ExamAI

KI-Audit in der Arbeitswelt

Ein integratives Framework zum Auditieren und Testen von KI-Systemen

Dr. Bernhard Waltl, Nikolas Becker

KI-Audit in der Arbeitswelt

Ein integratives Framework zum Auditieren und Testen von KI-Systemen

Dr. Bernhard Waltl, Nikolas Becker

INHALTSVERZEICHNIS

1.
Executive Summary

Mit den zunehmenden Erfolgen, die sich durch den Einsatz von KI-Systemen zeigen, werden auch Risiken und Herausforderungen sichtbar. KI-Systeme haben das Potential, bestehende Geschäftsprozesse zu verbessern oder umzugestalten. Für viele Bereiche zählt jedoch nicht nur die Geschwindigkeit und Präzision eines KI-Systems. In sensiblen Bereichen, wie Personal- und Talentmanagement, zählen auch Zuverlässigkeit, Erklärbarkeit, oder Transparenz. In der Vergangenheit standen diese Aspekte bei der Forschung und Entwicklung von KI-Systemen selten im Fokus. Die letzten Jahre haben deutlich gezeigt, dass ein unreflektierter Einsatz von KI-Systemen zu Problemen führt. Beispiele hierfür sind Diskriminierung durch KI-Systeme, fehlerhafte Entscheidungen eines KI-Systems, die nicht nachvollzogen werden können, oder mangelnde Widerspruchmöglichkeiten von betroffenen Personen.

Ein Grundproblem ist die unzureichende Verankerung von Anforderungen an KI-Systeme, welche die Zuverlässigkeit, Erklärbarkeit oder Transparenz eines KI-Systems erhöhen. Während sich die Forschung und Entwicklung der letzten Jahre stark auf Leistungsfähigkeit und Performance von KI-Systemen konzentrierte, wurden die Bereiche Zuverlässigkeit und Transparenz vernachlässigt. In der jüngsten Vergangenheit entstanden jedoch Anforderungen, die diese Lücke schließen können. Beispielsweise sei hier auf die sieben Kernanforderungen an Trustworthy AI der High-Level Expert Group AI (HLEG AI) verwiesen. Bis zum heutigen Zeitpunkt ist jedoch unklar, wie sich diese Anforderungen auf den Einsatz und die Entwicklung von KI-Systemen auswirken. Unklar ist auch, wie diese Anforderungen in der Praxis überprüft und deren Erfüllen sichergestellt werden kann.

Diese Arbeit beschreibt ein Framework zur ganzheitlichen Darstellung von KI-Systemen als sozio-technische Systeme. Das Framework könnte die Grundlage für die Regulierung von KI-Systemen darstellen und umfasst fünf Ebenen, die aufeinander aufbauen bzw. ineinandergreifen:

1. Einsatzszenario und Anwendungsfall
2. Lebenszyklus von software-intensiven KI-Systemen
3. ADM Systeme

4. Entscheidungsstrukturen
5. ADM Entscheidungen

Das Framework stellt einen Rahmen zur Anwendung von etablierten Qualitätssicherungsmethoden aus dem Bereich des Systems- und Softwareengineerings dar: Auditierung und Testen. Insbesondere das Audit wird ein wichtiger Baustein zur Analyse von KI-Systemen, wobei die Berücksichtigung von charakteristischen Eigenschaften des betrachteten Systems essentiell ist. Im Rahmen dieser Arbeit wird das Konzept eines KI-Audits als Synthese allgemeiner Auditprinzipien und dem Framework für KI-Systeme abgeleitet. Beispiele aus dem Bereich Personal- und Talentmanagement konkretisieren die Methode und zeigen Chancen und Potenziale für die Praxis auf.

Die Arbeit ist ein Plädoyer für ein gemeinsames und ganzheitliches Verständnis von KI-Systemen und zeigt darüber hinaus noch Handlungsbedarf in folgenden Bereichen auf:

- Evaluation und Praxistauglichkeit,
- Weiterentwicklung des KI-Audits,
- Standortattraktivität und Aus- und Weiterbildungsangebote,
- Harmonisierung von Initiativen und Regulierung,
- Internationaler Austausch und Communities,
- Forschung

