur mithilfe von KI zu realisieren. Mit der Zuweisung des Integritätslevels wird automatisch auch festgelegt in welchem Bereich die Fehlerrate oder die Fehlerwahrscheinlichkeit liegen sollte. Die Fehlerraten basieren auf dem Risikoakzeptanzprinzip „Minimale Endogene Mortalität (MEM)". MEM liefert ein fixes Maß für das akzeptierte (unvermeidliche) Risiko, dass durch die betreffende Technologie Personen zu Tode zu kommen. Im Hinblick auf Aufgaben die zuvor durch Menschen erledigt wurden könnte allerdings auch

ein vergleichendes Risikoakzeptanzprinzip vorrangig herangezogen werden. Nach diesem Prinzip würde es ausreichen, wenn die Funktion die Aufgabe deutlich sicherer erledigt als sie zuvor manuell erledigt wurde. In Kontext des automatisierten Fahrens auf der Straße wird dieses Risikoakzeptanzkriterium in ISO/TR 4804:2020 [24] als „positive Risikobilanz“ bezeichnet. Es ist verwandt mit dem Risikoakzeptanzprinzip „GAMAB-Prinzip (Globalement au moins aussi bon – Generell mindestens so gut)“ wo der Vergleich zu existierenden technischen Systemen im Vordergrund steht.

[24] ISO/TR 4804:2020 Road vehicles — Safety and cybersecurity for automated driving systems — Design, verification and validation.

Die Beispiele zeigen, dass es ganz unterschiedliche Bezugspunkte geben kann, um festzulegen welche Maßnahmen im Hinblick auf den Einsatz von KI ausreichen.

7.2. Lösungsansätze für sicherheitskritische KI

In diesem Abschnitt werden Lösungsansätze für die beiden Herausforderungen bezüglich sicherheitskritischer KI vorgestellt:

1. Herausforderung 1: Grenzen der Maschine im Kontext von IIoT festlegen
   a. Lösungsansatz: Größeres soziotechnisches Gesamtsystem betrachten
2. Herausforderung 2: Ausreichende Sicherheitsmaßnahmen für KI festlegen
   a. Lösungsansatz: Assurance Cases

Größeres soziotechnisches Gesamtsystem betrachten

Die zunehmende Vernetzung und das Industrielle Internet der Dinge (IIoT) ermöglicht viele Anwendungen von KI, weil große Datenmengen gesammelt werden können. Die Herausforderung die Grenzen der Maschine festzulegen steht somit in Beziehung zur Absicherung von KI-Anwendungen. Sie wurde aber nicht in der Übersicht in Abschnitt 2.5 aufgeführt und die zugehörige Lösung wird im Folgenden nur sehr grob skizziert, weil Sie primär die zunehmende Vernetzung der Systeme adressiert. Die Lösung ist die Sicherheit im Kontext eines größerem Gesamtsystems zu betrachten und die Systemgrenzen weiter zu fassen. Das Referenzmodell für eine vertrauenswürdige KI gemäß der Anwendungsregel VDE-AR-E 2842-61-1 [25] berücksichtigt dies bereits, indem eine neue Abstraktionsebene eingeführt wird. Auf dieser Ebene wird die komplexe Interaktion mit Menschen und Umgebung modelliert und analysiert. Sicherheitsbetrachtungen auf dieser Ebene sind in anderen Anwendungsbereichen wie dem Militär schon lange etabliert und könnten als Vorbild dienen. In dem Militärstandard „MIL-STD-882E“ [26] wird ein „System“ beispielsweise als “Organisation von Hardware, Software, Material,

[25] https://www.dke.de/de/news/2019/referenzmodell-vertrauenswuerdige-ki-vde-anwendungsregel

[26] http://everyspec.com/MIL-STD/MIL-STD-0800-0899/MIL-STD-882E_41682/

Einrichtungen, Personal, Daten, und Diensten, die benötigt wird um eine bestimme Funktion in einer bestimmten Umgebung mit spezifizierten Ergebnissen auszuführen“ betrachtet. Auch der IEEE Report „Safety in the future“ [27] zeichnet ein ähnliches Bild und empfiehlt ein dreigliedriges System (tripartite system), bestehend aus Menschen, Maschinen und Arbeitsumgebung zu betrachten. Die Einbeziehung der Einsatzumgebung ist generell sehr hilfreich, um für die Sicherheit von sicherheitskritischen KI-Anwendungen zu argumentieren, da KI immer nur in einem bestimmten Anwendungskontext (target application scope) funktionieren kann.

[27] https://www.iec.ch/basecamp/safety-future

Assurance Cases

Der Lösungsansatz um die zentrale Herausforderung zu adressieren, dass unklar ist wann welche Maßnahmen ausreichen um KI abzusichern, besteht darin zu erklären, warum die Maßnahmen „ausreichen“. Folgende Abbildung veranschaulicht den Ansatz an einem Beispiel.

