Masterarbeit

Erstellung eines Software-Tools für die automationsgestützte Beurteilung zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz

ausgeführt am

Fachhochschul-Masterstudiengang
Automatisierungstechnik-Wirtschaft

von
Sascha Steinkrauß, BSc
52101870

betreut und begutachtet von
FH-Prof. Dipl.-Ing. Dieter Lutzmayr

Graz, im November 2022 .....

Unterschrift

EHRENWÖRTLICHE ERKLÄRUNG

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benützt und die benutzten Quellen wörtlich zitiert sowie inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

.....

Unterschrift

DANKSAGUNG

Zuerst, als wichtigster Mensch in meinem Leben, danke ich meiner Frau Conny für die starke Unterstützung, das Verständnis während dieser fordernden Zeit und für den Rückhalt, den du mir gegeben hast – VIELEN DANK.

Mein Dank geht an FH-Prof. Dieter Lutzmayr für die FH-seitige Betreuung, an Boris Scharinger von Siemens für die Kontaktaufnahme, die Bereitstellung des Themas, die Unterstützung und das parallele Anpassen des Wizard-Frameworks während der Ausarbeitung. Ich bedanke mich außerdem bei Peter Seeberg vom Industrial AI Podcast für die Kontaktvermittlung.

Weiterer Dank geht an Lena und Diane Steinkrauß für das Erst-Review des Abstracts und das Korrekturlesen der gesamten Arbeit. Ich bedanke mich außerdem bei Mag. Melissa Maitz für den Feinschliff des Abstracts.

KURZFASSUNG

Künstliche Intelligenz (KI) ist heutzutage ein wichtiger Teil von vielen Software-Applikationen geworden. Sie ist in Handy Apps integriert und je komplexer KI wird, umso größer ist in diesem Zusammenhang der Einfluss, den sie auf die individuelle Privatsphäre hat. Potentielle Risiken, die durch KI entstehen können, haben die Europäische Union dazu gebracht, einen Vorschlag für eine KI-Verordnung zu erarbeiten. Um die darin geforderten Risikominderungsmaßnahmen zu adressieren, hat das Fraunhofer Institut einen Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz veröffentlicht.

Das Hauptziel dieser Masterarbeit ist es, diesen textbasierten Leitfaden in ein Software-Tool zu überführen, welches Entwicklern von KI-Applikationen hilft, den Prozess gemäß Leitfaden zu dokumentieren. Dies hilft, den Leitfaden zu vereinfachen, um einen niederschwelligen Zugang zu dem Leitfaden für viele Menschen zu ermöglichen. Weitere Ziele dieser Arbeit sind sowohl das Aufdecken von möglichen Optimierungen der Richtlinie als auch Verbesserungen des bestehenden Software-Frameworks.

Das Software-Tool wurde unter Zuhilfenahme dieses Frameworks innerhalb der Low-Code Plattform Mendix entwickelt.

Das Ergebnis ist ein Software-Tool, welches die Anforderungen an vertrauenswürdige KI vereinfacht, je nach dem Anwendungsbereich der gerade beurteilten KI-Applikation. Nach der Nutzung des Software-Tools kann ein Bericht erzeugt werden, um die erfüllten Anforderungen der Vertrauenswürdigkeit zu dokumentieren.

ABSTRACT

Artificial intelligence (AI) has become an essential part of various software-applications nowadays. It is embedded into mobile phone apps making an impact on individuals' privacy. Due to AI´s potential risks the European Union recently launched a proposed AI-Act. The Fraunhofer Institute released a guideline on how to design trustworthy AI in order to meet the according risk reduction measures.

The aim of the master´s thesis is to convert the Fraunhofer guideline into a software-tool. The tool shortens the requirements of the guideline depending on the scope of each AI-application and it enables developers of AI-applications to document the process as required in the guideline. An additional objective of the master´s thesis is to identify potential improvements of the guideline as well as of the already existing software-framework.

The software-tool was developed by using the software-framework within the low-code platform Mendix.

The tool simplifies the requirements of trustworthy AI depending on the actual scope of each assessed AI-application. As a result, a report can be generated to document the fulfilled requirements of trustworthiness. The master´s thesis serves as a simplified guideline leading to a more commonly utilized application.

INHALTSVERZEICHNIS

1 EINLEITUNG

Die vorliegende Arbeit befasst sich mit Künstlicher Intelligenz (KI), im Speziellen mit der vertrauenswürdigen Intelligenz. Dabei wird im theoretischen Teil zunächst geklärt, was KI im Sinne der vorliegenden Arbeit bedeutet und welchen Anteil die Vertrauenswürdigkeit trägt.

Aufgrund der aktuell bestehenden politischen Debatte zum Thema KI und der erkannten Komplexität, werden gesetzliche Vorgaben geschaffen. Die politisch geforderten Ziele und der aktuelle Stand der Gesetzgebung in Bezug auf dieses Themengebiet sollen ebenso betrachtet werden. Da in der Gesetzgebung häufig sehr breite Themengebiete abgedeckt werden, stellen sich in der Umsetzung konkrete Fragen, wie diese Ziele erreicht werden können. Dies führt direkt zu dem Basisdokument der vorliegenden Arbeit, dem Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz | KI-Prüfkatalog[1] des Fraunhofer IAIS. Der Leitfaden beschreibt eine konkrete Herangehensweise für die Bewertung von vertrauenswürdiger KI und wie Maßnahmen für eine dritte, möglicherweise zertifizierende, Stelle dokumentiert werden sollten.

Der Leitfaden dient ebenso als Basis für das Software-Tool, welches den praktischen Teil der Arbeit darstellt. Das Software-Tool wird auf einem bestehenden prototypischen Wizard-Framework entwickelt. Es soll die Anwendung des Leitfadens für jede*n Interessierte*n erleichtern und somit dazu beitragen, die geforderten Maßnahmen flächendeckend umzusetzen.

Dabei sollen im Rahmen der vorliegenden Arbeit sowohl Erkenntnisse für eine Optimierung des Leitfadens, als auch mögliche Funktionserweiterungen für eine intuitive und effiziente Verwendung des Wizards-Frameworks festgehalten werden.

1.1 Künstliche Intelligenz (KI) – Definitionen

Spricht man von KI, wird schnell deutlich, dass je nach Kontext eine große Anzahl von unterschiedlichen Definitionen existieren. Diese reichen von einfachen statistischen Methoden bis hin zu neuronalen Netzen.

Schon bei dem Begriff [Intelligenz] wird es schwierig, eine prägnante Definition zu finden, da verschiedene wissenschaftliche Disziplinen eine Vielzahl von unterschiedlichen Ansätzen einbringen. So ist „Intelligenz in der Psychologie ein hypothetisches Konstrukt (d.h. eine Erklärung für ein nicht direkt beobachtbares Phänomen), das die erworbenen kognitiven Fähigkeiten und Wissensbestände einer Person bezeichnet, die ihr zu einem gegebenen Zeitpunkt zur Verfügung stehen.“[2]

Es gibt darüber hinaus auch eine Unterscheidung in schwache und starke KI. Sämtliche aktuell existierenden KI-Systeme werden gemäß dem momentanen Entwicklungsstand im Jahr 2022 der schwachen KI zugeordnet. Schwache KI bedeutet zusammengefasst, dass Maschinen bzw. Computer so

[1] Poretschkin u.a. (2021), S.1.

[2] Maier (2018), S.1.

handeln, als ob sie intelligent wären.[3] Tatsächlich wurden sie jedoch nur über entsprechende Programmierung für bestimmte Aufgaben vorbereitet und besitzen lediglich Algorithmen zur Lösung der gestellten Aufgabe und keine tatsächliche Intelligenz. Von starker KI wird gesprochen, wenn die Maschine bzw. der Computer tatsächlich denken und handeln kann.

Aufgrund von immer leistungsfähigeren Rechnern und der großen Möglichkeiten der Anwendung von KI, ist es in den letzten Jahren zu einer weiten Verbreitung von KI gekommen, d.h. sowohl global als auch branchenübergreifend werden immer neue Anwendungsbereiche erschlossen. Aufgrund der unterschiedlichen Blickwinkel auf das Thema KI entstehen teilweise neue Definitionen[4] bzw. wird die KI in verschiedene Teilgebiete aufgeteilt[5] - siehe Abb. 1.

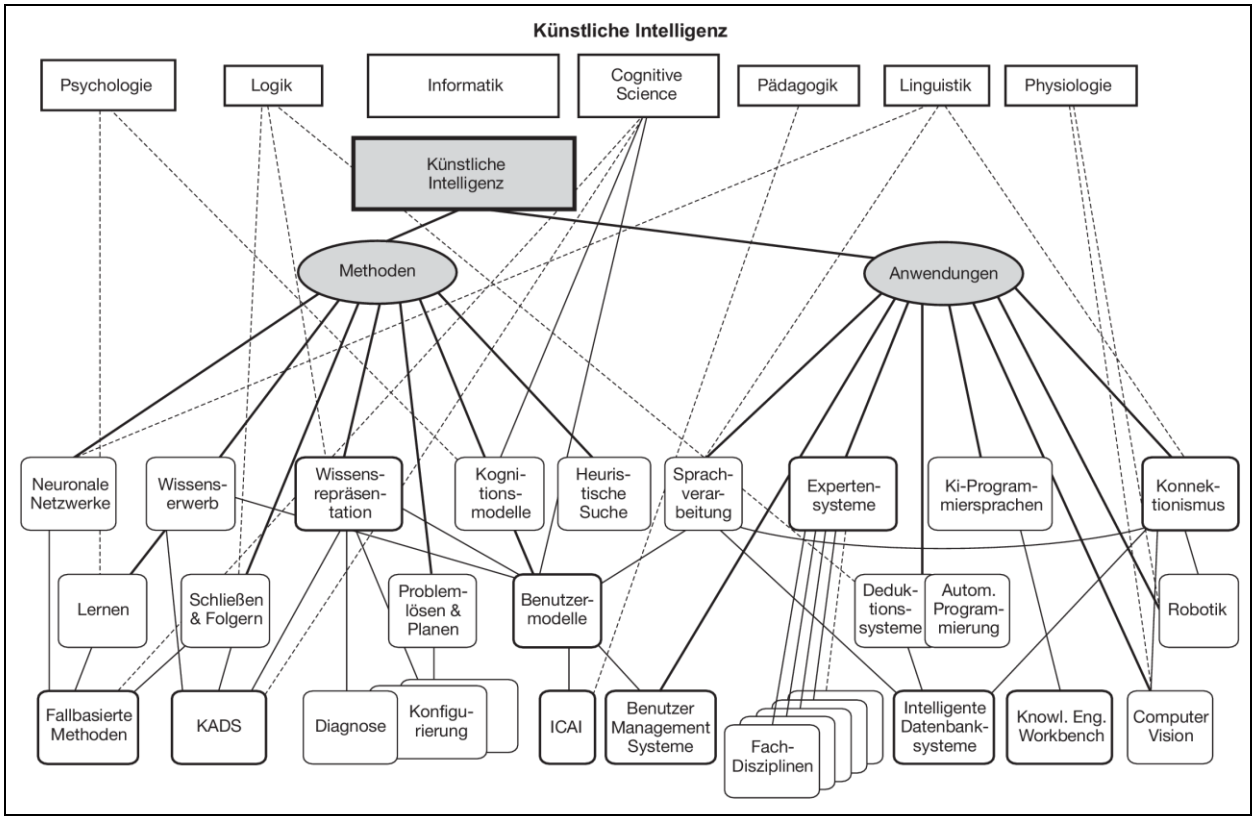


Abb. 1: Teilgebiete der KI, Quelle: Gabler Wirtschaftslexikon.[6]

Im Rahmen der vorliegenden Arbeit wird der weite Begriff KI, abgegrenzt von diversen Fachdisziplinen, auf seine momentan tatsächlich möglichen technischen Möglichkeiten begrenzt, betrachtet. Nachfolgend soll daher eine Übersicht über die wichtigsten Verfahren und Begriffe im Zusammenhang mit KI im technischen Kontext gegeben werden.

[3] Vgl. Russell/Norvig (2010), S.1020.

[4] Vgl. Wahlster/Winterhalter (Hrsg.) (2020), S.36.

[5] Lackes/Siepermann (2018), S.1.

[6] Lackes/Siepermann (2018), S.1.

1.1.1 Statistik

Die Statistik befasst sich, vereinfacht gesagt, mit der Erhebung, der Analyse und der Interpretation von Daten. In einfachen Fällen wird Statistik dazu verwendet eine gewisse Anzahl an, häufig unstrukturierten, Daten zu visualisieren. Dies wird beispielsweise zur Darstellung von quartalsweisen Geschäftszahlen oder von Jahresabschlüssen verwendet. So kann auf einfache, anschauliche Weise mit Vergleichszeiträumen und in Diagrammen dargelegt werden, wo sich der aktuelle Stand befindet. Bei diesen einfachen statistischen Vergleichen werden die Daten häufig durch den Menschen interpretiert, da es lediglich darum geht, festzustellen, ob der Zielwert in der aktuellen Periode *erreicht* oder *nicht erreicht* wurde. Geht es um komplexere Sachverhalte ist es sinnvoll für die Interpretation der Datensätze Software mit entsprechenden Algorithmen einzusetzen, die den Sachverhalt bestmöglich interpretieren.
Im industriellen Umfeld kann das z.B. eine Produktionslinie sein, die an der fünften Position in regelmäßigen Abständen einen Ausschuss produziert. Das erfahrene Produktionspersonal prüft sämtliche Einstellungen an der fünften und an der vorigen, der vierten, Position und kommt zu dem Schluss, dass sämtliche Einstellungen korrekt sind. Dennoch wird wieder Ausschuss produziert. Aufgrund von vielen hunderten bzw. tausenden Sensoren entlang der Linie können bei Auswertung des kompletten Datensatzes mittels *Machine Learning*[7] ggf. Rückschlüsse auf falsche Einstellungen in einer der vorgelagerten Positionen der Produktionslinie gezogen werden.

1.1.2 Machine Learning

Das sogenannte *Machine Learning* ist ein Überbegriff für verschiedene Arten, wie ein System aus Informationen, die in statistischen Datensätzen enthalten sind, Erfahrungen sammelt. Diese Erfahrungen und der damit verbundene Lernfortschritt werden in weiterer Folge für die Verarbeitung weiterer Daten verwendet. Je mehr Daten dabei verarbeitet werden, umso besser wird die Leistungsfähigkeit des Systems in Bezug auf die gewünschten Ausgabedaten (Outputs) werden – siehe Abb. 2. Das heißt die Datensätze werden durch die Software nicht nur interpretiert, sondern es entsteht durch die verwendeten Algorithmen ein lernendes System.

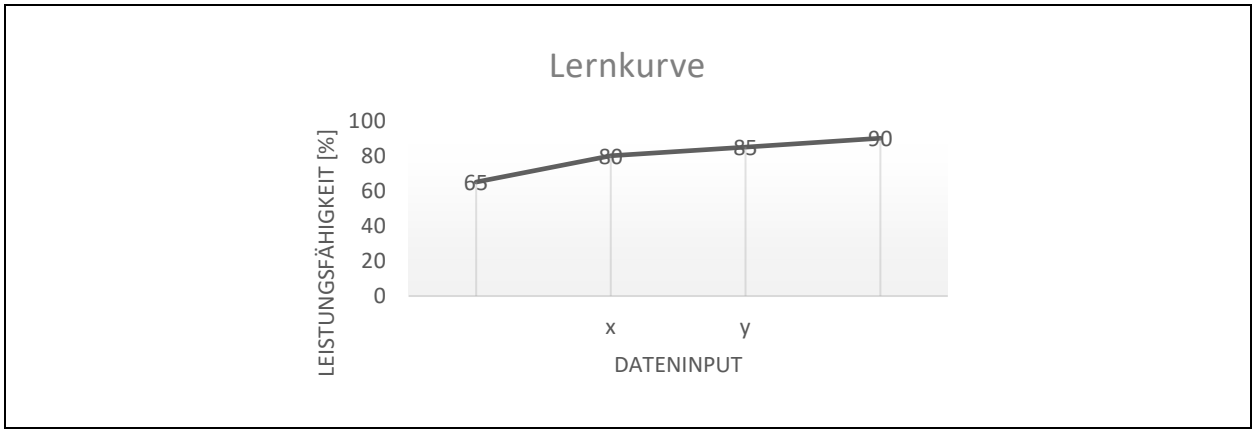


Abb. 2: Lernkurve Machine Learning, Quelle: Eigene Darstellung.

[7] Vgl. Kapitel 1.1.2.

Dabei können Eingabedaten (Inputs) unter Anwendung von unterschiedlichen Algorithmen auf verschiedene Arten verarbeitet werden. Je nach Komplexität und Anwendungsfall kann die Art der Verarbeitung auf eine Weise schneller zielführend sein oder zu besseren Outputs führen als bei einer anderen Art.

Diese verschiedenen Arten lassen sich wie folgt kategorisieren: *Unsupervised Learning*, *Supervised Learning* und *Reinforcement Learning* – siehe Abb. 3, wobei sich die Kategorien zu Deutsch am ehesten mit *Nicht Überwachtem Lernen*, *Überwachtem Lernen* und *Be- bzw. Verstärktem Lernen* übersetzen lassen.[8]

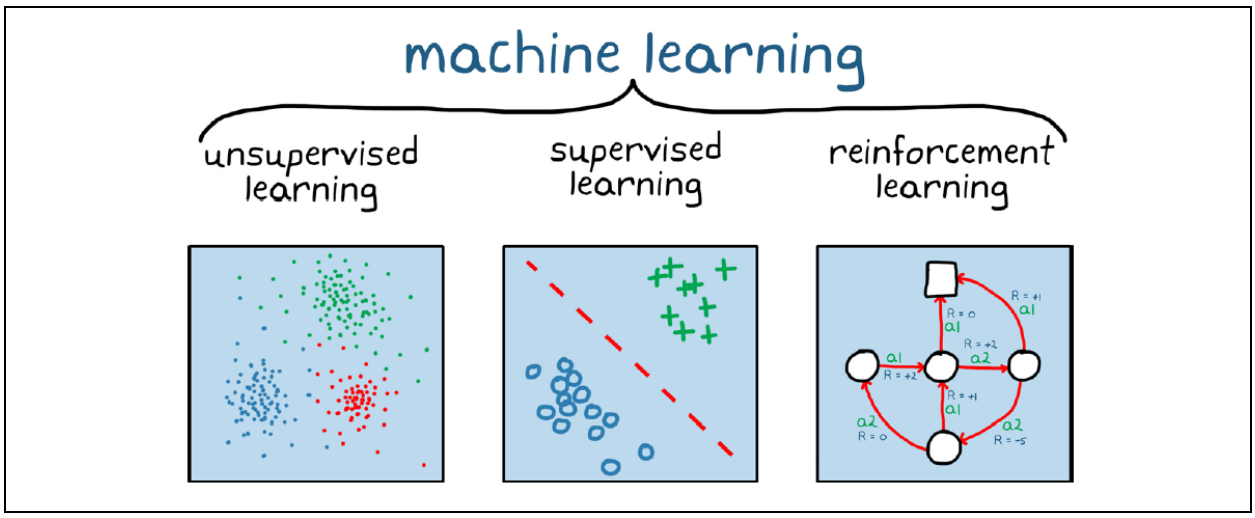


Abb. 3: Machine Learning Übersicht, Quelle: The MathWorks (2020).[9]

1.1.2.1 Unsupervised Learning

Beim *Unsupervised Learning* werden Daten analysiert und Muster bzw. Strukturen in den Daten erkannt, die in irgendeiner Art und Weise miteinander übereinstimmen. Im Gegensatz zum *Supervised Learning*[10] werden diese Informationen in den Daten dabei nicht durch eine Person bzw. ein Modell überwacht.

Eine Möglichkeit des *Unsupervised Learning* ist das sogenannte *Clustering*. Beim *Clustering* werden Gruppen aus ähnlichen Objekten gebildet. Die Ähnlichkeit kann dabei in Bezug auf verschiedene Attribute hin gebildet werden. Wird beispielsweise eine beliebige Menge Menschen an betrachtet und sollen diese Menschen in Gruppen aufgeteilt werden, so gibt es dafür verschiedene Möglichkeiten. Die Menschen können in Gruppen nach

- Geschlecht
- Haarfarbe
- Musikgeschmack

[8] Vgl. Kapitel 1.1.2.1, 1.1.2.2 und 1.1.2.3.

[9] The MathWorks (2020), S.9.

[10] Vgl. Kapitel 1.1.2.2.

- Fitnesslevel

oder anderen Gemeinsamkeiten aufteilen. Es wird allerdings anhand der unterschiedlichen Ebenen der Einteilungsmöglichkeiten klar, dass es dementsprechend unterschiedliche Gruppen geben kann, die beim Unsupervised Learning alle richtig sind. Es gibt keine überwachende Instanz, die sagt, welche Gruppeneinteilung richtig und welche falsch ist.

1.1.2.2 Supervised Learning

Beim *Supervised Learning* wird zunächst ein Modell erstellt. Dieses Modell wird mit Trainingsdaten je nach gewünschtem Ergebnis für seinen Anwendungszweck angelernt. Bei der *Klassifikation,* einer Technik des *Supervised Learning,* beinhalten die Trainingsdaten Objekte (Klassen) mit zugehörigen Attributen. Wird ein Modell für die Erkennung von Kraftfahrzeugen erstellt, so werden die Trainingsdaten, z.B. Bilder von verschiedenen Kraftfahrzeugen mit verschiedenen Attributen, klassifiziert. Dies kann z.B. folgendermaßen geschehen:

- Hersteller (A, B, C)
- Einsatzgebiet (Sportfahrzeug, Nutzfahrzeug, Familientransporter)
- Farbe (Rot, Grün, Blau)

Wird dem Modell anschließend ein neues Bild von einem Kraftfahrzeug zugeführt, kann es anhand der Trainingsdaten zuordnen, welches Fahrzeug abgebildet ist.

Wird diesem Modell jedoch ein Bild von einem Fahrrad als Input zugeführt, wird das Modell anhand der erkannten Attribute jedoch einen falschen Output generieren (Kraftahrzeug), da das Modell nicht für Fahrräder konzipiert wurde, jedoch nur diskrete Outputs liefert. Anhand des einzigen Attributs (Farbe), das erkannt wird, wird das Modell eine Einstufung vornehmen.

Bei der *Regression*, einer weiteren Technik des *Supervised Learnings,* können auf Basis des Modells und der vorhandenen Trainingsdaten Werte entlang eines Kontinuums ausgegeben werden. Das bedeutet, dass, wie bei der Klassifikation, nicht nur diskrete Outputs generiert werden, sondern es können Zwischenwerte abgebildet werden, die in den Trainingsdaten noch nicht enthalten waren. Dies kann z.B. für die Voraussage von kontinuierlichen Prozessoutputs in Abhängigkeit von diversen Inputs genutzt werden.

1.1.2.3 Reinforcement Learning

Beim Reinforcement Learning wird der KI eine zu lösende Aufgabe, eine Zieldefinition, gestellt. Der Weg dorthin, bzw. wie dieses Ziel erreicht wird, bleibt der KI dabei komplett überlassen. Werden zu Beginn noch zufällige Aktionen gesetzt, lernt und verbessert sich die KI durch Iterationen (trial an error), bis das Ziel erreicht ist. Die Verbesserung wird dadurch erreicht, dass die KI für positive Aktionen belohnt wird. Somit wird die KI bei dem nächsten Versuch, den Zielzustand zu erreichen, den Weg optimierten Weg wählen. Dabei sind je nach Komplexität der Aufgabe sehr viele Iterationen notwendig und es ist dementsprechend zeitintensiv.

1.1.3 Neuronale Netze

Vereinfacht gesagt wird bei den *Neuronalen Netzen* auf Software-Ebene versucht, das menschliche Gehirn nachzuempfinden, indem in einem Modell künstliche Neuronen miteinander vernetzt werden. Somit wird, je nach Vernetzung, eine große Anzahl an Informationen aus der Umgebung, d.h. aus dem Netzwerk, genutzt, um einen Output zu generieren – siehe Abb. 4.

Neuronale Netze müssen vor der tatsächlichen Anwendung mit einer sehr großen Anzahl an Daten trainiert werden, um einen akzeptablen Output zu bekommen.

1.1.4 Deep Learning

Beim Deep Learning werden in einem Modell künstliche *neuronale Netze*[11] in mehreren Schichten miteinander verknüpft. Je mehr Schichten verwendet werden, umso tiefer das Netzwerk des Modells, daher die englische Bezeichnung *Deep Learning*. Diese *neuronalen Netze* benötigen einerseits eine große Rechnerleistung, können dafür andererseits einen Input von sehr großen Datenmengen verarbeiten und durch den inneren Aufbau optimierte Outputs bei komplexeren Anforderungen erzeugen. In diesem Komplexitätsgrad wird es jedoch, selbst für Spezialisten in diesem Gebiet, schlicht unmöglich, sämtliche Funktionalitäten bzw. Outputs aus diesen intelligenten Netzwerken sofort nachzuvollziehen.

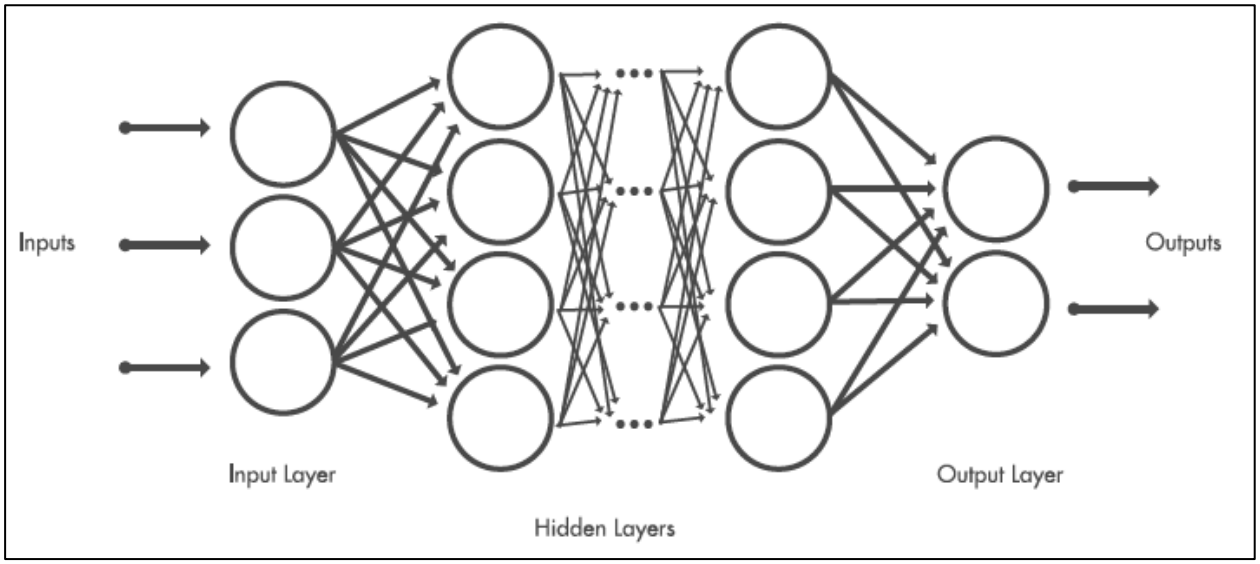


Abb. 4: Neuronale Netze Aufbau, Quelle: The MathWorks (2021).[12]

[11] Vgl. Kapitel 1.1.3.

[12] The MathWorks (2021), S.6.

1.1.5 Korrelation und Kausalität

Werden bei der Analyse von Daten Muster oder Strukturen erkannt, die einen Lerneffekt für die KI haben, ist besonders bei dem Einlernen von Trainingsdaten darauf zu achten, ob es sich dabei lediglich um *Korrelation* oder tatsächlich auch um *Kausalität* handelt.

Bei der *Korrelation* entstehen Muster in den Daten, die zufälligerweise auftreten und den Anschein erwecken, relevant zu sein, jedoch nichts mit dem zu untersuchenden Ereignis zu tun haben. Bei der *Kausalität* ist das Muster in den Daten der tatsächliche Grund für das Ereignis.

Wird beispielsweise in einer Produktionslinie zu einem Zeitpunkt y Ausschuss produziert, werden durch die KI mögliche Gründe dafür im zugrundeliegenden Datensatz analysiert. Erkennt die KI, dass zu einem vorherigen Zeitpunkt x, zu dem ebenfalls Ausschuss produziert wurde, die gleiche Hallentemperatur (29,3°C) geherrscht hat, so *korreliert* dieser Wert beider Zeitpunkte. Ist dieser Wert, d.h. die Hallentemperatur, jedoch auch *kausal* für den produzierten Ausschuss? Dies sollte nun vom erfahrenen Fachpersonal, mit der entsprechenden Domain-Expertise, kritisch hinterfragt werden. Es sollte eine Schnittstelle zur KI existieren, über die in solchen Fällen durch das Fachpersonal entweder bestätigt (*Kausalität*) oder abgelehnt (*Korrelation*) wird. Damit wird möglichen Fehlinterpretationen im Lernen der KI vorgebeugt.

1.1.6 KI in der Gesetzgebung

Aufgrund der vielfältigen Definitionen und der Notwendigkeit, KI für eine europäische Gesetzgebung klar zu definieren, stellte sich auch die europäische Kommission bereits im Jahr 2018 die Frage, „Was ist künstliche Intelligenz?“ und definiert sie folgendermaßen:
„Künstliche Intelligenz (KI) bezeichnet Systeme mit einem „intelligenten“ Verhalten, die ihre Umgebung analysieren und mit einem gewissen Grad an Autonomie handeln, um bestimmte Ziele zu erreichen.
KI-basierte Systeme können rein softwaregestützt in einer virtuellen Umgebung arbeiten (z. B. Sprachassistenten, Bildanalysesoftware, Suchmaschinen, Sprach- und Gesichtserkennungssysteme), aber auch in Hardware-Systeme eingebettet sein (z. B. moderne Roboter, autonome Pkw, Drohnen oder Anwendungen des „Internet der Dinge“).“[13]

In dem, 2021 folgenden, Vorschlag für eine KI-Verordnung wird die europäische Kommission im Artikel 3 Absatz 1 konkreter, was ein System der künstlichen Intelligenz (KI-System) sein soll:

„eine Software, die mit einer oder mehreren der in Anhang I aufgeführten Techniken und Konzepte entwickelt worden ist und im Hinblick auf eine Reihe von Zielen, die vom Menschen festgelegt werden,

[13] Europäische Kommission (2018), S.1.

Ergebnisse wie Inhalte, Vorhersagen, Empfehlungen oder Entscheidungen hervorbringen kann, die das Umfeld beeinflussen, mit dem sie interagieren;“[14]

Im Anhang I finden sich dann die folgenden Techniken und Konzepte:

„a) Konzepte des maschinellen Lernens, mit beaufsichtigtem, unbeaufsichtigtem und bestärkendem Lernen unter Verwendung einer breiten Palette von Methoden, einschließlich des tiefen Lernens (Deep Learning);

b) Logik- und wissensgestützte Konzepte, einschließlich Wissensrepräsentation, induktiver (logischer) Programmierung, Wissensgrundlagen, Inferenzmaschinen, Deduktionsmaschinen, (symbolischer) Schlussfolgerungs- und Expertensysteme;

c) Statistische Ansätze, Bayessche Schätz-, Such- und Optimierungsmethoden.“[15]

Dabei ist erkennbar, dass konkrete technische Umsetzungsformen von KI, die bereits bekannt sind und verwendet werden, adressiert werden sollen. Dies ist insofern nachvollziehbar, da die europäische Kommission die regulatorischen Spielregeln für die zulässigen Umsetzungsformen von KI festlegen möchte bzw. muss. Regulatorik sollte durch eindeutige Definitionen möglichst zielführend und für alle Betroffenen gleich ausgestaltet sein. Da die oben genannten Konzepte und Methoden schon etabliert und bekannt sind, werden diese als erstes von der Gesetzgebung aufgenommen und in die neu entstehende Gesetzgebung eingebunden. Es ist jedoch auch zu erwarten, dass die Gesetzgebung bei dieser, sich sehr schnell entwickelnden, Technologie neuen technischen Entwicklungen folgen wird und ggf. die gesetzliche Definition anpassen wird.

1.2 Anwendungsbereiche und Vorbehalte

Es besteht aus verschiedenen Gründen die Forderung, dass KI und die damit entstehenden Möglichkeiten als Werkzeug eingesetzt werden sollen. Gründe dafür können vielfältig sein. Genauso wie für die Robotik wird oft der Fachkräftemangel oder auch die alternde Gesellschaft genannt. Durch den gezielten Einsatz von KI soll dieser wirtschaftliche Nachteil und diese gesellschaftliche Problematik adressiert werden.

Vergleicht man KI mit einem herkömmlichen Handwerkzeug, z.B. einem Hammer, wird jedoch auch schnell klar, dass der mögliche Eingriff von KI in das Leben der Menschen viel tiefer gehen kann als der eines Hammers. Einen Hammer kann man in der Hand halten und ihn als Verlängerung des menschlichen Körpers, als Manipulator, verwenden. Legt man den Hammer wieder zur Seite, hat er von sich aus keine Möglichkeit mehr, in das Leben des/ der Nutzer*in einzugreifen. KI jedoch greift in bestimmten Bereichen ggf. viel tiefer in das Leben der Menschen ein und das teilweise ohne, dass den Nutzer*innen das zu jedem

[14] Europäische Kommission (2021a), S.47.

[15] Europäische Kommission (2021b), S.2.

Zeitpunkt bewusst ist. Ein Beispiel dafür sind KI-gestützte **Sprachassistenzsysteme**, die über das Handy oder andere Endgeräte permanent „mithören“ und so von den Gesprächen für zukünftige Interaktionen lernen. Daher ist die Abgrenzung von gewolltem Anwendungsbereich zu anderen Bereichen bei KI schwieriger als bei einem einfachen mechanischen Werkzeug, wie einem Hammer.

Wenn von KI gesprochen wird, bestehen aufgrund mangelnder Fachkenntnis, bzw. Aufklärung der breiten Masse der Gesellschaft, oft Vorbehalte gegenüber KI.[16] Wie bei vielen Technologien sind Science-Fiction-Filme häufig frühe Wegbereiter. Jedoch werden in der Gesellschaft die fiktionalen Anteile der Filme auf Ängste schürende Weise zu Vorurteilen gegen die Technologie. An dieser Stelle sei als Beispiel der Film *I, Robot*[17] genannt, in dem ein **Zentralcomputer** über Software-Aktualisierungen eine große Anzahl von Robotern unter seine Kontrolle bringt. Durch eine scheinbar **starke KI** ist der Zentralcomputer zu dem Ergebnis gekommen, die Menschen zu entmündigen. Die Roboter nutzt er, um den Menschen seinen Willen aufzuzwingen. Solche Szenarien sind aus momentaner Sicht allein aufgrund der mangelnden starken KI in der Realität nicht zu befürchten.

Allerdings gibt es auch heute schon umgesetzte Anwendungsbeispiele, welche bei Menschen in einer Demokratie, zu Unverständnis bzw. totaler Ablehnung führen. Das Demokratie-Verständnis beruht, auch in der europäischen Union, unter anderem auf der Freiheit des Individuums, die nur in besonderen Fällen eingeschränkt werden soll.

In anderen politischen Systemen bzw. Kulturkreisen gibt es dazu andere Ansätze. Die individuelle Freiheit ist in z.B. China dem Kollektiv untergeordnet. Dies ist sowohl der Politik als auch dem kulturellen Selbstverständnis geschuldet. So gibt es in chinesischen Städten eine sehr weite Verbreitung von Überwachungsmaßnahmen der einzelnen Personen mittels sogenanntem **Social Scoring**. Dabei wird das Handy genutzt, um eine Vielzahl an Tätigkeiten mit der sozialistischen Ideologie abzugleichen und einzustufen. Dies erfolgt über ein Punktevergabesystem in einer App. Vergleichbar mit einem Spiel kann man durch gute Aktionen in verschiedenen Kategorien Medaillen verdienen, z.B.: Geringer CO_2 - Fußabdruck durch die eigenen Tätigkeiten, Wohltätigkeitsspenden, Sport oder Freizeitaktivitäten. Werden die öffentlichen Verkehrsmittel verwendet, wird dies beispielsweise einem geringen Kohlenstoffverbrauch zugeordnet. Man sammelt Punkte und erhöht damit seinen Social Score. Diese Punkte können für kostenlose Angebote eingelöst werden. Begeht man jedoch Fehler im Sinne der Ideologie, kann man auf einer schwarzen Liste landen, die einem die Teilhabe in vielen Bereichen des gesellschaftlichen Lebens verwehrt.[18] Dies kann bereits beim Überqueren einer Straße bei roter Ampel erfolgen. Die Erkennung erfolgt dabei durch automatische Gesichtserkennung.