2.
KI-Systeme im Bereich Personal- und Talentmanagement

Künstliche Intelligenz dringt in alle Bereiche digitaler Gesellschaften vor. Dies gilt für den beruflichen und professionellen Kontext, aber auch für den privaten Bereich. Dabei zeigt sich immer häufiger, dass der Einsatz von KI bzw. KI-basierten Systemen zu einem Spannungsverhältnis führt. Dieses Spannungsverhältnis hat viele Facetten und Ebenen. Es reicht von unzulässiger Diskriminierung, Datenschutzverletzungen und unzureichender Kennzeichnung eines KI-Systems bis hin zu einer fehlenden Aufklärung von betroffenen Personen sowie mangelnden Alternativen, die ein „opt-out“ praktisch unmöglich machen. Für Verbraucher*innen bzw. Betroffene ergeben sich daher oftmals Risiken, die nicht kontrollierbar sind.
Aus unternehmerischer Sicht hingegen ist der Einsatz von KI-Systemen aus mehreren Gründen attraktiv. Zumeist stehen Einsatzszenarien rund um algorithmische Entscheidungsfindung im Vordergrund. Daten werden hierbei genutzt, um Entscheidungen zu optimieren bzw. zu automatisieren. Nun gilt nicht immer, dass größere Datenmengen auch notwendigerweise zu besseren Entscheidungen von KI-Systemen führen. Die Menge der zur Verfügung stehenden Daten erlaubt es jedoch, neue Technologien zu verwenden, die in spezifischen Einsatzgebieten sehr leistungsfähig sind und den Stand der Technik neu definieren. Gerade in komplexen Anwendungsfällen, in denen der Einsatz von KI-Systemen bislang nicht oder kaum erfolgte, ergeben sich neue und spannende Möglichkeiten. Ein Bereich hierbei ist Personal- und Talentmanagement. Repräsentative Anwendungsfälle wurden identifiziert und im Rahmen des Forschungsprojekts ExamAI aufgezeigt:

1. Automatisierte Vorschlagssysteme auf Personalplattformen
2. Persönlichkeitsbewertung per Lebenslauf
3. KI-basierte Background-Checks
4. Chatbot der HR-Abteilung
5. Internes Jobprofil-Matching
6. Vorhersage der Jobkündigungsbereitschaft
7. Automatisierte Arbeitszeitzuweisung bei Gig-Workern

Für jeden dieser Anwendungsfälle wurden die Chancen, aber auch die Risiken, die beim Einsatz von KI-Systemen entstehen aufgezeigt. Es zeigt sich, dass Strategien und Methoden zur Minimierung der Risiken durch KI-Systeme nicht in der gleichen Geschwindigkeit entwickelt werden, wie sich die KI-Systeme entwickelt haben. Umso wichtiger ist es, dass die Forschung im Bereich aufholt. An dieser Schnittstelle soll diese Arbeit einen Beitrag leisten.

Die Arbeit beschreibt ein Framework, dass zu Bewertung und zur Analyse von KI-Systemen hinsichtlich der Risiken verwendet werden kann. Das Framework geht auf den Aufbau und die Bestandteile von KI-Systemen auf 5 Ebenen ein und erlaubt es, einen ganzheitlichen Blick auf KI-Systeme einzunehmen und gleichzeitig Schwerpunkte auf spezifische Aspekte, wie zum Beispiel Diskriminierung oder Daten, zu legen. Ausgehend von dem Framework wird die Wichtigkeit von Audits und Testen unterstrichen und der Begriff des KI-Audits eingeführt. Dem KI-Audit liegt ein Software Audit zugrunde. Es geht jedoch auf die Charakteristiken eines KI-Systems im Speziellen ein. Besonders wichtig ist der Hinweis, dass es sich bei einem KI-System um ein software-intensives und sozio-technisches System handelt. Mit Blick auf so manche Diskussion der vergangenen Jahre könnte der Eindruck entstehen, dass es sich bei KI um etwas Undurchsichtiges und „magisches“ handelt. Das ist nicht der Fall. Mit den gleichen Methoden, mit denen Forscher*innen und Entwickler*innen KI-Systeme erforschen und entwickeln, lassen sich auch Fortschritte im Bereich Erklärbarkeit und Transparenz erzielen. Die interdisziplinäre Zusammenarbeit und der Austausch über einzelne Fachkreise hinweg sind dabei entscheidend. Erklärbarkeit und Transparenz von KI-Systemen sind Forschungsbereiche mit komplexen Fragestellungen, die nicht nur durch Informatiker*innen und Ingenieur*innen beantwortet werden können.