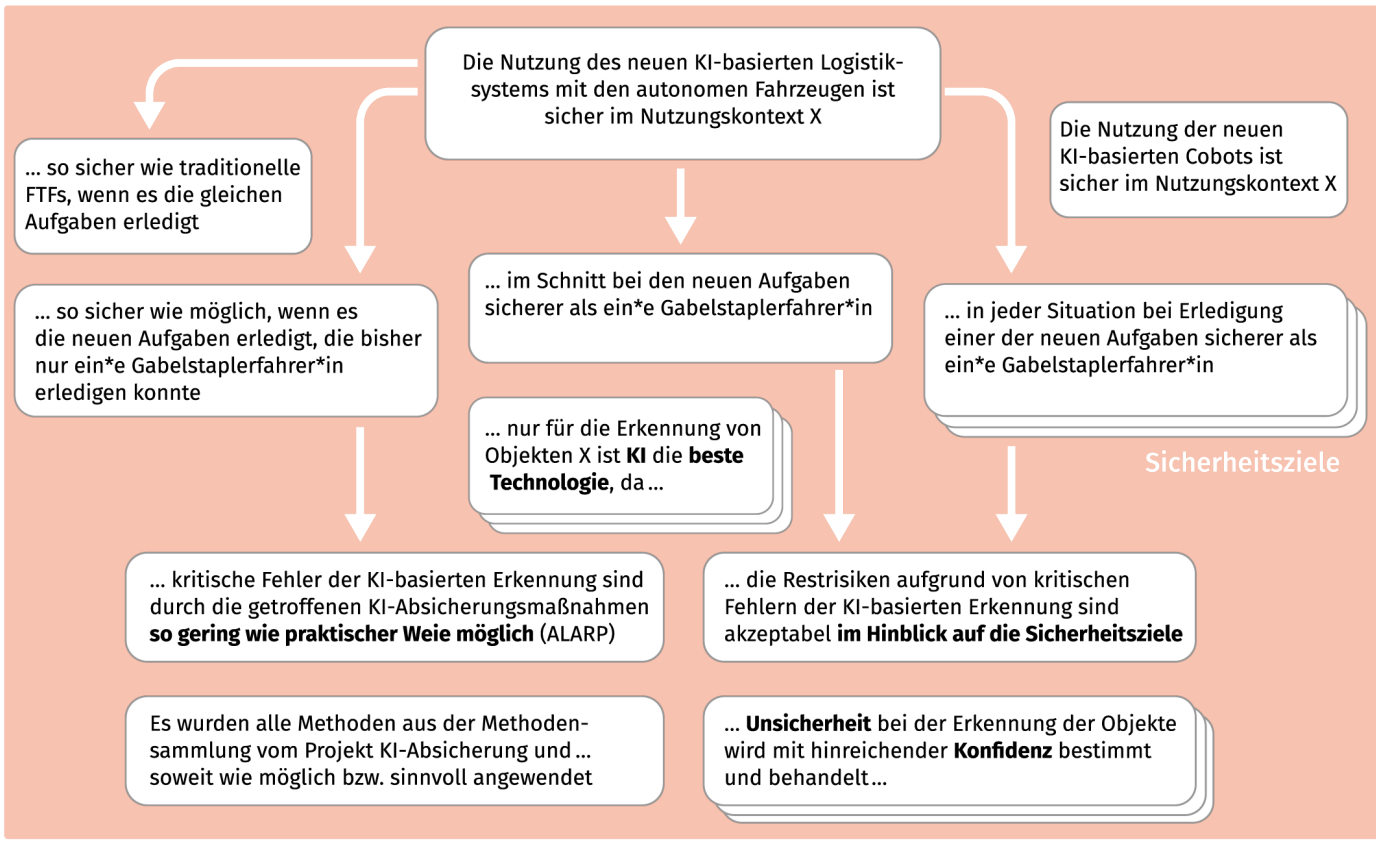

Abbildung 3 – Beispiel für den Aufbau einer Sicherheitsargumentation

In dem Beispiel geht es um ein KI-basiertes Logistiksystem mit autonomen Fahrzeugen. Um zu erklären, warum dieses System sicher ist wird in diesem Beispiel zunächst geklärt, was Sicherheit in diesem Kontext genau bedeutet. Gemäß dem GAMAB-Prinzip sollten Aufgaben, die auch von traditionellen Fahrerlosen Transportsystem (FTFs) erledigt werden können, mindestens genauso sicher erledigt werden wie von diesen traditionellen FTFs. Für die neuen Aufgaben, die bisher nur mithilfe von manueller Steuerung erledigt werden konnten, wird das Prinzip der positiven Risikobilanz herangezogen, das heißt, die Sicherheit mit manuell gesteuerten Fahrzeugen verglichen. Die Ziele bezüglich diesem werden explizit gemacht. Ein Ziel könnte sein, dass die Aufgaben im Schnitt sicherer erledigt werden, wobei sich der Schnitt auf die möglichen Situationen bezieht. Ein anspruchsvolleres Ziel könnte sein, dass es in jeder Situation sicherer ist. Aus diesen Zielen lassen sich probabilistische Zielwerte ableiten. Dies liefert eine Grundlage um zu argumentieren, dass die Unsicherheit der Ausgabe von einer KI-Komponente immer hinreichend abgesichert ist. In dem Projekt wurde dazu ein Ansatz [28] entwickelt. Der Ansatz basiert auf einer Überwachungskomponente („Uncertainty Wrapper“), die die Unsicherheit der Ausgaben von der KI-Komponente bestimmt. Um sicherzustellen, dass die Überwachungskomponente die Unsicherheit nicht zu niedrig abschätzt, kann die Konfidenz bezüglich der Unsicherheit vorgegeben werden. In diesem Ansatz werden die gemessenen Unsicherheiten verwendet, um das sicherheitsrelevante Verhalten anzupassen.

Die Erfüllung der quantitativen Ziele mithilfe von Unsicherheitsbetrachtungen ist aber nur ein Teilaspekt der skizzierten Sicherheitsargumentation in Abbildung 3. Neben den quantitativen Zielen gibt es noch das Ziel, dass das System so sicher wie möglich sein soll. Dazu zählt beispielsweise die Berücksichtigung des 3 Stufenmodells (s. „Aktuelle Vorgehensweise – „Maschinensicherheit“ (DIN EN ISO 12100)“). Wenn eine Gefährdung oder ein zugehöriges Risiko einfach mittels inhärent sicheren Designs eliminiert werden kann, dann hat dies Vorrang gegenüber einer Risikominderung durch eine (KI-basierte) Sicherheitsfunktion. Weiterhin sollte KI nur dann eingesetzt werden, wenn klassische Software weniger geeignet ist. Dies trifft beispielsweise häufig zu für die Erkennung und Klassifikation von Objekten auf Kamerabildern. Die Argumentation sollte darlegen, warum es auch aus der Sicherheitsperspektive heraus sinnvoll ist KI einzusetzen. Danach sollte gezeigt werden, dass durch die gewählten KI-Absicherungsmaßnahmen sicherheitskritische Fehler der KI-Komponente so gut wie möglich minimiert wurden. Dies betrifft in erster Linie Maßnahmen bezüglich der Konstruktion und der Analyse des datengetriebenen Modells, aber auch in anderen Phasen können KI-spezifische Fehlerursachen wie geringe Datenqualität oder Übertraining mit Maßnahmen adressiert werden. Welche Maßnahmen notwendig sind hängt auch von der Art der

[28]

Kläs, M., Adler, R., Sorokos, I., Jöckel, L., Reich, J., „Handling Uncertainties of Data-Driven Models in Compliance with Safety Constraints for Autonomous Behaviour,“ Proceedings of European Dependable Computing Conference (EDCC 2021), Munich, Germany, IEEE, 2021.

KI-Komponente ab. Es gibt beispielsweise Analysen, die spezifisch sind für Neuronale Netze oder spezifisch für bayessche Netze. Die VDE-AR-E 2842-61 führt einen KI-blueprint ein, um mit dieser Vielfalt umzugehen. Der KI-blueprint ist ein generischer Ansatz, der einer Prozess-FMEA ähnelt und bei dem für jeden Schritt der Entwicklung der KI-Komponente überlegt wird, welche Maßnahmen getroffen werden müssen.

Im Beispiel in Abbildung 3 gibt es das qualitative Ziel „so sicher wie möglich“, das dem ALARP Prinzip Rechnung trägt und ein quantitatives Ziel, das aus anderen Risikoakzeptanzprinzipien abgeleitet ist. Beide Ziele sind hilfreich, um zu entscheiden, welche Maßnahmen ausreichen im Hinblick auf KI und um dies in einer nachvollziehbaren Argumentation für ein Audit darzulegen. Für konstruktive Maßnahmen während der KI-Entwicklung ist es schwer abzuschätzen wie effektiv sie sind und ob sie ausreichen. An dieser Stelle kann nach dem Prinzip „so sicher wie möglich“ vorgegangen werden, wobei Bedingungen wie verfügbare Daten, Zeit und Kosten gute Argumente liefern, dass die effektivsten Maßnahmen ausgewählt wurden und nicht weiter optimiert werden kann. Andere Verifikationsmaßnahmen wie das Testen zur Entwicklungszeit und das Monitoring während des Betriebs liefern konkrete Ergebnisse, die auf die Effektivität der konstruktiven Maßnahmen schließen lassen. Basierend auf diesen Ergebnissen kann die Sicherheitsargumentation zeigen, dass das quantitative Ziel erreicht wurde. Abbildung 4 skizziert den Aufbau einer Sicherheitsargumentation für datengetriebene Modelle. Der Ansatz ist in der ExamAI Publikation in [29] beschrieben.

[29]

Kläs, M., Adler, R., Sorokos, I., Jöckel, L., Reich, J., „Handling Uncertainties of Data-Driven Models in Compliance with Safety Constraints for Autonomous Behaviour,“ Proceedings of European Dependable Computing Conference (EDCC 2021), Munich, Germany, IEEE, 2021.

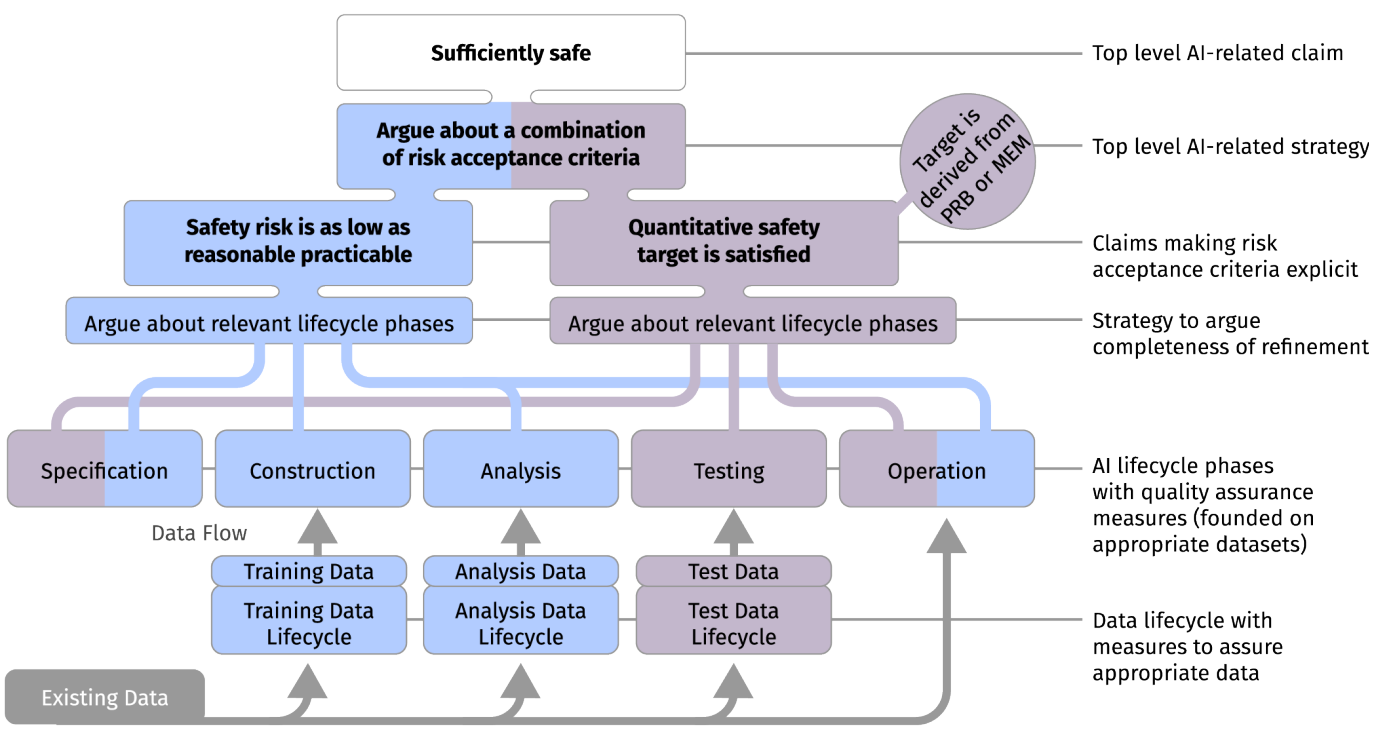

Abbildung 4 – Skizze einer generische Sicherheitsargumentation für datengetriebene Modelle