Diese Beispiele zeigen, dass eine solche Umsetzung mit dem Wertekonsens der europäischen Union nicht vereinbar ist. Daher und mit dem Wissen, dass andere wirtschaftsstarke Nationen, wie die Vereinigten

[16] Vgl. Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK) (2021), S. 36.

[17] Vgl. Proyas u.a. (2004), S.1.

[18] Vgl. Gaylor (2020), S.1.

Staaten von Amerika und die Volksrepublik China bereits große Schritte in der Umsetzung von KI gemacht haben, hat die europäische Kommission eine Initiative in Bezug auf KI gestartet. Dabei soll KI aus der europäischen Union den Menschen und der Gesellschaft zugutekommen unter Wahrung eines hohen Niveaus des Datenschutzes, der digitalen Rechte und der ethischen Standards.[19] Wie später dargelegt wird, sind dies Forderungen, welche in die vertrauenswürdige KI überführt wurden.

In Österreich gibt es, zum Zeitpunkt des Verfassens der vorliegenden Arbeit, eine aktuelle Pressemeldung des Bundesministeriums Finanzen, dass eine KI, genauer Machine Learning, vom Predictive Analytics Competence Center (PACC) des Ministeriums verwendet wird, um **Steuerhinterziehung und Zollbetrug** aufzudecken. Beziffert wird der Erfolg des KI-Einsatzes in diesem Bereich mit 540 Mio. Euro, die sich der Steuerzahler in Österreich allein seit Anfang des Jahres 2020 gespart haben soll.[20] Das zeigt, dass bestimmte Anwendungsbereiche aus europäischer Sicht direkt ausgeklammert bzw. verboten werden. Andere Bereiche, die dem Erhalt und der Durchsetzung der Gesetze dienen, können oder sollen sich aber durchaus der Möglichkeiten bedienen, welche die KI bietet.

Darüber hinaus gibt es schon zahlreiche Anwendungsfälle z.B. bei **selbstfahrenden Fahrzeugen**, in der **Regelungstechnik**, der **Übersetzung von Sprachen** oder der **Erkennung von Objekten**. Im Rahmen der vorliegenden Arbeit soll hier nur ein kleiner Einblick gegeben werden. In weiterer Folge werden jedoch die Besonderheiten der vertrauenswürdigen KI genauer betrachtet.

[19] Vgl. Europäische Kommission (2018), S.3.

[20] Vgl. Bundesministerium Finanzen (2022), S.1.

1.3 KI-Strategie der EU

Die europäische KI-Initiative dient dabei einerseits der Sicherstellung der Wettbewerbsfähigkeit im Bereich KI, andererseits soll sie: „einen KI-Ansatz fördern, der den Menschen und der Gesellschaft insgesamt zugutekommt.“[21] Dies lässt den menschenorientierten Ansatz erkennen, der dem Ansatz aus den zuvor genannten, anderen politischen Systemen entgegensteht. Es wird damit nicht versucht, den Entwicklungsvorsprung anderer Länder aufzuholen, sondern einen Teilbereich, die Vertrauenswürdige KI, aus der gesamten KI herauszunehmen und diesen eigenständig auszugestalten.

Die europäische Union ist schon seit vielen Jahren stark in der Regulatorik, d.h. in der Festschreibung der einzuhaltenden Verordnungen, Gesetze und Normen. Auch wenn die Regulatorik teilweise als entwicklungshemmend gesehen wird, zeigt sich anhand der Datenschutzgrundverordnung[22], die seit 2018 in Kraft getreten ist, dass sie auch positive Seiten haben kann. So hat jeder Europäer die unmittelbaren Auswirkungen kennen gelernt, als z.B. bei Ärzt*innen, bei Friseur*innen oder in Hotels plötzlich dafür unterschrieben werden musste, dass die persönlichen Daten gespeichert werden durften. Doch wurden sich dadurch nicht nur die Menschen der Tatsache bewusst, dass es eine große Anzahl an persönlichen gespeicherten Daten gibt. Es hebt sich auch die damit verbundene Freiheit des Europäers beispielsweise von der eines US-Amerikaners ab. In den USA ist es speziell den Geheimdiensten möglich, auf diverse persönliche Daten zuzugreifen, da es nationale gesetzliche Regelungen, wie den Patriot Act[23] oder den Cloud Act[24] gibt.

1.3.1 Ethische Leitlinien

Die Europäische Kommission hat daher eine unabhängige hochrangige Expertengruppe für Künstliche Intelligenz (High Level Expert Group on AI - HLEG)[25] eingesetzt und diese damit beauftragt, Ethik-Leitlinien für eine vertrauenswürdige KI zu erarbeiten. Die HLEG hat in diesen Leitlinien zunächst drei Komponenten definiert, die eine vertrauenswürdige KI auszeichnen:

Sie sollen **rechtmäßig**, **ethisch** und **robust** sein.[26] Bei der weiteren Ausführung wird der rechtmäßige Anteil nicht weiter behandelt. Stattdessen sollen die ethischen und robusten Grundsätze bei der tatsächlichen Umsetzung von KI-Systemen konkretisiert werden.

In dem zweiten Kapitel der Leitlinien werden dann unter Berücksichtigung dieser Grundsätze konkrete Anforderungen definiert, die für eine Umsetzung von vertrauenswürdiger KI notwendig sind – siehe Abb. 5.

[21] Europäische Kommission (2018), S.3.

[22] Vgl. Europäische Kommission (2016), S.1 ff.

[23] Vgl. U.S. Government (2001), S.1 ff.

[24] Vgl. U.S. Government (2018), S.1 ff.

[25] Vgl. High-Level Expert Group on AI (2022), S.1.

[26] Vgl. High-Level Expert Group on AI (2019), S.2.

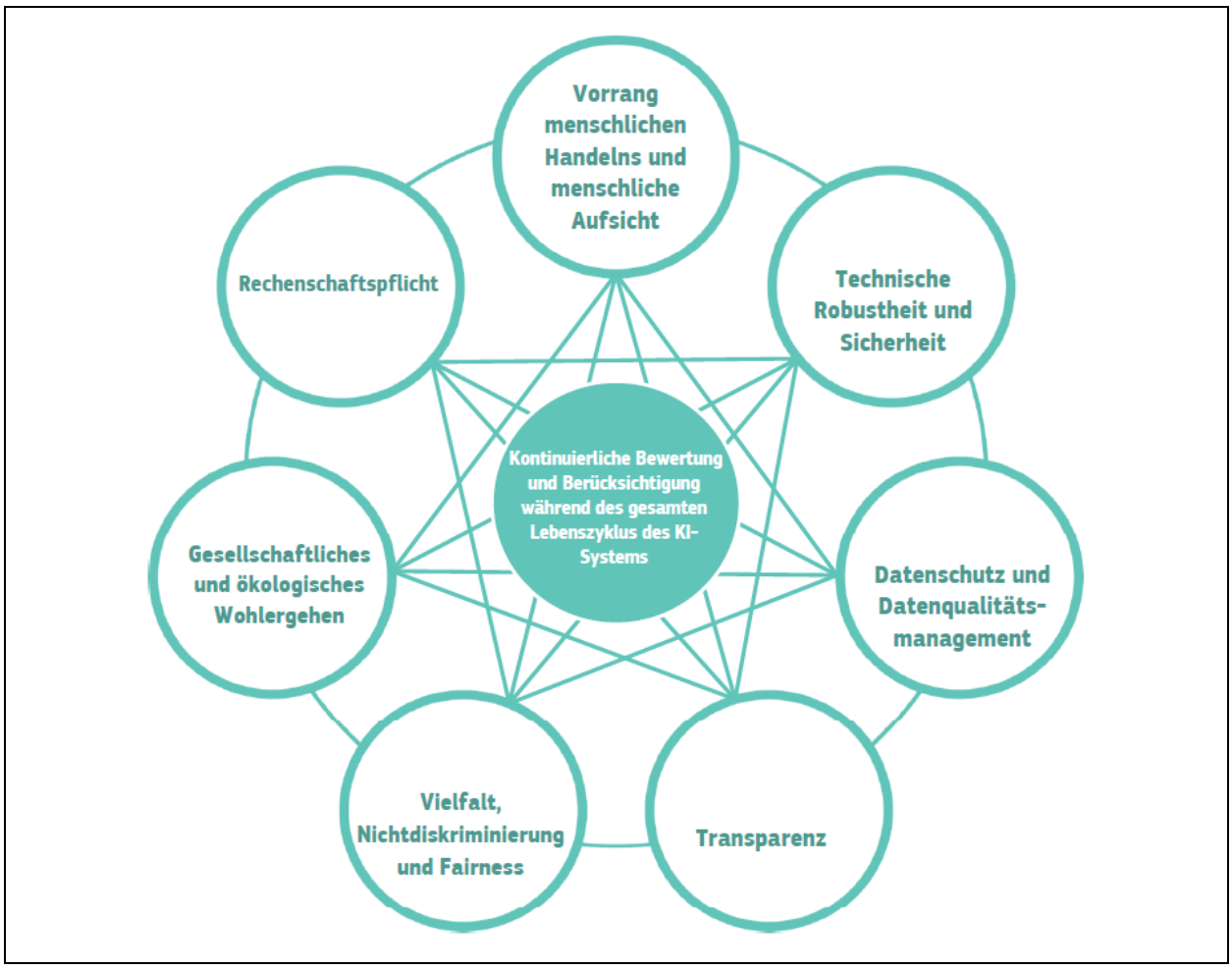


Abb. 5: Anforderungen an vertrauenswürdige KI, Quelle: Ethics Guidelines on trustworthy AI.[27]

In diesen konkreten Anforderungen der HLEG sind schon die wichtigen Punkte genannt, die anschließend auch Eingang in den Leitfaden des Fraunhofer IAIS gefunden haben und dort die Dimensionen der Vertrauenswürdigkeit genannt werden.[28]

1.3.2 Investitionen

Eine wesentliche Forderung der europäischen Kommission im Rahmen der KI-Initiative ist eine deutliche Steigerung von Investitionen in diesem Bereich:

„Ziel dieser Investitionen ist es, Forschung und Innovationen im Bereich der KI zu konsolidieren, Tests und Versuche zu fördern, die Exzellenzzentren im Bereich der KI-Forschung zu stärken und Initiativen einzuleiten, mit denen KI allen potenziellen Nutzern und insbesondere kleinen und mittleren Unternehmen zugänglich gemacht werden soll.“[29]

[27] High-Level Expert Group on AI (2019), S.18.

[28] Vgl. Poretschkin u.a. (2021), S. 22.

[29] Europäische Kommission (2018), S.8.

Ein Ziel der vorliegenden Arbeit soll diesen Punkt aufnehmen und die Inhalte des Leitfadens des Fraunhofer IAIS in leicht verständlicher Weise auch für KMU verfügbar machen.

Schaut man sich die bisherigen Investitionen im weltweiten Vergleich in den Jahren 2013 bis 2021 an, wird auch deutlich, dass eine sehr große Differenz zwischen den USA, China und dem Rest der Welt besteht und besonders Europa einen deutlichen Nachholbedarf hat – siehe Abb. 6.

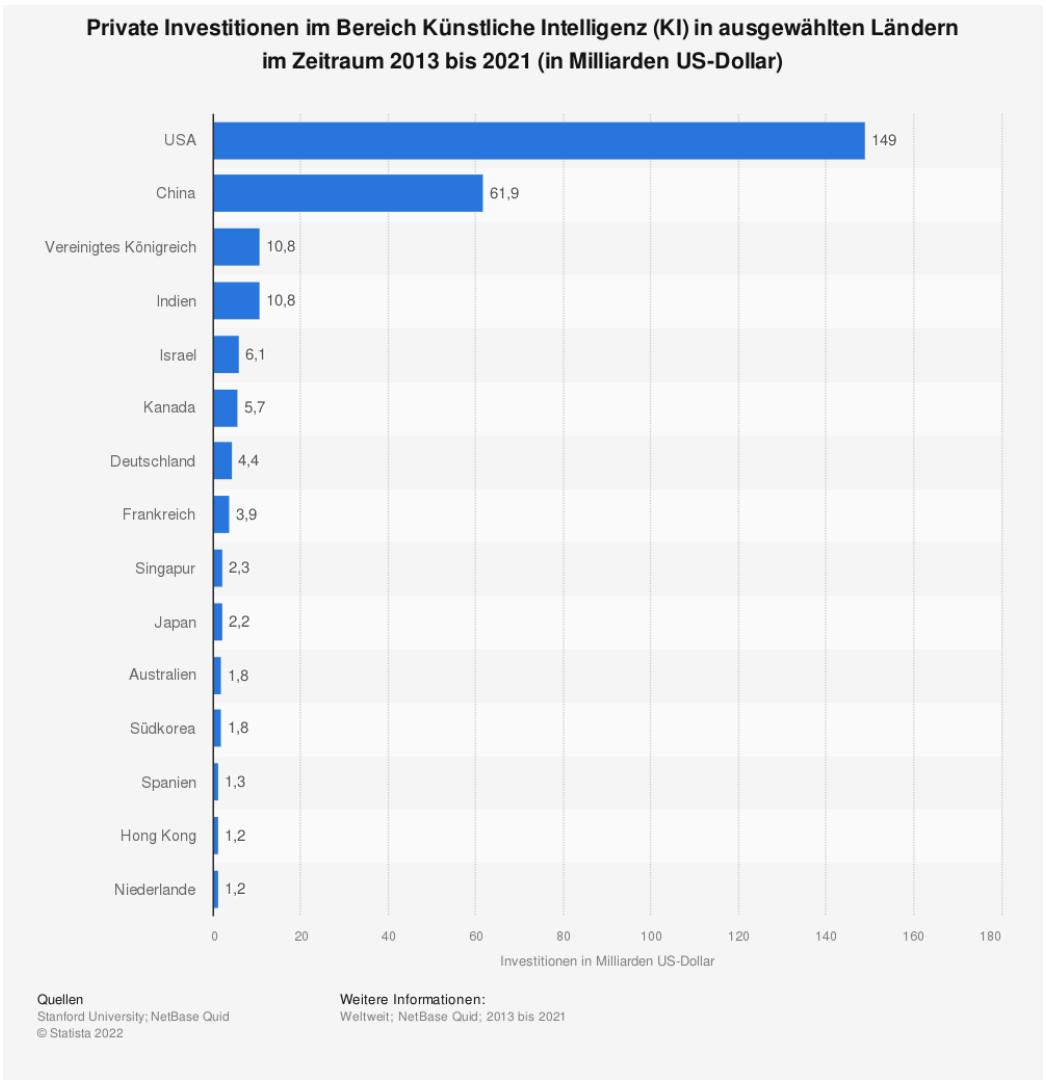


Abb. 6: Private Investitionen im Bereich KI, Quelle: The AI Index 2022 Annual Report.[30]

Politisch getrieben wird beispielsweise in China jedoch weit darüber hinaus investiert. So hat der chinesische Staatsrat bereits im Jahr 2017 eine Strategie veröffentlicht, nach der China in drei Schritten bis in das Jahr 2030 die weltweite Vormachtstellung erreichen soll und beziffert die Vorgabe mit einem Wert für die KI-Industrie mit 130 Milliarden Euro und dem damit verbundenen Industriezweig mit 1,2 Billionen Euro.[31]

[30] Zhang u.a. (2022), S. 155.

[31] Vgl. Groth u.a. (2018), S. 18.

Dies ist der europäischen Kommission durchaus bewusst, weshalb die KI-Initiative auch ein investitionsfreundliches Umfeld fördern und private Investitionen durch öffentliche Mittel stützen soll.[32]

1.4 Strategie KI der Bundesregierung in Deutschland

Der KI-Initiative der europäischen Kommission folgend, wurden national ebenfalls Strategien ausgearbeitet. In diesem Zusammenhang hat die deutsche Bundesregierung zwölf Handlungsfelder definiert, auf denen der nationale Fokus liegen soll. Für die vorliegende Arbeit sind zwei dieser Handlungsfelder interessant und sollen daher kurz erläutert werden.

1.4.1 Ordnungsrahmen anpassen

Im Handlungsfeld *Ordnungsrahmen anpassen* wird sinngemäß festgelegt, dass KI in allen Anwendungsbereichen immer unter Berücksichtigung der ethischen Grundwerte und des Schutzes des Individuums umgesetzt werden soll: „Ziel der Bundesregierung wird sein, […] die Grundwerte der freiheitlich-demokratischen Grundordnung der Bundesrepublik Deutschland ebenso wie den verfassungsrechtlich verankerten Schutz der Grundrechte – insbesondere des Rechts auf allgemeine Handlungsfreiheit, auf Privatsphärenschutz und informationelle Selbstbestimmung – unangetastet bleiben.“[33] Das heißt auch in der nationalen Umsetzung wird dezidiert auf die Wahrung der ethischen Leitlinien der HLEG beharrt und das Thema als ein europäisches Kernthema beansprucht.

1.4.2 Standards setzen

In diesem Handlungsfeld besinnt sich die deutsche Bundesregierung auf die bestehende Stärke in der Normierung.[34] Durch das Entwickeln von Standards sollen auch andere Länder in eine Richtung gebracht werden, welche von Europa gelenkt werden kann. Dafür wird eine Normungsroadmap KI in Zusammenarbeit mit dem Deutschen Institut für Normung (DIN) erstellt.[35] Die Normungsroadmap KI soll dabei Handlungsempfehlungen für die Entwicklung und Überprüfung von Normen in Bezug auf Themengebiete aus der gesamten KI geben. Eine dieser Handlungsempfehlungen besteht in der Erstellung einer horizontalen KI-Basis-Sicherheitsnorm, die gleichermaßen für alle KI-Systeme anwendbar ist und Themen wie Security, Safety und Privacy abdeckt. Damit soll die Akzeptanz der Normung in diesem Bereich gesteigert werden.[36] Die weiteren vier Handlungsempfehlungen zielen außerdem darauf ab, dass wiederum die ethischen und demokratischen Wertvorstellungen bei Anwendung jeglicher KI-Systeme gewahrt bleiben. Mittels Normung soll jedoch nicht der Fortschritt gedrosselt werden, es sollen lediglich die

[32] Vgl. Europäische Kommission (2018), S.5.

[33] Deutsche Bundesregierung (2018), S. 38.

[34] Siehe auch Kapitel 1.3.

[35] Vgl. Deutsche Bundesregierung (2018), S. 41.

[36] Vgl. Wahlster/Winterhalter (Hrsg.) (2020), S.24.

kritischen Systeme, die einen Einfluss auf individuelle Grundrechte haben, ermittelt und über einen Zertifizierungsprozess abgesichert werden.[37] Die Handlungsempfehlungen sind die folgenden:

1. Datenreferenzmodelle für die Interoperabilität von KI-Systemen umsetzen
2. Horizontale KI-Basis-Sicherheitsnorm erstellen
3. Praxisgerechte initiale Kritikalitätsprüfung von KI-Systemen ausgestalten
4. Nationales Umsetzungsprogramm „Trusted AI“ zur Ertüchtigung der europäischen Qualitätsinfrastruktur initiieren und durchführen
5. Use Cases auf Normungsbedarf analysieren und bewerten

1.5 Strategie der Bundesregierung für KI in Österreich (AIM AT 2030)

In der österreichischen Umsetzung der Strategie werden zunächst drei Ziele genannt:

1. Die Bundesregierung strebt für Österreich einen auf das Gemeinwohl orientierten Einsatz von KI an.
2. Österreich soll ein international anerkannter Forschungs- und Innovationsstandort für KI werden.
3. KI soll helfen, die Wettbewerbsfähigkeit des österreichischen Technologie- und Wirtschaftsstandorts zu sichern.

Dabei liegt auch hier wieder ein menschenzentrierter Ansatz zugrunde, somit ist eine Durchgängigkeit von europäischer auf die nationale Ebene klar erkennbar. Auch dem Thema vertrauenswürdige KI ist ein eigenes Kapitel gewidmet, wobei hier keine weiteren inhaltlichen Konkretisierungen im Vergleich zu den ethischen Leitlinien auf europäischer Ebene zu erkennen sind. Es soll jedoch ein gesellschaftlicher Dialog zur KI angeregt werden, um die Akzeptanz von KI-Systemen auf einer breiten Basis zu schaffen.[38]

Der Einsatz von KI für das Gemeinwohl, wird in Österreich durch die Anwendung beim Bundesministerium Finanzen, bereits durch die Regierung in Teilen umgesetzt.[39]

1.6 Relevante EU-Verordnungen

Um den gesetzgebenden Prozess zu beschleunigen und um die europäische Gesetzgebung und deren Ziele auf nationaler Ebene möglichst gleichförmig zu gestalten, ist die europäische Kommission in den letzten Jahren vermehrt dazu übergegangen, anstelle von EU-Richtlinien sogenannte EU-Verordnungen zu verabschieden. Dies hat aus europäischer Sicht den Vorteil, dass die Verordnungen direkt gelten, d.h. sie richten sich an den Inverkehrbringer des Produkts, welches vom Anwendungsbereich der jeweiligen Verordnung erfasst wird. Die, zuvor und auch jetzt noch teilweise vorhandenen, EU-Richtlinien richten sich an die Mitgliedsstaaten und müssen, in Übereinstimmung mit der nationalen Verfassung, zunächst in

[37] Vgl. Wahlster/Winterhalter (Hrsg.) (2020), S.25.

[38] Vgl. Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK) (2021), S. 22 ff.

[39] Vgl. Kapitel 1.2.

nationales Recht überführt werden. Dabei gibt es Rahmen-Richtlinien, die ein relativ breites Thema abdecken, wie z.B. den Arbeitnehmer*innenschutz. Die Rahmen-Richtlinien werden in nationale Gesetze überführt, im Beispiel auf Abb. 7 ist das österreichische Arbeitnehmer*innenschutzgesetz (ASchG) zu sehen. Außerdem gibt es sogenannte Einzel-Richtlinien, die etwas enger gefasste Themen behandeln, z.B. aus dem übergeordneten Bereich Arbeitnehmerschutz der Teilbereich der Arbeitsmittel, wie z.B. Werkzeuge, Maschinen und Anlagen.

Diese Umsetzung von europäischer Richtlinie zu nationaler Regelung führt zu mehreren möglichen nachgeschalteten Problemstellungen:

- Die Richtlinien werden in Teilaspekten unterschiedlich ausgelegt
    - aufgrund von unterschiedlichen Übersetzungen je nach Mitgliedsstaat
    - aufgrund von unterschiedlichen nationalen Ausformulierungen der Richtlinien
- Die Umsetzung in nationales Recht führt zu Verzögerungen des Inkrafttretens der jeweiligen EU-Richtlinie

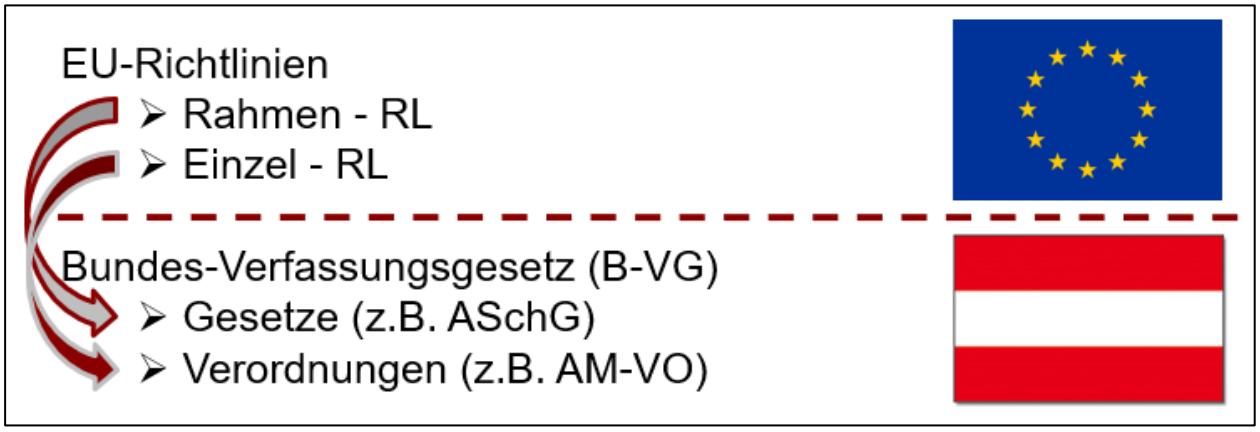


Abb. 7: Umsetzung EU-Richtlinien in nationales Recht, Quelle: Eigene Darstellung.

Für den neuen Bereich der KI sind aus europäischer Sicht dementsprechend EU-Verordnungen zu bevorzugen. Dies wurde auch in dem Vorschlag zur KI-Verordnung mit einer „Notwendigkeit einer einheitlichen Anwendung der neuen Vorschriften“[40] argumentiert. Nachfolgend sollen daher einige, für KI relevante, Verordnungen betrachtet werden.

1.6.1 DSGVO

Bedenkt man die bereits besprochenen Ethik-Leitlinien für eine vertrauenswürdige KI der HLEG, so ist eine der konkret formulierten Anforderungen der **Datenschutz und das Datenqualitätsmanagement**.[41] Wird nun der ersten Erwägungsgrund für die Ausarbeitung der Datenschutzgrund-Verordnung (DSGVO)

[40] Europäische Kommission (2021a), S.8.

[41] Vgl. Abb.5, S.11.

betrachtet, wird klar, dass die seit 2018 in Europa verbindlich anzuwendende Verordnung bereits dieses konkrete Thema adressiert:

„Der Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten ist ein Grundrecht.“[42]

Die DSGVO definiert verschiedene Verantwortlichkeiten, je nachdem, wie personenbezogene Daten verwendet werden und beschreibt auch, wann es überhaupt rechtmäßig ist, diese Daten zu verarbeiten, z.B. nach Einwilligung oder für die Erfüllung eines Vertrags.[43] Daraus geht deutlich hervor, dass personenbezogene Daten immer nur zu einem bestimmten Zweck verwendet werden dürfen. Somit soll niemand weder versehentlich, noch beabsichtigt, in seinen individuellen Rechten durch die fehlerhafte Verwendung von Daten eingeschränkt werden.

Seit dem Jahr 2018 legt die DSGVO damit schon den Grundstein für einen Teil der Vertrauenswürdigkeit, wie sie gemäß HLEG gefordert wird.

1.6.2 EU-Maschinenprodukteverordnung

Die, zum Zeitpunkt des Verfassens der vorliegenden Arbeit, aktuelle Maschinenrichtlinie 2006/42/EG befindet sich in der Überarbeitung und es existiert bereits ein Vorschlag für die neue Version, die fortan als EU-Verordnung[44] veröffentlicht werden soll.

Als ein Grund für die Überarbeitung der Maschinenrichtlinie werden gleich zu Beginn mögliche Risiken genannt, die sich aus aufstrebenden Technologien ergeben und nicht ausreichend von der bestehenden Maschinenrichtlinie abgedeckt werden. Im Detail werden z.B. autonome Maschinen genannt, aber auch ganz konkret künstliche Intelligenz. Es wird speziell auf Hochrisiko-KI-Systeme hingewiesen, die in eine Maschine integriert sind und in Zukunft von einer eigenen KI-Verordnung abgedeckt werden sollen.[45] Ebenso wird auf Sicherheitskomponenten verwiesen, die den Anhang II des neuen Vorschlags der EU-Maschinenprodukteverordnung, die nicht erschöpfende Liste der Sicherheitskomponenten, um folgenden Punkt erweitern sollen:

„18. Software, die Sicherheitsfunktionen wahrnimmt, einschließlich KI-Systeme.“[46] Das heißt, KI-Systeme werden in der Verordnung dezidiert aufgenommen und müssen von Maschinenherstellern zukünftig bei der Risikobeurteilung berücksichtigt werden, wenn sie in einer Maschine eingesetzt werden und sicherheitstechnisch relevant sind. Mögliche, daraus resultierende, Sicherheitsmaßnahmen inklusive Tests und Validierung sind dementsprechend auch umzusetzen. Das Thema der Sicherheit, welches zuvor schon

[42] Europäische Kommission (2016), S.1.

[43] Vgl. Europäische Kommission (2016), S.36.

[44] Vgl. Kapitel 1.6.

[45] Vgl. Europäische Kommission (2021c), S.2 ff.

[46] Europäische Kommission (2021d), S.3.

in den Ethischen Leitlinien der HLEG definiert wurde, findet sich im Leitfaden des Fraunhofer IAIS als eigene Dimension der Vertrauenswürdigkeit.

Die **funktionale Sicherheit** ist heutzutage schon eines der Kernthemen in der Maschinensicherheit und wird in der Praxis durch steuerungstechnische Sicherheitsfunktionen, wie beispielsweise mit Sicherheitslichtvorhängen (berührungslos trennende Schutzeinrichtung), Schutztürüberwachungen (beweglich trennende Schutzeinrichtungen mit Verriegelung und ggf. Zuhaltung) oder auch Not-Halt-Einrichtungen realisiert. Die Ausfallwahrscheinlichkeit dieser Sicherheitsfunktionen wird rechnerisch durch Einhaltung der EN ISO 13849-1 und EN ISO 13849-2 bzw. der EN IEC 62061 nachgewiesen. Damit existieren schon harmonisierte Normen, die ein weiteres Thema abdecken, welches ebenfalls ein Risikogebiet im Leitfaden des Fraunhofer IAIS darstellt.[47]

Die sogenannte KI-Verordnung[48], bzw. im Englischen AI-Act genannt, wird parallel zur EU-Maschinenprodukteverordnung neugestaltet und weitergehende, über die Anforderungen der EU-Maschinenprodukteverordnung, Inhalte detaillieren.

1.6.3 KI-Verordnung

Die KI-Verordnung liegt, zum Zeitpunkt des Verfassens der vorliegenden Arbeit, als Vorschlag vor. Daher ist nicht klar, welche Inhalte in der tatsächlichen Verordnung verbleiben und welche nicht. In weiterer Folge wird auf den IST-Stand des Vorschlags Bezug genommen und es wird, der besseren Lesbarkeit geschuldet, von der KI-Verordnung gesprochen, auch wenn die tatsächliche Verordnungsbezeichnung eine andere sein wird.

Auch in dem Vorschlag zur KI-Verordnung wird als ein Ziel wieder inhaltlich das aufgenommen, was zuvor im Rahmen der KI-Strategie der EU[49] bereits definiert wurde:
„Es muss gewährleistet sein, dass die auf dem Unionsmarkt in Verkehr gebrachten und verwendeten KI-Systeme sicher sind und die bestehenden Grundrechte und die Werte der Union wahren.“[50]

Die KI-Verordnung folgt einem risikobasierten Ansatz und unterteilt in vier verschiedene Gruppen von KI-Systemen, die unterschiedlich zu beurteilen sind bzw. denen unterschiedliche Anforderungen zugeteilt werden – siehe Abb. 8. Sind Anwendungen unvereinbar mit den europäischen Werten, stellt dies ein unannehmbares Risiko dar und ist verboten. Als Beispiel sei hier das Verbot von KI-Systemen für die soziale Bewertung von natürlichen Personen genannt, das in direktem Gegenteil zu Praktiken in anderen Kulturkreisen steht.[51] Bei den sogenannten Hochrisiko-KI-Systemen ist ein verpflichtendes Konformitätsbewertungsverfahren durchzuführen, da mögliche Auswirkungen auf die Grundrechte zu

[47] Vgl. Kapitel 1.7.2.5.

[48] Vgl. Europäische Kommission (2021a), S.1 ff.

[49] Vgl. Kapitel 0.

[50] Europäische Kommission (2021a), S.3.

[51] Vgl. Social Scoring Kapitel 1.2.

befürchten sind. Bei KI-Systemen mit geringem oder minimalem Risiko wird in der Bezeichnung unterschieden, maßnahmenseitig wird jedoch bei beiden Anwendungen auf die freiwillige Einhaltung der Konformitätsbewertung abgezielt.[52]

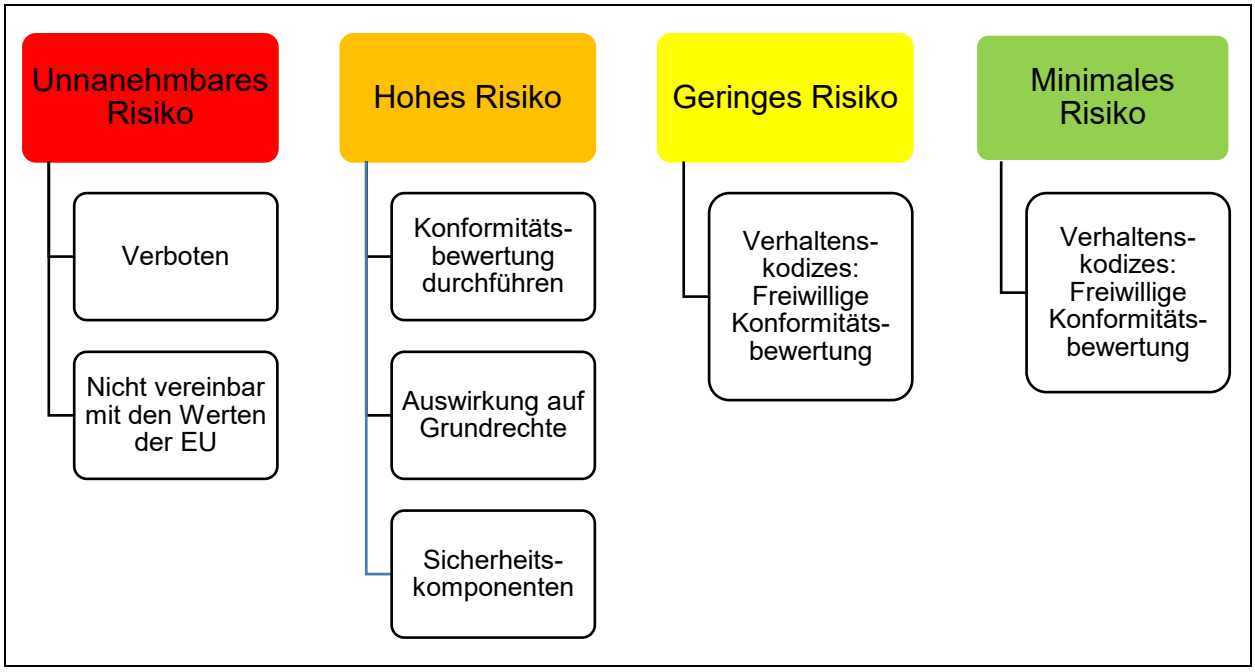


Abb. 8: Risikobasierte Einstufung von KI-Systemen, Quelle: Eigene Darstellung.

Wie von anderen europäischen Richtlinien bereits bekannt, werden auch in der KI-Verordnung sowohl dem Hersteller, als auch weiteren Wirtschaftsakteuren, wie Einführern, Händlern und Bevollmächtigten, entsprechende Pflichten zugedacht.[53] Dies resultiert aus der Erfahrung, dass es bei dem System der reinen Herstellerhaftung verschiedene Möglichkeiten für beteiligte Wirtschaftsakteure gab, sich trotz Profit an einem Produkt oder einer Dienstleistung, einer rechtlichen Verantwortung zu entziehen, z.B. wenn der Hersteller außerhalb der EU ansässig und damit für die europäische Justiz nicht greifbar gewesen wäre.

In der KI-Verordnung wird im verfügenden Teil sowohl die Durchführung einer Risikobeurteilung als auch die entsprechende Dokumentation als Vorbereitung für eine dritte Stelle gefordert.[54] An dieser Stelle setzt der Leitfaden des Fraunhofer IAIS konkret an und definiert die inhaltlich notwendigen Anwendungsgebiete und die Funktionsweise des KI-Systems. Darüber hinaus werden Maßnahmen für bestimmte Teile - Daten, KI-Komponente, Einbettung und Maßnahmen für den Betrieb - des KI-Systems definiert.[55] Dies ist in diesem Detailierungsgrad bisher an keiner anderen Stelle, z.B. in EN bzw. EN ISO Normen zu finden, da solche Normen erst entwickelt werden müssen.

[52] Vgl. Europäische Kommission (2021a), S.14 ff.

[53] Vgl. Europäische Kommission (2021a), S.16.

[54] Vgl. Europäische Kommission (2021a), S.54 ff.

[55] Vgl. Poretschkin u.a. (2021), S.32 ff.