3.
Aufbau und Bestandteile von KI-Systemen

3.1. Übersicht

Um die Funktionsweise von KI-Systemen vollständig zu erfassen und zu verstehen ist ein ganzheitlicher Ansatz erforderlich. Ganzheitlich bedeutet in diesem Kontext, dass alle technischen Komponenten, aber auch die die sozio-technischen Komponenten des Entwicklungsprozesses des KI-Systems betrachtet werden müssen [1]. Da es sich bei einem KI-System um ein software-intensives System handelt, unterliegt es ähnlichen Gesetzmäßigkeiten wie andere Softwaresysteme, die in der Informatik schon seit Jahrzenten analysiert und entwickelt werden [2]. Von dieser Gesetzmäßigkeit werden insbesondere auch umfasst:

- Analyse des Einsatzszenarios und Beschreibung des Anwendungsfalls
- Lebenszyklus von software-intensiven Systemen
- Aufbau und die Struktur des ADM-Systems
- Entstehende Entscheidungsstrukturen
- Getroffenen Entscheidungen

Diese unterschiedlichen Bereiche eines KI-Systems müssen transparent gemacht werden, um das Verhalten des KI-Systems verstehen, nachvollziehen und bewerten zu können. Das im Rahmen dieser Arbeit erarbeitete Framework bildet diese fünf Ebenen von KI-Systemen ab. Aus diesen Ebenen lassen sich konsequenterweise auch Hinweise und Kriterien ableiten, um ein potenzielles Fehlverhalten, also eine Abweichung von dem zu erwartenden Verhalten, zu erkennen.

Bei näherer Betrachtung zeigt sich, dass die unterschiedlichen Bereiche voneinander abhängig sind bzw. ineinandergreifen. Dies lässt sich beispielsweise an der von einem KI-System getroffenen Entscheidung veranschaulichen: die Entscheidung ist in der Regel von den bereitgestellten Daten (Input) und der zugrundeliegenden Entscheidungsstruktur (z. B. das trainierte ML-Modell) abhängig. Die Daten werden vorverarbeitet und in eine Form überführt, sodass diese von der Entscheidungsstruktur verarbeitet wer-

[1] B. Waltl, R. Vogl, Explainable Artificial Intelligence – the New Frontier in Legal Informatics, in: Jusletter IT 22. Februar 2018

[2] M. Broy, M. Kuhrmann, Einführung in die Softwaretechnik, Springer Berlin Heidelberg, 2021

den können. Die Ausgabe der Entscheidungsstruktur entsteht also auf Basis der Eingabedaten und der Entscheidungsstruktur. Dies muss, neben weiteren Faktoren (siehe unten) berücksichtigt werden. Diese Abhängigkeiten sind in Abbildung 1 als Übersicht dargestellt.

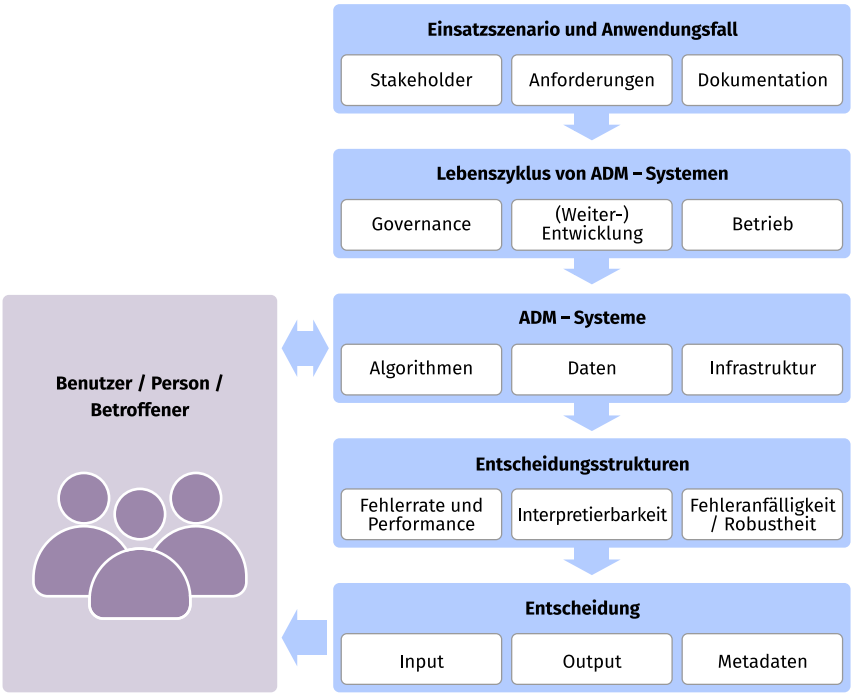


Abbildung 1: Ganzheitliche Übersicht des Entstehungsprozesses, der Komponenten und Sub-Komponenten, eines KI-Systems.