Eine strukturierte Sicherheitsargumentation zu Nutzen um schwierige Sicherheitsnachweise zu führen wird bereits in einer Dissertation [30] von 2001 vorgeschlagen. In den letzten 20 Jahren ist dieser Ansatz immer populärer geworden. Inzwischen gibt es einen System- und Software Engineering Standard [31] und Spezifikationen um sie modelbasiert zu repräsentieren [32], die angewendet werden, um mithilfe einer Argumentation dazulegen, dass für bestimmte Evidenzen eine Bedingung wahr oder eine Zielsetzung erfüllt ist. Wie in Abbildung 5 gezeigt, gibt es also 3 wesentliche Elemente:

- eine Zielstellung/Bedingung/Behauptung
- eine Argumentation, die zeigt, dass die Zielstellung erfüllt bzw. die Bedingung/Behauptung wahr ist
- Evidenzen oder Fakten auf denen die Argumentation aufbaut

[30] Kelly, T. (2001). Arguing Safety - A Systematic Approach to Managing Safety Cases.

[31] ISO/IEC/IEEE 15026

[32] https://www.omg.org/spec/SACM/2.1/About-SACM/

Alle Elemente zusammen bilden den Assurance Case. Um die Argumentation darzustellen gibt es verschiedene Notationen, die eine baumartige Struktur darstellen. Wie in Abbildung 5 gezeigt wird, besteht die baumartige Struktur aus Teilschritten, die in ihrer Summe die Zielstellung ergeben. Die Teilziele werden solange zerlegt bis sie direkt durch Evidenzen nachgewiesen werden können. Die se Evidenzen können Ergebnisse sein, die bei der Anwendung von Maßnahmen aus traditionellen Sicherheitsstandards herauskommen. Beispielsweise Testergebnisse oder das Ergebnis einer Fehlerbaumanalyse. Da die Maßnahmen aus traditionellen Sicherheitsstandards für KI-Komponenten nicht ausreichen, werden auch Evidenzen von KI-spezifischen Absicherungsmaßnahmen benötigt. Es gibt viele dieser Maßnahmen und aufgrund der intensiven Forschung im Bereich KI werden es stetig mehr. Die Kunst besteht darin, die besten Maßnahmen für den speziellen Fall auszuwählen, sie richtig anzuwenden und basierend auf den Ergebnissen darzulegen, dass eine ausreichende Sicherheit erreicht ist. Wenn es einfache Regeln gäbe, die genau sagen welche Maßnahmen wie anzuwenden sind, um Sicherheit zu erreichen, dann könnte man sich die Sicherheitsargumentation sparen und die Regeln einfach in eine Norm gießen. In absehbarer Zeit wird es aber kein gutes Kochrezept für Sicherheit geben. Solange es kein gutes Kochrezept gibt, könnte die Sicherheitsargumentation als Hauptgegenstand des Audits dienen. Im Kontext des automatisierten Fahrens wird dieser Ansatz bereits im mit dem Standard UL 4600 und voraussichtlich auch im PAS 8800 verfolgt. Diese Vorgehensweise ermöglicht es, aus Felderfahrungen zu lernen. Wie in Abbildung 5 dargestellt, können Felderfahrungen genutzt werden, um neue Evidenzen zu generieren, die Annahmen in der Sicherheitsargumentation bestätigen und Behauptungen unterstützen. Während des Aufbaus der Sicherheitsargumentation werden Safety Performance Indikatoren identifiziert und entsprechende Marktüberwachungsmechanismen geschaffen,

um diese Indikatoren zu sammeln. Weiterhin fördern Assurance Cases die Bildung eines guten Konsenses darüber was wann ausreicht. Es wird forciert, dass der Konsens über die Qualität der Argumente gebildet wird und wenn sich Argumente über die Zeit als standhaft zeigen, dann können sie in Normen verankert werden. Sicherheitsstandards im Automobilbereich verfolgen dementsprechend diesen Ansatz. Auch die sektorübergreifende Anwendungsregel VDE-AR-E 2842-61 verfolgt den Assurance Case-Ansatz. In der VDE-AR-E 2842-61 heißt er Trustworthiness Assurance Case, da neben Sicherheit auch andere Aspekte der Vertrauenswürdigkeit adressiert werden. Unter Einbeziehung aller Vertrauenswürdigkeitsaspekte passt dieser horizontale (sektorübergreifende) Ansatz zum europäischen Vorschlag zur Regulierung von KI. Von einer deutschen Anwendungsregel bis zu einer europäischen Norm, die mit dem AI Act harmonisiert ist, ist es allerdings noch ein langer Weg und die Anwendungsregel müsste sich erst in der Praxis bewähren.