1.7 Vertrauenswürdigkeit

Spricht man von Vertrauenswürdigkeit als Eigenschaft, so beinhaltet dies mitunter ein Geben von etwas, in der Hoffnung, dass dieses Geben auf positive Art und Weise erwidert wird. Dabei können verschiedenste Dinge (preis-)gegeben werden, z.B. Geld, Gefühle aber auch Informationen. Wird dieses Geben negativ oder gar nicht erwidert, so das Vertrauen entzogen und es wird sich so etwas nicht wiederholen. Zudem reicht es mitunter aus, dass ein einmaliges diesbezügliches negatives Erlebnis zu einem Vertrauensentzug führt, da dies womöglich auf Naivität oder Gutgläubigkeit schließen lassen könnte.

Sagt Person A zu Person B, die sich als Freunde bezeichnen, beispielsweise, dass sie eine dritte Person nicht mag, so wird innerhalb dieser Freundschaft darauf vertraut, dass Person B diese Information nicht an die dritte Person weitergeben wird. Macht sie das dennoch, wird Person A in Zukunft keine solchen vertraulichen Informationen an Person B geben. Im schlimmsten Fall kann dies sogar zu einer Beendigung der Freundschaft führen.

Analog dazu verhält es sich mit Informationen, die automatisiert von einer Software - im Rahmen der vorliegenden Arbeit KI-Systeme - verarbeitet werden und im Fehlerfall diese Informationen einem nicht autorisierten Empfänger weitergibt oder diese Informationen falsch verarbeitet. Im schlimmsten Fall hat dies größere Konsequenzen für diejenige Person, welche die Informationen gegeben hat. Daraus resultierend wird diese Person, das KI-System zukünftig gar nicht mehr verwenden. Aufgrund der breiten Anwendungsmöglichkeit von KI-Systemen können die Konsequenzen durchaus beachtlich sein. Daher sind die entwickelten Anforderungen aus dem Leitfaden bezüglich vertrauenswürdiger KI zwar sehr umfangreich, aber durchaus nachvollziehbar und für eine dementsprechende Risikominimierung sicher zielführend.

1.7.1 Vertrauenswürdige KI und mögliche Problemstellungen

Problematisch in Bezug auf Vertrauenswürdigkeit von KI-Systemen sind, wie zuvor schon genannt, negative Vorstellungen, die durch fiktionale Filme erweckt werden.[56] Allerdings gibt es auch schon konkrete Erfahrungen mit bestehenden KI-Systemen, die auf verschiedene Art und Weise negativ auf sich aufmerksam gemacht haben. So gab es ein Beispiel, bei dem ein Chatbot, aufgrund seiner KI-gestützten Lernfähigkeit, rassistische, sexistische und antisemitische Äußerungen von sich gab.[57]

Ebenso existierte ein Fall, bei dem eine KI-gestützte Software von Google Personen mit schwarzer Hautfarbe auf Fotos als Gorillas eingestuft hatte. Weitere bekannte Beispiele, bei denen Personen durch die Verwendung von KI offensichtlich diskriminiert wurden, liegen in Bereichen der Kreditvergabe oder auch bei automatisiert verarbeiteten Bewerbungen für Arbeitsstellen.

[56] Vgl. Kapitel 1.2.

[57] Vgl. Beck u.a. (2019), S.6.

Selbst bei dem, auf den ersten Blick vermeintlich harmlosen, KI-gestützten Tool DALL-E, ist den Erstellern nach der Veröffentlichung der ersten Version klar geworden, dass Maßnahmen vorgesehen werden müssen, damit das Tool nicht für negative Zwecke verwendet wird. Das Tool bietet folgende Funktionalität: Der/die Anwender*in gibt einen beschreibenden Text ein und der Algorithmus erstellt ein Bild, in mehreren Variationen, welches den beschreibenden Text darstellt. In der zweiten Version des Tools haben die Ersteller Maßnahmen gegen die Erzeugung von gewaltverherrlichenden Bildern und anderen negativen Resultaten gesetzt:

„Preventing Harmful Generations - We've limited the ability for DALL·E 2 to generate violent, hate, or adult images. By removing the most explicit content from the training data, we minimized DALL·E 2's exposure to these concepts. We also used advanced techniques to prevent photorealistic generations of real individuals' faces, including those of public figures. "[58]

1.7.2 Vertrauenswürdigkeit im Leitfaden des Fraunhofer IAIS

Im Leitfaden werden sechs Dimensionen der Vertrauenswürdigkeit definiert. Basierend auf den Ethischen Leitlinien[59] der HLEG wurden diese sechs Kernthemen dazu bestimmt, Vertrauenswürdigkeit von KI-Systemen zu bestimmen. Den sechs Dimensionen wiederum sind Risikogebiete zugeordnet, welche in weiterer Folge konkrete Risikobetrachtungen und Maßnahmen fordern. Für eine Übersicht der Dimensionen und der zugehörigen Risikogebiete – siehe Tab. 1.

[58] Vgl. https://openai.com/dall-e-2/.

[59] Vgl. High-Level Expert Group on AI (2019), S.10 ff.

Dimension	Risikogebiete
Fairness (FN)	Fairness (FN)
	Beherrschung der Dynamik (BD)
Autonomie und Kontrolle (AK)	Angemessene und verantwortungsvolle Gestaltung der Aufgabenverteilung zwischen Mensch und KI-Anwendung (GE)
	Sicherstellung der Informiertheit und Befähigung von Nutzer*innen und Betroffenen (IB)
Transparenz (TR)	Transparenz gegenüber Nutzer*innen und Betroffenen (NB)
	Transparenz für Expert*innen (EX)
	Auditfähigkeit (AF)
	Beherrschung der Dynamik (BD)
Verlässlichkeit (VE)	Verlässlichkeit im Regelfall (RE)
	Robustheit (RO)
	Abfangen von Fehlern auf Modellebene (AF)
	Einschätzung von Unsicherheit (UN)
	Beherrschung der Dynamik (BD)
Sicherheit (SI)	Funktionale Sicherheit (FS)
	Integrität und Verfügbarkeit (IV)
	Beherrschung der Dynamik (BD)
Datenschutz (DS)	Schutz personenbezogener Daten (PD)
	Schutz geschäftsrelevanter Information (GI)
	Beherrschung der Dynamik (BD)

Tab. 1: Übersicht Dimensionen und Risikogebiete, Quelle: Eigene Darstellung.

Um ein Verständnis der Dimensionen der Vertrauenswürdigkeit und der zugeordneten Risikogebiete zu bekommen, sollen diese nachfolgend in einer kurzen Zusammenfassung beschrieben werden.

1.7.2.1 Fairness (FN)

Die Dimension **Fairness** setzte sich aus zwei Risikogebieten zusammen, der gleichlautenden Fairness und der Beherrschung der Dynamik. Kernziel dieser Dimension ist es, dass es durch die KI-Anwendung nicht

zu Diskriminierung kommt, weil z.B. bestimmte Gruppen in den Trainingsdaten unterrepräsentiert sind.[60] Die KI-Anwendung soll ethisch und rechtskonform den Gleichbehandlungsgrundsatz einhalten, indem sie nicht gleiche soziale Sachverhalte ungleich oder ungleiche gleichbehandelt, außer wenn ein Abweichen davon sachlich gerechtfertigt ist.[61]

Im Risikogebiet **Fairness** innerhalb dieser Dimension sind die Bewertung und entsprechende Maßnahmen gegen ein mögliches Lernen von unfairem oder diskriminierendem Verhalten der KI-Anwendung während der Entwicklung enthalten.[62]

Das zweite Risikogebiet **Beherrschung der Dynamik** folgt in allen Dimensionen als eigenes Risikogebiet, außer bei der Autonomie und Kontrolle. Dies zeigt die Wichtigkeit dieses Risikogebietes und die besondere Schwierigkeit der Risikobewertung, die KI-Anwendungen innewohnt. Es ist ein mitunter sehr dynamisches Verhalten des Gesamtsystems, welches durch die KI ausgelöst wird. Daher ist es wichtig, speziell die Dynamik zu beherrschen. Innerhalb dieser Dimension soll das Risikogebiet **Beherrschung der Dynamik** dazu dienen, Risiken zu erkennen und zu minimieren, die sich durch geänderte Rahmenbedingungen oder geändertes Nutzer*innenverhalten ergeben können.[63]

1.7.2.2 Autonomie und Kontrolle (AK)

In der Dimension **Autonomie und Kontrolle** wird sowohl die Autonomie der KI-Anwendung selbst, als auch die Autonomie des Menschen beurteilt. Dahingehend wird hier betrachtet, inwiefern die KI-Anwendung Möglichkeiten bietet, einzugreifen bzw. zu interagieren.[64] Dies kann z.B. dann notwendig werden, wenn durch die KI-Anwendung aufgrund des Algorithmus neue Regeln gelernt werden, die ursprünglich nicht vorgesehen waren und durch den/die Anwender*in wieder rückgängig gemacht werden sollen oder als nicht korrekt im Modell hinterlegt werden müssen.
Wird das Risikogebiet **Angemessene und verantwortungsvolle Gestaltung der Aufgabenverteilung zwischen Mensch und KI-Anwendung** betrachtet, so sollen dabei die Risiken adressiert werden, „die sich aus Einschränkungen der Nutzerautonomie bzw. unangemessener Autonomie der KI-Anwendung ergeben.“[65] Wie bei anderen Risikogebieten liegt im Rahmen der Beurteilung auch hier das Spannungsfeld zwischen dem Grad an Systemautonomie und dem notwendigen Benutzer*inneneingriff.[66] Ziel einer KI-Anwendung ist in den meisten Fällen, den/die Nutzer*in zu entlasten. Ist die KI-Anwendung jedoch so autonom in ihren Entscheidungen und haben diese Entscheidungen einen, womöglich negativen, Einfluss auf die Nutzer*innen, ist genau dies die sprichwörtliche rote Linie, die von der KI-Anwendung nicht

[60] Vgl. Poretschkin u.a. (2021), S. 23.

[61] Vgl. Poretschkin u.a. (2021), S. 37.

[62] Vgl. Poretschkin u.a. (2021), S. 38.

[63] Vgl. Poretschkin u.a. (2021), S. 38.

[64] Vgl. Poretschkin u.a. (2021), S. 24.

[65] Poretschkin u.a. (2021), S. 49.

[66] Vgl. Poretschkin u.a. (2021), S. 51.

überschritten werden darf. Diese Bewertung und notwendige Maßnahmen werden in diesem Risikogebiet behandelt.

Im zweiten Risikogebiet der Dimension **Autonomie und Kontrolle**, der **Sicherstellung der Informiertheit und Befähigung von Nutzer*innen und Betroffenen**, geht es darum, dass Nutzer*innen und mögliche Betroffene ausreichende Informationen über die KI-Anwendung bekommen. Dabei geht es inhaltlich um ausreichende Erklärung der Risiken und der korrekten Bedienung der KI-Anwendung.[67]

1.7.2.3 Transparenz (TR)

Im Rahmen der Dimension **Transparenz** sieht der Leitfaden vor, sich mit der Nachvollziehbarkeit, der Reproduzierbarkeit und der Erklärbarkeit der KI-Anwendung auseinander zu setzen.[68]

Dafür werden im Risikogebiet **Transparenz gegenüber Nutzer*innen und Betroffenen** Risiken betrachtet, die „dadurch entstehen, dass Entscheidungen und Auswirkungen der KI-Anwendung gegenüber Nutzer*innen und Betroffenen nicht hinreichend erklärt werden können."[69]

Ergänzt werden diese Anforderung mit dem zweiten Risikogebiet **Transparenz für Expert*innen** insofern, als dass das Verhalten der KI-Anwendung zumindest durch Experten nachvollzogen werden kann.[70] Ist dies nicht der Fall, müssen entsprechend weitere Maßnahmen zur Risikominderung gesetzt werden.

Über das dritte Risikogebiet **Auditfähigkeit** soll sichergestellt werden, dass die Entwicklung und der Betrieb der KI-Anwendung derart dokumentiert ist, so dass es möglich wird, ein Audit über die Applikation anhand der Dokumentation durchzuführen.[71]

Mit der **Beherrschung der Dynamik**, dem vierten Risikogebiet, ist an dieser Stelle gemeint, „dass sich Anforderungen an die Transparenz oder die implementierten Transparenzverfahren selbst ändern"[72] und die diesbezüglich möglich entstehenden Risiken zu betrachten sind. So könnte sich während des Betriebs z.B. das Machine Learning-Modell verändern, ein sogenannter Model Drift, und dies dazu führen, dass es undurchsichtig wird, oder es könnten externe Anforderungen, wie gesetzliche Änderungen, neu entstehen, die zu berücksichtigen sind.[73]

1.7.2.4 Verlässlichkeit (VE)

In der Dimension **Verlässlichkeit** wird im Wesentlichen die Robustheit und Performanz der KI-Komponente betrachtet, also die Konsistenz der Ausgaben bei veränderten Eingabedaten. Dies wird erfolgt durch

[67] Vgl. Poretschkin u.a. (2021), S. 57.

[68] Vgl. Poretschkin u.a. (2021), S. 24.

[69] Poretschkin u.a. (2021), S. 57.

[70] Vgl. Poretschkin u.a. (2021), S. 64.

[71] Vgl. Poretschkin u.a. (2021), S. 64.

[72] Poretschkin u.a. (2021), S. 64.

[73] Vgl. Poretschkin u.a. (2021), S. 83.

Bewertung der Risikogebiete **Verlässlichkeit im Regelfall**, in dem fehlerhafte Vorhersagen betrachtet werden, durch die Risikogebiete **Robustheit** und **Abfangen von Fehlern auf Modellebene**, die sich mit störungsbehafteten und manipulierten Eingaben befassen, ebenso wie mit Eingabedaten, die nicht Teil der bestimmungsgemäßen Verwendung der KI-Anwendung sind.[74]

Außerdem werden im Risikogebiet **Einschätzung von Unsicherheit**, die möglichen Risiken durch Unsicherheiten der Ausgabedaten bewertet. In dem bekannten Risikogebiet **Beherrschung der Dynamik** wird einerseits der bereits erwähnte Model Drift und andererseits der sogenannte Concept Drift, d.h. Veränderungen des Anwendungskontexts der KI-Anwendung betrachtet.[75]

1.7.2.5 Sicherheit (SI)

Sicherheit als Dimension beinhaltet Risikobewertungen von Personen- und Sachschäden im Risikogebiet **Funktionale Sicherheit,** die aufgrund von Fehlfunktionen oder Ausfall der KI-Anwendung entstehen können.[76] Ergänzend dazu werden im Risikogebiet **Integrität und Verfügbarkeit** mögliche Verfälschungen und Manipulation von verwendeten Daten bewertet. Die **Beherrschung der Dynamik** im Kontext der Sicherheit beschäftigt sich mit möglichen neuen Gefährdungen oder auch einem Concept Drift, bei dem bestehende Maßnahmen nicht mehr ausreichend sind.[77]

1.7.2.6 Datenschutz (DS)

Bei der Dimension **Datenschutz** geht es um den Schutz sensibler Daten, die inhaltlich aufgeteilt werden in das Risikogebiet **Schutz personenbezogener Daten** und **Schutz geschäftsrelevanter Daten**, wobei Erstgenanntes sich auf die DSGVO[78]-konforme Nutzung bezieht und Letztgenanntes auf Geschäftsgeheimnisse abzielt, die z.B. bei Speicherung in einer Cloud ebenfalls nicht fremdverwertet werden sollen.[79]

Das Risikogebiet **Beherrschung der Dynamik** dient in diesem Zusammenhang der Aufrechterhaltung des Datenschutzes während des gesamten Betriebs der KI-Anwendung, d.h. auch bei Änderungen in den gesetzlichen Anforderungen oder möglicher Identifizierbarkeit von z.B. personenbezogenen Daten aufgrund von technologischen Fortschritten.[80]

[74] Vgl. Poretschkin u.a. (2021), S. 25.

[75] Vgl. Poretschkin u.a. (2021), S. 87.

[76] Vgl. Poretschkin u.a. (2021), S. 26.

[77] Vgl. Poretschkin u.a. (2021), S. 26; S. 138.

[78] Vgl. Europäische Kommission (2016), S. 1 ff.

[79] Vgl. Poretschkin u.a. (2021), S. 26; S. 152.

[80] Vgl. Poretschkin u.a. (2021), S. 26; S. 158.

1.7.3 Der Prüfablauf gemäß Leitfaden

Der Prüfablauf gemäß Leitfaden besteht aus zwei Phasen, die nacheinander zu erfolgen haben. In der ersten Phase wird für jede Dimension eine Schutzbedarfsanalyse durchgeführt, die bei mittlerem bzw. hohem Schutzbedarf zu einer genaueren Betrachtung, d.h. zu einer Risikobeurteilung der Risikogebiete der jeweiligen Dimension führt. Nachdem die risikomindernden Maßnahmen beschrieben werden, ist zuerst eine Bewertung jeder einzelnen Dimension durchzuführen. Sind sämtliche anwendbaren Dimensionen bewertet, ist eine abschließende dimensionsübergreifende Beurteilung der KI-Anwendung durchzuführen, um die Vertrauenswürdigkeit zu prüfen.[81]

[81] Vgl. Poretschkin u.a. (2021), S. 27 ff.

2 IMPLEMENTIERUNG DES LEITFADENS

Im folgenden Kapitel wird die Umsetzung des Leitfadens in ein Software-Tool, einen sogenannten Wizard, beschrieben. Als Plattform wird die Mendix Low-Code Entwicklungsumgebung verwendet. Mit dieser Entwicklungsumgebung wurde ein Wizard-Framework erstellt, auf dessen Basis das Software-Tool, der Wizard, entstanden ist.

Es wird zunächst die Motivation für die Umsetzung erläutert. Anschließend werden die Funktionen des bestehenden Wizard-Frameworks beschrieben, um zu verdeutlichen, welche Möglichkeiten für die Umsetzung bestehen. Danach wird der konzeptionelle Aufbau und die tatsächliche Umsetzung des Wizards beschrieben.

2.1 Motivation für die Umsetzung

Ein wesentliches Ziel des Wizards ist es, die im Leitfaden beschriebenen potentiellen Gefährdungen und risikomindernden Maßnahmen zu dokumentieren. Nach Beantwortung der Fragen soll automatisiert ein Report erstellt werden können, um die jeweils spezifische KI-Anwendung für Dritte nachvollziehbar zu machen. Dies kann z.B. eine externe Stelle für die Zertifizierung der KI-Anwendung sein, falls dies aufgrund gesetzlicher Vorgaben notwendig sein sollte. Wird dieser Wizard dann einer breiten Nutzer*innenschaft verfügbar gemacht, so sind die Schutzziele und Maßnahmen, die im Leitfaden beschrieben sind, über den Wizard leichter zugänglich, d.h. der Zugang wird niederschwelliger. Eine weiter verbreitete Anwendung wird somit wahrscheinlicher.

Mit dem bestehenden Wizard-Framework wurden bereits andere Wizards für KI-Anwendungen erstellt, z.B. für einen *Requirements Engineering Test* – siehe Abb. 9. Der Vorteil eines weiteren Wizards innerhalb einer Bewertungsumgebung liegt in der Kombinierbarkeit.

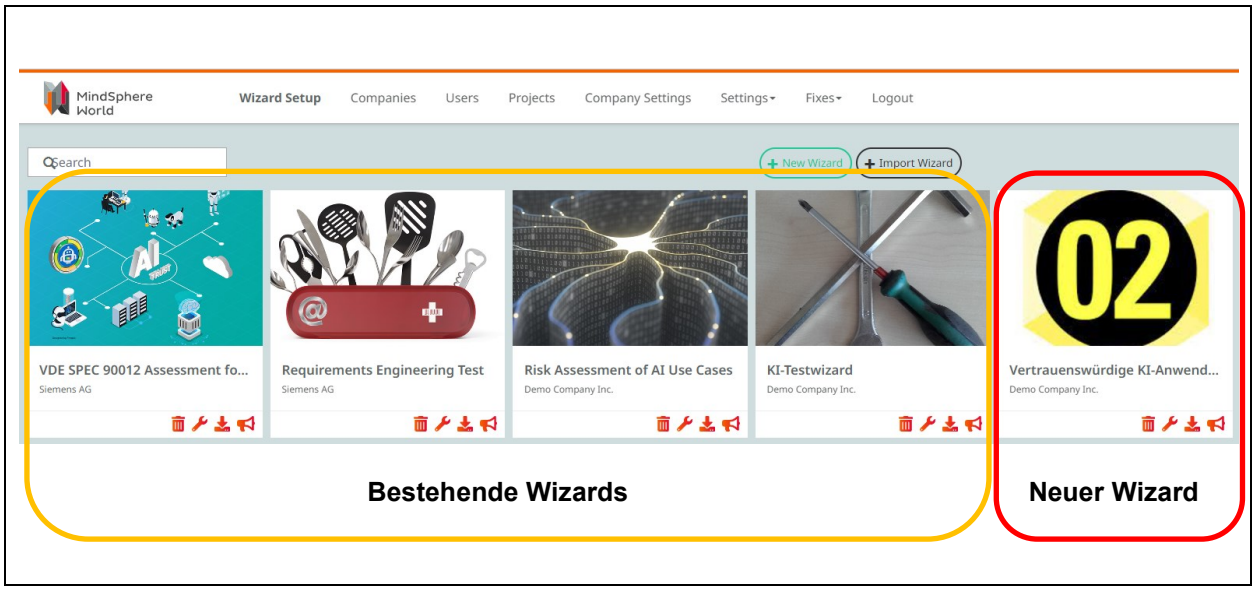


Abb. 9: Bestehende und neuer Wizard(s), Quelle: Eigene Darstellung.

So können KI-Vorhaben unter Anwendung mehrerer unterschiedlicher Wizards betrachtet und bewertet werden. Die verschiedenen Blickwickel können außerdem Optimierungspotential veranschaulichen, z.B.

durch Betrachtung von Rentabilität und Risiko der jeweiligen Anwendung. So können potentielle Maßnahmen und Funktionen vor Umsetzung bewertet werden.

Ein weiteres Ziel der vorliegenden Arbeit ist es, für das bestehende Framework, welches für die Erstellung von Wizards vorab erstellt wurde, durch Erkenntnisse während der Ausarbeitung, Feedback zu geben, um diese Eingaben in das Wizard-Framework Siemens-seitig zu implementieren.[82]

Außerdem sollten aufgrund der intensiven Beschäftigung mit dem Leitfaden auch mögliche Optimierungen in Bezug auf die Inhalte des Leitfadens erarbeitet und als weiteres Ergebnis in dieser Arbeit dargestellt werden.[83]

2.2 Mendix Low-Code Entwicklungsumgebung

Das Wizard-Framework wird in der Mendix Low-Code Entwicklungsumgebung erstellt. Dabei handelt es sich um eine Multiplattform, die aufgrund des Low-Code Ansatzes ein schnelleres Entwickeln von Apps ermöglichen soll. Die Mendix Entwicklungsumgebung erlaubt ein visuelles Programmieren und damit ein schnelles Generieren von Logik. Dennoch ist es möglich durch React Widgets und Java / Javascript Code mit eigener Programmierung eigene Funktionen zu implementieren.

Die Multiplattform ermöglicht es außerdem, eine Applikation für diverse Endgeräte zu erstellen, ohne dass zusätzlicher Programmieraufwand für die Darstellung erfolgen muss. Im Rahmen der vorliegenden Arbeit liegt der Fokus auf einer Web-Browser App.

Ein Vorteil dieser Entwicklungsumgebung liegt in der Kollaboration während der Entwicklung. Mittels integriertem Projektmanagement und einer direkten User-Feedback-Umgebung ist es möglich, den Wizard zu erstellen, Feedback zum Framework zu geben und parallel können Anpassungen und Ergänzungen im Framework vorgenommen werden, so dass ein schnelles teamübergreifendes Entwickeln der Applikationen ermöglicht wird. Diese Anpassungen können jederzeit mittels Update-Funktion übernommen werden – siehe Abb. 10. So werden während der Erstellung des Wizards parallel bereits diverse Funktionen angepasst und verfügbar gemacht.[84]

[82] Vgl. Kapitel 4.2.

[83] Vgl. Kapitel 4.1.

[84] Siehe Kapitel 4.2.

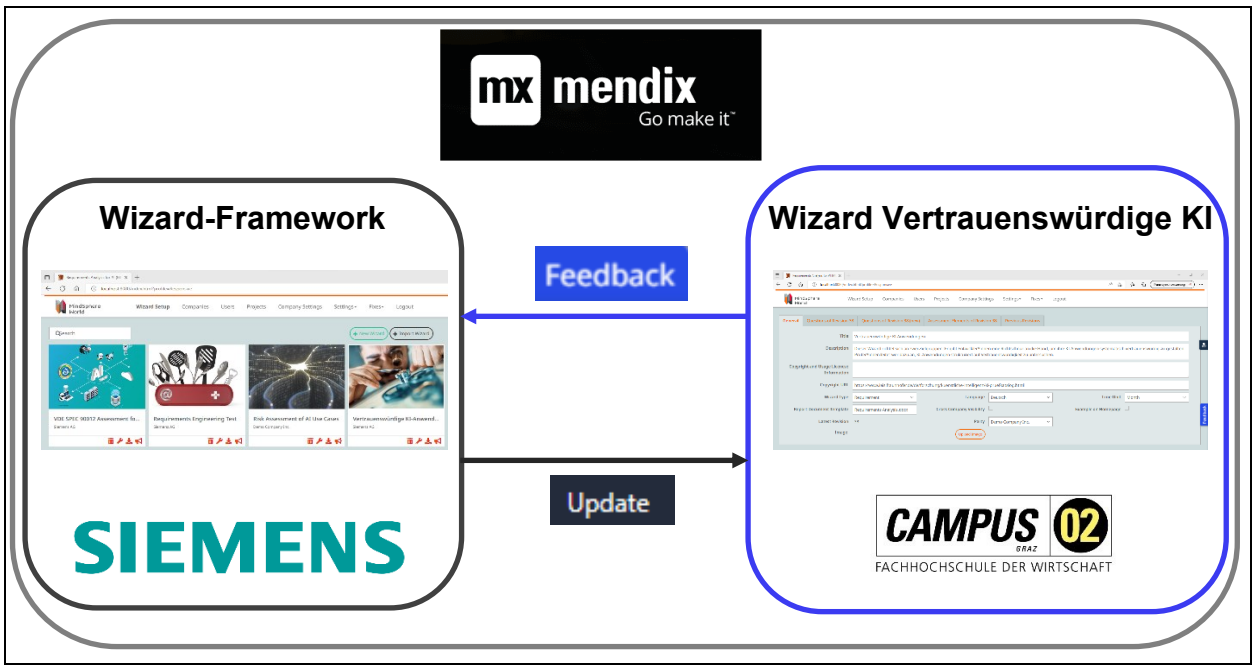


Abb. 10: Entwicklung in der Mendix Umgebung, Quelle: Eigene Darstellung.

2.3 Bestehendes Framework

Das Framework, welches für die Erstellung des Wizards verwendet wird, ist zuvor von Siemens erstellt worden. Inhaltlich umfasst dieses Framework eine erste Version eines Konfigurators für Wizards. Die Wizards sind eine Erweiterung einer Marktplatz-Komponente und können zur Konfiguration von Fragebögen verwendet werden. Nach Beantwortung der Fragen können zudem Berichte erzeugt werden, welche der Dokumentation der entsprechenden Wizard-Inhalte dienen.

2.4 Kürzel des Leitfadens

Gemäß Leitfaden werden sowohl den Dimensionen der Vertrauenswürdigkeit als auch den zugeordneten Tätigkeiten, wie z.B. der Schutzbedarfsanalyse (S) Kürzel zugeordnet, um die entstehende Komplexität bei jedem einzelnen Schritt bei der Beurteilung zu vereinfachen und dem Anwender zu verdeutlichen, an welcher Stelle des Beurteilungsprozesses er sich befindet. Dazu werden die Abkürzungen an der jeweiligen Stelle genannt. Mit Kenntnis der Abkürzungen kann direkt auf den momentanen Bearbeitungsschritt geschlossen werden:

„**Beispiel 2:** In der Dimension Fairness **[FN]** hat im Risikogebiet **[R]** Beherrschung der Dynamik **[BD]** das erste Kriterium **[KR]** die Nummer **[01]**: **[FN-R-BD-KR-01]**“[85]

[85] Poretschkin u.a. (2021), S. 29.

2.5 Umsetzung des Wizards

Um einen Wizard anzulegen, wird zunächst auf die Schaltfläche *+New Wizard* geklickt. Hat man den Titel, den Wizard-Typ, die Sprache und eine kurze Beschreibung angegeben, kann man über die Bestätigung der Schaltfläche *Start Configuration* den neuen Wizard erstellen – siehe auch Abb. 11.

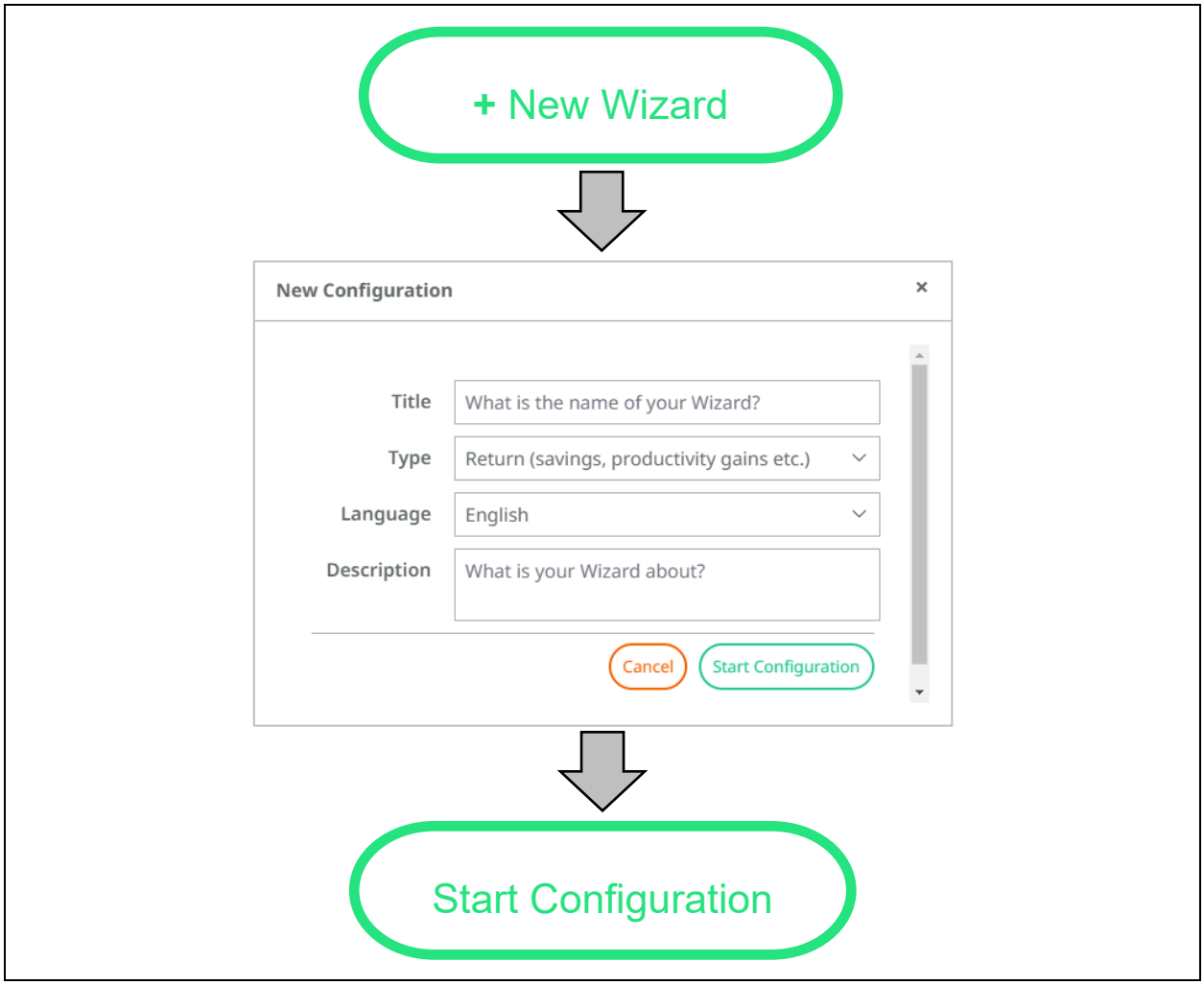


Abb. 11: Erstellen eines Wizards, Quelle: Eigene Darstellung.

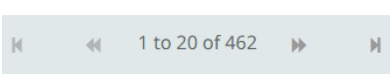
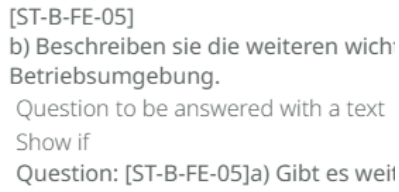
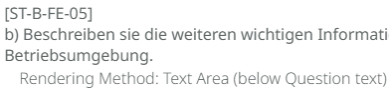
Der Wizard wird anschließend angelegt und man kann mit der Konfiguration beginnen. Der gewählte Wizard-Typ ist *Requirement*, da es sich im KI-Leitfaden um Anforderungen handelt. Die anderen möglichen Wizard-Typen sollen an dieser Stelle daher nicht weiter erläutert werden.

2.5.1 Kartenreiter des Wizards

Zu Beginn der Wizard-Konfiguration werden auf dem ersten Kartenreiter ***General*** die zuvor eingegebenen allgemeinen Daten angezeigt. Auf dieser Seite können weitere Informationen hinterlegt werden oder die bestehenden Informationen ergänzt werden.

Außerdem gibt es vier weitere Kartenreiter, in denen der Wizard entwickelt werden kann. Es gibt zwei Kartenreiter mit verschiedenen Darstellungen der Fragen, ***Questions of Revision XX*** und ***Questions of Revision XX (new)***. Unabhängig davon, in welchem der beiden Kartenreiter man arbeitet, in dem jeweils anderen werden die Fragen übernommen. In der Anwendung kann demnach variiert werden.

Folgende Unterschiede existieren jedoch in den beiden Anzeigearten – siehe Tab. 2.

	Questions of Revision XX	***Questions of Revision XX (new)***
Anzahl an dargestellten Elementen	20 je Seite	Unbeschränkt
Navigation	Scrollen bis Element 20, dann Skip auf die nächsten 20 Elemente 	Endlos scrollbar, ab 200 Elementen nötig, weitere zu laden; Step skip möglich 
Zusätzliche Anzeige unter Frage	Frageart, ggf. Konditionen 	Antwort, Anzeigeart 
Funktionen je Element	Edit – Bearbeiten Delete – Löschen Add element below – neues Element unterhalb hinzufügen 	Edit – Bearbeiten Duplicate – Kopieren Delete – Löschen Make Substep – Substep erstellen Move element – Position des Elements verändern Add element below – neues Element unterhalb hinzufügen 

Tab. 2: Unterschiedliche Darstellung Kartenreiter, Quelle: Eigene Darstellung.

Je nach Erstellungsfortschritt und Fragestellung kann es so sinnvoll sein entweder den einen Kartenreiter zu verwenden oder den anderen. Die Kopierfunktion (duplicate) z.B. wird für die Erstellung einer neuen

Dimension notwendig[86]. Da die Dimensionen der Vertrauenswürdigkeit immer gleich strukturiert sind, wird es somit einfacher, die Struktur für die nächste Dimension vorzubereiten. Erst nach dem Kopieren sämtlicher strukturellen Elemente der vorigen Dimension werden die inhaltlichen Anpassungen der jeweiligen Dimension vorgenommen.

Im vierten Kartenreiter ***Assessment Elements of Revision XX*** können die Assessment Elements erstellt werden. Die verschiedenen Konfigurationsmöglichkeiten der Assessment Elements werden im nächsten Kapitel der Wizard-Funktionen erläutert.

Der fünfte und letzte Kartenreiter ***Previous Revisions*** zeigt die Versionierung des Wizards, d.h. welche Version zu welchem Zeitpunkt geändert wurde. Eine neue Version wird immer dann erzeugt, wenn man die Schaltfläche *Publish* betätigt.

2.5.2 Wizard-Funktionen

Innerhalb eines Wizards gibt es verschiedene Funktionen für die Strukturierung und die inhaltliche Ausgestaltung. Beginnend bei einer Sektionierung geht es über verschiedene Möglichkeiten der Fragenart und zu erwartenden Antwortarten. Des Weiteren gibt es sogenannte *Assessment Elements*, die mittels Wenn-Dann-Logik nach der Beantwortung aller notwendigen Wizard-Fragen erzeugt werden können.

Die genauen Konfigurationsmöglichkeiten für die Strukturierung eines Wizards und die Funktionen der Fragen, Antworten und der Assessment Elements werden in den folgenden Kapiteln betrachtet, für eine Übersicht siehe Tab. 3.

Funktion	Elementbezeichnung
Strukturierung	Section Start
	Section End
Fragen	Question to be answered by selecting options
	Question to be answered with an amount
	Question to be answered with a date
	Question to be answered with a text

Tab. 3: Funktionen des Frameworks, Quelle: Eigene Darstellung.

2.5.2.1 Section Start / Section End

Das bestehende Framework stellt verschiedene Funktionen für die Wizard-Erstellung und Ausgestaltung bereit. Für die Strukturierung besteht die Möglichkeit, mittels *Section Start* und *Section End* Fragen zu gruppieren und in dieser Form den Anwender*innen eine bessere Übersichtlichkeit zu gewährleisten. Dies

[86] Vgl. Kapitel 4.2.6.

leistet einen Beitrag zur Akzeptanz des Wizards und zur effizienten Bearbeitung eines Projekts mit dem Wizard.

Im Element *Section Start* kann der Name und ein zugehöriger Text angegeben werden. Außerdem kann ein Bild hochgeladen werden und eine Variable für diese Section definiert werden. Als Konfigurationsmöglichkeit für dieses Element kann die Sichtbarkeit[87] eingestellt werden.

Im Element *Section End* kann ebenfalls die Sichtbarkeit eingestellt werden. Die einzige weitere Konfigurationsmöglichkeit dieses Elements ist die Aktivierung einer Schaltfläche *Next Button?*, die mit einem frei wählbaren Namen (*Button caption*) versehen werden kann und deren Funktion das Weiterklicken auf die nächste *Section* ist.

2.5.2.2 Fragen

Die nächste Funktion beinhaltet unterschiedliche Konfigurationsmöglichkeiten von Fragen. Wird ein neues Element eingefügt (*Add element below*[88]), besteht die Möglichkeit eine der möglichen Fragen gemäß Tab. 3 einzufügen. Es öffnet sich ein Pop-Up-Fenster, in welchem die Fragen eingegeben und konfiguriert werden können.

Die Fragen selbst, können in einem freien Textfeld (Abb. 12 - A, *Question*) auf dem ersten Kartenreiter eingegeben werden. In diesem Textfeld bestehen außerdem einige Formatierungsmöglichkeiten für den Text:

- Text **fett** formatieren
- Text *kursiv* formatieren
- Text unterstreichen
- Nummerierung (1., 2., 3., …)
- Aufzählungszeichen (Bullet points)

Des Weiteren ist es möglich einen Hilfetext (Abb. 12 - D, *Help Text*) anzeigen zu lassen. Dafür muss ein eigener Haken gesetzt werden (Abb. 12 - C, *Display Help Text*). Ist der Haken nicht gesetzt, wird auch das Textfeld für den Hilfetext nicht angezeigt.

Außerdem besteht auch hier wieder die Möglichkeit, die Frage, in Abhängigkeit von zuvor beantworteten Fragen, sichtbar zu schalten (Abb. 12 - E, *Visibility*) und bei allen Varianten kann man auch ein Bild (F - *Upload Image*) hochladen.

Bei den beiden Varianten *Question to be answered with an amount* und *Question to be answered with a date*, gibt es ergänzend die Möglichkeit den Inhalt in eine Variable (Abb. 12 – B, *Stored in variable*) zu speichern.

[87] Vgl. Kapitel 2.5.3.

[88] Vgl. Tab. 2.

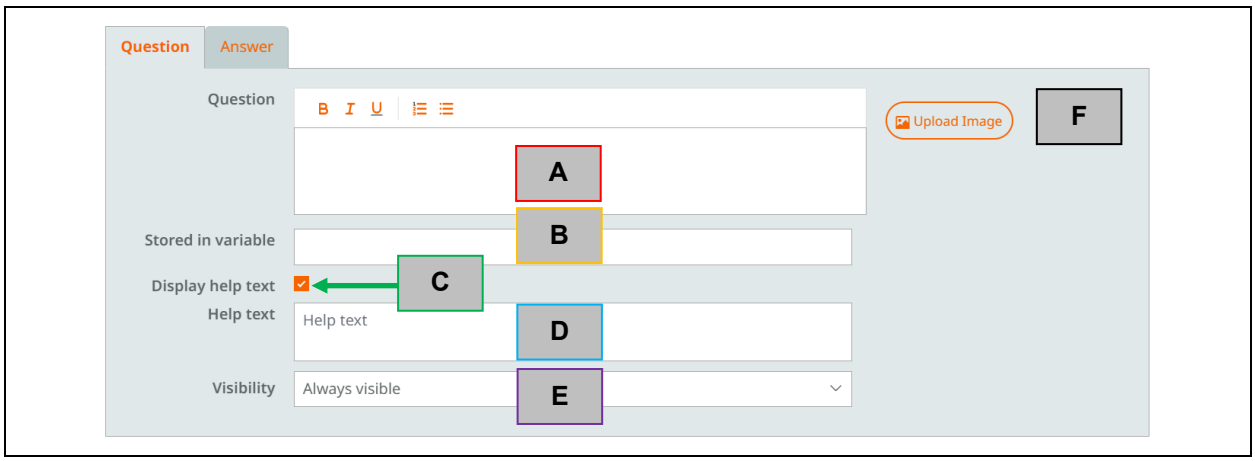


Abb. 12: Konfigurationsfenster Fragen, Quelle: Eigene Darstellung.

Wurde die jeweilige Frage entsprechend konfiguriert, ist im nächsten Schritt der Nächste Kartenreiter *Answer* auszuwählen und die gewünschte Antwort zu konfigurieren.

2.5.2.3 Antworten

Bei den Antworten sind die Konfigurationsmöglichkeiten noch stärker unterschiedlich ausgeführt als die Fragen. Daher werden hier die unterschiedlichen Varianten nacheinander beschrieben.

Question to be answered by selection options

Bei dieser gewählten Antwortart gibt es drei Checkboxen – siehe Tab. 4.

Checkboxbezeichnung	Funktion
Required	Wird diese Checkbox ausgewählt, so muss der/die Anwender*in diese Frage zwingend beantworten. Andernfalls lässt der Wizard eine weitere Bearbeitung nicht zu.
Multiselect	Wenn der/die Anwender*in mehrere der vorgegebenen Antwortmöglichkeiten auswählen darf, ist diese Checkbox anzuwählen. Bei nur einfacher Antwortmöglichkeit, z.B. Ja oder Nein, ist diese Checkbox nicht anzuwählen.
Percentage Value	Wird diese Checkbox ausgewählt, öffnet sich ein eigenes Textfeld, in dem ein Wert in Prozent angegeben werden kann.

Tab. 4: Checkboxen bei Antworten, Quelle: Eigene Darstellung.

Je nachdem, ob die *Multiselect*-Checkbox angewählt wird oder nicht, gibt es unterschiedlichen Anzeigearten, die über das Feld *Rendermode* ausgewählt werden können. In der Tab. 5 werden sämtliche Rendermode-Möglichkeiten beschrieben. Die, im Wizard verwendeten, Möglichkeiten werden zudem als beispielhafte Abbildung in Kapitel 2.6.1 dargestellt.

Multiselect	Rendermode-Möglichkeiten
Nicht ausgewählt 	Dropdown
	Selection popup
	Radiobutton(s) shown vertically – below
	Radiobutton(s) shown vertically – right side
	Radiobutton(s) shown horizontally - right side
Ausgewählt 	Checkboxes
	Selection Popup
	Checkboxes & Percentage Sliders
	Checkboxes & Percentage Textboxes

Tab. 5: Rendermöglichkeiten Multiselect, Quelle: Eigene Darstellung.

Im Kartenreiter ***Questions of Revision XX (new)*** wird der ausgewählte Rendermode eines jeden Elements in der unteren Zeile angezeigt. Dies kann bei der Prüfung während der Konfiguration hilfreich sein, da mit dieser Funktion direkt ersichtlich ist, um welche Anzeigeart es sich handelt. Es muss nicht ein jedes Element im Bearbeitungsmodus geöffnet werden, sondern es kann mittels Scrollens durch die Elemente navigiert werden.

Question to be answered with an amount

Wird diese Antwortart gewählt, kann man folgende Antwortmöglichkeiten konfigurieren. Es gibt die Funktion *Required* (Abb. 13 - A) in bekannter Form als Checkbox - vgl. Tab. 4. Über die Funktion *Number* (Abb. 13 - B) kann man angeben, ob eine ganze Zahl oder eine Dezimalzahl als Antwort gegeben werden soll. Wenn eine Währung als Antwort erwartet wird, kann dies über die Checkbox *Currency* (Abb. 13 - C) konfiguriert werden. Im Textfenster *Unit* (Abb. 13 - D) ist die entsprechende Einheit, z.B. [mm], anzugeben. Der Rendermode ist immer Text box (Abb. 13 - E). Ein Dropdown-Menü ist vorgesehen, andere Konfigurationsmöglichkeiten werden jedoch bei einer Angabe einer Zahl nicht benötigt. Im Dropdown-Menü *Range* (Abb. 13 - F) besteht die Möglichkeit zwischen einer beliebigen Zahl (Infinite) oder eine Zahl innerhalb einer, zu definierenden Bandbreite (Specific intervals) auszuwählen.

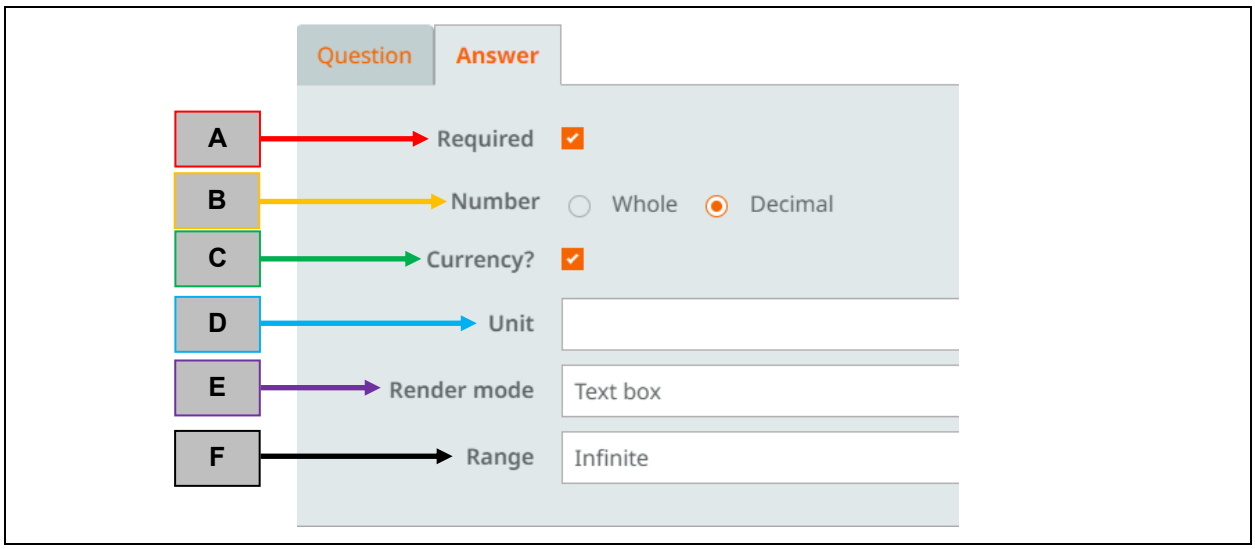


Abb. 13: Konfigurationsmöglichkeiten Amount, Quelle: Eigene Darstellung.

Question to be answered with a date

Bei der Antwortart Question to be answered with a date gibt es auch die Konfigurationsmöglichkeiten *Required* und, wie bei der Antwortart Question to be answered with an amount, die *Range*, d.h. einen oder mehrere Zeiträume, aus denen die Antwort gegeben werden muss.

Question to be answered with a text

Bei dieser gewählten Antwortart gibt es zwei Checkboxen. Das, von den anderen Antwortarten bekannte, *Required* und als zweite Checkbox gibt es *allow File Uploads*. Diese Funktion erlaubt es dem/der Anwender*in, ganze Dateien hochzuladen. Dies kann z.B. verwendet werden, wenn die Beantwortung der jeweiligen Frage umfangreicher ausfällt und bereits in einem anderen Rahmen dokumentiert wurde. Wenn beispielsweise technische Daten von einer verwendeten Komponente zu dokumentieren sind, könnte man mit dieser Funktion auch ein Datenblatt hochladen.

Auch bei dieser Antwortart gibt es wieder verschiedene Möglichkeiten der Darstellung (Rendermode). Es gibt die Varianten *Text box, Text area (right-side), Text area (below Question text)* und *Email.* Die *Text box* und die *Text area (right-side)* sind aufgrund der kompakten Darstellungsform für sehr kurze Antworten geeignet, während *Text area (below Question text)* auch für ausführlichere Antworten zu bevorzugen ist. Bei der Funktion *Email* ist eine Emailadresse als Antwort anzugeben.

Die Textantworten können über die Angabe eines Variablennamens im Testfeld *Stored in Variable* zudem in eine Variable gespeichert werden.

2.5.2.4 Assessment Elements

Den sogenannten Assessment Elements ist in der Konfigurationsoberfläche ein eigener Kartenreiter, unabhängig von den Fragen, gewidmet. Die Assessment Elements können dafür genutzt werden, dem/der Anwender*in offene, noch durchzuführende Aktionen in einer eigenen Liste anzuzeigen. Damit können umfangreiche Inhalte, die im Leitfaden abgefragt werden, ausgelagert werden und in einem eigenen Dokument beschrieben werden, z.B. Tests, die mit der KI-Anwendung durchgeführt und gesondert protokolliert wurden. So kann der Wizard zum einen durchgearbeitet werden, ohne zwingend alle Punkte

vorab erfüllen zu müssen. Zum anderen werden aber auch keine Anforderungen vergessen, da diese in der Übersichtsliste angezeigt werden.

Über die Schaltfläche *New* kann ein neues *Assessment Element* angelegt werden. Es öffnet sich wieder ein Konfigurationsfenster mit den drei Kartenreitern *General, Conditions* und *Sources*.

Auf dem Kartenreiter *General* kann eine ID, ein Name, ein Variablenname und eine Beschreibung / Erklärung (Explanation) eingegeben werden. Der Assessment Element Type kann aus den folgenden Auswahlmöglichkeiten bestimmt werden:

- Return (savings, productivity gains etc.)
- Cost (invest, effort, …)
- Risk (Dataset)
- Risk (Project)
- Rating
- Requirement

Da das Wizard-Framework für verschiedene Themengebiete erstellt wurde, sind hier übergreifende Auswahlmöglichkeiten enthalten, die für den, in der vorliegenden Arbeit beschriebenen Wizard, nicht notwendig oder sinnvoll sind. Der Vollständigkeit halber werden diese Möglichkeiten hier dennoch erwähnt. Für den Wizard werden als Assessment Elements daher diejenigen vom Typen *Requirement* gewählt – siehe Abb. 14. Für die genaue Auswahl, die Unterkategorien (Requirement Type) und Bezeichnung der Assessment Elements, siehe auch Kapitel 2.6.3.

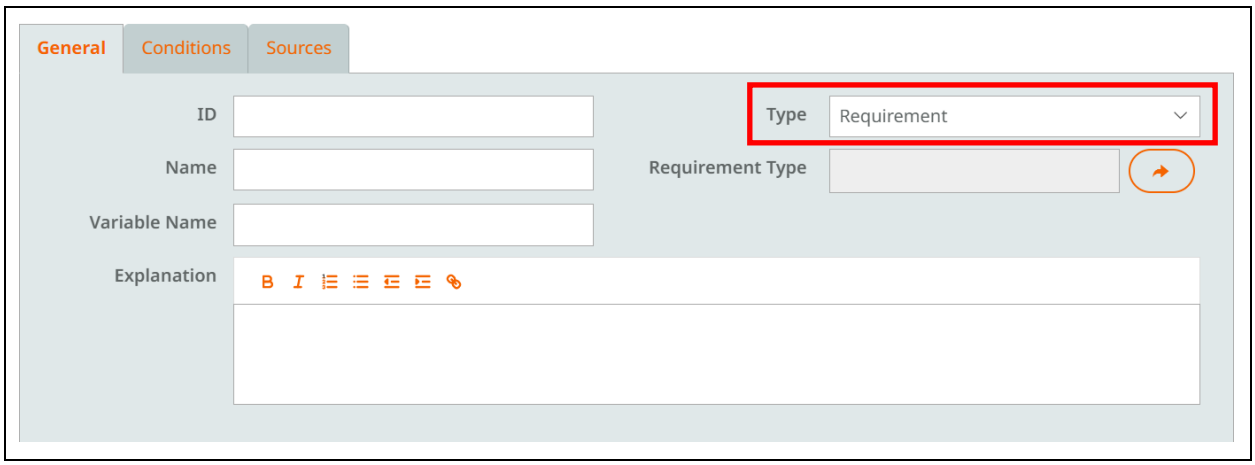


Abb. 14: Konfigurationsfenster Assessment Elements General, Quelle: Eigene Darstellung.

Auf dem Kartenreiter *Conditions* können die notwendigen Bedingungen eingegeben werden, die erfüllt sein müssen, um das zugehörige Assessment Element anzuzeigen. Das können, wie bei der Sichtbarkeit der Fragen, verschiedene Antworten sein, die zuvor von dem/der Anwender*in gegeben wurden. Um eine Bedingung zu hinterlegen, sind die Schaltflächen *Add Condition* und *Add answers* nacheinander zu betätigen. In dem sich öffnenden Fenster, werden sämtliche Elemente mit den zugehörigen Antwortmöglichkeiten angezeigt – siehe auch folgendes Kapitel 2.5.3.

Im dritten und letzten Kartenreiter *Sources* können Quellen angegeben werden, die entweder ein Dokument, oder auch ein Link zu einer Website sind.

2.5.3 Visibility

Die Sichtbarkeit (Visibility) der Fragen und der Assessment Elements ist eine der wichtigsten Funktionen für die Vereinfachung des Leitfadens, d.h. um den Inhalt des Leitfadens für den Ersteller eines KI-Systems auf die, für den Anwendungsfall notwendigen, Anforderungen zu reduzieren.[89] Diejenigen Fragen und Assessment Elements, die für die Applikation nicht relevant sind, können in Abhängigkeit der zuvor beantworteten Fragen ausgeblendet werden. Dabei gibt es mehrere Möglichkeiten, die Abhängigkeiten, d.h. die Bedingungen auszuwählen – siehe Tab. 6.

Auswahlmöglichkeit	Funktion
Always visible	Das Element ist immer sichtbar, unabhängig von zuvor beantworteten Fragen
Visibility depends on previous answer(s)	Das Element ist nur dann sichtbar, wenn eine zuvor gestellte Frage, mit einer bestimmten Antwort beantwortet wurde
Visibility depends on Section Visibility	Das Element ist nur dann sichtbar, wenn das Element Section Start, in dem dieses Element enthalten ist, auch sichtbar ist
Visibility depends on Section Visibility plus Conditions	Das Element ist nur dann sichtbar, wenn eine zuvor gestellte Frage, mit einer bestimmten Antwort beantwortet wurde UND das Element Section Start, in dem dieses Element enthalten ist, auch sichtbar ist

Tab. 6: Sichtbarkeit der Elemente und Bedingungen, Quelle: Eigene Darstellung.

Bei der zweiten Auswahlmöglichkeit *Visibility depends on previous answer(s)* besteht die Möglichkeit, auch mehrere Antworten als Bedingungen zu setzen. Dabei können diese Antworten sowohl UND-, als auch ODER-verknüpft werden. Damit ergibt sich eine Vielzahl an möglichen Kombinationen.

Die Auswahl der Bedingungen erfolgt über die Schaltfläche *Add condition*, welche direkt unter der Auswahl für die Visibility positioniert ist. Nach Betätigung öffnet sich ein eigenes Fenster. In diesem Fenster kann man mittels Betätigung der Schaltfläche *Add Answers* ein weiteres Fenster (Select Dependencies) öffnen, in welchem sämtliche Fragen dargestellt werden, die zuvor konfiguriert wurden. Jede Frage erscheint mehrfach, d.h. jeweils einmal mit jeder zugehörigen Antwortmöglichkeit. Besteht z.B. die Möglichkeit eine Frage mit Ja oder Nein zu beantworten, wird die Frage zwei Mal dargestellt. In der ersten Zeile mit einem Ja in der zweiten Spalte, in der zweiten Zeile mit einem Nein in der zweiten Spalte.[90]

[89] Vgl. Kapitel 2.1.

[90] Siehe auch Abb. 25.

2.6 Konzept und Umsetzung

Da der Wizard den Nutzer*innen den Vorteil bieten soll, nach der Verwendung eine zugehörige Dokumentation, d.h. einen Bericht, zu erstellen, können im Wizard-Framework zunächst über den Kartenreiter General[91] allgemeine Projektdaten eingegeben werden, die nachher auch auf dem Bericht hinterlegt werden. Im Wizard selbst wird zunächst, ebenso wie im Leitfaden vorgesehen, der allgemeine Teil, d.h. der KI-Steckbrief zuerst erstellt. Der KI-Steckbrief wird als eigene *Section* erstellt, was aus Anwender*innensicht daran zu erkennen ist, dass in der Projektansicht eine große umrundete Ganzzahl (1) links neben dem angezeigten Text dargestellt wird. Die beiden Teilbereiche des KI-Steckbriefs, die **grundlegende Funktionalität und der vorgesehene Einsatzkontext** und **die Struktur der KI-Anwendung** werden darunter liegend als *Subsection* angelegt. Die Subsections werden ebenfalls durchnummeriert dargestellt. Dabei setzt sich die Zahl immer aus der Section-Nummer und einer fortlaufenden Nummer nach dem Punkt innerhalb der Section zusammen, beginnend bei 1.1 – siehe Abb. 15.

Als zweite *Section* wird im Wizard die Schutzbedarfsanalyse vorgesehen. Im Leitfaden folgt an dieser Stelle eine Dimension der Vertrauenswürdigkeit der Nächsten, d.h. zuerst ist die Schutzbedarfsanalyse der jeweiligen Dimension durchzuführen und gleich daran anschließend die Risikoanalyse [RI], die Definition der Zielvorgaben, das Festlegen von Kriterien [KR] und die Definition der Maßnahmen [MA] und die Bewertung [BW] des ersten Risikogebiets. Ist dieser Prozess für sämtliche Risikogebiete der jeweiligen Dimension durchlaufen, wird eine zusammenfassende Betrachtung [Z] der gesamten Dimension gefordert. Erst dann wird zur Schutzbedarfsanalyse der nächsten Dimension übergegangen.

Abweichend davon werden im Wizard sämtliche Schutzbedarfsanalysen der einzelnen Dimensionen in dieser *Section* vorab zusammengefasst.

[91] Vgl. Kapitel 2.5.1.

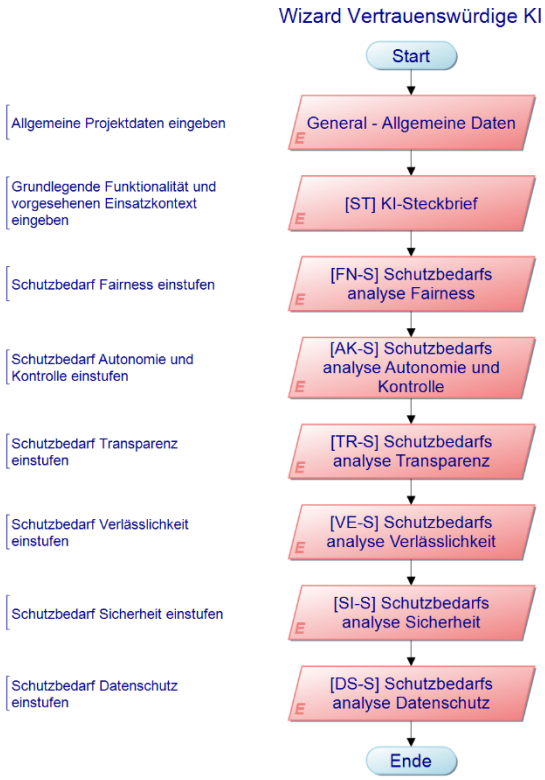
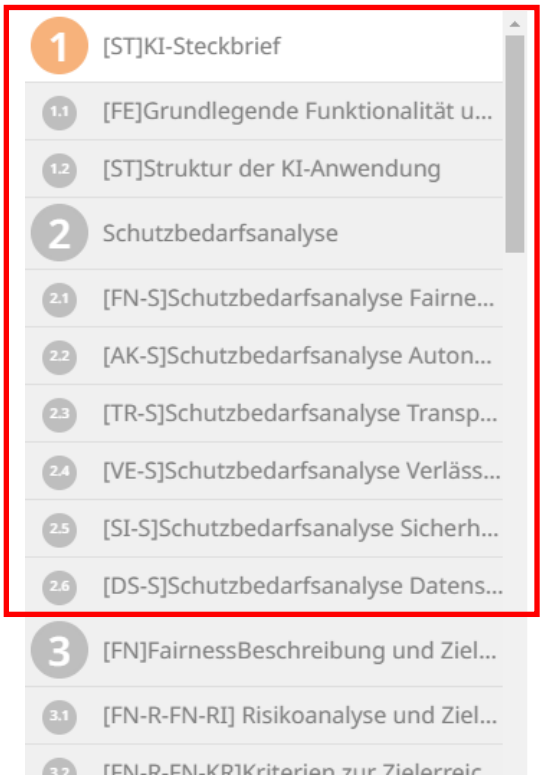
Geplanter Programmablauf bis inklusive Schutzbedarfsanalyse	Darstellung im Wizard (Projektstruktur) aus Anwender*innensicht
	

Abb. 15: Wizard-Ablauf inklusive Schutzbedarfsanalyse, Quelle: Eigene Darstellung.

Dies hat den Grund, dass eine Reduzierung der inhaltlichen Anforderungen für die Nutzer*innen eine der Hauptaufgaben des Wizards ist. Es sollen nur diejenigen Anforderungen abgefragt bzw. dargestellt werden, die für die spezifische KI-Applikation tatsächlich nötig sind und erfüllt werden müssen. Durch die Abfrage des Schutzbedarfs der einzelnen Dimensionen können bei geringem Schutzbedarf gesamte Dimensionen ausgeblendet werden. Dies hat eine deutliche Reduzierung der Anforderungen zufolge, was wiederum den Nutzer*innen hilft, sich leichter zurecht zu finden und sich auf die notwendigen Anforderungen zu konzentrieren.

Nach der Schutzbedarfsanalyse werden die Anforderungen der Dimensionen angezeigt, bei denen die Schutzbedarfsanalyse einen mittleren oder hohen Schutzbedarfs ergeben hat. Dabei wird wieder in der Reihenfolge, wie im Leitfaden definiert, vorgegangen – siehe Abb. 16.

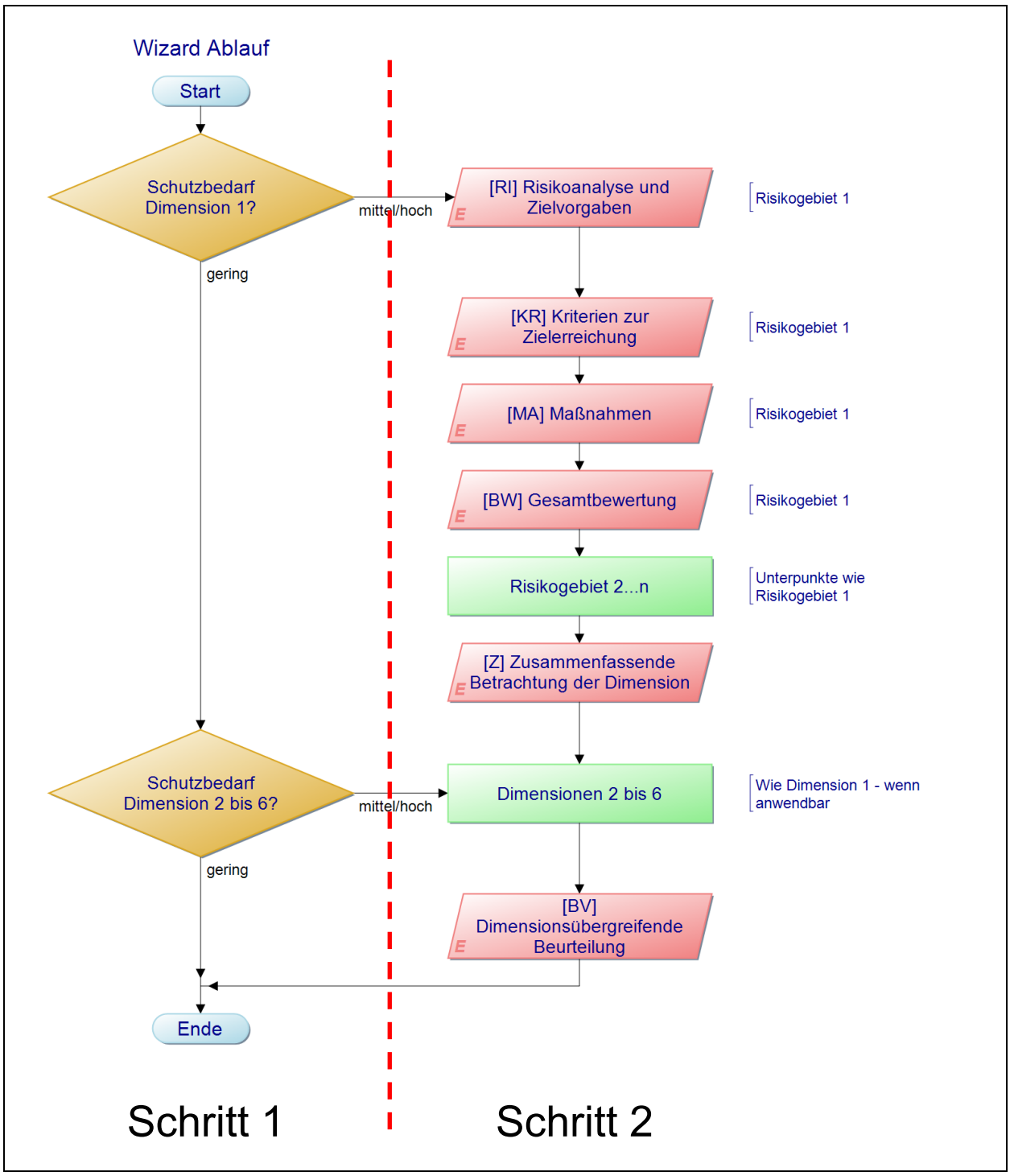


Abb. 16: Wizard-Ablauf gesamt, Quelle: Eigene Darstellung.

Für jedes Risikogebiet wird die Reihenfolge RI-KR-MA-BW eingehalten. Sind sämtliche Risikogebiete bearbeitet, ist die Dimension zusammenfassend zu betrachten – siehe Abb. 17. Der gleiche Ablauf erfolgt

für die weiteren anwendbaren Dimensionen, bevor zuletzt eine dimensionsübergreifende Beurteilung[92] vorgenommen wird.

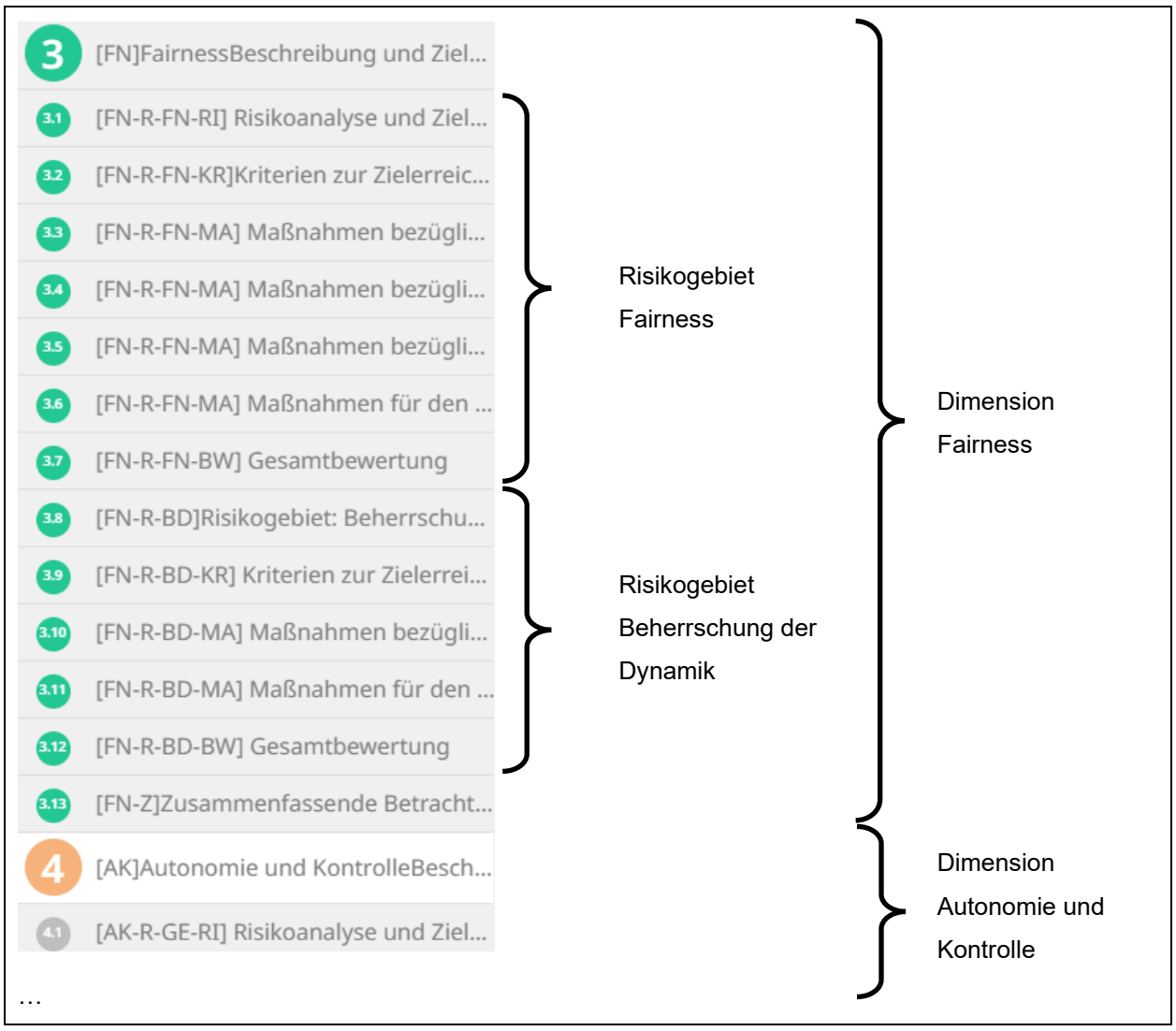


Abb. 17: Projektstruktur aus Nutzer*innensicht, Quelle: Eigene Darstellung.

2.6.1 Eingabeoberfläche

Aus Nutzer*innensicht gibt es eine Eingabeoberfläche, in der man die Anforderungen aus dem Leitfaden dokumentieren kann. Die Anforderungen wurden als einzelne Fragen formuliert, um es den Nutzer*innen einfach zu machen, die zu erfüllenden Anforderungen Stück für Stück zu bearbeiten.

Auf der linken Seite der Eingabeoberfläche wird die Projektstruktur angezeigt – siehe Abb. 18 - A. Auf der rechten Seite befinden sich die einzelnen Fragen und Eingabemöglichkeiten der jeweiligen *Section*, bzw. *Subsection*, in der man sich gerade befindet – siehe Abb. 18 - B. Die Eingaben werden übernommen und

[92] Vgl. Poretschkin u.a. (2021), S. 161 f.

können in einem Report ausgegeben werden. Im rechten unteren Bereich befinden sich mehrere Schaltflächen – siehe Abb. 18 - C.