Erstellung und Prüfung vor Markteinführung durch ein **Audit**

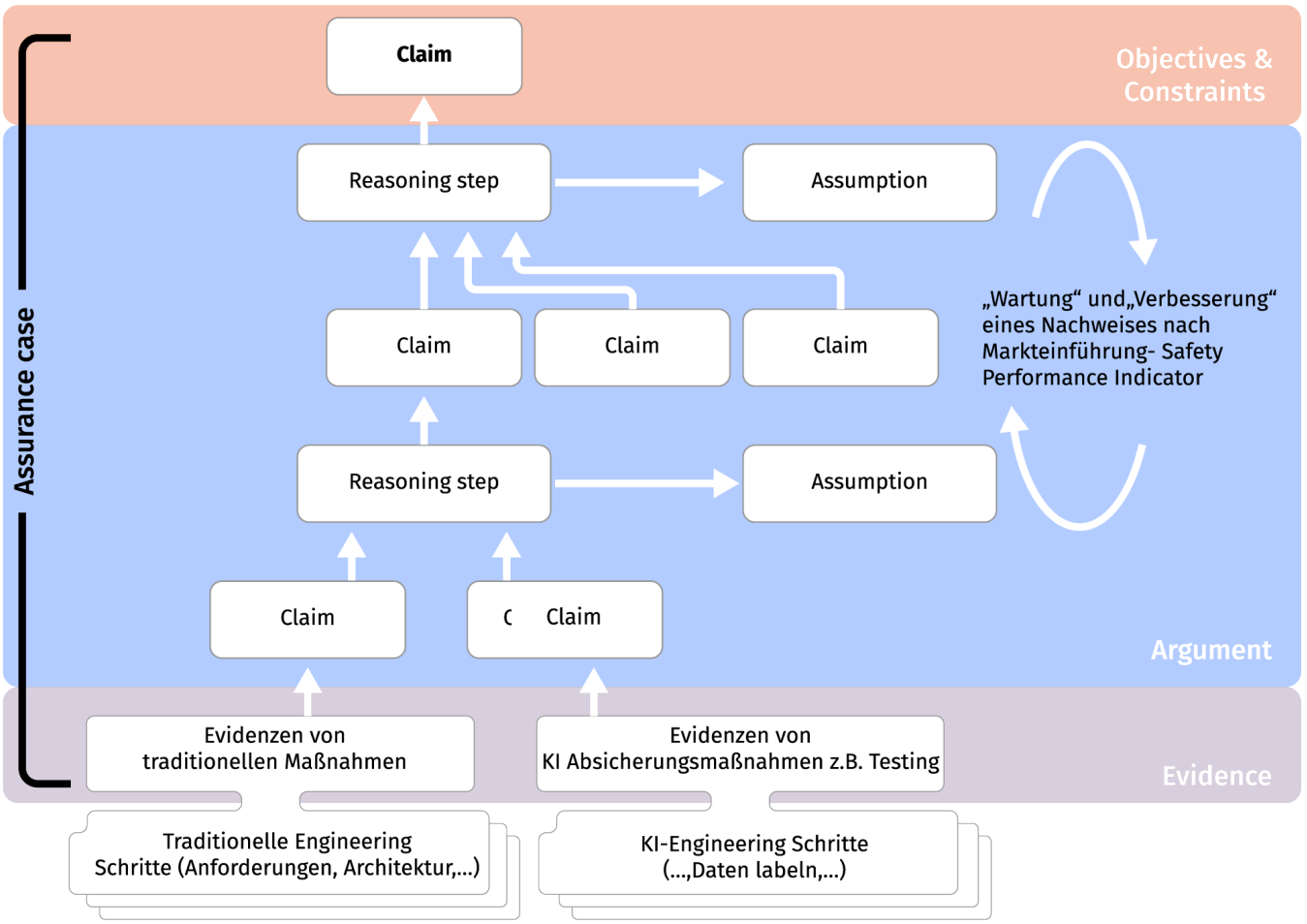

Abbildung 5 - Assurance Cases im Kontext von Testing und Auditing von KI

7.3. Herausforderungen fairnesskritischer KI

Bei Fairnesskritischen Anwendungen von KI besteht die erste wesentliche Herausforderung darin, den Begriff der Fairness im Hinblick auf ein Testorakel zu schärfen. Wenn es keinen Konsens darüber gibt was eine faire Entscheidung ist, dann fehlt auch die Grundlage, um überhaupt testen oder bewerten zu können, ob ein algorithmisches Entscheidungssystem fair entscheidet. Es gibt zwar viele Metriken, um Fairness zu messen aber keinen Konsens darüber, welche Metrik in welchen Fall geeignet ist. Um zu entscheiden, ob das Ergebnis eines Testfalls erfolgreich war (pass) oder nicht (fail) braucht man ein Testorakel. Wenn es keinen Konsens darüber gibt was eine faire Entscheidung ist, dann kann es auch kein Testorakel geben.

Eine weitere Herausforderung sind KI-spezifische Ungleichbehandlungen, die mit einem perfekten Testorakel zwar gefunden aber nicht komplett vermieden werden können. Beim Lernprozess werden Korrelationen in Daten erkannt und in dem datengetriebenen Modell verankert. Da es nur bedingt möglich ist herauszufinden welche Korrelationen gelernt wurden, kann auch nicht gut überprüft werden, ob es sich um kausale Abhängigkeiten handelt. Wenn in einem Trainingsdatensatz zufälligerweise alle guten Bewerber*innen im Vornamen den Buchstaben „A“ hatten, dann könnte sich diese Korrelation im datengetriebenen Modell manifestieren. Ein Algorithmus für die Auswahl der besten Bewerber*innen sollte aber bei der Entscheidungsfindung offensichtlich nicht danach gehen, ob ein*e Bewerber*in den Buchstaben „A“ im Vornamen hat. Dies ist so offensichtlich, dass der Konsens bezüglich des Begriffs „Fairness“ hier keine Hürde darstellt. Hier zeigt sich ein Unterschied zwischen der umgangssprachlichen Bedeutung von „Fairness“ und der rechtlichen Bedeutung von „Diskriminierung“, da in diesem Fall an kein verpöntes Merkmal angeknüpft werden würde und so trotz eines „unfairen“ Geschehens keine Diskriminierung im rechtlichen Sinne vorliegt. Trotzdem ist es ein schwerwiegender Fehler, man vermeiden möchte. In dem Beispiel könnte man den Fehler einfach vermeiden indem man den Vornamen erst gar nicht beim Trainingsdatensatz berücksichtigt. Diese Art von Fehlern kann aber nicht komplett vermeiden werden mit so einfachen Maßnahmen. Bei relevanten Bewertungszahlen könnten beispielsweise die Ziffermuster gelernt werden, die nichts mit der Höhe der Zahl zu tun haben und somit auch nicht mit der relevanten Kausalität.

Lösungsansätze für fairnesskritische KI

Ein Lösungsansatz um damit umzugehen, dass das Ziel „Fairness“ nicht eindeutig festgelegt werden kann, ist die Akzeptanztestgetriebene Entwicklung (Acceptance Test

Driven Development, kurz ATDD). Akzeptanztestgetriebene Entwicklung dient als Kommunikationswerkzeug zwischen den Kund*innen bzw. den Anwender*innen, den Entwickler*innen und den Tester*innen. Die Akzeptanztests sollen die Akzeptanz bei den Kund*innen bzw. den Anwender*innen sicherstellen und tragen zur Klarheit der Anforderungen bei. Dazu müssen die Tests auch für Nicht-Entwickler*innen lesbar sein. Dieser Aspekt ist hilfreich, um den Begriff „Fairness“ aus verschiedenen nicht-technischen Perspektiven zu schärfen. Dabei geht es nicht nur um die Perspektive der Anwender*innen, sondern auch um die Perspektive von Rechtswissenschaftler*innen, Ethiker*innen und denjenigen, die von der algorithmischen Entscheidung betroffen sind. Entsprechend wurde der generelle ATDD-Ansatz im Rahmen des Projekts zugeschnitten, um die Bedeutung von „Fairness“ in einem bestimmten Anwendungsfall anhand von Akzeptanztests zu formalisieren.

Über den ATTD-Ansatz kann auch die zweite Herausforderung adressiert werden und die Frage geklärt werden, welche Maßnahmen ausreichend sind, um offensichtlich unfaire Entscheidungen zu adressieren. In den ATDD-Prozess sind auch KI-Expert*innen und Anwendungsexpert*innen involviert, die Maßnahmen gegen offensichtlich unfaire Entscheidungen auswählen und anderen Stakeholdern die Restrisiken nach Anwendung der Maßnahmen erläutern können. Dies bietet die Grundlage um gemeinsam die Restrisiken zu bewerten und für inakzeptable Risiken weitere Maßnahmen zu ergreifen.

Um die Ergebnisse des ATDD-Prozesses zu dokumentieren und basierend auf den Ergebnissen der Akzeptanztests zu argumentieren, dass das System hinreichend fair ist, bieten sich Assurance Cases an. Assurance Cases werden aktuell zwar primär im Safety Engineering verwendet, sie sind jedoch auch für die Zusicherung von anderen Eigenschaften als Sicherheit geeignet. Für die Anwendung von Assurance Cases ist eher die Stringenz und die Sorgfalt mit der die Behauptung dargelegt werden soll relevant, als die Art der Behauptung. Einen Assurance Case zu erstellen lohnt sich nur wenn es um wichtige Behauptungen beziehungsweise Systemeigenschaften geht. Die Eigenschaft Sicherheit hat eine hohe Priorität im Vergleich zu anderen Eigenschaften, wie an dem Spruch „Safety first“ erkennbar ist. Es ist somit nicht verwunderlich, dass Assurance Cases vorwiegend im Safety Engineering zum Einsatz kommen. Das Schadenspotential von fairnesskritischer KI ist allerdings in einigen Anwendungen ebenfalls sehr groß und rechtfertig die Anwendung von Assurance Cases. Außerdem kann der Aufwand beziehungsweise die Sorgfalt der Erstellung und Überprüfung des Assurance Cases an die Kritikalität angepasst werden.

7.4. Gemeinsamkeiten und Unterschiede

Im Folgenden wird auf die Gemeinsamkeiten und Unterschiede bezüglich der Herausforderungen und Lösungsansätze für das Testen und Auditieren von sicherheitskritischer KI und fairnesskritischer KI eingegangen.

Viele Gemeinsamkeiten und Unterschiede gehen direkt aus den zuvor genannten Herausforderungen und Lösungsansätzen hervor. Beispielsweise die gemeinsame Herausforderung, dass unklar ist welche Maßnahmen ausreichen, um ein bestimmtes Ziel (Sicherheit bzw. Fairness) zu erreichen und der zugehörige Lösungsansatz, die Angemessenheit der Maßnahmen in einem Assurance Case zu argumentieren.

Die Gemeinsamkeiten bei den Herausforderungen und Lösungsansätzen hängen mit KI-spezifischen Themen zusammen. Die Herausforderung in einem datengetriebenen Modell das Konzept „Fairness“ zu implementieren hat letztendlich etwas damit zu tun, dass das Konzept „Fairness“ sehr unscharf ist und aktuell nicht in eindeutigere Konzepte heruntergebrochen werden kann. Es ist als würde man versuchen einem Fahrzeug das Konzept „sicher fahren“ beizubringen, anstatt eine Kollisionsvermeidung zu implementieren. Die Herausforderung, dass sehr unscharfe Konzepte gelernt werden müssen, kann es im Bereich Safety auch geben; typischerweise aber eher in der Medizintechnik als in der Produktionsautomatisierung. Beispielsweise beim Versuch das Wissen und die Intuition von Ärzt*innen mithilfe eines datengetriebenen Modells zu erfassen.