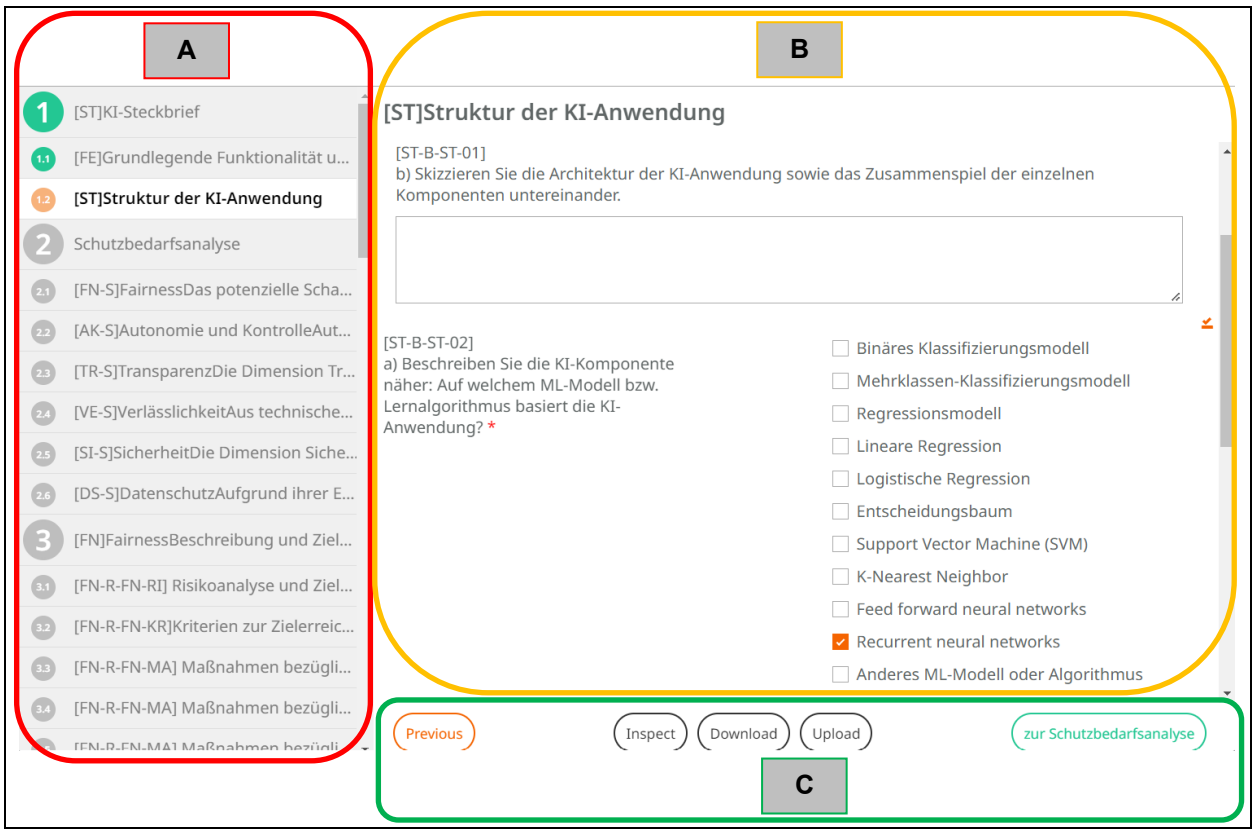


Abb. 18: Bedienoberfläche aus Nutzer*innensicht, Quelle: Eigene Darstellung.

Die Schaltflächen von links nach rechts haben folgende Funktionen:

- Previous – Navigation zur vorigen (Sub-)Section
- Inspect – Hier können die Variablenwerte des Projekts angezeigt werden
- Download – Export der bereits eingegebenen Antworten
- Upload – Import von bereits gegebenen Antworten
- *Unterschiedlicher Text* – Navigation zur nächsten (Sub-)Section

Bei der letzten Schaltfläche wird die jeweils nächste (Sub-)Section textlich auf den Schaltflächen beschrieben. Im Beispiel in Abb. 18 ist der Text **zur Schutzbedarfsanalyse** zu sehen, da die Schutzbedarfsanalyse die nächste Section darstellt.

2.6.2 Fragen und Antworten

Da der KI-Prüfkatalog eine nachvollziehbare Dokumentation der KI-Anwendung fordert, ist ein Großteil der Fragen direkt mit der Variante *Question to be answered with a text* zu beantworten – siehe Abb. 19. Somit werden die Angaben im späteren Report, der generiert werden kann, gespeichert und dokumentiert.

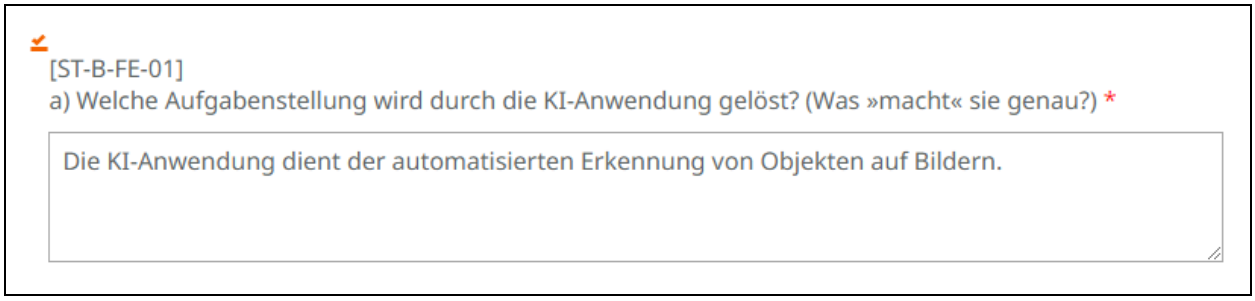


Abb. 19: Darstellung Textfenster unterhalb der Frage, Quelle: Eigene Darstellung.

Werden längere Antworten erwartet, ist es sinnvoller, diese auszulagern. In einem solchen Fall empfiehlt es sich, die Assessment Elements zu verwenden und den Nutzer*innen eine eigene diesbezügliche Anforderung zu definieren, die gesondert von dem Report zu erfüllen ist, z.B. sollte bei vorgeschriebenen Tests ein eigenes Prüfprotokoll erstellt werden, um diese Anforderung zu erfüllen.

2.6.2.1 Dropdown

Die Funktion Dropdown wurde für kurze Antworten, d.h. einzelne Wörter, die jedoch mehr als nur zwei Antwortmöglichkeiten zulassen, ausgewählt. Dies lässt eine übersichtliche, aber dennoch kompakte Darstellungsform aus Anwender*innensicht entstehen – siehe Abb. 20.

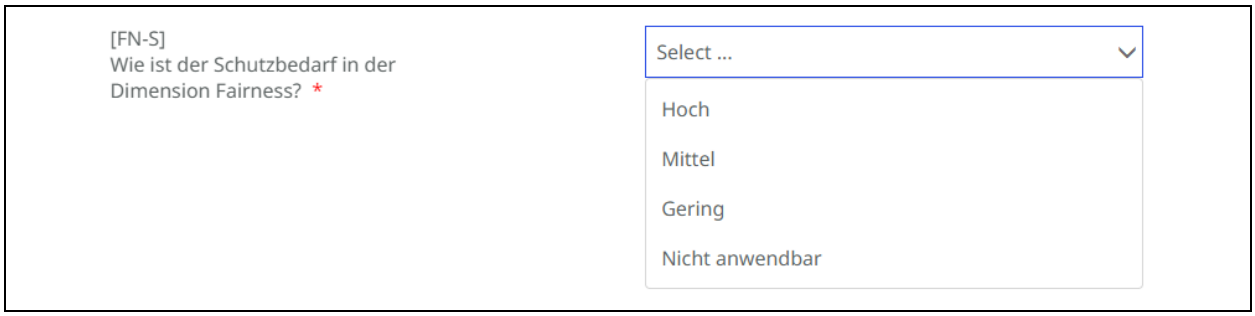


Abb. 20: Darstellung Dropdown aus Anwender*innensicht, Quelle: Eigene Darstellung.

Verwendet wird diese Darstellungsart z.B. bei der direkten Einstufung in der Schutzbedarfsanalyse, bei der es möglich ist, zwischen Hoch, Mittel, Gering und Nicht anwendbar[93] einzustufen. Diese direkte Einstufung ist für erfahrene Anwender*innen nützlich, da diese nicht über diverse Einstufungskriterien zu einer Bewertung des Schutzbedarfs gehen müssen. Dies setzt eine gewisse Expertise in diesem Bereich voraus. Nach der entsprechenden Einstufung werden die Einstufungskriterien neben der Einstufung als Information angezeigt. Dies soll eine mögliche Fehleinstufung verhindern und dem/der Anwender*in direkt ein

[93] Vgl. Kapitel 4.1.1.

Feedback geben, so dass eine ggf. nötige Korrektur direkt an dieser Stelle vorgenommen werden kann und nicht erst im Nachhinein erfolgen muss.

2.6.2.2 Radiobutton(s) shown vertically – below

Für die Anzeige von einfachen Fragen, die lediglich mit Ja oder Nein beatwortet werden sollen, ist die Anzeigeart *Radiobutton(s) shown vertically – below* gewählt worden. Diese Anzeigeart wurde gewählt, weil nur wenig Text und damit wenig Platz im Browserfenster benötigt wird. Die volle Anzeige der beiden Antworten dient somit auch der Übersichtlichkeit – siehe Abb. 21.

Abb. 21: Darstellung Radiobutton(s) shown vertically - below aus Anwender*innensicht, Quelle: Eigene Darstellung.

2.6.2.3 Radiobutton(s) shown vertically – right side

Die Darstellungsart *Radiobutton(s) shown vertically – right side* wurde getestet. Es wurde jedoch kein durchgängig sinnvoller Einsatz im Rahmen des erstellten Wizards offensichtlich. Denkbar wäre ein Einsatz bei sehr kurzen Fragen, bei denen auch so viele Antworten zur Auswahl stehen, dass sie unter der Frage nicht mehr nebeneinander dargestellt werden können. Sobald es mehrere Antwortmöglichkeiten gibt, die einen längeren Text beinhalten, zieht sich die gesamte Frage jedoch optisch in die Länge – siehe Abb. 22. Dadurch wird die Darstellung insgesamt unübersichtlicher, was zu vermeiden ist.

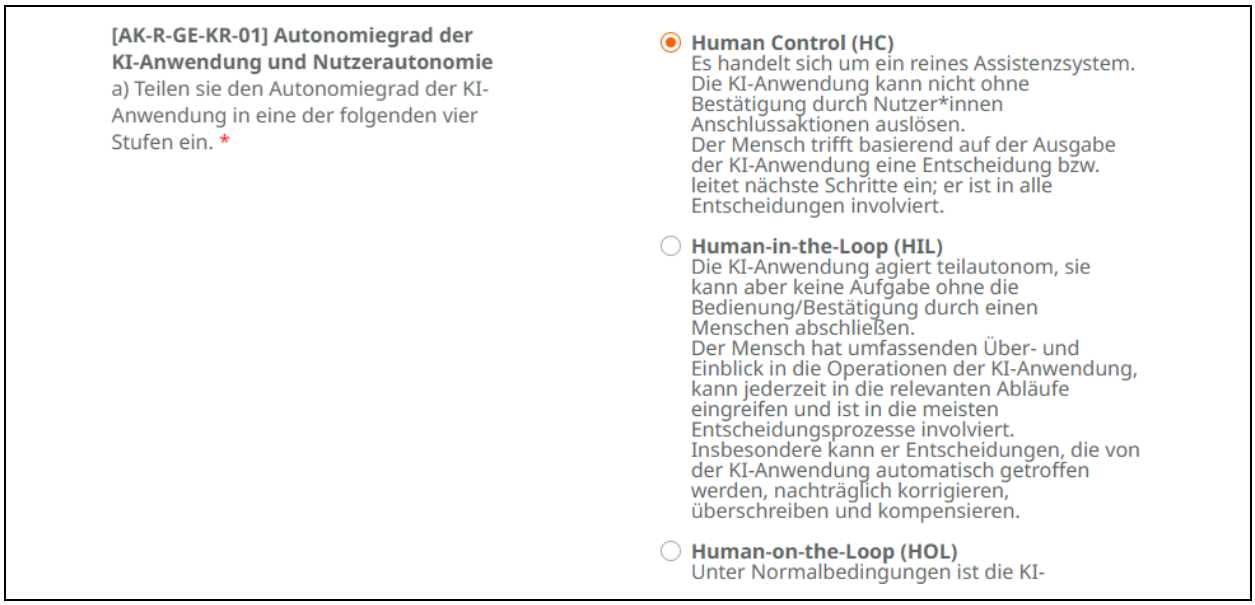


Abb. 22: Darstellung Radiobutton(s) shown vertically – right side aus Anwender*innensicht, Quelle: Eigene Darstellung.

Eine mögliche diesbezügliche Optimierung wäre, das Darstellungsfenster in der Breite einstellbar zu machen.[94]

2.6.2.4 Checkboxes

Sind im Leitfaden mehrere Antwortmöglichkeiten vordefiniert, so wird die Darstellungsart Check Boxes ausgewählt. Dies hat den Vorteil für die Nutzer*innen, dass unnötige Schreibarbeit gespart wird und die Fragen stattdessen mit wenigen Klicks beantwortet werden können.

Im Steckbrief z.B. werden unter [ST-B-ST-02] auf die Frage nach dem ML-Modell bzw. Algorithmus verschiedene, bereits bekannte Modelle bzw. Algorithmen vorgeschlagen. Die vordefinierten Beispiele werden als Checkboxes hinterlegt und können von den Anwender*innen angeklickt werden. Für eine beispielhafte Betätigung einer Checkbox – siehe Abb. 23 – A.

Die Liste wird um eine Checkbox erweitert (siehe Abb. 23 – B), die gewählt werden soll, falls das verwendete Modell oder der Algorithmus nicht in den Vorgaben enthalten sind. Diese Checkbox wird für eine nächste Frage als Bedingung genutzt, da es bei einem anderen verwendeten ML-Modell oder Algorithmus notwendig ist, dieses bzw. diesen zu beschreiben.

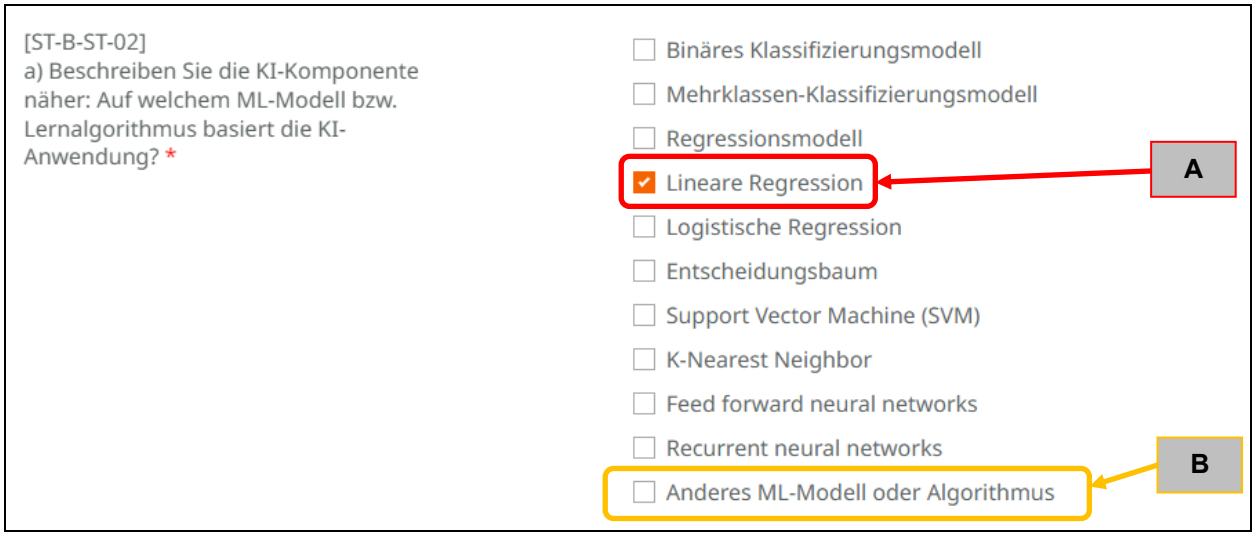


Abb. 23: Darstellung Checkboxes aus Anwender*innensicht, Quelle: Eigene Darstellung.

Auch bei dieser Darstellungsart ist wieder darauf zu achten, dass die Antworten keine gesamten Sätze beinhalten, da es auch hier andernfalls zu einer unübersichtlichen Darstellung führen würde.

Die Antwortmöglichkeit Checkboxes wird aufgrund der, im Leitfaden vordefinierten, Fairness-Definitionen unter [FN-R-FN-KR-01] z.B. auch für dieses Kriterium gewählt.

[94] Vgl. Kapitel 4.2.9.

2.6.2.5 Nutzung der Variablen

Bei Auswahl der Checkboxes als Antwort, werden die Antwortmöglichkeiten im Kartenreiter Answer(s) vordefiniert.[95] Da es bei der Konfiguration der Antwort-Elemente die Möglichkeit gibt, neben der textlichen Antwort auch einen frei wählbaren Variablenwert im Feld *Value* zu hinterlegen, wird diese Funktion zum leichteren Auffinden der entsprechenden Antwort für eine nachgeschaltete Bedingung verwendet. Während der Erstellung des Wizards wird der immer größer werdende Umfang und die damit verbundene große Anzahl an Elementen zu einem Faktor, der ein zügiges Bearbeiten zunehmend schwieriger gestaltet. Speziell bei Änderungen in den Bedingungen der Sichtbarkeit von Fragen wird das Anpassen immer zeitaufwändiger.

Die Sichtbarkeit der Dimensionen als Hauptvereinfachung muss daher schnell auffindbar und anpassbar gestaltet werden. Hierfür werden die Variablenwerte genutzt. Sämtliche Eigenschaften, die zu einer hohen Einstufung des Schutzbedarfs führen, werden mit einem Wert von 100.00 deklariert. Die Eigenschaften, die zu einer mittleren Einstufung führen, werden mit einem Wert von 50.00 deklariert und die verbleibenden Eigenschaften, die zu einer geringen Einstufung in Bezug auf den Schutzbedarf der jeweiligen Dimension führen, werden mit einem Wert von 0.00 deklariert – siehe Abb. 24. Die Auswahl der Zahlenwerte erfolgt in Anlehnung an Prozent.

Seq ▲	Option	Applicability	Value	
1	Die KI-Anwendung regelt den Zugang zu essenziell die Persönlichkeit betreff...	Always visible & appli...	100.00	hoch
2	Die KI-Anwendung dient der Vergabe eines Visums	Always visible & appli...	100.00	
3	Die KI-Anwendung dient der Zulassung zu Schulen/Universitäten	Always visible & appli...	100.00	
4	Die KI-Anwendung dient der automatisierten Kreditvergabe	Always visible & appli...	100.00	
5	Die KI-Anwendung dient der Entscheidung über die Art der medizinischen Beha...	Always visible & appli...	100.00	
6	Die Ausgabe der KI-Anwendung steht, wenn auch nur im weiteren Sinne, im Zus...	Always visible & appli...	50.00	mittel
7	Die KI-Anwendungen gibt eine Entscheidung über eine Person aus bzw. kategor...	Always visible & appli...	50.00	
8	Die KI-Anwendung, verarbeitet personenspezifische Eingaben (z. B. eine Spr...	Always visible & appli...	50.00	
9	Die Ausgabe der KI-Anwendung ist weder sensibel noch hat sie maßgebliche Au...	Always visible & appli...	50.00	
10	Die KI-Anwendung gibt Empfehlungen für Gesichtserkennung auf Fotos in den S...	Always visible & appli...	50.00	
11	Die KI-Anwendung dient der Klassifikation des Alters einer Person basierend...	Always visible & appli...	50.00	
12	Die KI-Anwendung ist ein Spracherkennungssystem	Always visible & appli...	50.00	
13	Die KI-Anwendung verarbeitet keine personenbezogenen Daten, die Aufschluss ...	Always visible & appli...	0.00	gering
14	Die Funktion/Ausgabe der KI-Anwendung ist nicht in einen Prozess oder eine ...	Always visible & appli...	0.00	
15	Die KI-Anwendung dient der Empfehlung personalisierter Werbung	Always visible & appli...	0.00	

Abb. 24: Konfiguration Checkboxes inklusive Variablenwert für den Schutzbedarf Fairness, Quelle: Eigene Darstellung.

Um anschließend die Sichtbarkeit der Dimensionen zu konfigurieren (siehe auch Kapitel 2.6.4.2), kann zuerst über *Add Condition* und *Add Answer*[96] das Fenster Select Dependencies mit der Anzeige der

[95] Vgl. Kapitel 2.5.2.3.

[96] Vgl. Kapitel 2.5.3.

vorherigen Fragen und Antworten geöffnet werden. Durch die Nutzung der Kürzel[97] aus dem Leitfaden, im Beispiel [FN-S] für die Schutzbedarfsanalyse der Dimension Fairness, und die Verwendung der Textsuche in dem geöffneten Fenster, kann zunächst vorselektiert werden. Es werden nur Fragen angezeigt, in deren Text auch [FN-S] enthalten ist. Da es anfangs schwierig ist zu erkennen, welche der gegebenen Antworten zu einer mittleren oder hohen Einstufung führen, wird durch das Ablesen der Variablenwerte neben den textlichen Antworten eine einfache Zuordnung realisiert – siehe Abb. 25. In der Spalte A werden die Variablenwerte angezeigt und in den Zeilen B werden die ersten beiden Antworten angezeigt, die mit einem Variablenwert von 100.00 zu einem hohen Schutzbedarf der entsprechenden Dimension führt.

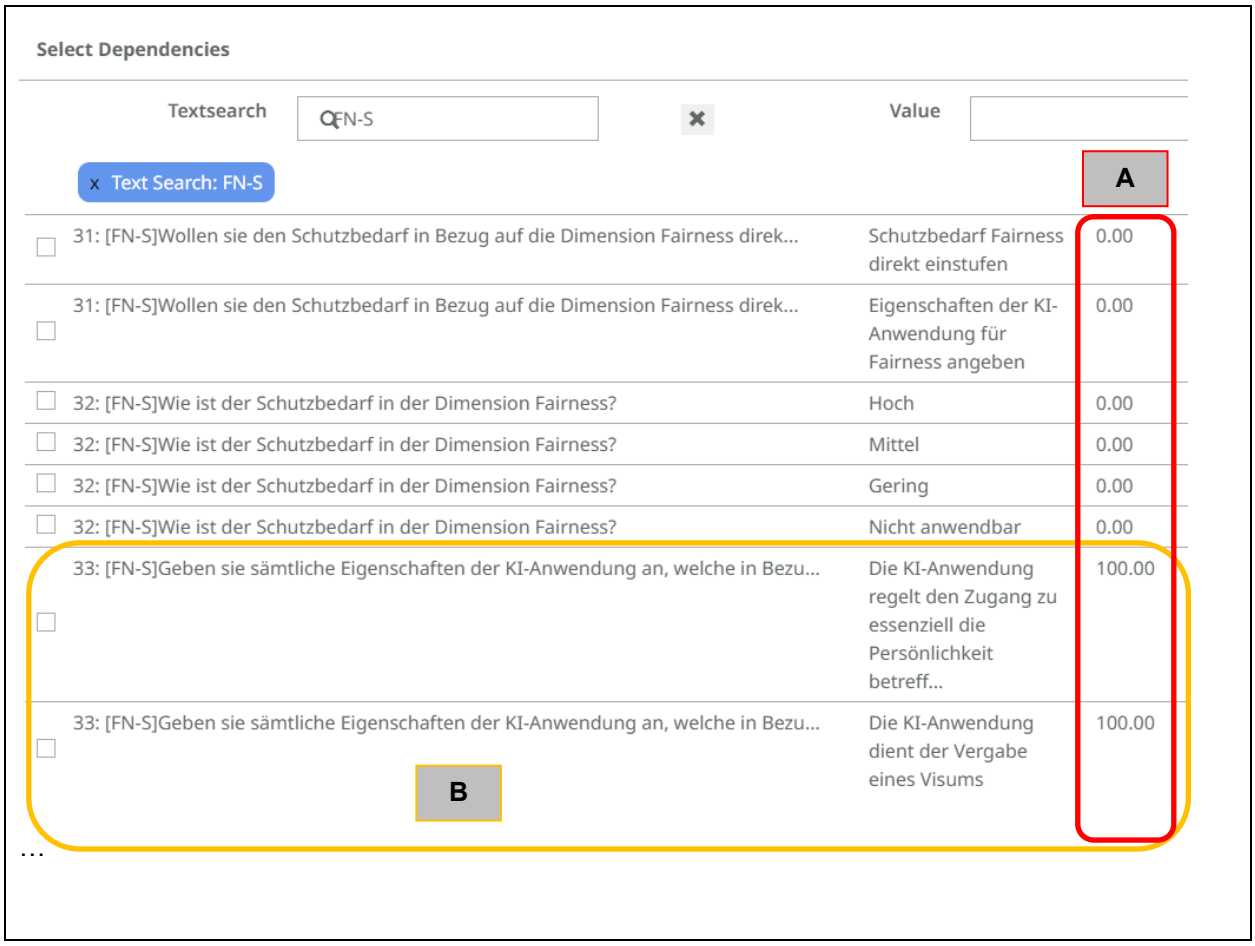


Abb. 25: Variablenanzeige bei der Auswahl der Bedingungen, Quelle: Eigene Darstellung.

[97] Vgl. Kapitel 2.6.5.1.

2.6.3 Assessment Elements – Umsetzung und Bezeichnung

Die Assessment Elements wurden vom Typ *Requirement* gewählt. Da der Leitfaden Anforderungen stellt, die zu dokumentieren sind, ist dies der bevorzugte Typ. Innerhalb des Assessment Elements *Requirement* gibt es noch die Unterkategorie *Requirement Types.* Diese *Requirement Types* können folgender Art sein: *Documentation, Process / Organization* und *Test*. Diese drei verschiedenen Anforderungen ergeben sich auch aus dem Leitfaden, daher wurden diese in der ID zur leichten Erkennbarkeit gesondert benannt. An der ersten Stelle der Abkürzungen der Assessment Elements findet sich die jeweilige Anforderung:

- D – Documentation
- P – Process / Organization
- T – Test

So wird im Rahmen des Projektmanagements eine spätere Zuordnung von einzelnen *Assessment Elements* zu bestimmten Projektmitgliedern leichter möglich, da leicht ersichtlich ist, um welche Anforderung es sich handelt.

Für die gesamte Kodierung der Bezeichnungen siehe Beispiel in Tab. 7.

Beispiel	T	-	FN	-	FN	-	MA	-	05
Bedeutung	Test	-	Fairness	-	Fairness	-	Maßnahme	-	05
Allgemein	Requirement Type	-	Dimension	-	Risikogebiet	-	Detailan-forderung gemäß Leitfaden	-	Ggf. Unterpunkt gemäß Leitfaden

Tab. 7: Bezeichnungen Assessment Elements, Quelle: Eigene Darstellung.

Hat der/die Anwender*in den Wizard bearbeitet und die entsprechenden Fragen beantwortet, werden diese *Requirements* in der Liste angezeigt. Hier liegt der Vorteil wiederum darin, dass bei Verwendung mehrerer Wizards des Frameworks zur Beurteilung der KI-Anwendung, sämtliche Assessment Elements in einer Liste angezeigt werden – siehe auch Abb. 26.[98]

Die einzelnen Assessment Elements, im folgenden Beispiel *Überwachung der Trainingsdaten* – siehe Abb. 26 - A, sind mit Überschriften angezeigt. Wird das Element angeklickt, öffnet sich ein weiteres Fenster, in dem weiterführende Hinweise gegeben werden – siehe Abb. 26 – B, so dass die Nutzer*innen wissen, was bei ihrer KI-Anwendung noch an Maßnahmen zu treffen ist, um sie als vertrauenswürdig einstufen zu können.

[98]Vgl. Kapitel 2.1.

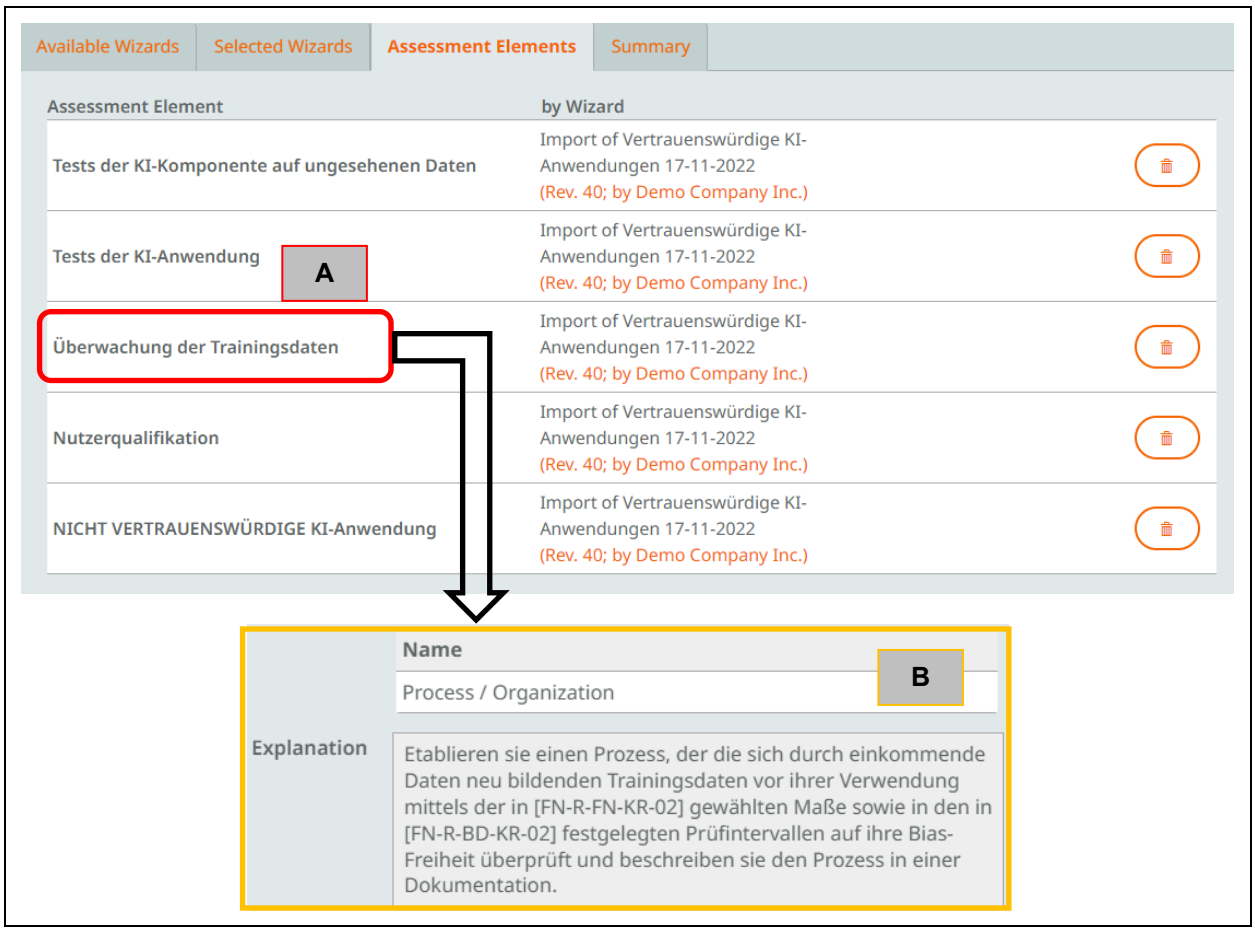


Abb. 26: Übersichtsliste Assessment Elements für den/die Anwender*in, Quelle: Eigene Darstellung.

2.6.4 Vereinfachung aus Nutzer*innensicht

Für eine Vereinfachung aus Anwender*innensicht wird der Leitfaden auf diejenigen Fragen und Anforderungen hin untersucht, die an mehreren Stellen in ähnlicher Weise existieren und einen Einfluss auf die Beurteilung der jeweiligen Dimension der Vertrauenswürdigkeit hat.

2.6.4.1 KI-Steckbrief

Der KI-Steckbrief, der im ersten Teil eine grobe Übersicht über die KI-Anwendung fordert, bietet dafür speziell eine Anforderung, die als Abhängigkeit in mehreren Dimensionen dienen kann. In der Struktur der Anwendung unter [ST-B-ST-02] findet sich die Frage, ob die KI-Komponente im Betrieb kontinuierlich, in regelmäßigen zeitlichen Abständen oder nach Initiation von Neutrainings weiterlernt.[99] Diese Anforderung kann als Bedingung für das Anzeigen der Risikogebiete *Beherrschung der Dynamik* verwendet werden, da sich die Dynamik immer auch aus möglichen Veränderungen aufgrund des Lernens der KI-Anwendung ergibt. Überall dort, wo externe Einflüsse wie z.B. Gesetzesänderungen[100], als eigene Anforderung

[99] Vgl. Poretschkin u.a. (2021), S. 36.

[100] Vgl. Poretschkin u.a. (2021), S. 112.

vorhanden sind, ist diese Bedingung jedoch nicht für das komplette Risikogebiet anzuwenden, sondern nur für den Teil des Lernens der KI-Anwendung.

2.6.4.2 Schutzbedarfsanalyse

Für die Schutzbedarfsanalyse wird zuerst die Abfrage erstellt, ob der jeweilige Schutzbedarfs der Dimension direkt oder über Angabe der Eigenschaften der KI-Anwendung eingestuft werden soll. Die direkte Einstufung erlaubt den erfahrenen Nutzer*innen eine schnellere und damit effiziente Einstufung.

Da es auf Seiten der Nutzer*innen des Wizards mit Sicherheit viele Personen geben wird, die noch keine Erfahrung mit dem Leitfaden bzw. der darin definierten Vorgehensweise bei der Einstufung des Schutzbedarfs haben, werden die Erläuterungen der verschiedenen Einstufungen **Gering**, **Mittel** und **Hoch** genauso wie die weiteren Anforderungen des Leitfadens als Fragen formuliert. Dies hat den Hintergrund, dass somit eine möglichst hohe Akzeptanz und eine Praktikabilität bei der Anwendung des Leitfadens gefördert werden soll. Außerdem sollen die Nutzer*innen somit, unabhängig vom diesbezüglichen Vorwissen, zu einem Ergebnis bei der Einstufung des Schutzbedarfs kommen.

Die Eigenschaften wurden in verschiedene Sätze aufgeteilt, welche die Nutzer*innen mit einer Checkbox je Eigenschaft bestätigen können.[101] Wird eine Eigenschaft gewählt, die zu einer Einstufung von **Mittel** oder **Hoch** führt, ist der Schutzbedarf gegeben und die jeweilige Dimension mit ihren Risikogebieten im Detail zu betrachten. Dies werden daher auch Bedingungen für die Anzeige der Dimensionen – siehe nächstes Kapitel 2.6.4.3.

2.6.4.3 Bedingungen für die Anzeige der Dimensionen

Jede Dimension wird als eigene *Section* des Wizards ausgeführt. Wird die jeweilige Dimension in der Schutzbedarfsanalyse als *gering* eingestuft, wird der Inhalt dieser *Section* im Wizard nicht angezeigt. Um dies zu erreichen, wird das *Section Start* – Element[102] der jeweiligen Dimension mit Bedingungen für die Sichtbarkeit belegt. Um auf das Beispiel der Dimension Fairness zurückzukommen: Die *Section* mit der Dimension Fairness wird nur dann angezeigt, wenn der Schutzbedarf mittel oder hoch ist. Da es bei der Einstufung des Schutzbedarfs sowohl die Möglichkeit gibt, direkt einzustufen, als auch über eine Einstufung der einzelnen Beispiele zu einem Ergebnis zu kommen, sind hier 14 Antworten mit einer ODER-Verknüpfung als Bedingungen zu konfigurieren – für einen Ausschnitt der Konfiguration siehe Abb. 27. Das sind sowohl die beiden Antworten **Hoch** und **Mittel** bei Direkteinstufung, als auch die zwölf Antworten mit den Eigenschaften, die zu einer mittleren oder hohen Einstufung führen.[103]

[101] Eigenschaften der Dimension Fairness – siehe Abb. 24.

[102] Siehe auch Kapitel 2.5.2.1.

[103] Vgl. Abb. 24.

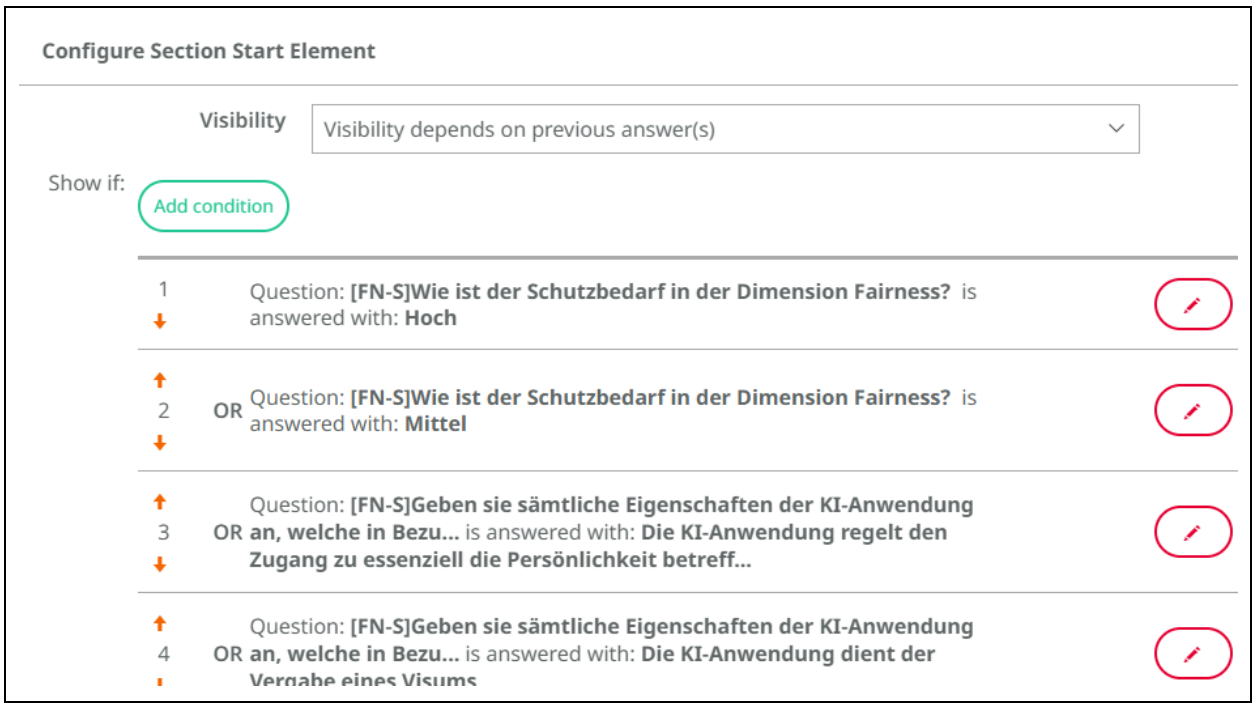


Abb. 27: Konfiguration Section Fairness, Quelle: Eigene Darstellung.

2.6.4.4 Bedingungen für die Anzeige von Fragen innerhalb der Dimensionen

Innerhalb der Dimensionen gibt es diverse Anforderungen, die je nachdem, was die Nutzer*innen erfüllt haben oder nicht, weiterführende Anforderungen bedingen.

Beispielhaft soll eine Anforderung aus dem Leitfaden beschrieben werden, der für die KI-Komponente in Bezug auf die Faire Adaption und Nachverarbeitung zu treffen ist. Es gibt unter [FN-R-FN-MA-04] eine Anforderung, dass die getroffenen Maßnahmen beschrieben und begründet werden, falls die Ergebnisse des ML-Modells unfair werden.[104] Diese Anforderung wird mittels *Ja/Nein*- Abfrage im Wizard umgesetzt und für die Nutzer*innen insofern vereinfacht, dass je nach Antwort nur die jeweils anwendbaren weiterführenden Anforderungen angezeigt werden – siehe Abb. 28. Das heißt, wird die Frage mit *Ja* beantwortet, sind die angewandten Maßnahmen zu beschreiben, wird die Frage mit *Nein* beantwortet, so ist zu begründen, warum keine Maßnahmen angewandt werden.

[104] Vgl. Poretschkin u.a. (2021), S. 43.

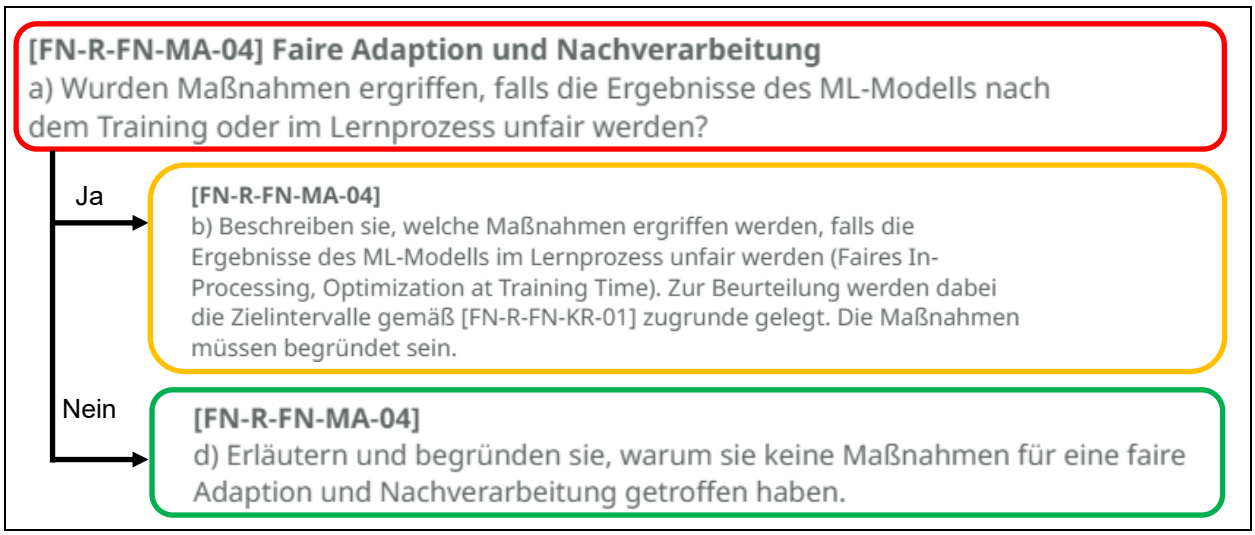


Abb. 28: Bedingungen für Fragen in den Dimensionen, Quelle: Eigene Darstellung.

Dieser Logik folgend, werden Anforderungen identifiziert, die je nach umgesetzter Basis zu unterschiedlichen Folgemaßnahmen führen und in gleicher Weise im Wizard umgesetzt. Diese Vorgehensweise folgt immer dem Ziel der Vereinfachung der Anforderungen aus dem Leitfaden für die Nutzer*innen.

2.6.5 Weitere Umsetzungsdetails

2.6.5.1 Verwendung der Kürzel

Die Kürzel, die im Leitfaden verwendet werden, um den jeweiligen Schritt in der Bewertung zu beschreiben, werden auch im umgesetzten Wizard verwendet und bei jedem Abschnitt bzw. vor jeder Frage vor dem zugehörigen Inhalt genannt. Dies hat zwei Vorteile:

1. Suchfilter

Bei der Erstellung der Elemente, d.h. sowohl der Fragen als auch bei den Assessment Elements konnte für die Konfiguration der Anzeigebedingungen jeweils die Suchfunktion (Textsearch - siehe Abb. 29 - A) verwendet werden und somit leicht vorgefiltert werden. Nach der Eingabe des Kürzels werden unterhalb nur die vorherigen Elemente angezeigt, die mit dem eingegebenen Text übereinstimmen. In der zweiten Spalte werden die zur Frage gehörenden und möglichen Antworten in je einer Zeile dargestellt (siehe Abb. 29 - B). Die jeweilige Anzeigebedingung wird durch das Setzen eines Hakens auf der linken Seite der Zeile ausgewählt und mittels Betätigung der Schaltfläche *Save* wird diese Bedingung für das Element übernommen.

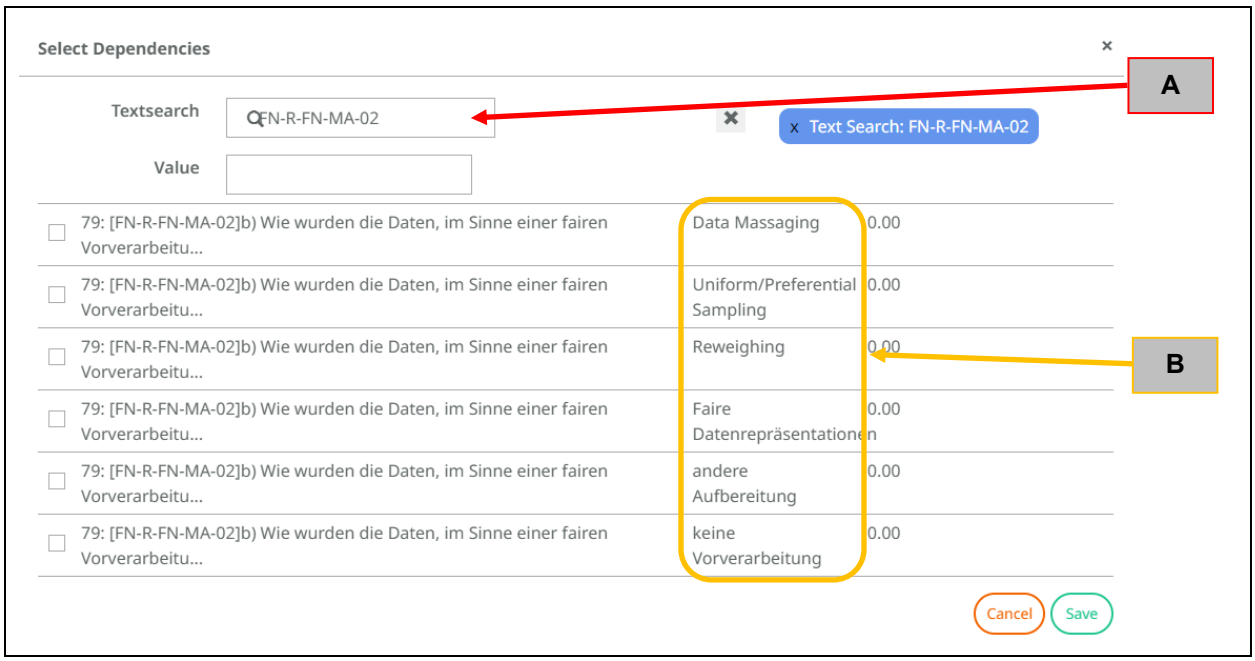


Abb. 29: Kürzel als Hilfe für die Elementkonfiguration, Quelle: Eigene Darstellung.

2. Leitfaden bekannt – Wizard nachvollziehbar.

Für diejenigen Personen, die sich zuvor mit dem rein textbasierten Leitfaden bereits auseinandergesetzt haben und daher Kenntnis der Kürzel haben, wird es leicht sein, sich im Wizard zurecht zu finden, da die Kürzel bei jedem Schritt sichtbar und damit leicht nachvollziehbar sind.

Lediglich bei den *Assessment Elements* werden nicht direkt die Kürzel des Leitfadens verwendet, sondern eine Abwandlung in Anlehnung daran.[105]

2.6.5.2 Symbole / Icons

Um eine einfache optische Zuordnung der Dimensionen im Wizard zu erreichen, werden Symbole für die Dimensionen verwendet. Da bereits vom Fraunhofer Institut Symbole dafür vorgesehen wurden, werden genau diese auch für den Wizard verwendet– siehe Tab. 8. Dabei wurden die Icons zunächst bei den *Section Start* – Elementen der jeweiligen Dimension vorgesehen, um den Wizard nicht unnötig mit Dateivolumen zu füllen und ggf. die Performance zu verschlechtern.

[105] Vgl. Kapitel 0.

Symbol / Icon					
					
Fairness	Autonomie und Kontrolle	Transparenz	Verlässlichkeit	Sicherheit	Datenschutz
Dimension					

Tab. 8: Symbole und Dimensionen, Quelle: Vertrauenswürdiger Einsatz von KI.[106]

[106] Vgl. Cremers u.a. (2019), S.15.

3 DAS TOOL IN DER PRAKTISCHEN ANWENDUNG

Um das Software-Tool, den Wizard, zu testen, sollen Eingaben von möglichen Use Cases, d.h. existierende KI-Anwendungen, in den Wizard eingepflegt werden und die Ausgaben, also auch die Anzeigen der Fragen, die an Bedingungen geknüpft sind, geprüft werden. Es existieren zwar diverse öffentlich verfügbare KI-Applikationsbeschreibungen, jedoch fehlen häufig Detailinformationen, die für eine Beantwortung vieler Fragen aus dem Leitfaden nötig wären. Auf die Umsetzung des Wizards hat dies jedoch keinen Einfluss, da ein Großteil der Eingaben zu den KI-Anwendungen reine textliche Beschreibungen für eine übergeordnete Dokumentation sind. Unabhängig von tatsächlichen umgesetzten Applikationen wurde die Anzeige von Fragen, die an Bedingungen geknüpft sind, getestet, um sicherzustellen, dass diese Abhängigkeiten tatsächlich zu einer Vereinfachung für die Nutzer*innen führen.

3.1 Use Case

Ein tatsächlich realisierter Use Case wird von Siemens für die Bewertung bereitgestellt. Es werden konkrete Teilangaben zu dem Use Case gegeben. Diese Angaben werden für den KI-Steckbrief verwendet und dort eingegeben. Bei fehlenden Angaben werden für den Test im Wizard Ersatzangaben getroffen. Eine Bewertung des Use Case mit den tatsächlichen Angaben kann daher von den Ergebnissen des Tests abweichen.

3.1.1 Anwendungsbereich

In dem Projekt *Development and implementation of application scenarios of machine learning in electronics manufacturing* wird eine KI-Anwendung als Assistenzsystem für das Anlagenbedienpersonal einer Montageanlage für Elektronikkomponenten entwickelt. Die KI-Anwendung dient der Aufdeckung von Pseudofehlern, die bei der bestehenden Prüfung der Elektronikkomponenten mittels Röntgenprozess häufig auftreten. Durch den Einsatz von automatisierter Bilderkennung mittels Neuronaler Netze soll der Prüfaufwand durch das Bedienpersonal reduziert werden. Bei erkannten Pseudofehlern bekommt das Bedienpersonal einen Hinweis und muss den Fehler bestätigen, bevor es zu einer weiteren Verarbeitung der Komponenten kommt – siehe Abb. 30.

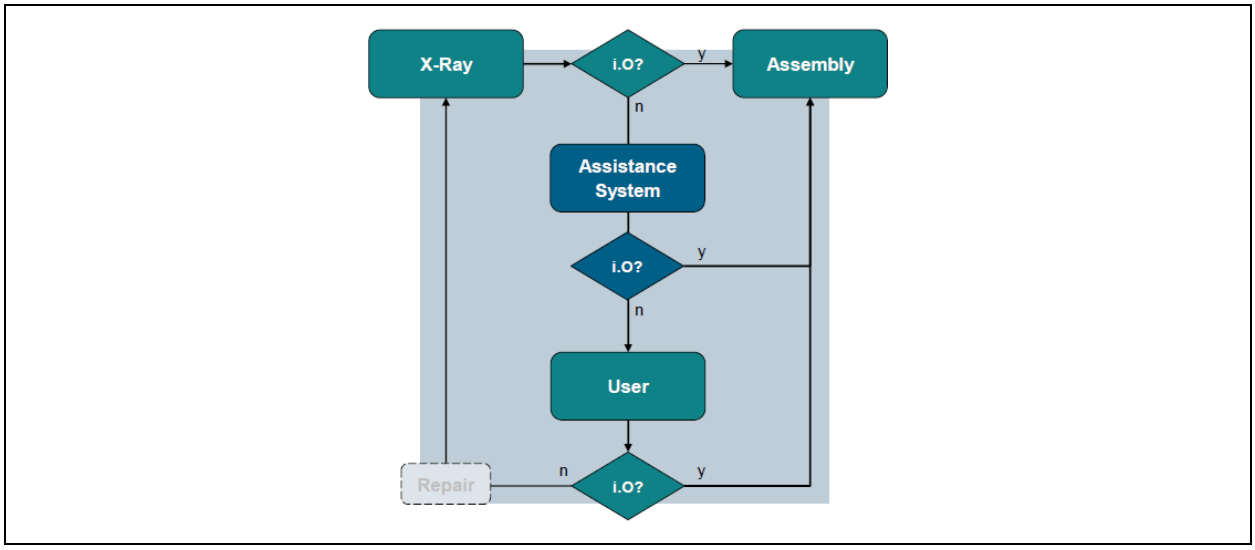


Abb. 30: Ablauf Use Case, Quelle: Siemens.

3.1.2 Umsetzung im Tool

Aufgrund der inhaltlich sehr umfangreichen Anforderungen von textlichen Beschreibungen, die sich aus dem Leitfaden ergeben, wird die Beantwortung der Fragen für den Test des Wizards vereinfacht. Der KI-Steckbrief wird mit Antworten des tatsächlichen Use Case beschrieben. Bei dem ersten Risikogebiet der ersten Dimension werden die Teilsätze beschrieben, um zu verifizieren, dass sie an der entsprechenden Stelle im Report übernommen werden. Bei den weiteren Risikogebieten wird lediglich ein beispielhaftes Wort zum Testen eingefügt, um die gleiche Funktionalität zu bestätigen.

Die Funktionen werden daher auf drei verschiedenen Ausgabearten verifiziert. Dies erfolgt über die Anzeige der Fragen während der Eingabe, über die Anzeige der Assessment Elements und über die Kontrolle der Ausgaben im generierten Report – siehe folgende Kapitel 3.1.2.1, 3.1.2.2 und 3.1.2.3.

3.1.2.1 Fragenanzeige während der Eingabe

Um zu verifizieren, dass die Fragen in der Bedienoberfläche den Nutzer*innen abhängig von den vorherigen Eingaben angezeigt werden, werden jeweils die unterschiedlichen Antwortmöglichkeiten ausgewählt. Im folgenden Beispiel in Abb. 31 wird als Antwort zunächst Ja (A) und anschließend Nein (B) ausgewählt.

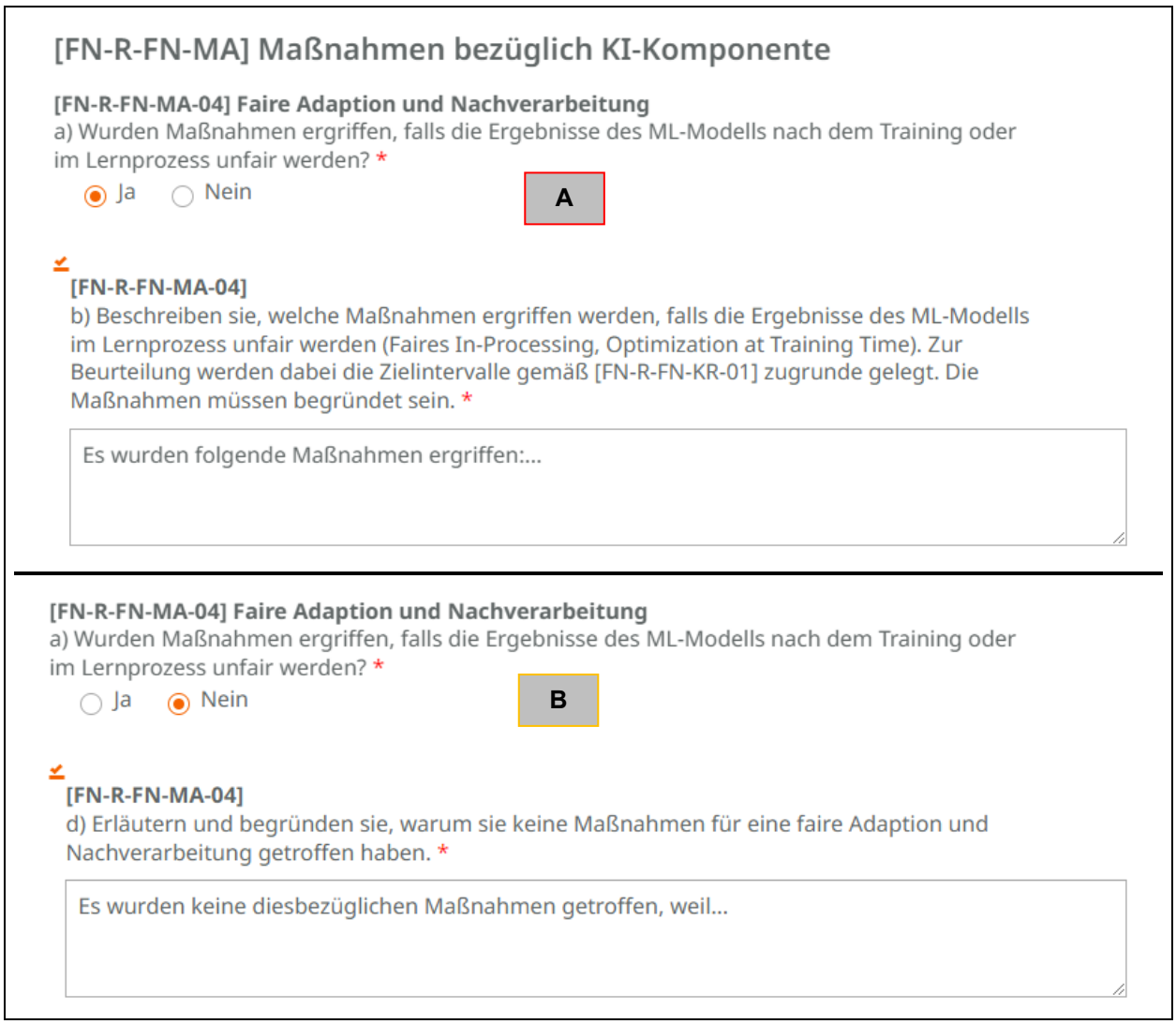


Abb. 31: Fragenanzeige während der Eingabe aus Nutzer*innensicht, Quelle: Eigene Darstellung.

Es ist zu erkennen, dass sich die nachfolgende Frage dementsprechend verändert und das Verhalten des Wizards somit den Erwartungen gemäß Konfiguration entspricht.[107]

3.1.2.2 Anzeige der Assessment Elements

Die zweite Art der Verifizierung erfolgt durch Prüfung der Anzeige der Assessment Elements.[108] Dafür werden die Fragen, welche Bedingungen für die Anzeige der Assessment Elements sind, derart beantwortet, dass die Elemente angezeigt werden müssten. Nach Beantwortung sämtlicher Fragen des Wizard werden nun auf dem Kartenreiter Assessment Elements die entsprechenden Elemente angezeigt - siehe Abb. 32.

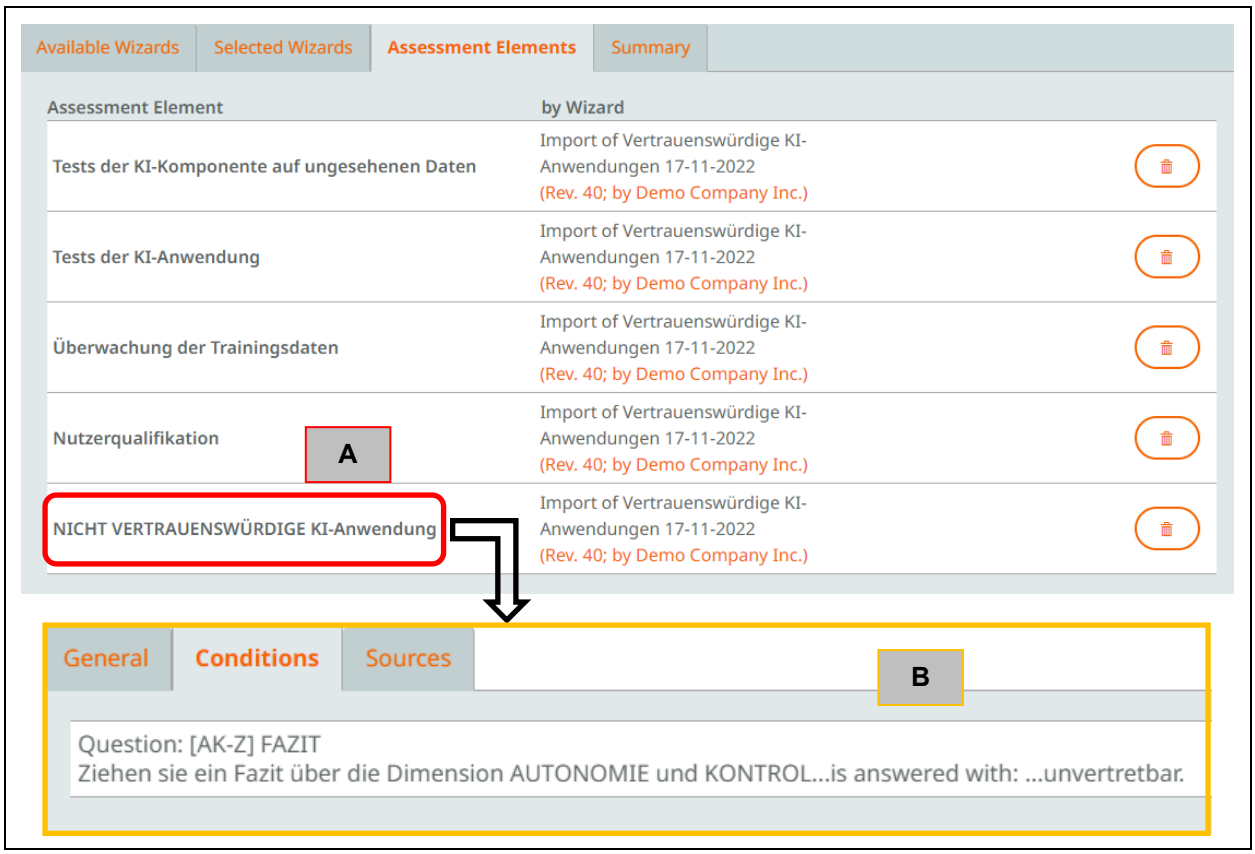


Abb. 32: Dimensionsübergreifende Beurteilung in den Assessment Elements, Quelle: Eigene Darstellung.

Der letzte Schritt für die Bewertung der Vertrauenswürdigkeit von KI-Anwendungen ist gemäß Leitfaden, nach der Detailbetrachtung der einzelnen Dimensionen, eine dimensionsübergreifende Beurteilung. Diese Beurteilung wird ebenfalls über die Assessment Elements erfasst. Am Ende einer jeden Dimension ist ein Fazit zu ziehen. Ergibt dieses Fazit ein oder mehrere unvertretbare Restrisiken, so ist die KI-Anwendung NICHT vertrauenswürdig. Diese Gesamtbewertung wird auch in der Liste des Assessment Elements angezeigt – siehe Abb. 32 – A. Damit wird den Nutzer*innen direkt visualisiert, dass noch weitere Maßnahmen zu treffen sind. Klickt man das entsprechende Assessment Element an, öffnet sich ein

[107] Vgl. Abb. 28.

[108] Siehe auch Kapitel 2.6.3.

weiteres Fenster, in dem unter dem Kartenreiter *General* ein erläuternder Text dargestellt wird.[109] Auf dem Kartenreiter *Sources* wird der Link auf die Website des Fraunhofer IAIS zum Download des Leitfadens hinterlegt. Öffnet man als Nutzer*in den Kartenreiter *Conditions*, so werden diejenigen Bedingungen (Fragen mit den Antworten) angezeigt, welche Voraussetzungen für die Anzeige des Assessment Elements sind. In dem Beispiel auf Abb. 32 ist die Bedingung dafür, dass die KI-Anwendung nicht vertrauenswürdig ist, das Fazit aus der Dimension Autonomie und Kontrolle [AK-Z] mit der Antwort **…unvertretbar**. Erst wenn sich diese Einstufung ändert, kann die KI-Anwendung vertrauenswürdig sein.

3.1.2.3 Ausgabe im generierten Report

Sind sämtliche Daten eingeben, kann ein Report für die Dokumentation generiert werden. Betätigt man die Schaltfläche *Download Assessment Report*, kann man die erzeugte Datei im PDF-Format speichern. Dies war die dritte Form der Verifizierung der konfigurierten Wizard-Funktionen.

Nach der Eingabe der Daten wird ein Report erzeugt, der die Eingaben übernommen hat. Dabei werden zuerst die Fragen und die Antworten angezeigt. Im Anschluss daran werden außerdem die Assessment Elements wiedergegeben. Dargestellt werden zunächst der Titel und die Erklärung des Assessment Elements (Abb. 33 – A). Darunter folgen die Bedingungen (Abb. 33 – B) für die Anzeige und zuletzt die Quellen (Abb. 33 – C).

Abb. 33: Darstellung der Assessment Elements im Report, Quelle: Eigene Darstellung.

Erkenntnisse, die während der Tests für die Verifizierung entstanden sind und mögliche Optimierungen werden im Kapitel 4.2 genauer behandelt.

[109] Siehe auch Abb. 26.

4 ERKENNTNISSE UND OPTIMIERUNGSMÖGLICHKEITEN

Im folgenden Kapitel sollen die Erkenntnisse aus der Umsetzung des Wizards und daraus folgende mögliche Weiterentwicklungen angeregt werden. Dabei wird in zwei Teilgebiete unterteilt. Zum einen werden mögliche Optimierungen für den Leitfaden selbst beschrieben und zum anderen werden Optimierungen für das verwendete Wizard-Framework erläutert. Diese Vorschläge haben unterschiedliche Hintergründe. Primär wird eine einfache Anwendbarkeit des Leitfadens für Nutzer*innen angestrebt. Außerdem soll der Wizard die Nutzer*innen dabei unterstützen, sämtliche Angaben richtig und so ausführlich wie nötig zu machen. Dennoch wird versucht, den Aufwand für die Bearbeitung mit dem Wizard so gering wie möglich zu halten, indem Eingaben nicht doppelt gemacht werden müssen.

4.1 Hinweise für den Leitfaden

Folgende Optimierungsansätze werden für den Leitfaden festgehalten:

4.1.1 Einstufung des Schutzbedarfs

Im Leitfaden wird im Rahmen der Schutzbedarfsanalyse eine mögliche Einstufung des Schutzbedarfs je Dimension mit *gering* (außer bei der Dimension Verlässlichkeit), *mittel* und *hoch* vorgegeben.[110] Diesbezüglich stellen sich folgende Fragestellungen:

1. Gering - Sollte es nicht eine ergänzende Einstufungsmöglichkeit geben?

   Eine Einstufung der jeweiligen Dimension als *gering,* bezeichnet diese womöglich als unbedeutend bzw. unbeträchtlich. Dennoch ist bei dieser Wortbedeutung davon auszugehen, dass in kleinen Teilen trotzdem ein möglicher Schutzbedarf besteht.

2. Mittel/hoch – Warum gibt es keine Unterscheidung bei der weiterführenden Bewertung?

   Es besteht die Möglichkeit, bei der Einstufung zwischen mittel und hoch zu unterscheiden. Die nachgelagerten Risikoanalysen und Maßnahmen unterscheiden sich jedoch nicht. Daher ist nicht ersichtlich, warum bei der Einstufung während der Schutzbedarfsanalyse unterschieden werden soll.

Bekannt sind solche Einstufungsmöglichkeiten z.B. aus Risikobeurteilungen von Maschinen. Gemäß EN ISO 12100 – Sicherheit von Maschinen – Allgemeine Gestaltungsleitsätze – Risikobeurteilung und Risikominderung ist jedes, an der Maschine auftretende, Risiko einzuschätzen. Dabei werden die zu berücksichtigen Faktoren für diese Einschätzung, Schadensausmaß und Eintrittswahrscheinlichkeit, genannt und auch inhaltlich teilweise benannt.[111] In der tatsächlichen Anwendung, d.h. in der Praxis zeigt sich jedoch, dass genau dieser Teil der Risikobeurteilung einem subjektiven Einfluss derjenigen Person

[110] Vgl. Poretschkin u.a. (2021), S. 30 ff.

[111] Vgl. Austrian Standards Institute/Österreichisches Normungsinstitut (ON) (2010), S. 23 ff.

unterliegt, welche die Risikobeurteilung durchführt. Es zeigt sich jedoch auch, dass es selbst bei unterschiedlichen Risikoeinschätzungen häufig zu den gleichen risikomindernden Maßnahmen kommt und die Notwendigkeit dieser Einstufung daher fragwürdig erscheinen lässt.

Zur ersten Fragestellung sei daher erwähnt, dass für eine sinnvolle Durchgängigkeit eine weitere Einstufungsmöglichkeit vorgesehen werden sollte: **Nicht anwendbar**. Damit wäre deutlich, dass in Bezug auf die derart eingestufte Dimension tatsächlich **kein** Schutzbedarf besteht.

Zur zweiten Fragestellung:

Es sollte verdeutlicht werden, dass die Einstufung lediglich der Priorisierung dient. Ein diesbezüglicher Hinweis findet sich im Leitfaden: „Umgekehrt kann ein hoher Schutzbedarf ein besonderes Augenmerk auf die entsprechende Dimension im Rahmen der Prüfung erforderlich machen.“[112] Dieser kann jedoch von Anwender*innen leicht überlesen werden bzw. zeigt sich, dass dies nur eine Möglichkeit sein **kann**.

Eine andere Variante, um auf die Einstufungen einzugehen, wäre es, unterschiedliche Maßnahmen je nach Einstufungshöhe zu definieren. Ist dies nicht gewünscht, wäre es einfacher, auf eine unterschiedliche Einstufung zu verzichten und bei der Schutzbedarfsanalyse lediglich anzugeben, ob die Dimension betrachtet werden muss oder nicht (Ja/Nein).

4.1.2 Doppelte Kürzel

Die verwendeten Kürzel können hilfreich sein, um sich innerhalb des Leitfadens zurecht zu finden. Es gibt jedoch zwei Kürzel, die doppelt verwendet werden und damit mögliche Verwechslungen entstehen lassen können. Die Abkürzung FN steht sowohl für die Dimension *Fairness*, als auch für das Risikogebiet *Fairness*, welches sich ausschließlich in derselben Dimension findet. Des Weiteren existiert das Kürzel AF als Risikogebiet an zwei Stellen mit unterschiedlicher Bedeutung. Es wird sowohl ***A**uditf**ä**higkeit* als Risikogebiet der Dimension Transparenz (TR), als auch ***A**bfangen von **F**ehlern auf Modellebene* als Risikogebiet der Dimension Verlässlichkeit (VE) zugeordnet. Auch wenn die Logik nachvollziehbar ist, wäre es sinnvoll an dieser Stelle für die Eindeutigkeit[113] weitere Kürzel zu verwenden, auch um Verwechslungen zu verhindern - siehe Tab. 9.

[112] Poretschkin u.a. (2021), S. 31.

[113] Siehe auch Kapitel 4.1.3.

Bezeichnung	Dimension / Risikogebiet	Kürzel im Leitfaden	Kürzel neu
Fairness	Dimension	FN	FN
Fairness	Risikogebiet	FN	FA
Auditfähigkeit	Risikogebiet	AF	AU
Abfangen von Fehlern auf Modellebene	Risikogebiet	AF	AF

Tab. 9: Kürzel neu, Quelle: Eigene Darstellung.

4.1.3 Abkürzung -R-

Die Abkürzungen und Kürzel, die im Leitfaden verwendet werden, sind, wie zuvor schon genannt, durchaus sinnvoll. Das verwendete -R- für das Risikogebiet erscheint jedoch wenig hilfreich für eine Unterscheidung der einzelnen Kürzel, da es nach dem Steckbrief in jedem weiteren folgenden Kürzel an der immer gleichen Stelle vorhanden ist. Da die Dimension der Vertrauenswürdigkeit immer vor dem jeweiligen Risikogebiet steht, ist das -R- entbehrlich, d.h. allein die Reihenfolge beschreibt schon eindeutig, dass entweder die Dimension oder das Risikogebiet gemeint ist. Würde die Maßnahme gemäß vorangegangenem Kapitel 4.1.1 umgesetzt, wäre aufgrund der Einzigartigkeit der jeweiligen Abkürzung ein weiteres eindeutiges Merkmal vorhanden. Dies würde zu einer Vereinfachung der Kürzel führen und damit zu einem leichteren Zugang zu dem Leitfaden im Allgemeinen, was zu begrüßen wäre.

4.1.4 [ST-B-FE-03]

Die Funktion des Steckbriefs ist durchaus nachvollziehbar. An der Stelle [ST-B-FE-03] werden folgenden Beispiele für bestimmte Anforderungen an die KI-Anwendung gegeben: Funktionale Sicherheit, IT-Sicherheit, Persönlichkeitsrechte, …[114]

Diese drei Beispiele sind eine Möglichkeit, im Wizard weiter verarbeitbare Informationen für die entsprechend anwendbaren Dimensionen der Vertrauenswürdigkeit zu erzeugen, d.h. wenn der/die Nutzer*in an dieser Stelle ein oder mehrere dieser drei Beispiele auswählt, dann könnte man innerhalb des Wizards bereits festhalten, dass die entsprechende Dimension(en) im Detail zu bewerten sind, da aufgrund der Anwendbarkeit offensichtlich eine mittlere oder hohe Einstufung in den Dimensionen erfolgen wird.