Die Unterschiede bei den Herausforderungen und Lösungsansätzen ergeben sich aus den Eigenschaften Sicherheit und Fairness. Um herauszufinden ob eine Entscheidung sicher ist, schaut man sich typischerweise die Konsequenzen an. Wenn kein Personenschaden entstehen kann oder die Wahrscheinlichkeiten von Personenschäden hinreichend gering sind, dann ist die Entscheidung sicher. Um herauszufinden ob eine Entscheidung fair war kommt es stattdessen vielmehr auf die Begründung als auf die Konsequenz an. Wenn man nur die Entscheidung betrachtet, die eine Software für eine Eingabe getroffen hat, dann kann man generell nicht entscheiden, ob dies fair war, weil man kein Verständnis darüber hat, wie die Software entscheidet. Wenn die Software nicht verständlich ist, wie es bei KI generell der Fall ist, dann bleibt nur die Möglichkeit sich anzuschauen wie die Software in anderen Fällen entscheiden würde. Dies kann man als abstrakte Begründung für eine Entscheidung interpretieren um damit zu bestimmen, ob die Entscheidung in dem betrachteten Fall fair war.

8. Handlungsempfehlungen

Um auf Basis der gewonnenen Erkenntnisse konkrete politische Handlungsempfehlungen zu formulieren, wurden zwei Expert*innen-Workshops zu KI in den Anwendungsbereichen Industrieproduktion und Personalmanagement mit unterschiedlichen Stakeholdern wie Herstellern, Zulieferern, Gewerkschaftsvertreter*innen, und Versicherern (DGUV) durchgeführt.

Der erste Workshop konzentrierte sich inhaltlich auf die Anwendung von KI im Industriekontext und fand am 30. Juni 2021 unter Leitung der Stiftung Neue Verantwortung statt. Im Zentrum des Workshops stand die Frage, wie die Sicherheit von KI-Komponenten dargelegt und überprüft werden kann. Aktuell sehen sich Herstellerfirmen, Prüforganisationen und Marktüberwachungsbehörden diesbezüglich mit enormer Rechtsunsicherheit konfrontiert. Entsprechend war die Diskussion konkreter Ansätze und Methoden von zentraler Bedeutung für die beteiligten Stakeholder. Im Laufe des Projektes wurden Assurance Cases (AC) als eine vielversprechende Methode identifiziert, um die aktuelle Rechtsunsicherheit zu verringern. Die Workshopteilnehmenden diskutierten die Potenziale und Herausforderungen von Assurance Cases als zielbasierten Ansatz für die Überprüfung von KI-Safety und leiteten daraus politische Handlungsempfehlungen ab.

Der zweite Workshop widmete sich der Anwendung von KI im Personalmanagement und fand am 6. Oktober 2021 unter Leitung der Gesellschaft für Informatik e.V. (GI) statt. Hier lag der Fokus auf der Frage, wie die Fairness bzw. Diskriminierungsfreiheit von KI-Systemen sichergestellt werden kann. Zwar gibt es in der Informatik verschiedene Fairnessmaße, die versuchen, Fairness zu operationalisieren. Aber da es keinen gesellschaftlichen Konsens darüber gibt, was genau eine Diskriminierung darstellt, gibt es auch aus rechtlicher Sicht kein klar quantifizierbares Modell zur Messung von Fairness. Im Rahmen des Projekts wurde eine Kombination aus der Methode des Acceptance Test-Driven Development (ATDD) und dem Konzept der Assurance Cases (AC) als ein sinnvoller zielbasierter Ansatz identifiziert, um diesem Problem zu begegnen. Die Workshopteilnehmenden diskutierten auch in diesem Fall die Potenziale und Herausforderungen des Ansatzes und definierten politische Handlungsempfehlungen.

Die Ergebnisse der beiden Workshops ergeben die folgenden Handlungsempfehlungen für Politik, Industrie, Wissenschaft, Normung, Konformitätsbewertung und Marktüberwachung:

8.1. Experimentierräume schaffen

Industrie

Die Erarbeitung individueller und auf Use-Cases zugeschnittener Assurance Cases ist mit hohem finanziellem und personellem Aufwand sowie dem Risiko verbunden, dass die Projekte am Ende keine Machbarkeit aufweisen. Es fehlen also Anreize im Bereich der Absicherung von KI, Pionierarbeit zu leisten. Projekte mit öffentlicher Förderung fokussieren häufig die Untersuchung des wirtschaftlichen Potenzials von KI und der Entwicklung neuer Methoden bzw. ihrer Anwendung und weniger Safety-Aspekte. Der Regulierungsvorschlag der EU-Kommission sieht die Einrichtung von „regulatory sandboxes“ zur Erprobung von KI insbesondere mit Blick auf deren Einhaltung gesetzlicher Vorgaben explizit vor. Ein Experimentierraum könnte beispielsweise eine Lagerhalle oder eine Fabrikhalle sein, in der autonome mobile Roboter oder mobile Cobots eingesetzt werden, mit Menschen interagieren und die Sicherheitskonzepte teilweise auf KI basieren. Idealerweise bringt ein Experimentierraum neben direktem Bezug zum Praxisbetrieb, Safety-Expert*innen, Entwickler*innen von Safety-Maßnahmen und Expert*innen der Konformitätsbewertung zusammen. Neben der notwendigen interdisziplinären und engen Zusammenarbeit zwischen Wissenschaftler*innen und Praktiker*innen aus der Industrie, Sozialpartnern und Akteuren der Qualitätsinfrastruktur ist es essenziell, die zwei Welten KI und Safety stärker zusammen zu bringen.

Personalmanagement

In einer Situation, in der zahlreiche Unsicherheiten hinsichtlich Prüfmethoden, Prüfkriterien sowie den notwendigen Qualifikationen und Kompetenzen der am Prüfprozess – d. h. auch in der ATDD Phase und bei der Aufstellung eines Assurance Case – beteiligten Personen herrschen, ist es zuallererst opportun, Experimentierräume zu schaffen, die dem Aufbau wichtiger Kompetenzen und der Klärung von Unsicherheiten dienen. Dies betrifft insbesondere Fälle, in denen wie bei KI-Anwendungen im Personalbereich großer inhaltlicher Klärungsbedarf hinsichtlich Fairnessmaßen besteht. Auch wenn sich Fairness – im Unterschied beispielsweise zu Experimenten und Simulationen bei Industrieanwendungen – weniger dafür eignet, experimentell getestet zu werden und Diskriminierung auch in einem Experiment unbeachtet bleiben kann, so erlaubt das Durchspielen möglichst vieler verschiedener Fälle, dass auch unkonventi-

onelle Situationen Betrachtung finden und besondere Formen von Diskriminierung sowie unerwartete Zwischenfälle bei der Anwendung eines KI-Systems erkannt werden. Mit Blick auf anwendende Akteure scheinen für die Experimentierphase insbesondere Start-ups interessant, da sich diese besonders häufig mit der Anwendung von KI-Systemen im Personalmanagement befassen.

8.2. Standardisierung angehen und fördern

Industrie

Für einen verlässlichen Rechtsrahmen sind Normen und Standards als objektive Kriterien, anhand derer die Qualität und Sicherheit von KI bewertet werden kann, unabdingbar. Vorstellbar ist etwa eine Norm, die Anforderungen und Prüfkriterien für einen Assurance Case festlegt (zum Beispiel zentrale Aspekte, die ein Assurance Case adressieren muss). Außerdem könnten eine solche Norm die Ziele des Assurance Case beschreiben (zum Beispiel eine bestimmte Fehlerrate mit einer bestimmten Konfidenz nachzuweisen) und benennen, welche Möglichkeiten und Lösungsansätze grundsätzlich bestehen, um zu zeigen, dass die definierten Ziele erfüllt sind. Die Anwendung und Entwicklung der Normen könnte zunächst in Experimentierräumen getestet werden und, um die Langwierigkeit der Entwicklung von Normen entgegenzuwirken, durch die Förderung von kleineren Initiativen (z.B. Technische Spezifikationen wie DIN Specs), beschleunigt werden. Grundsätzlich kann die Normung nur durch ein aktives Engagement der entsprechenden Expert*innen vorankommen. Dafür müssten insgesamt die Bedingungen für eine Teilnahme an Standardisierungsaktivitäten weiter verbessert werden, unter anderem durch angemessene finanzielle Förderung.

Personalmanagement

Bei KI-Anwendungen im Personalmanagement stellt sich Zertifizierung und Normung als besonders herausforderungsreich dar. Wenn es mit Bezug auf Fairness keine generalisierbaren Fälle gibt und sich Einzelfälle nicht übertragen lassen, ist es äußerst fraglich, wie Zertifizierung praktisch gestaltet werden, und ob es etwa zahlreiche individuelle Zertifikate für jeden Anwendungsfall geben müsste. Zusätzlich gibt es weder Normen noch vergangene Zertifizierungsprozesse, die zur Orientierung dienen könnten. Methoden wie Assurance Cases in Normen zu überführen, würde hier einen ersten Ansatzpunkt darstellen. Zwar ließe sich auch dadurch nicht die Fairnessfrage lösen, allerdings kann die Aufstellung eines Assurance Cases dabei helfen, bestimmte Eigenschaften eines KI-Systems zu beurteilen. Auch aus der anfangs notwendigen Betrachtung zahlreicher Einzelfälle können mit der Zeit Gemeinsamkeiten extrahiert werden,

die ein gewisses Maß an Generalisierung und verallgemeinerbaren Prüfkriterien zulassen. Zeitgleich muss sich die Relevanz von KI-Prüfung wesentlich stärker im öffentlichen Diskurs etablieren und mehr (finanzielle) Anreize geschaffen werden, Richtlinien und Prüfmethoden durch die Erstellung von Studien anzuregen.