Problematisch erscheinen an dieser Stelle zwei Sachverhalte:

4.1.4.1 Unkenntnis der Nutzer*innen

Aufgrund der Komplexität der europäischen Regulatorik und der persönlichen Erfahrung des Autors im Bereich der Regulatorik zur Maschinensicherheit, wird im Rahmen der vorliegenden Arbeit davon ausgegangen, dass der Großteil der Nutzer*innen keine bis wenig Kenntnis über die anwendbare

[114] Vgl. Poretschkin u.a. (2021), S. 36.

Regulatorik in Bezug auf die KI-Anwendung hat. Es besteht daher die Gefahr, dass Nutzer*innen aufgrund der Unkenntnis davon ausgehen, dass es keine Anforderungen gibt. In diesem Fall wäre nicht nur der Steckbrief inhaltlich falsch ausgefüllt, es würde bei einer automatisierten Weiterverarbeitung dieser Eingabe im Wizard zu falschen Folgeschritten führen, so würden gegebenenfalls einige Dimensionen ausgeblendet. Da der Leitfaden und der Wizard sich aber an Nutzer*innen wendet, wird hier empfohlen, entweder diese Frage zu streichen oder mit Beispielen für sämtliche Dimensionen bzw. Risikogebiete zu erweitern – siehe auch folgendes Kapitel 4.1.4.1.

4.1.4.2 Willkürliche Beispiele

Die gewählten Beispiele *Funktionale Sicherheit, IT-Sicherheit, Persönlichkeitsrechte* erscheinen an dieser Stelle willkürlich ausgewählt. Die anderen Dimensionen bzw. Risikogebiete sind an dieser Stelle nicht erwähnt und rücken somit aus dem Fokus. Daher sollten diese Beispiele entweder gestrichen werden oder es sind Beispiele aus sämtlichen Risikogebieten zu ergänzen. So bleibt einerseits eine gleichgewichtige Betrachtung gewahrt und andererseits könnten diese Beispiele z.B. in Form einer Auswahltabelle im Wizard direkt für eine erste Einstufung des Schutzbedarfs sämtlicher Dimensionen verwendet werden. Alle nicht anwendbaren Dimensionen bzw. Risikogebiete könnten direkt ausgeblendet werden. Dies führt zu einer einfacheren Bearbeitung für die Nutzer*innen.

4.1.5 Schutzbedarfseinstufungen weitere Beispiele

Im Rahmen der Schutzbedarfsanalysen werden für die verschiedenen Dimensionen der Vertrauenswürdigkeit und die jeweilige Einstufung (gering / mittel / hoch) Beispiele genannt, d.h. unter welchen Umständen eine bestimmte Einstufung zu treffen ist. Diese anwendungsbezogenen Beispiele kommen vermutlich aus Erfahrungen der Autor*innen des Leitfadens bzw. ist es bei einigen Einstufungen offensichtlich, warum diese Beispiele zu einer bestimmten Einstufung führen. So führt die „Entscheidung über die Art der medizinischen Behandlung“[115] durch eine KI-Anwendung offensichtlich zu einem hohen Schutzbedarf in der Dimension Fairness.

Je nach zukünftiger Erfahrung wären weitere konkrete anwendungsbezogene Beispiele im Leitfaden hilfreich. Die Beispiele könnten auch im Wizard übernommen werden. Dies führt zu einer einfacheren Anwendung für die Nutzer*innen des Leitfadens und damit voraussichtlich zu einer größeren Akzeptanz und einer weiter verbreiteten Anwendung, was zu begrüßen wäre.

4.1.6 [FN-R-FN-KR-01] Sich widersprechende Fairness-Definitionen

Im Leitfaden findet sich unter dem Punkt [FN-R-FN-KR-01] die Frage danach, wie die Nutzer*innen mit sich widersprechenden Fairness-Definitionen umgehen. Zuvor werden verschiedene mögliche Definitionen, wie z.B. Group Fairness oder Conditional Statistical Parity, vorgeschlagen.[116] Um eine weit verbreitete

[115] Poretschkin u.a. (2021), S. 38.

[116] Vgl. Poretschkin u.a. (2021), S. 41.

Anwendung des Leitfadens zu erreichen, wäre es hilfreich, wenn bei diesen konkreten Beispielen bereits im Leitfaden dargestellt würde, welche der angegebenen Definitionen widersprüchlich sein können. Andernfalls sind die Anwender*innen bei der Recherche nach möglichen Widersprüchen wieder sich selbst überlassen.

4.2 Hinweise für das Wizard-Framework

In Bezug auf das Wizard-Framework werden die Hinweise in Form von Funktionsbeschreibungen erfolgen, da es hierbei weniger um die konkrete Anwendung des Leitfadens geht, sondern um die gewünschte Funktionalität. Die entsprechende Funktionalität kann in weiterer Folge auch für andere Wizards verwendet werden. Diese notwendigen Funktionen für die Optimierung sind teilweise bei der Erstellung des Wizards erkannt worden und teilweise während der Testdurchläufe aus Sicht der Anwender*innen entdeckt worden.

Durch die Bearbeitung des Wizards, dem wöchentlichen Feedback während der Ausarbeitung und der schnellen Reaktion und Anpassung innerhalb des Wizard-Frameworks wurden viele Funktionen schon während der Ausarbeitung der vorliegenden Arbeit implementiert. Dies wird bei den jeweiligen Funktionen kurz erwähnt.

4.2.1 Tooltip / Pop-Up-Fenster

Da innerhalb des Leitfadens an mehreren Stellen auf andere, bereits erfolgte, Eingaben der Nutzer*innen Bezug genommen wird, indem das jeweilige Kürzel genannt wird, ist aus der jeweiligen Frage nicht direkt ersichtlich, was sie beinhaltet. Als Beispiel sei an dieser Stelle [ST-B-FE-04] genannt:

„[…] Und in welchen zu **[ST-B-FE-02]** verwandten Einsatzkontexten bzw. Betriebsumgebungen sollte von der KI-Anwendung abgesehen werden?“[117]

Hier wäre es hilfreich, durch Mouse-Klick auf das genannte Kürzel ein Pop-up-Fenster mit der entsprechenden Information zu öffnen oder diese Information in Form eines Tooltips, d.h. bei Fahren mit dem Mouse-Pfeil über das Kürzel wird die Information angezeigt.

4.2.2 Listenerstellung und -erweiterung durch Nutzer*innen

Im Rahmen des KI-Steckbriefs findet sich unter [ST-B-ST-01] die Anforderung, die wichtigsten Komponenten der KI-Anwendung und deren Spezifikationen und Funktionalitäten aufzulisten.[118] Dafür wäre es hilfreich, den Nutzer*innen die Möglichkeit zu geben, eine Liste anzulegen und dieser Liste eigene Komponenten hinzuzufügen. Speziell bei den verwendeten Komponenten könnte man diese in weiterer Folge als Bedingungen für bestimmte Fragen innerhalb der Dimensionen verwenden. Werden beim Steckbrief z.B. KI-Komponenten als vorhanden angegeben, so sind ggf. bestimmte Prüfungen oder Tests in Bezug auf diese Komponenten notwendig und daher in der entsprechenden Dimension zu erfragen.

[117] Poretschkin u.a. (2021), S. 36.

[118] Vgl. Poretschkin u.a. (2021), S. 36.

Bei einigen Fragen sind konkrete Beispiele aus dem Leitfaden als Checkboxes für mögliche Antworten ausgeführt, so z.B. unter [FN-R-FN-KR-01] bei den Fairness-Definitionen. Falls jedoch keine der möglichen vorgegebenen Antworten der KI-Anwendung entspricht, ist es notwendig eine nächste Frage zu stellen, um von den Nutzer*innen zu erfragen, welche sonstigen Definitionen verwendet wurden. Dies könnte durch eine Nutzer*innen-seitige Erweiterung der Liste vereinfacht werden, z.B. über eine zusätzliche Schaltfläche **+ weitere Definition**. Außerdem wäre es hilfreich, wenn man diese Eingaben für eine weitere Verwendung im Wizard speichern könnte. Ggf. wäre sogar eine Aufnahme von häufig verwendeten Nutzer*inneneingaben in eine zukünftige Version des Leitfadens sinnvoll.

4.2.3 Help Text / Explanation bei Radiobutton(s)

Sollen Radiobuttons verwendet werden, können diese nur für kurze Antworten verwendet werden, da die Bedienung andernfalls unübersichtlich wird.[119] Möchte man Radiobuttons verwenden, die einen längeren Text beinhalten, ist es bei der Konfiguration möglich, eine textliche Erklärung (Explanation) der jeweiligen Antwortmöglichkeit zu hinterlegen. Ist diese Erklärung auch für die Nutzer*innen von Bedeutung, sollte diese dementsprechend auch in der Nutzer*innenansicht zu sehen sein. Ähnlich einem Help Text sollte dieser z.B. als Tooltip für die Nutzer*innen verfügbar gemacht werden.

Diese Funktion wurde bereits während der Umsetzungsphase des Wizards ergänzt – siehe Abb. 34.

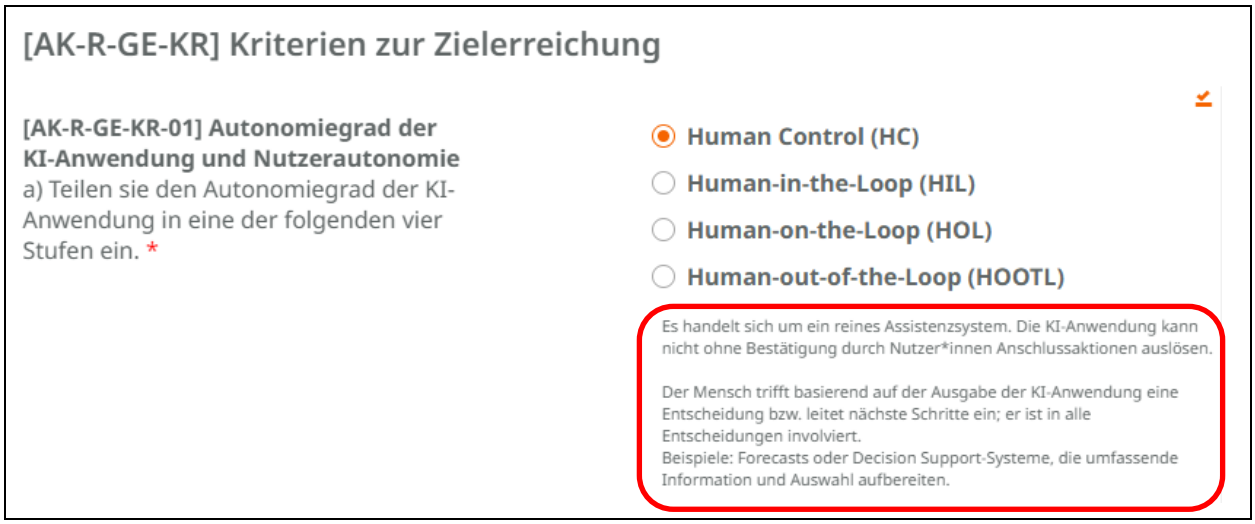


Abb. 34: Hilfetext bei Radiobuttons, Quelle: Eigene Darstellung.

Der Hilfetext wird nach Betätigen der jeweiligen Schaltfläche, d.h. des Radiobuttons, in kleiner Schrift unter der Antwort dargestellt.

4.2.4 Ausblenden der nicht anwendbaren Dimensionen

Zu Beginn der Erstellung des Wizards, wurden diejenigen Fragen, die aufgrund von fehlenden Bedingungen für die Sichtbarkeit ausgeblendet werden sollten, nicht angezeigt. Für einzelne Fragen

[119] Vgl. Abb. 22.

innerhalb einer Dimension ist dies auch ausreichend und durchaus zielführend. Werden jedoch gesamte Dimensionen aufgrund der Nutzer*inneneingabe nicht anwendbar, so sollen diese Dimensionen daher logischerweise auch nicht angezeigt werden. In der Erstversion des Wizard-Frameworks wurden jedoch sämtliche Subsections der nicht anwendbaren Dimensionen angezeigt, ohne jedoch auch nur eine einzige Frage zu beinhalten. Es war notwendig, sich durch die einzelnen Subsections zu klicken, um zur nächsten anwendbaren Dimension zu gelangen.

Optimiert wurde das Framework diesbezüglich schon während der Ausarbeitung. Nicht anwendbare Dimensionen werden in der linksseitigen Projektstruktur nun ausgegraut dargestellt und können nicht mehr angeklickt werden.

4.2.5 Checkboxes – Variablen für Anzahlauswertung

Gemäß Kapitel **[FN-R-FN-KR-01] Quantifizierung von Fairness im Output** des Leitfadens ist zu dokumentieren, wie mit möglichen widersprüchlichen Definitionen von Fairness umgegangen wird.

Im Wizard wurden dazu die, Im Leitfaden zuvor aufgelisteten, Fairness-Definitionen mittels Checkboxes für den/die Anwender*in abgefragt. Für die Anzeigenotwendigkeit der Frage auf widersprüchliche Definitionen gibt es zwei notwendige Voraussetzungen:

1. Es wurde mehr als eine Definition ausgewählt
2. Die ausgewählten Definitionen sind widersprüchlich

Zu 1.) Es wäre hilfreich, die Visibility der Fragen mittels Angabe der entsprechenden *Conditions* auch davon abhängig machen zu können, wenn in einer Checkbox-Liste eine bestimmte Anzahl von Antworten ausgewählt wurde. Dafür könnten z.B. die Variablen verwendet werden. In diesem speziellen Fall könnte man jeder Antwort, d.h. jedem gesetzten Haken in einer Checkbox einen Zahlenwert (1) zuordnen. Werden zwei Haken gesetzt, wird in der Variable der Wert 2 gespeichert. Die Visibility der nächsten Frage hat dann als *Condition*, dass die Variable größer 1 ist.

Zu 2.) Diese Widersprüchlichkeiten sollten direkt im Leitfaden benannt sein, so dass die Anwendung des Leitfadens für den/die Anwender*in ohne weitere Dokumente möglich ist. Dieser Hinweis ist dem Kapitel 4.1 zuzuordnen, verbleibt wegen einer besseren Übersichtlichkeit jedoch an dieser Stelle.

4.2.6 Kopieren und Einfügen ganzer Abschnitte

Der Ablauf der Beurteilung erfolgt strukturell in jeder Dimension gleich. So wird beispielsweise zu Beginn der Schutzbedarfsanalyse den Nutzer*innen im Wizard die Frage gestellt, ob sie eine direkte Einstufung des Schutzbedarfs treffen wollen oder ob die Einstufung mittels Angabe von Eigenschaften erfolgen soll – siehe Abb. 35.

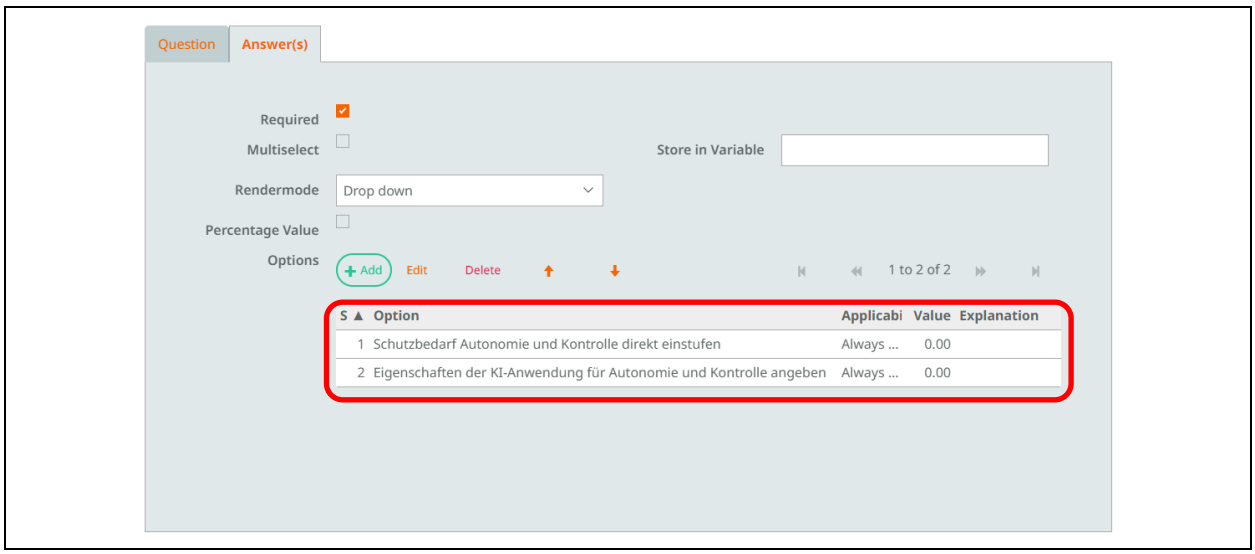


Abb. 35: Konfiguration Einstufungsoptionen AK, Quelle: Eigene Darstellung.

Der einzige Unterschied ist die Bezeichnung der jeweiligen Dimension. Daher wäre es für eine Verbesserung der Effizienz hilfreich, wenn man diesen gesamten Teilabschnitt (im Beispiel: List question) kopieren, an anderer Stelle einfügen und den Text, d.h. die jeweilige Dimensionsbezeichnung, bei der Frage und den möglichen Antworten adaptieren könnte. Gleiches wäre sinnvoll bei den anderen Objekten, wie Step, Substep oder Section.

Diese Funktion wurde bereits während der Ausarbeitung der vorliegenden Arbeit realisiert und derart verfügbar gemacht, dass im Wizard-Framework bei jedem Element, das angelegt wird, auf der rechten Seite der Zeile ein Symbol *kopieren* (duplicate) vorgesehen wurde – siehe Abb. 36.

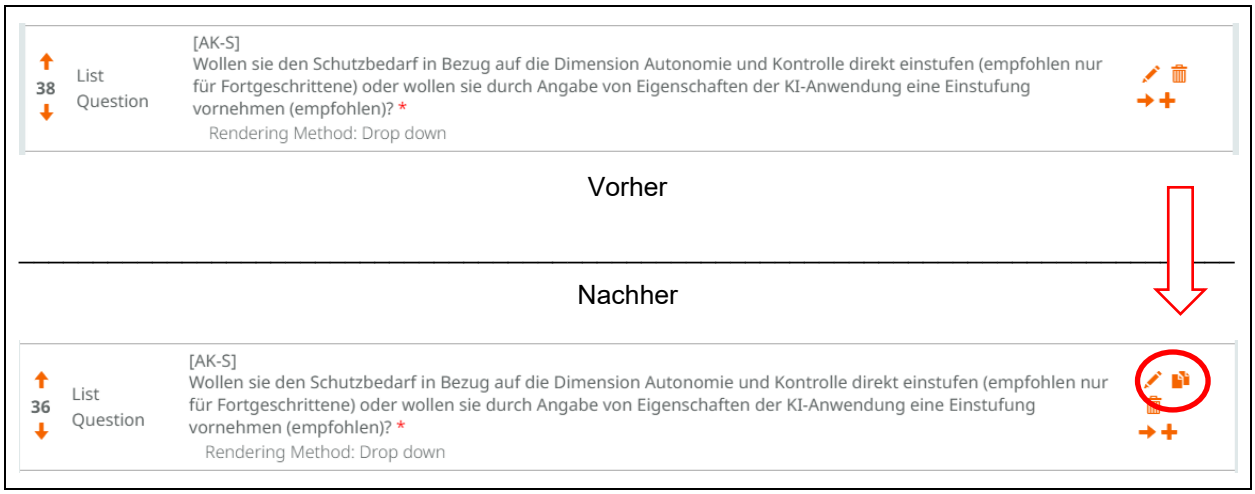


Abb. 36: Kopierfunktion im Wizard-Framework, Quelle: Eigene Darstellung.

Wird auf das Symbol geklickt, öffnet sich ein Fenster, in welchem angeben werden kann, an welcher Stelle die Kopie des jeweiligen Elements eingefügt werden soll. Hier ist es im ersten Schritt möglich, das Element vor (before) einem anderen bestehenden Element einzufügen. In weiterer Folge wäre es hilfreich, eine Kopie auch nach einem anderen Element einzufügen, z.B. wenn das Element ganz am Ende eingefügt werden soll.

In der aktuellen Version ist diese Positionierung jedoch folgendermaßen möglich:

1. Einfügen vor der gewünschten Position
2. Einzelpositionsverschiebung durch Klick auf den Pfeil links unter der Positionsnummer

4.2.7 Export und Import der Antworten

Bei jeder Frage, die man innerhalb des Wizards konfigurieren kann, gibt es die Möglichkeit, die jeweilige Frage als verpflichtend zu beantworten (required) zu kennzeichnen. Dies erfolgt durch setzen eines Hakens auf dem Kartenreiter Answer(s) – siehe Abb. 37.[120]

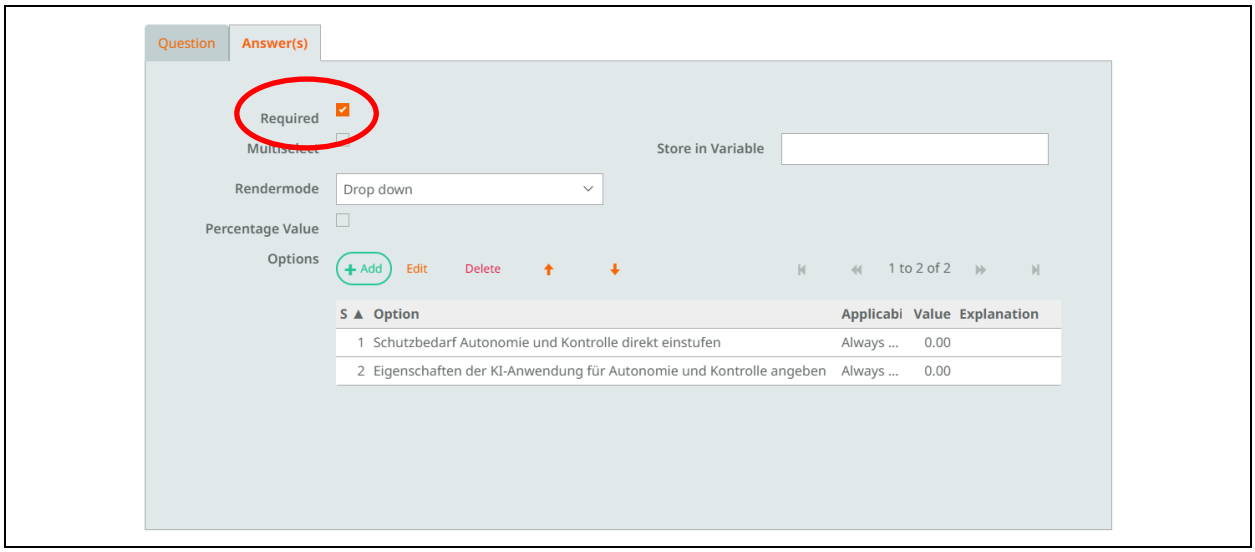


Abb. 37: Verpflichtende Antworten Konfiguration, Quelle: Eigene Darstellung.

Wird dieser Haken gesetzt, ist es nun notwendig, bei Ausführung des Wizards aus Anwender*innensicht, die jeweilige Frage zu beantworten. Andernfalls versagt der Wizard das Weiter-klicken auf die nächste Seite, so lange die entsprechende Frage noch nicht beantwortet ist. Dies wird dargestellt, indem unter der nicht beantworteten Frage der Hinweis in einem roten Kästchen *An answer is required.* angezeigt wird – siehe Abb. 38.

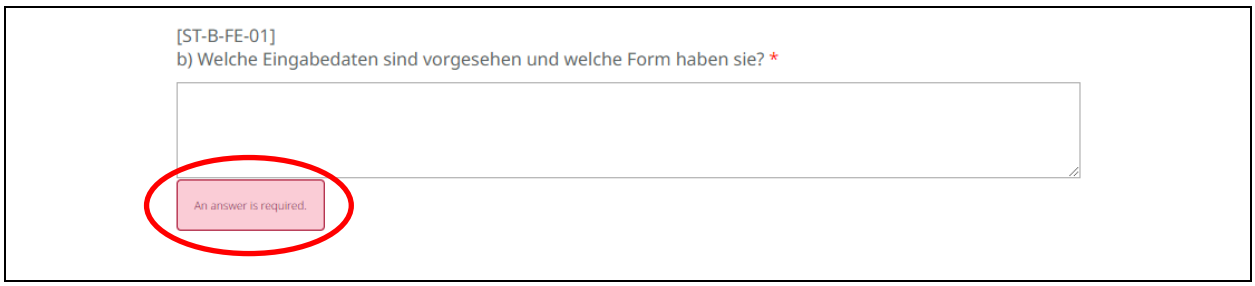


Abb. 38: Hinweis verpflichtende Antworten bedienseitig, Quelle: Eigene Darstellung.

Da gemäß KI-Prüfkatalog durch den/die Anwender*in sämtliche Antworten zu geben sind, bei denen die Schutzbedarfsanalyse ergeben hat, dass die zugehörige Dimension zu betrachten ist, wurden im Wizard sämtliche Antworten als *Required* gesetzt. In der Testphase wurde jedoch schnell deutlich, dass dadurch

[120] Vgl. Kapitel 2.5.2.3.

bei jedem Testlauf sämtliche Fragen zu beantworten sind, weil es andernfalls zu dem automatischen Hinweis gekommen ist, dass diese Fragen noch zu beantworten sind und ein Überspringen von Fragen somit nicht möglich war. Für jede neue Version des Wizards muss man erneut veröffentlichen (publish). Nach jeder neuen Version war es dann nötig, sämtliche Fragen erneut zu beantworten.

Eine Möglichkeit wäre es, die Konfiguration wieder zurückzusetzen und erst am Ende wieder auf *Required* zu setzen. Da dies konfigurationsseitig jedoch keine zufriedenstellende Lösung ist, muss im Wizard-Framework die Möglichkeit geschaffen werden, einmal beantwortete Fragen zu exportieren und bei einer neuen Version des Wizards wieder zu importieren.

Diese Funktion wurde bereits während der Ausarbeitung der vorliegenden Arbeit realisiert und derart verfügbar gemacht, dass auf der Bedienoberfläche im linken unteren Bereich zwei Schaltflächen vorgesehen wurden, die sowohl einen Export (Download), als auch einen Import (Upload) der Fragen erlauben – siehe Abb. 40.

Diese Funktion hat in weiterer Folge auch den Vorteil für den/die Anwender*in, dass bei begonnener Ausarbeitung für ein tatsächliches Projekt zu jeder Zeit unterbrochen werden kann, die bis dahin dokumentierten Antworten gesichert werden können und zu einem späteren Zeitpunkt wieder fortgesetzt werden können.

Während des Testens dieser Funktion wurde entdeckt, dass bei einem Import/Upload der Fragen die zuvor konfigurierten Bedingungen für die Sichtbarkeit der Fragen nicht übernommen wurden. Dadurch, dass sämtliche Bedingungen nicht mehr vorhanden waren, wurden daraufhin sämtliche Fragen angezeigt, unabhängig davon, ob die Bedingungen erfüllt waren oder nicht. Dieses Feedback wurde ebenfalls für die Entwicklung des Wizard-Frameworks zurückgemeldet und bearbeitet.

4.2.8 Scrollen Projektübersicht und Arbeitsfenster

Aus Anwender*innensicht ist der Scrollbereich in der bestehenden Bedienoberfläche immer abhängig von der linksseitigen Projektstruktur ist, d.h. der rechtsseitige Arbeitsbereich hat je nach Anzahl der zu beantwortenden Fragen einen mehr oder weniger großen Leerbereich unterhalb der Fragen, welcher bei den Anwender*innen zu unnötigem Scrollen nach unten führen kann – siehe Abb. 39.

Abb. 39: Leerbereich Anwender*innenoberfläche, Quelle: Eigene Darstellung.

Als Optimierung wären mehrere Möglichkeiten denkbar. Die verschiedenen Varianten sollen nachfolgend dargestellt werden.

4.2.8.1 Unabhängiges Scrollen

Eine Möglichkeit wäre, die beiden Bereiche links, d.h. den Projektbaum, und rechts, also das eigentliche Arbeitsfenster, voneinander zu entkoppeln und ein unabhängiges Scrollen je Bereich vorzusehen. Dabei sollte der Arbeitsbereich nur genau so groß sein, wie die jeweils in dem entsprechenden Kapitel enthaltenen Fragen Platz benötigen. Es weiteres Scrollen nach unten würde damit verhindert.

Diese Funktion wurde bereits während der Ausarbeitung der vorliegenden Arbeit realisiert und derart verfügbar gemacht, dass zwei unabhängige Scrollbereiche vorgesehen wurden – siehe Abb. 40.

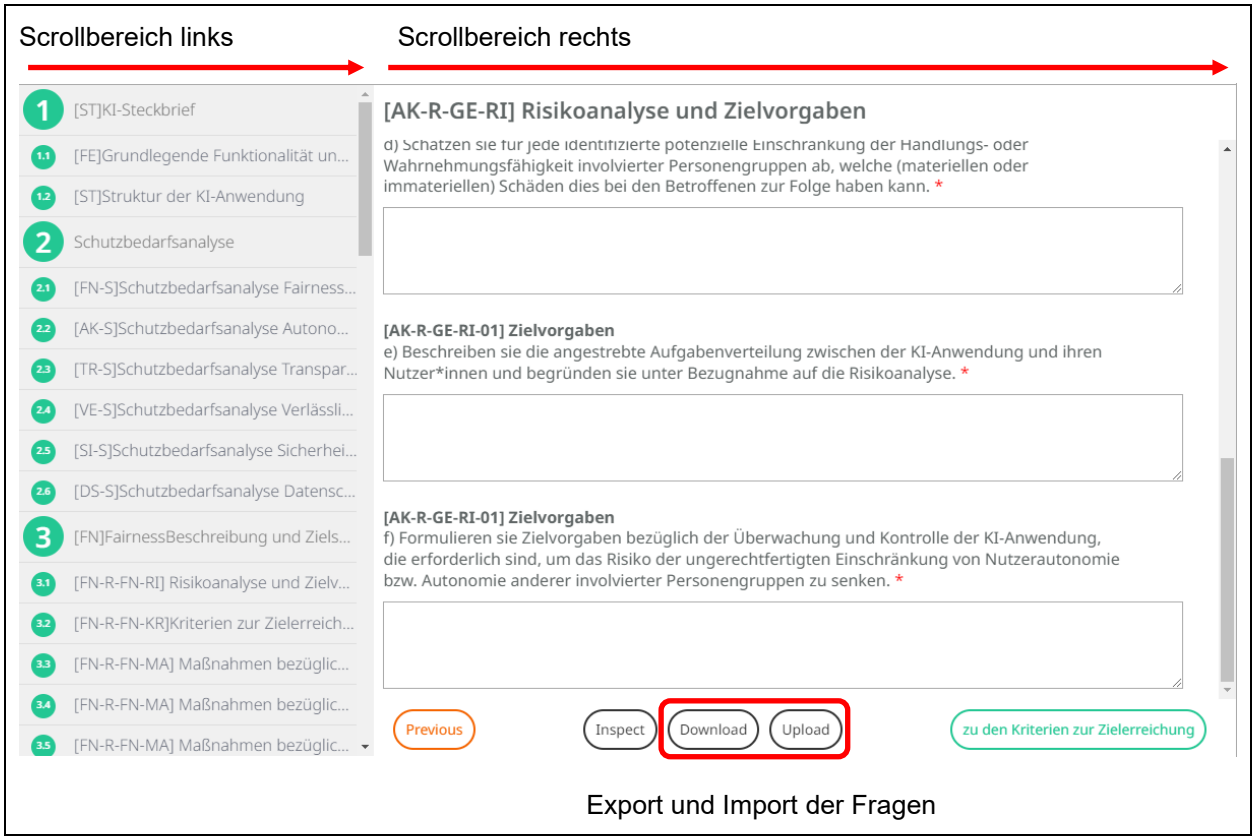


Abb. 40: Unabhängiges Scrollen umgesetzt, Quelle: Eigene Darstellung.

4.2.8.2 Projektstruktur einklappbar

Eine weitere Möglichkeit wäre, den linken Bereich, d.h. die Projektstruktur einklappbar zu gestalten, z.B. so, dass die nicht bearbeiteten Kapitel eingeklappt werden und dadurch in der linken Spalte keine unnötige Länge entsteht.

4.2.8.3 Bearbeitung anzeigen

Eine dritte denkbare Variante besteht darin, in der linken Spalte nur den jeweils bearbeiteten Unterpunkt anzuzeigen und den Rest auszublenden.

Generell, d.h. unabhängig von der zuvor genannten, gewählten Lösung, könnten für eine weitere Reduzierung der Projektstruktur auch direkt nach der Schutzbedarfsanalyse diejenigen Dimensionen ausgeblendet werden, bei denen der Schutzbedarf nicht hoch oder mittel ist. Nur für diese Dimensionen sind gemäß KI-Prüfkatalog auch entsprechende Maßnahmen zu setzen und zu dokumentieren. Bei den anderen reicht die Dokumentation der Schutzbedarfsanalyse, weil kein Schutzbedarf besteht und dies über die entsprechende Einstufung ausreichend dokumentiert ist.

4.2.9 Projektübersicht automatisch mitscrollen

Die zuvor bereits genannte, linksseitige Projektübersicht wird nach jedem Klick vor oder zurück immer wieder nach oben gesetzt, d.h. man sieht im Arbeitsfenster die neuen Fragen, aber die Projektübersicht steht wieder ganz am Anfang bei dem KI-Steckbrief. Durch die grüne Markierung wird zwar klar, dass dieser Punkt schon bearbeitet wurde, aber in der Projektübersicht ist der aktuelle Bearbeitungsschritt nicht sichtbar. Dazu muss auf jeder neuen Seite manuell hinuntergescrollt werden. Hilfreich wäre es, je nach Bearbeitungsfortschritt, immer an der gerade zu bearbeitenden Stelle in der Projektübersicht zu stehen.

4.2.10 Fenstergröße / Überschriften Nutzer*innenansicht

Die Fenstergröße in der Nutzer*innenansicht ist fest definiert, d.h. nicht einstellbar. Dies führt zu mehreren Darstellungsproblemen. So wird bei kleineren Bildschirmen der untere Rand der Eingabeoberfläche abgeschnitten – siehe auch Abb. 18. Bei größeren Bildschirmen hingegen wird im unteren Bereich des Fensters ein leerer Bereich angezeigt. Außerdem ist die Breite der linksseitigen Projektstruktur fest definiert, was zu schlechter Lesbarkeit der Überschriften der Subsections führt, da bei längeren Wörtern nur der Anfang lesbar ist. Des Weiteren wird der informative Textteil, der bei den Start-Elementen als Normal-Text konfiguriert wurde, direkt an die Überschrift angehängt und nicht als nachfolgender Fließtext dargestellt – siehe Abb. 41.