Ausblick

Der Entwurf des AI-Act unternimmt bereits die ersten Schritte, um mit „gemeinsamen Spezifikationen“ eine Alternative zu harmonisierten Normen bereitzustellen. Hierbei handelt es sich nicht um Normen, sondern um technische Lösungen, die zeigen sollen, wie die Anforderungen des AI Acts an Hochrisiko-KI-Systeme eingehalten werden können. Dabei löst ihre Anwendung ebenfalls eine Konformitätsvermutung aus. Die gemeinsamen Spezifikationen werden nach einem bestimmten Verfahren durch die Europäische Kommission erlassen, wenn harmonisierte Normen fehlen, unzureichend sind oder Bedenken ausgeräumt werden müssen. Insofern könnten hier Lösungen zu fehlenden harmonisierten Normen im Bereich der KI durch die Europäische Kommission geschaffen werden.

8.3. Kulturwandel

Industrie

Die Anwendung von Assurance Cases erfordert von allen Beteiligten eine neue Denkweise. Statt der traditionellen Herangehensweise zum Nachweis von Sicherheit, die auf konkrete, standardisierte Gestaltungsvorgaben setzt, geht es nun um einen flexibel gestaltbaren Argumentationsprozess, der von allen Stakeholdern einen Kulturwandel erfordert. Für die Industrie bedeutet das z.B. mithilfe von Schulungen, einen größeren Fokus auf Safety Engineering zu legen. In den Normungsgremien braucht es einen stärkeren Fokus auf das Zusammenspiel zwischen KI und Safety – das Augenmerk sollte auf der Zusammensetzung der Gremien liegen und diese sollten mehr ausgewiesene KI-Expert*innen enthalten. Zuletzt sollten auch die Behörden der Marktüberwachung Verständnis für die Assurance-Case-Methodik entwickeln. Dafür müssen sie früh, z.B. mithilfe von Schulungen, in den Prozess integriert werden, um die Akzeptanz dafür zu erhöhen. Auch die Europäische Kommission geht davon aus, dass europäischen Marktüberwachungsbehörden die Ressourcen, Expertise und technische Ausstattung fehlen, um die von im Einsatz befindlichen KI-Systemen ausgehenden Risiken zu überwachen oder KI-Komponenten zu untersuchen [33].

[33] Europäische Kommission (2021): Impact Assessment of the Regulation on Artificial intelligence, S.22. Abgerufen am 24.08.2021

8.4. Wissens- und Kompetenzaufbau

Der Wissens- und Kompetenzaufbau ließe sich durch eine größere Offenheit auf Seiten der Unternehmen weiter beschleunigen, wenn diese die Assurance Cases der Forschung zugänglich machen würden. Ein anderer wichtiger Aspekt betrifft den Umgang mit sicherheitskritischen Ereignissen während des Betriebs von KI-Komponenten: Auch hier kann mehr Transparenz, zum Beispiel durch eine Veröffentlichung von Vorfällen im Bereich sicherheitskritischer KI und dem Austausch von Best Practices, andere Stakeholder in die Lage versetzen, aus den Erfahrungen zu lernen und die Sicherheit von KI-Anwendungen insgesamt zu verbessern. Staatliche Stellen könnten hier als vermittelnde Akteure fungieren. Abschließend ist auch zu betonen, dass es weiterer öffentlicher Förderung bedarf, um kontinuierliche Grundlagenforschung zum Thema Prüfwerkzeuge zu betreiben.

Rechtssicherheit in Bezug auf Diskriminierung schaffen

Fairness als Qualitätsmaß für ein KI-System kann durch die rechtlich geforderte Diskriminierungsfreiheit bestimmt werden. Hierzu bedarf es der Entwicklung von Maßstäben, anhand derer eine Diskriminierung durch ein KI-System festgestellt werden kann. Dabei gilt es, sofern notwendig, zwischen verschiedenen Anwendungsbereichen zu differenzieren. Letztlich muss jedoch der Gesetzgeber tätig werden, um die konkreten Anforderungen an die Diskriminierungsfreiheit von KI-Entscheidungen in das Recht zu überführen und Rechtssicherheit zu schaffen.

9. Fazit

Das 20-monatige Forschungsprojekt ExamAI – KI Testing & Auditing untersuchte aus einer interdisziplinären Perspektive, wie sinnvolle Kontroll- und Testverfahren für KI-Systeme aussehen können. Die Ergebnisse leisten einen Beitrag zur Gestaltung effektiver Test-, Kontroll- und Zertifizierungspraktiken für den Einsatz von KI auf nationaler Ebene. Die konkreten Ansätze für die rechtliche und technische Umsetzung dieser Praktiken wurden anhand von zwei Anwendungsszenarien (Mensch-Maschine-Kooperation in der Industrieproduktion / KI-Systeme im Personal- und Talentmanagement sowie im Recruiting) entwickelt.

9.1. Fehlende technische Normen für KI

Die umfassende Analyse der technischen Normen und Standards für KI machte zunächst große Leerstellen in der Normierungslandschaft in Bezug auf Safety sowie Fairness bei KI-Anwendungen sichtbar. Im Safety-Bereich (Anwendungsbereich Industrieproduktion) liegt die Problematik in erster Linie darin, dass bereits vorhandene Safety-Normen für herkömmliche Systeme nicht kompatibel für vergleichbare Systeme mit KI-Anwendungen sind. Die vorhandenen Normen bilden entsprechend nicht den aktuellen Stand der Technik ab. Für den Einsatz von KI-gestützter Software im Personal- und Talentmanagement in Bezug auf Fairness-Aspekte fehlt es fast vollständig an geeigneten harmonisierten Normen für die Anwendung. Damit fehlt eine entsprechende Orientierung für Softwarehersteller und Anwender. Zwar existieren bereits Normen und Standards, die sich mit ethischen Aspekten und Testverfahren von KI- und autonomen Systemen im Allgemeinen beschäftigen, für die Hersteller liefern sie aber keine konkrete Orientierung.

9.2. Rechtsunsicherheit für herstellende und anwendende Unternehmen

Auch auf der rechtlichen Ebene konnten im Laufe des Projekts deutliche Leerstellen identifiziert werden. Im Bereich Safety (Anwendungsbereich Industrieproduktion) setzt

sich der rechtliche Rahmen durch das Produktsicherheitsrecht und das Haftungsrecht zusammen. Das Produktsicherheitsrecht erfasst de lege lata noch keine spezifischen KI-Risken. Das Haftungsrecht kann dies hingegen tun, wobei im Einzelfall weiterer Forschungsbedarf besteht. Beide Rechtsgebiete benötigen zudem technische Normen, um den „Stand der Technik" zu bestimmen. Im Bereich Fairness (Personal- und Talentmanagement) stellt das Allgemeine Gleichbehandlungsgesetz die Basis für Softwareanwendungen in diesem Umfeld dar. Problematisch ist hier, dass die Kausalität von für Ungleichbehandlung ausschlaggebenden Kriterien nicht sicher nachgewiesen oder ausgeschlossen werden kann, da die Systeme an Korrelationen anknüpfen. Und auch die Rechtsdurchsetzung stellt für Betroffene eine Hürde da: Durch die fehlende Transparenz ist weder erkennbar ob KI-Software eingesetzt wurde, noch anhand welcher Kriterien diese entschieden hat.

9.3. Bessere Rahmenbedingungen und Transparenz schaffen

Grundsätzlich bedarf es beim Thema Sicherheit und Fairness von KI eine neue Denkweise. Statt wie bisher auf Nachweise von Sicherheit zu setzen, die sich auf konkrete, standardisierte Gestaltungsvorgaben stützen, sollte das Ziel einer neuen Regulierung ein flexibel gestaltbarer Argumentationsprozess sein. Auf Seiten der Industrie ist es erforderlich einen größeren Fokus auf Safety Engineering zu legen. In der Arbeit der Normungsgremien sollte stärker auf den Zusammenhang zwischen KI und Safety/Fairness und auf die personelle Zusammensetzung der Gremien - diese sollten mehr ausgewiesene KI-Expert*innen enthalten - geachtet werden. Auf der politischen Ebene müssen die Institutionen (z.B. Behörden) ein besseres Verständnis für die Assurance-Case-Methodik entwickeln. Zum Zwecke der Marktüberwachung und zur Förderung der Akzeptanz bedarf es hier Schulungen und die Einbindung in Prozesse. Dafür müssten auf europäischer Ebene Ressourcen, Expertise und technische Ausstattung verbessert werden.

Ziel sollte es auch sein, eine transparente Zusammenarbeit mit Unternehmen zu forcieren, um die Forschung im Themengebiet voranzutreiben. Mithilfe von Forschungsdaten und Einblicken in sicherheitskritische Vorgänge in der Industrie könnte der wissenschaftliche Rahmen verbessert und konstant weiterentwickelt werden. Bei diesem Austausch zwischen den Stakeholdern könnten staatliche Stellen als vermittelnde Akteure fungieren.

Darüber hinaus bedarf es weiterer öffentlicher Förderung, um kontinuierlich Grundlagenforschung im Bereich der Prüfwerkzeuge zu betreiben.

9.4. Assurance Cases als Grundstein der Prüfung

Um die im Safety-Bereich notwendigen Sicherheitsmaßnahmen für den Einsatz von KI-Komponenten in sicherheitskritischen Kontexten festzulegen haben sich im Projektverlauf Assurance Cases (strukturierte Sicherheitsargumentationen) als geeignete Methodik herauskristallisiert. Diese sind für herkömmliche Safetykritische Anwendungen bereits etabliert und mit ihrer Hilfe könnte auch für KI-Systeme ein Sicherheitsniveau erreicht werden, welches mindestens genauso hoch ist, wie das von klassischer Software. Sicherheitsnormen können dabei helfen, die Struktur und das Vorgehen der Sicherheitsargumentation sowie den Auditierungsablauf festzulegen.

Im Umfeld des Personal- und Talentmanagements (Fairnesskritisch) besteht die Herausforderung zunächst darin, den Begriff “Fairness” im Hinblick auf den Umgang mit algorithmischen Entscheidungssystemen zu schärfen. Insbesondere gilt dies bei der Anwendung von Methoden zur Bestimmung von Akzeptanzkriterien für Fairnesskritische Anwendungen. Ein möglicher Lösungsansatz ist die Kombination der Akzeptanztestgetriebenen Entwicklung (Acceptance Test Driven Development – ATDD) mit Assurance Cases. Beim ATDD-Ansatz würde ein interdisziplinäres Team aus Anwendenden, KI-Expert*innen, Ethiker*innen, Rechtswissenschaftler*innen und anderen Stakeholdern so früh wie möglich festlegen, welche Akzeptanzkriterien und Akzeptanztests für ein konkretes Fairnesskritisches KI-System relevant sind. Zusätzlich können auch in diesem Bereich die Assurance Cases zur Anwendung kommen, um eine Argumentation dafür zu liefern, warum Fairness hinreichend garantiert ist.

Als grundsätzliches Zertifizierungs- und Auditierungselement weisen Assurance Cases aus Sicht des Projektkonsortiums das größte Potential auf, um mittel- bis langfristig die Anwendung von KI in kritischen Bereichen sicherer bzw. fairer zu gestalten. Mithilfe von Assurance Cases können Erfahrungen gesammelt werden, die als Grundlage für die Entwicklung anwendungsspezifischer Normen verwendet werden können.

10. Anhang

10.1. Glossar

KI bezieht sich auf Software, die in ihrer Funktionalität nicht durch Regeln spezifiziert ist, welche von extern, beispielsweise mittels Programmierung, festgelegt wurden, sondern durch Regeln, die auf einer Datenbasis anhand eines Lernverfahrens algorithmisch bestimmt wurden.

KI-Methode: Um im Rahmen des Projektes einen klaren Schwerpunkt auf solche KI-Methoden zu setzen, welche eine besondere Relevanz im Kontext von Testing, Auditing und Zertifizierung haben, legen wir den Fokus auf KI-Methoden, die traditionell dem Forschungsfeld des Maschinelles Lernen zuzuordnen sind, das ein klar definiertes Untergebiet der Künstlichen Intelligenz darstellt. Das Maschinelle Lernen stellt hierbei aktuell das mit Abstand prominenteste und in der industriellen Anwendung am weitesten verbreitete Teilgebiet der KI dar. Beim Maschinellen Lernen wird ein Modell auf Basis von Daten mittels KI-Methoden erlernt, hierbei spricht man häufig vom Trainieren des Modells.

Als KI-Komponente sehen wir Softwarekomponenten an, die auf einer KI-Methode basieren. Darunter verstehen wir, dass ihr Verhalten (in Teilen) durch den Einsatz von KI-Methoden bestimmt ist.

KI-Systeme verwenden mindestens eine KI-Komponente, können jedoch beliebig viele weitere Komponenten umfassen, wie zum Beispiel Hardwarekomponenten und weitere Softwarekomponenten, die keinen Bezug zu KI haben.

Ein System ist generell das was im Fokus des Engineerings steht. Das kann neben den technischen Komponenten, auch die physikalische Umgebung sein oder das Verhalten von Menschen. Wenn die Menschen ein Teil des Systems sind, dann spricht man von einem soziotechnischen (KI-)System.

KI-Komponenten und Systeme können danach differenziert werden, ob sie im Entwicklungsprozess trainiert und statisch ausgeliefert werden („offline“ Lernen), oder im Einsatz zur Laufzeit weiterlernen („online“ Lernen).

Fehler bezieht sich auf den technischen Fehlerbegriff gemäß Laprie [34] und nicht auf den juristischen Fehlerbegriff. Laprie definiert die Begriffe „fault“(Fehlerursache), „error“(Fehlerzustand) und „failure“ (Fehlverhalten). In dem Bericht ist mit dem Begriff „Fehler“ je nach Kontext eine der drei Bedeutungen gemeint.

Die Funktion (oder Aufgabe) von einer (KI-)Komponente oder einem (KI-)System beschreibt das, was die Komponente oder das System tun soll [35]. Ein System und eine Komponente können mehrere Funktionen haben. Jede Funktion eines Systems kann auf die Funktionen ihrer Komponenten abgebildet werden. Die Funktion „Kollisionsvermeidung“ eines fahrerlosen Transportsystems kann beispielsweise auf Funktionen von Sensoren, Steuerungskomponente und Aktuatoren abgebildet werden. Die Funktion des Sensors ist beispielsweise Objekte zu erkennen mit denen das Fahrzeug kollidieren könnte.

[34] Laprie, J. C. (Hrsg.): Dependability: Basic Concepts and Terminology. Springer-Verlag, 1992. Laprie, J. C. (Hrsg.): Dependability: Basic Concepts and Terminology. Springer- Verlag, 1992.

[35] A. Avizienis, J. -. Laprie, B. Randell and C. Landwehr, „Basic concepts and taxonomy of dependable and secure computing,“ in IEEE Transactions on Dependable and Secure Computing, vol. 1, no. 1, pp. 11-33, Jan.-March 2004, doi: 10.1109/TDSC.2004.2.

10.2. Publikationen aus dem Projekt

Adler, R., Heidrich, J., Jöckel, L., Kläs, M. (2020). Möglichkeiten und Grenzen von Anwendungen künstlicher Intelligenz in der Produktionsautomatisierung, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

Adler, R., Heidrich, J., Jöckel, L., Kläs, M. (2020). Anwendungsszenarien: KI-Systeme in der Produktionsautomatisierung, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

Arbeitspapier (2021): Diskriminierende KI? Risiken algorithmischer Entscheidungen in der Personalauswahl, ExamAI – KI Testing & Auditing. Gesellschaft für Informatik e.V. (GI).

Becker, N., Junginger, P., Martinez, L., Krupka, D. (2021). KI in der Arbeitswelt: Übersicht einschlägiger Normen und Standards, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

Becker, N., Junginger, P., Martinez, L., Krupka, D., Beining, L. (2021). Mitigating safety and discriminatory risk with technical standards, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V. https://arxiv.org/ftp/arxiv/papers/2108/2108.11844.pdf

Beining, L. (2020). Vertrauenswürdige KI durch Standards? Herausforderungen bei der Standardisierung und Zertifizierung von Künstlicher Intelligenz, in: Impulse. Berlin: Stiftung Neue Verantwortung https://www.stiftung-nv.de/de/publikation/vertrauenswuerdige-ki-durch-standards.

Hauer, M. P., Adler, R., & Zweig, K. (2021, April). Assuring Fairness of Algorithmic Decision Making. In 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW) (pp. 110-113). IEEE. https://ieeexplore.ieee.org/document/9440188

Hilpisch, S. T., Kreutzer, A., Sesing, A. (2021). KI-Systeme in der Produktionsautomatisierung – Rechtsfragen im Überblick, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V., in Veröffentlichung.

Hoffman, R., Sesing, A., Borges, G. (2021). KI-Systeme im Personal- und Talentmanagement – Rechtsfragen im Überblick, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V., in Veröffentlichung.