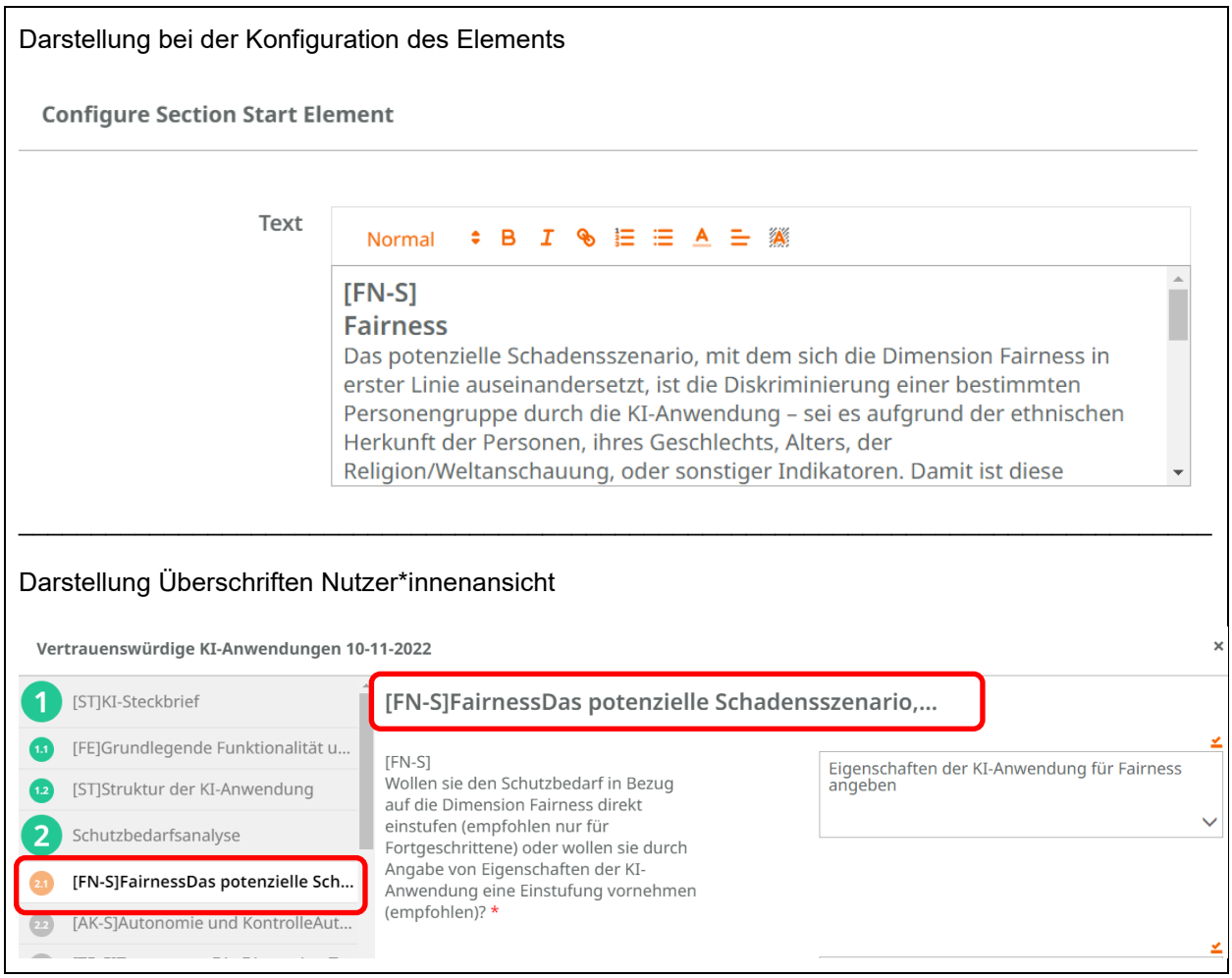


Abb. 41: Darstellung Überschriften Nutzer*innenansicht, Quelle: Eigene Darstellung.

Eine erste Lösung dieses Problems wäre, ein eigenes Element für den informativen Fließtext zu erstellen. Dies wiederum führt zu einer Erhöhung des Dateivolumens des Wizards. Daher wäre eine Anpassung im Wizard-Framework zu begrüßen.

4.2.11 Assessment Elements neu positionieren

Die Assessment Elements werden der Reihe nach untereinander angelegt. Soll diese Positionierung im Wizard verändern, besteht die Möglichkeit mit den linksseitig angeordneten Pfeilen je ein Assessment Element um eine Position nach oben oder nach unten zu verschieben.

In der Version des Wizard-Frameworks, welche zu Beginn der Wizard-Erstellung verwendet wurde, ist diese Funktion nicht durchgängig möglich gewesen. Über das integrierte Feedbacksystem wurde dieser Sachverhalt zurückgemeldet, nach kurzer Zeit im Wizard-Framework optimiert und über eine Aktualisierung des Softwarestands in der Konfigurationsumgebung verfügbar gemacht – siehe Abb. 42.

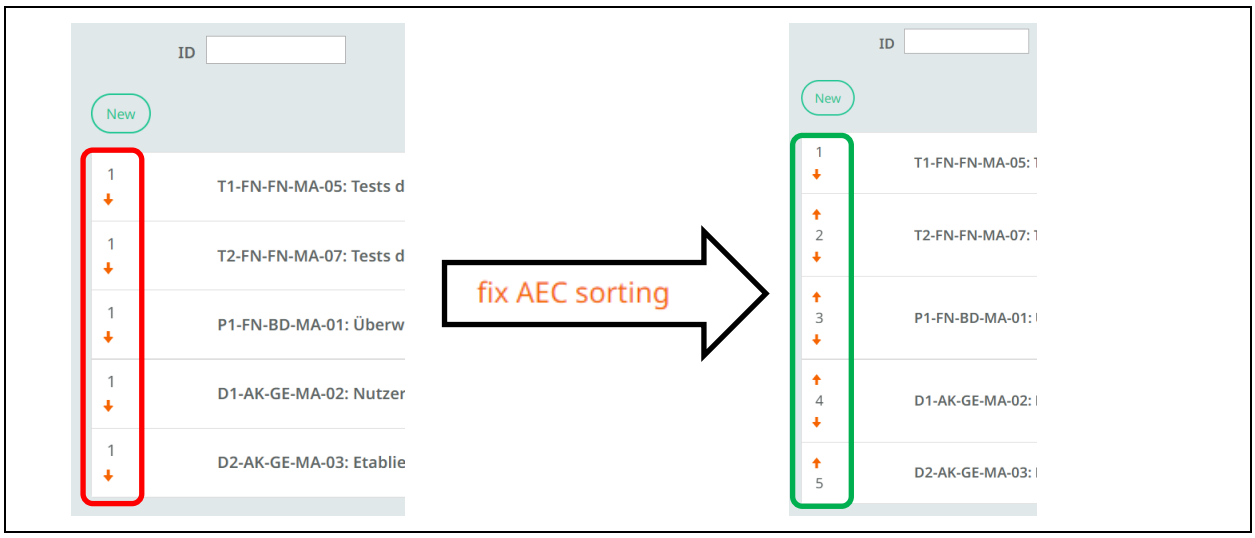


Abb. 42: Optimierung Sortierfunktion Assessment Elements, Quelle: Eigene Darstellung.

4.2.12 Assessment Elements – Sources

Da für die vorliegende Arbeit und für die Erstellung des Wizards der Leitfaden des Fraunhofer IAIS als wesentliche Quelle gedient hat, wurde bei sämtlichen Assessment Elements die Website hinterlegt, von der man den Leitfaden herunterladen kann – siehe Abb. 43.

Abb. 43: Website des Fraunhofer IAIS als Source, Quelle: Eigene Darstellung.

Nach dem Anlegen der *Sources* wurde eine neue Version des Wizards mittels *Publish* erstellt. Durch einen anschließenden Export des Wizard in ein JSON-file wurde aufgedeckt, dass diese Änderungen einen deutlichen Einfluss auf die Dateigröße hatten. Lag die Dateigröße vor dem Anlegen der *Sources* bei den *Assessment Elements* noch bei ca. 6,5 MB, ist die Größe bei gerade einmal 15 *Sources* auf ca. 74 MB angewachsen.

Nach einem Feedback wurde das Framework kurzfristig angepasst und ein nächster Export in ein JSON-file wurde zunächst schon auf ca. 28 MB reduziert. Weitere mögliche dahingehende Optimierung ist auch nach dieser Anpassung ratsam.

4.2.13 Sichtbarkeit - Visibility

Im Rahmen der vorliegenden Arbeit sind einige Funktionen in Bezug auf die Sichtbarkeit der Fragen bzw. Assessment Elements erkannt worden, die für eine einfachere Bearbeitung notwendig wären. Diese Funktionen sollen in den folgenden Unterkapiteln aufgezählt werden.

4.2.13.1 Next button

Die Schaltflächen (Next buttons) am Ende jeder Section, die zum Weiter-klicken dienen, werden selbst dann angezeigt, wenn die gesamte Dimension nicht anwendbar ist und ausgeblendet wird. So muss man sich dennoch durch die einzelnen Unterkapitel (Sub sections) klicken, obwohl darin keine Fragen angezeigt werden. Die Next buttons werden angezeigt, obwohl sie bei der Sichtbarkeit als abhängig von der Sichtbarkeit der Section (Visibility depends on Section Visibility) konfiguriert wurden. Daher ist darauf zu schließen, dass innerhalb des Frameworks Anpassungen zu treffen sind.

4.2.13.2 Variablenwert anzeigen

Die Sichtbarkeit der Dimensionen der Vertrauenswürdigkeit hängt von der Schutzbedarfsanalyse ab. Für einen einfachen Zugang wurde die Einstufung im Wizard in zwei Varianten realisiert. Bei der zweiten Möglichkeit, der Einstufung mittels Eigenschaften der KI-Anwendung, gibt es eine große Anzahl an möglichen Antworten, die zu einer Einstufung eines hohen oder mittleren Schutzbedarfs führen. Jede dieser Antworten muss als Bedingung für die Sichtbarkeit der entsprechenden Dimension angegeben werden. Um während der Auswahl die richtigen Antworten leicht zu erkennen, wurden die vorhandenen Variablen dafür genutzt, einen eindeutigen Wert hineinzuschreiben. Dabei wurden sämtliche Antworten, die zu einer hohen Einstufung des Schutzbedarfs führen, mit dem Wert 100 versehen. Die Antworten, die zu einer mittleren Einstufung führen, wurden mit einem Wert von 50 versehen. Um diese Variablenwerte bei der Auswahl als Hilfsmittel zu sehen, wurde bereits während der Ausarbeitung ein diesbezügliches Feedback gegeben.

Diese Funktion wurde daraufhin bereits während der Ausarbeitung der vorliegenden Arbeit realisiert und derart verfügbar gemacht, dass direkt neben der Antwort der Variablenwert angezeigt wird – siehe Abb. 44.

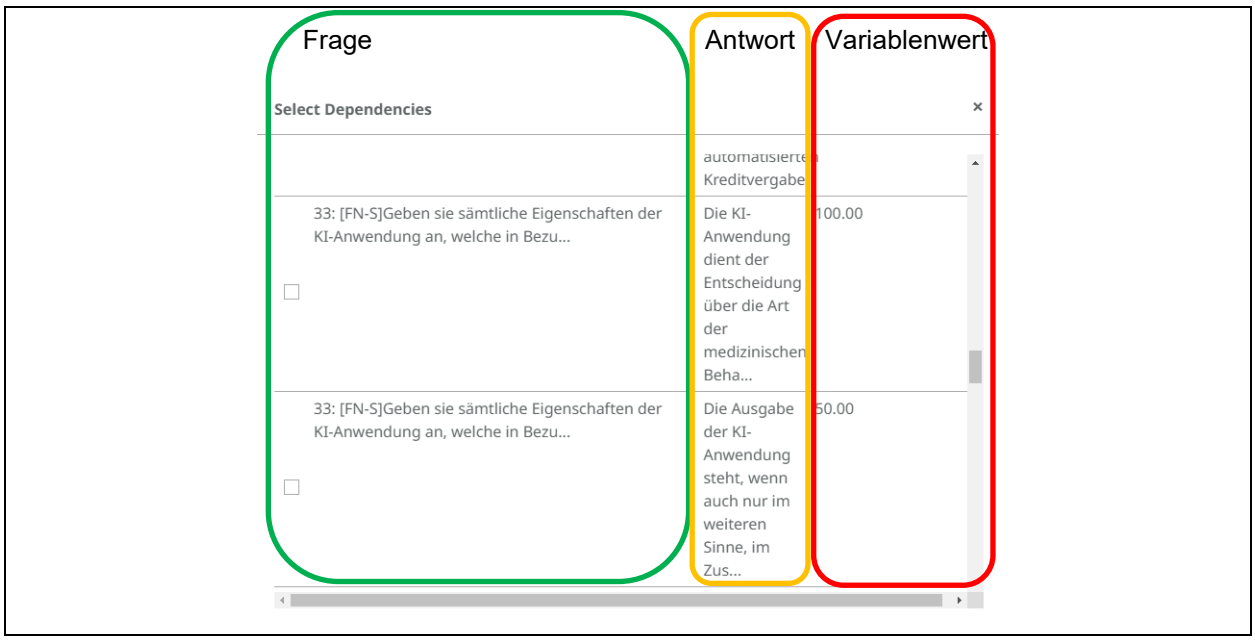


Abb. 44: Variablenwert in Auswahltabelle der Abhängigkeiten, Quelle: Eigene Darstellung.

So ist es leicht möglich, zunächst über die vorhandene Textsuche (Textsearch) die gewünschte Frage vorzuselektieren und anschließend über den Variablenwert sämtliche anwendbaren Fragen auszuwählen.[121]

4.2.13.3 Section Visibility plus Conditions

Zu Beginn der Arbeit waren als Auswahlmöglichkeiten der Sichtbarkeit lediglich *Always visible, Visibility depends on previous answer(s)* und *Visibility depends on Section Visibility* verfügbar. Wenn innerhalb einer Section jedoch einzelne Fragen bereits von einer Antwort aus einer vorherigen Section abhängig sind, war es nicht möglich, zusätzlich die Section Visibility als Bedingung anzugeben. Als Resultat gab es mögliche Fehlerzustände in Bezug auf die Sichtbarkeit – siehe Tab. 10.

Visibility depends on	Mögliche Fehlerquelle	Fehlerhafte Sichtbarkeit
previous answer(s)	Vorherige Antwort positiv, aber Section Visibility negativ	Aufgrund der Auswahl previous answer(s) wird die Frage angezeigt, obwohl die Section ausgeblendet wird.
Section Visibility	Section Visibility positiv, aber vorherige Antwort negativ	Aufgrund der Auswahl Section Visibility wird die Frage angezeigt, obwohl die vorherige Frage negativ beantwortet wurde.

Tab. 10: Fehler bei Visibility, Quelle: Eigene Darstellung.

[121] Vgl. Kapitel 2.6.2.5.

Aufgrund des erfolgten Feedbacks während der Ausarbeitung wurde als weitere Auswahlmöglichkeit für die Sichtbarkeit die *Visibility depends on Section Visibility plus Conditions* vorgesehen. Da diese Option noch nicht ganz durchgängig funktioniert hat, wurde die Konfiguration der Fragen, auf die solche Bedingungen zutreffen, folgendermaßen angepasst. Es wurde sowohl die Bedingung der zuvor notwendigen Einzelantwort (siehe Abb. 45 - C), als auch die Anforderungen für die Sichtbarkeit der Section als Bedingung konfiguriert. Dabei wurde die Bedingung für die Einzelantwort mit sämtlichen Bedingungen der Section (siehe Abb. 45 - D) jeweils UND-verknüpft (siehe Abb. 45 - A). Die einzelnen UND-Verknüpfungen wiederum wurden mittels ODER-Verknüpfung (siehe Abb. 45 - B) als Bedingungen aufgelistet. Das folgende Beispiel stammt aus [TR-R-AF-MA-03] Verfügbarkeit von Lerndaten aus dem Betrieb.

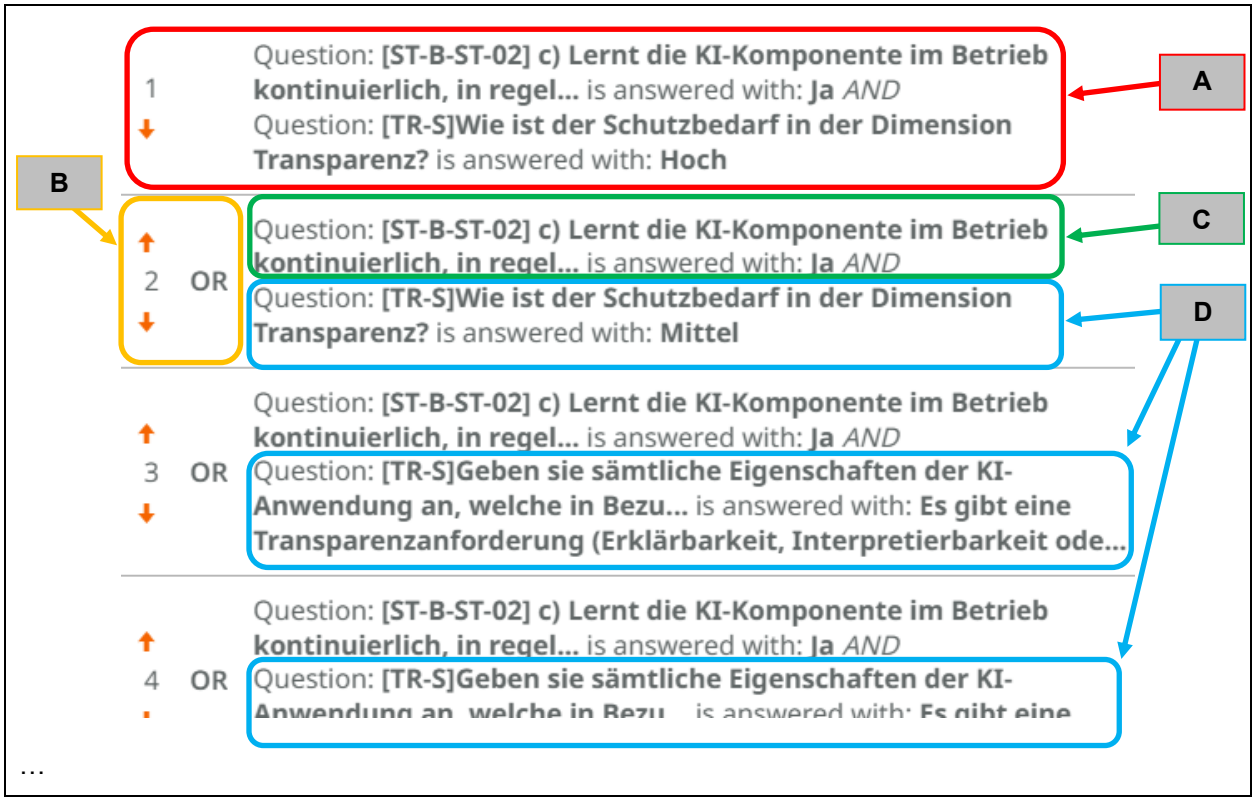


Abb. 45: Hilfskonfiguration Section Visibility plus Conditions, Quelle: Eigene Darstellung.

4.2.14 Report Layout

Nach der Eingabe der Antworten im Wizard kann ein Report generiert werden. Dieser Report, der als PDF-Datei ausgegeben wird, ist jedoch speziell bei längeren Antworten nicht sehr übersichtlich. Dies hat mehrere Gründe.

4.2.14.1 Spalte Section / Nummerierung

In der linken Spalte wird durchgängig die Bezeichnung *Section* und darunter eine Nummerierung der Fragen angegeben – siehe Abb. 46 - A. Beides scheint für den Report an dieser Stelle entbehrlich. Da die Fragen mit den Kürzeln aus dem Leitfaden versehen wurden, ist dadurch schon deutlich, an welcher Stelle man sich befindet. Des Weiteren nimmt diese Spalte dabei einen erheblichen Teil der Seitenbreite in Anspruch. Daher werden die Fragen und Antworten, die den inhaltlich wichtigen Teil darstellen, auf sehr engem Raum dargestellt.

Für eine übersichtlichere Darstellung wäre es hilfreich, diese Spalte ersatzlos zu streichen und den Platz für die Fragen und Antworten zu verwenden.

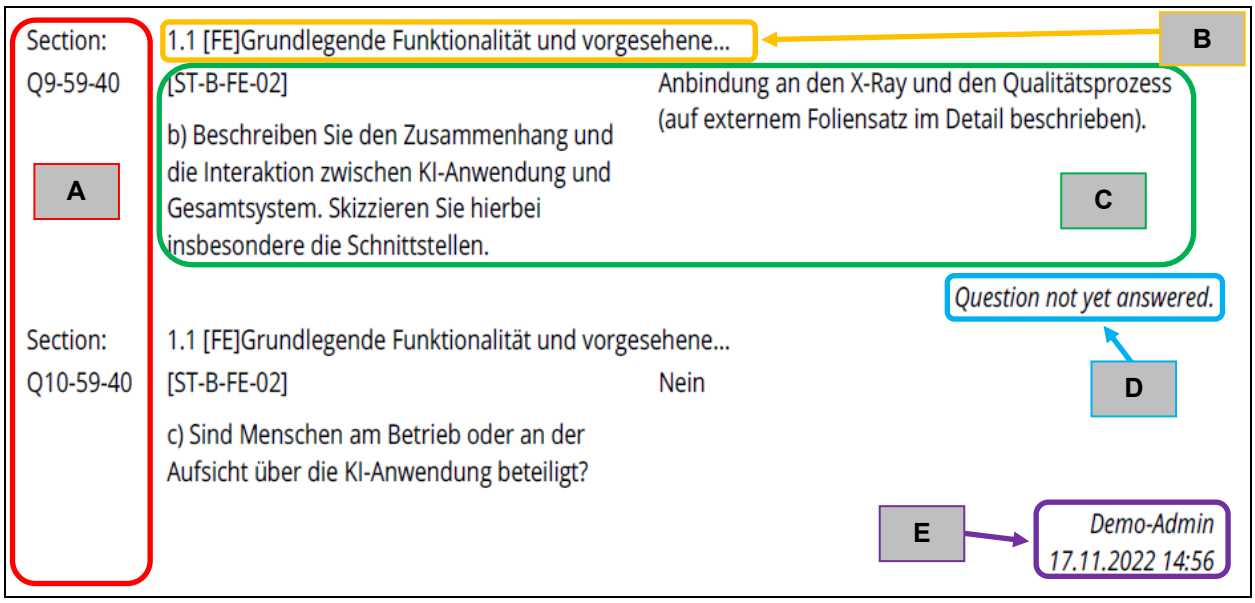


Abb. 46: Layout des Reports, Quelle: Eigene Darstellung.

4.2.14.2 Überschriften

Die Section-Überschriften selbst werden nur abgekürzt dargestellt. Bei längeren Texten wird die jeweilige Überschrift gekürzt und mit drei Punkten […] beendet – siehe Abb. 46 - B. Daher ist teilweise nicht klar, was in der entsprechenden Section behandelt wird.

Für eine übersichtlichere Darstellung wäre es hilfreich, wenn die gesamten Überschriften ausgeschrieben dargestellt würden.

4.2.14.3 Fragen und Antworten nebeneinander

Die Fragen und Antworten sind im gesamten Report immer nebeneinander dargestellt – siehe Abb. 46 - C. Daher passen in jede Zeile nur wenige Wörter. Das Lesen dieser ist daher nicht sehr komfortabel.

Für eine übersichtlichere Darstellung wäre es hilfreich, die Fragen über die gesamte Seitenbreite auszuführen und die Antworten in gleicher Weise direkt darunter darzustellen.

4.2.14.4 Question not yet answered

Nach vielen Antworten wird unten rechts die Angabe *Question not yet answered* dargestellt – siehe Abb. 46 - D. Bei einigen Antworten findet sich hier die Angabe derjenigen Person, welche im System angemeldet war, d.h. diejenige Person, welche die Antwort eingegeben hat – siehe Abb. 46 - E. *Question not yet answered* wird auch bei beantworteten Fragen ausgegeben. Dies lässt darauf schließen, dass an dieser Stelle noch Anpassungen bei der Generierung des Reports zu treffen sind.

5 AUSBLICK

Im Rahmen des Projekts wurden viele Erkenntnisse gesammelt, die zu großen Teilen in der Optimierung des Wizard-Frameworks lagen. Aufgrund der direkten Feedback-Funktionen und der parallelen Bearbeitung des Frameworks durch Siemens wurden diverse Funktionen bereits umgesetzt.[122]

War zu Projektbeginn noch nicht klar, wie umfangreich die Anforderungen des Leitfadens sein würden, wurde während der Umsetzung deutlich, dass aufgrund der entstandenen Wizard-Größe noch Anpassungen in Bezug auf die Performance zu treffen sind. Somit sollten zukünftig Ladezeiten z.B. für den Ex- und Import des Wizards oder auch der bereits beantworteten Fragen (Upload und Download)[123] verkürzt werden.

Der inhaltliche Umfang des Leitfadens führte zudem dazu, dass die Umsetzung im Wizard bis inklusive zum ersten Risikogebiet der Dimension Verlässlichkeit realisiert wurde, da zu Beginn des Projekts auch nicht klar war, wie zeitintensiv das Einpflegen der Fragen in das Tool werden würde. Um den gesamten Leitfaden abzubilden, sind die restlichen verbleibenden Risikogebiete zu integrieren. Dies kann, dem vorliegenden Konzept folgend, in gleicher Weise geschehen. Ist dies umgesetzt, kann der Wizard allen Interessierten verfügbar gemacht werden, um die Anwendung weiter zu verbreiten. Eine solche Veröffentlichung würde in weiterer Folge helfen, um ein Feedback von vielen Nutzer*innen in Bezug auf die Anwendung zu bekommen und dahingehende Optimierungen vorzunehmen. Ggf. sollten dafür eigene zusätzliche Feedback-Fragen erstellt werden, um die Anzahl der Eingaben zu erhöhen.

Die graphische Darstellung des Wizards bzw. einige der darin befindlichen Anforderungen könnten in Zukunft ausgebaut werden. So wäre es sinnvoll, z.B. die Beschreibung des Gesamtsystems oder die Auflistung der wichtigsten Komponenten im KI-Steckbrief[124] graphisch aufzubereiten, um den Aufbau und die Struktur darzustellen. Dafür sollten zunächst Regeln festgelegt werden, um eine einheitliche Darstellung in allen Anwendungsfällen zu haben und den Aufbau für Dritte nachvollziehbar zu gestalten. Denkbar wäre eine Darstellung wie in einem sicherheitsgerichteten Blockdiagramm.

Weitere graphische Ergänzungen wären Symbole bzw. Icons bei sämtlichen Elementen, um unabhängig von den Kürzeln anzuzeigen, bei welchem Schritt der Bewertung sich die Nutzer*innen gerade befinden. So könnten Symbole ergänzend zu den Dimensions-Icons für jedes einzelne Risikogebiet erstellt und eingebunden werden.

Aufgrund des derzeitigen Entwicklungs- und Erfahrungsstadiums im Bereich der KI im Allgemeinen und der vertrauenswürdigen KI im Speziellen ist davon auszugehen, dass es zukünftig diesbezügliche Entwicklungen von normativen Anforderungen geben wird. Dies können Normen für bestimmte Risikogebiete, aber auch für bestimmte Technologien, die für KI-Anwendungen genutzt werden, sein. Diese

[122] Vgl. Kapitel 4.2.

[123] Vgl. Kapitel 4.2.7.

[124] Vgl. Poretschkin u.a. (2021), S. 36.

normativen Entwicklungen sind zu beobachten und bei Anwendbarkeit bestimmter Anforderungen sollten diese in den Wizard integriert werden. Sind es sehr umfangreiche Anforderungen, könnten diese auch in einem eigenen Wizard abgebildet werden. Wird jedoch das Wizard-Framework genutzt, können diese Anforderungen jedoch z.B. über die Assessment Elements auf einer Bedienoberfläche angezeigt und parallel in einem Projekt von den Nutzer*innen bearbeitet werden. Dieser Teil der Weiterentwicklung wird voraussichtlich die meisten Ressourcen benötigen, da eine permanente Weiterentwicklung der Technologien auch neue Möglichkeiten in der Umsetzung von KI-Anwendungen und deren Anwendungsbereichen ermöglichen. Damit einhergehend wird es auch immer den politischen Willen bzw. die Notwendigkeit von Regulierung dieser Technologien geben.

LITERATURVERZEICHNIS

Gedruckte Werke

Austrian Standards Institute/Österreichisches Normungsinstitut (ON) (2010): *Sicherheit von Maschinen — Allgemeine Gestaltungsleitsätze — Risikobeurteilung und Risikominderung (ISO 12100:2010)*, Ausgabe: 2011-03-15

Russell, Stuart J.; Norvig, Peter (2010): *Artificial Intelligence, A Modern Approach*, 3. Auflage, Pearson Education Inc., New Jersey

Wittpahl, Volker (Hrsg.) (2019): *iit-Themenband: Künstliche Intelligenz, Technologie, Anwendung, Gesellschaft*, 1. Auflage, Springer Vieweg, Berlin

Online-Quellen

Beck, Susanne et al.; Lernende Systeme – Die Plattform für Künstliche Intelligenz (Hrsg.) (2019): *Künstliche Intelligenz und Diskriminierung Herausforderungen und Lösungsansätze*, https://www.plattform-lernende-systeme.de/publikationen.html [Stand: 16.09.2022]

Bundesministerium Finanzen (2022): *BMF/Spezialeinheit PACC: 2021 rund 6 Mio. Arbeitnehmerveranlagungen und 1,4 Mio. Anträge auf COVID-19 Hilfszahlungen überprüft*, https://www.bmf.gv.at/presse/pressemeldungen/2022/September/pacc-bilanz.html [Stand: 13.09.2022]

Bundesamt für Sicherheit in der Informationstechnik (2021): *Sicherer, robuster und nachvollziehbarer Einsatz von KI,* https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf?__blob=publicationFile&v=5 [Stand: 12.08.2022]

Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie (BMK) (2021): *Strategie der Bundesregierung für Künstliche Intelligenz, Artificial Intelligence Mission Austria 2030(AIM AT 2030),* https://www.bmk.gv.at/themen/innovation/publikationen/ikt/ai/strategie-bundesregierung.html [Stand: 27.08.2022]

Cremers, Armin B.; Englander, Alex; Gabriel, Markus; Hecker, Dirk; Mock, Michael; Poretschkin, Maximilian; Rosenzweig, Julia; Rostalski, Frauke; Sicking, Joachim; Volmer; Julia; Voosholz, Jan; Voss, Angelika; Wrobel, Stefan (2019): *Vertrauenswürdiger Einsatz von Künstlicher Intelligenz,* Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS, https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf [Stand: 27.08.2022]

Deutsche Bundesregierung (2018): *Strategie Künstliche Intelligenz der Bundesregierung,* https://www.ki-strategie-deutschland.de/home.html [Stand: 27.08.2022]

Europäische Kommission (2016): *Verordnung (EU) 2016/679 des europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-*

Grundverordnung), https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679&from=DE [Stand: 27.08.2022]

Europäische Kommission (2018): *Mitteilung der Kommission an das europäische Parlament, den europäischen Rat, den Rat, den europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen Künstliche Intelligenz für Europa {SWD(2018) 137 final},* https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52018DC0237 [Stand: 12.08.2022]

Europäische Kommission (2021a): *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 {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final},* https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52021PC0206 [Stand: 12.08.2022]

Europäische Kommission (2021b): *Anhänge des Vorschlags 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 {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final},* https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:52021PC0206 [Stand: 12.08.2022]

Europäische Kommission (2021c): *Vorschlag für eine Verordnung des europäischen Parlaments und des Rates über Maschinenprodukte,* https://eur-lex.europa.eu/legal-content/DE/TXT/DOC/?uri=CELEX:52021PC0202&from=EN [Stand: 13.09.2022]

Europäische Kommission (2021d): *Anhänge des Vorschlags für eine Verordnung des europäischen Parlaments und des Rates über Maschinenprodukte,* https://eur-lex.europa.eu/legal-content/DE/TXT/DOC/?uri=CELEX:52021PC0202&from=EN [Stand: 13.09.2022]

Gaylor, Brett (2020): *The Internet of Everything: How Smart Homes Can Be a Threat to Our Lives | ENDEVR Documentary,* https://www.youtube.com/watch?v=PYEAVC5Ifes [Stand: 22.08.2022]

Groth, Olaf J.; Nitzberg, Mark; Zehr, Dan; Straube, Tobias; Kaatz-Dubberke, Toni (2018): *Vergleich nationaler Strategien zur Förderung von Künstlicher Intelligenz,* Konrad-Adenauer-Stiftung e.V., Sankt Augustin/Berlin, https://www.kas.de/documents/252038/3346186/Vergleich+nationaler+Strategien+zur+F%C3%B6rderung+von+K%C3%BCnstlicher+Intelligenz.pdf/46c08ac2-8a19-9029-6e6e-c5a43e751556?version=1.0&t=1542129691776 [Stand: 27.08.2022]

High-Level Expert Group on AI (2019): *Ethics Guidelines on trustworthy AI,* Veröffentlicht von der Europäischen Kommission, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Stand: 26.08.2022]

High-Level Expert Group on AI (2020): *The Assessment List for Trustworthy Artificial Intelligence (ALTAI),* Veröffentlicht von der Europäischen Kommission, https://digital-

strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment [Stand: 26.08.2022]

High-Level Expert Group on AI (2022): *High-Level Expert Group on AI,* Veröffentlicht von der Europäischen Kommission, https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai [Stand: 26.08.2022]

Proyas, Alex (Regie); Vintar, Jeff (Drehbuch), Goldsman, Akiva (Drehbuch) (2004): *I, Robot,* https://www.imdb.com/title/tt0343818/?ref_=fn_al_tt_1 [Stand: 22.08.2022]

Lackes, Richard; Siepermann, Markus (2018): *Gabler Wirtschaftslexiko*n, https://wirtschaftslexikon.gabler.de/definition/kuenstliche-intelligenz-ki-40285/version-263673 [Stand: 21.08.2022]

Maier, Günter W. (2018): *Gabler Wirtschaftslexiko*n, https://wirtschaftslexikon.gabler.de/definition/intelligenz-37696/version-261129 [Stand: 21.08.2022]

Poretschkin, Maximilian; Schmitz, Anna; Akila, Maram; Adilova, Linara; Becker, Daniel; Cremers, Armin B.; Hecker, Dirk; Houben, Sebastian; Mock, Michael; Rosenzweig, Julia; Sicking, Joachim; Schulz, Elena; Voss, Angelika; Wrobel, Stefan (2021): *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz* | *KI-Prüfkatalog,* Sankt Augustin: Fraunhofer-Institut für intelligente Analyse- und Informationssysteme IAIS, https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html [Stand: 12.08.2022]

The MathWorks Inc. (2021): *Introducing Deep Learning with MATLAB*, https://de.mathworks.com/content/dam/mathworks/ebook/gated/80879v00_Deep_Learning_ebook.pdf [Stand: 13.09.2022]

The MathWorks Inc. (2020): *Reinforcement Learning with MATLAB*, https://de.mathworks.com/content/dam/mathworks/ebook/gated/reinforcement-learning-ebook-all-chapters.pdf [Stand: 09.09.2022]

U.S. Government (2001): *Uniting and strengthening America by appropriate tools required to intercept and obstruct terrorism (USA PATRIOT ACT) act of 2001,* https://www.congress.gov/107/plaws/publ56/PLAW-107publ56.pdf [Stand: 26.08.2022]

U.S. Government (2018): *Clarifying Lawful Overseas Use of Data Act / CLOUD Act,* https://www.congress.gov/115/bills/hr4943/BILLS-115hr4943ih.pdf [Stand: 26.08.2022]

Wahlster, Wolfgang; Winterhalter, Christian (Hrsg.) (2020): *Deutsche Normungsroadmap Künstliche Intelligenz*, Deutsches Institut für Normung e. V. und Deutsche Kommission Elektrotechnik, https://www.dke.de/de/arbeitsfelder/core-safety/normungsroadmap-ki [Stand: 12.08.2022]

Zhang, Daniel; Maslej, Nestor; Brynjolfsson, Erik; Etchemendy, John; Lyons, Terah; Manyika, James; Ngo, Helen; Niebles, Juan Carlos; Sellitto, Michael; Sakhaee, Ellie; Shoham, Yoav; Clark, Jack; Perrault, Raymond (2022): *The AI Index 2022 Annual Report,* AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University,

https://de.statista.com/statistik/daten/studie/1321674/umfrage/private-investitionen-in-ki-nach-laendern/ [Stand: 27.08.2022]

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

AF	Auditfähigkeit (Risikogebiet von TR); Abfangen von Fehlern auf Modellebene (Risikogebiet von VE)
AI	Artificial Intelligence
AK	Autonomie und Kontrolle (Dimension)
BD	Beherrschung der Dynamik (Risikogebiet)
DS	Datenschutz (Dimension)
EX	Transparenz für Expert*innen (Risikogebiet)
FE	Grundlegende Funktionalität und vorgesehener Einsatzkontext
FN	Fairness (Dimension und Risikogebiet)
FS	Funktionale Sicherheit (Risikogebiet)
GE	Angemessene und verantwortungsvolle Gestaltung der Aufgabenverteilung zwischen Menschen und KI-Anwendung (Risikogebiet)
GI	Schutz geschäftsrelevanter Information (Risikogebiet)
HLEG	High Level Expert Group on AI
IB	Sicherstellung der Informiertheit und Befähigung von Nutzer*innen und Betroffenen (Risikogebiet)
IV	Integrität und Verfügbarkeit (Risikogebiet)
KI	Künstliche Intelligenz
KMU	Kleine und mittlere Unternehmen
ML	Machine Learning (maschinelles Lernen)
NB	Transparenz gegenüber Nutzer*innen und Betroffenen (Risikogebiet)
PD	Schutz personenbezogener Daten (Risikogebiet)
RE	Verlässlichkeit im Regelfall (Risikogebiet)
RO	Robustheit (Risikogebiet)
SI	Sicherheit (Dimension)
ST	Struktur der KI-Anwendung
TR	Transparenz (Dimension)
UN	Einschätzung von Unsicherheit (Risikogebiet)
VE	Verlässlichkeit (Dimension)

ANHANG 1: WIZARD IM JSON-FORMAT

ANHANG 2: REPORT USE CASE