Jöckel L., Kläs M. (2021) Could We Relieve AI/ML Models of the Responsibility of Providing Dependable Uncertainty Estimates? A Study on Outside-Model Uncertainty Estimates. In: Habli I., Sujan M., Bitsch F. (eds) Computer Safety, Reliability, and Security. SAFECOMP 2021. Lecture Notes in Computer Science, vol 12852. Springer, Cham. https://doi.org/10.1007/978-3-030-83903-1_2

Jöckel, L., Bauer, T., Kläs, M., Hauer, M., Groß, J., (2021). Towards a Common Testing Terminology for Software Engineering and Artificial Intelligence Experts. Akzeptiert auf Profes 2021. Preprint: https://arxiv.org/abs/2108.13837.

Kläs, M., Adler, R., Sorokos, I., Jöckel, L., Reich, J., „Handling Uncertainties of Data-Driven Models in Compliance with Safety Constraints for Autonomous Behaviour," Proceedings of European Dependable Computing Conference (EDCC 2021), Munich, Germany, IEEE, 2021. www.researchgate.net/profile/Ioannis-Sorokos/publication/351659571_Handling_Uncertainties_of_Data-Driven_Models_in_Compliance_with_Safety_Constraints_for_Autonomous_Behaviour/links/60a39746299bf1d21d6ee26f/Handling-Uncertainties-of-Data-Driven-Models-in-Compliance-with-Safety-Constraints-for-Autonomous-Behaviour.pdf

Kläs, M., Adler, R., Jöckel, L., Gross, J., Reich, J., „Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components," AISaftey 2021 at International Joint Conference on Artifical Intelligence (IJCAI), Montreal, Canada, 2021. http://ceur-ws.org/Vol-2916/paper_9.pdf

Krafft, T. D., Zweig, K. A., König, P. D. (2020). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. Regulation & Governance. https://doi.org/10.1111/rego.12369

Krafft, T. D., Hauer, M. P., Zweig, K. A. (2020). Why Do We Need to Be Bots? What Prevents Society from Detecting Biases in Recommendation Systems In: Boratto L., Faralli S., Marras M., Stilo G. (eds) Bias and Social Aspects in Search and Recommendation. BIAS 2020. Communications in Computer and Information Science, vol 1245. Springer, Cham. https://doi.org/10.1007/978-3-030-52485-2_3.

Zweig, K., Hauer, M., Raudonat, F. (2020). Anwendungsszenarien: KI-Systeme im Personal- und Talentmanagement, ExamAI – KI Testing & Auditing. Berlin: Gesellschaft für Informatik e.V.

10.3. Projektpartner

Die Gesellschaft für Informatik e.V. (GI) ist die größte Fachgesellschaft für Informatik im deutschsprachigen Raum. Seit 1969 vertritt sie die Interessen der Informatikerinnen und Informatiker in Wissenschaft, Gesellschaft und Politik und setzt sich für eine gemeinwohlorientierte Digitalisierung ein. Mit 14 Fachbereichen, über 30 aktiven Regionalgruppen und unzähligen Fachgruppen ist die GI Platt-form und Sprachrohr für alle Disziplinen in der Informatik. Weitere Informationen finden Sie unter www.gi.de.

Das Algorithm Accountability Lab an der TU Kaiserslautern untersucht wie sich Softwaresysteme und Entwicklungsprozesse verantwortungsvoll gestalten lassen. Die naturwissenschaftlichinformatisch ausgebildete Leiterin Professor Katharina Zweig und Ihre Arbeitsgruppe engagieren unter anderem in mehreren Ministerialprojekten, Industriekooperationen, bei der DIN und in der Politikberatung. Weitere Informationen unter http://www.aalab.informatik.uni-kl.de.

Das Institut für Rechtsinformatik an der Universität des Saarlandes ist eine der führenden deutschen Forschungseinrichtungen auf dem Gebiet des IT-Rechts und der

Rechtsinformatik. Unter der Leitung von Professor Georg Borges forschen und lehren die Mitarbeiterinnen und Mitarbeiter in der gesamten Breite rund um rechtliche und technische Aspekte der IT-Sicherheit, des Datenschutzes, elektronischer Gerichtsverfahren (eJustice), autonomer Systeme und vieler weiterer Themen. Das Institut engagiert sich in besonderer Weise in der Ausbildung junger Juristen auf dem Gebiet des IT-Rechts und der Rechtsinformatik, etwa durch den Schwerpunktbereich »IT-Recht und Rechtsinformatik« im rechtswissenschaftlichen Studium an der Universität des Saarlandes. Weitere Informationen unter www.rechtsinformatik.saarland/de.

Das Fraunhofer-Institut für Experimentelles Software Engineering IESE beschäftigt sich mit Software als dem Herzstück innovativer Systeme. Seit 20 Jahren forscht und arbeitet das Fraunhofer IESE mit seinen Partnern an richtungsweisenden Schlüsseltechnologien für morgen. Ein Kernthema der letzten Jahre stellt dabei die Digitale Transformation und damit verbundene Entstehung Digitaler Ökosysteme dar, die mittlerweile in allen Bereichen entstehen: von Industrie 4.0 bis hin zu Smart Health, Smart Mobility und Smart Rural Areas. Als Mitglied der Fraunhofer-Allianz Big Data und KI unterstützt das IESE nicht nur Unternehmen darin, sich die notwendigen Engineering-Kompetenzen anzueignen, sondern engagiert sich auch in der Weiterbildung von Data Scientists. Ein Forschungsschwerpunkt liegt auf verlässlicher KI und autonomen Systemen mit besonderem Fokus auf funktionaler Sicherheit. Weitere Informationen unter https://www.iese.fraunhofer.de.

Die Stiftung Neue Verantwortung (SNV) unter der Leitung von Stefan Heumann versteht sich als Think-Tank zu aktuellen politischen und gesellschaftlichen Fragen im Kontext der Digitalisierung und neuer Technologien. Dafür bringt die SNV technisches Fachwissen und Expertise zu gesellschaftlichen und politischen Zusammenhängen in einer Organisation zusammen. Die SNV erarbeitet und veröffentlicht Analysen, entwickelt Handlungsempfehlungen für politische EntscheidungsträgerInnen, führt Expert:innen-Workshops durch, lädt zu öffentlich zugänglichen Fachdiskussionen ein und erklärt Zusammenhänge und Hintergründe in den Medien. Die SNV setzt in ihrer Arbeit auf den Austausch zwischen Expert:innen aus Zivilgesellschaft, Wissenschaft, Wirtschaft und Politik. Weitere Informationen unter http://www.stiftung-neue-verantwortung.de.

Impressum

Eine Veröffentlichung aus dem Projekt „ExamAI – KI Testing & Auditing“ https://testing-ai.gi.de

November 2021

Autor*innen

Rasmus Adler, Nikolas Becker,
Georg Borges, Marc Hauer,
Jens Heidrich, Sven Hilpisch,
Robert Hoffmann, Pauline Junginger,
Lisa Jöckel, Michael Kläs, Daniel Krupka,
Lukas Martinez, Andreas Sesing,
Katharina Zweig

Herausgeberin

Gesellschaft für Informatik e.V. (GI)
Spreepalais am Dom
Anna-Louisa-Karsch-Straße 2
10178 Berlin

Projektleitung

Nikolas Becker
nikolas.becker@gi.de

Gestaltung

Gabriela Kapfer
http://smileinitial.plus

ExamAI – KI Testing & Auditing

Dieses Arbeitspapier erscheint als Teil des Forschungsprojekts „ExamAI – KI Testing und Auditing“, das sich der Erforschung geeigneter Test- und Auditierungsverfahren für KI-Anwendungen widmet. Es steht unter der Leitung der Gesellschaft für Informatik e. V. und wird von einem interdisziplinären Team bestehend aus Mitgliedern der TU Kaiserslautern, der Universität des Saarlandes, des Fraunhofer-Instituts für Experimentelles Software Engineering IESE und der Stiftung Neue Verantwortung getragen und im Rahmen des Observatoriums Künstliche Intelligenz in Arbeit und Gesellschaft (KIO) der Denkfabrik Digitale Arbeitsgesellschaft des Bundesministeriums für Arbeit und Soziales (BMAS) gefördert.
Informationen zum Projekt und weitere Veröffentlichungen finden Sie unter: https://testing-ai.gi.de/

Projektpartner*innen:

GESELLSCHAFT FÜR INFORMATIK

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

INSTITUT FÜR RECHTSINFORMATIK
UNIVERSITÄT DES SAARLANDES

Fraunhofer IESE

Stiftung Neue Verantwortung

Gefördert durch:

DENK FABRIK DIGITALE ARBEITS-GESELLSCHAFT
im Bundesministerium für Arbeit und Soziales

Im Rahmen des:

ki_o
Observatorium Künstliche Intelligenz in Arbeit und Gesellschaft