Abbildung 1 stellt die fünf Ebenen mit jeweils drei, für KI-Systeme bedeutsamen, Subsystemen bzw. Teilbereichen dar. In den nachfolgenden Abschnitten werden die jeweiligen Ebenen, sowie deren Subsysteme im Detail vorgestellt und diskutiert. Die Abbildung zeigt auch die Rolle der typischen Benutzenden bzw. interagierenden und betroffenen Personen. Diese interagieren faktisch nur mit ausgewählten Teilbereichen des KI-Systems, nämlich mit Ebene 3 „ADM – Systeme“ und Ebene 5 „Entscheidung“. Die Pfeilrichtungen deuten dabei auch den vorrangigen Informationsfluss an. Während die Benutzenden mit dem ADM-System bilateral kommunizieren, ist der Informationsfluss auf der Entscheidungsebene in der Regel einseitig. Der Nutzende hat bereits seine Daten bereitgestellt und erhält die resultierende Entscheidung nach entsprechender Vorverarbeitung und Verarbeitung durch die Entscheidungsstruktur.

3.2. Ebene 1: Einsatzszenario und Anwendungsfall

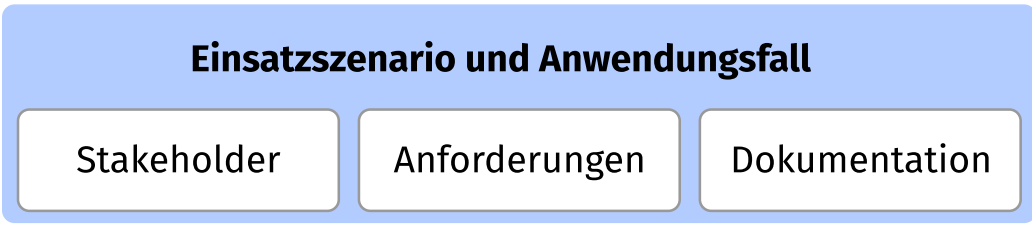


Abbildung 2: Stakeholder, Anforderungen und Dokumentation als entscheidende Elemente der Ebene 1 von KI-Systemen

Einsatzszenario und Anwendungsfall: Dem Einsatz von jedem KI-System liegen grundsätzliche Entscheidungen und Analysen in Bezug auf das Einsatzszenario und dem Anwendungsfall zugrunde. Das Einsatzszenario wird hierbei beschrieben und wichtige Weichenstellungen sowie Motivationen und Beweggründe für den Einsatz von KI identifiziert, die ökonomischer oder technischer Natur sein können. [3] [4] Für den Einsatz von KI – auch im HR- und Talent-Management sprechen im Allgemeinen Aspekte aus den folgenden drei Bereichen [5]:

1. Datenintensive Entscheidungen
2. Konsistente Entscheidungen
3. Kosten-günstige Entscheidungen

Der Anwendungsfall (Use Case) stellt die Beschreibung eines Systems aus Sicht der Akteur*innen dar. Die Gruppe der Akteur*innen umfasst dabei nicht nur die betroffenen Personen, sondern auch Mitarbeitende, die mit dem System arbeiten, Administrator*innen, die das System überwachen und konfigurieren, und weitere Benutzende des KI-Systems. Anwendungsfälle können auch grafisch in Diagrammen beschrieben sein. Der Standard im Bereich Modellierung (UML) sieht hierfür den Diagrammtyp „Anwendungsfalldiagramm“ vor. Der Anwendungsfall ist von entscheidender Bedeutung und kann der Haupteinstiegspunkt für die Analyse des Gesamtsystems sein. Bei der Analyse sind die Beschreibungen der folgenden Aspekte von zentraler Bedeutung:

1. Stakeholder
2. Anforderungen
3. Dokumentation

[3]

Why AI is the future of growth, Accenture, 2016. The economic impact of the automation of knowledge work, robots and self-driving vehicles could reach between EUR 6.5 and EUR 12 trillion annually by 2025 (including improved productivity and higher quality of life in ageing populations). Quelle: Disruptive technologies: Advances that will transform life, business, and the global economy, McKinsey Global Institute, 2013.

[4]

AI is part of the Commission‘s strategy to digitise industry (COM(2016) 180 final) and a renewed EU Industrial Policy Strategy (COM(2017) 479 final).

[5]

S. Russel, P. Norvig. “Artificial intelligence: a modern approach.” (2002).

Stakeholder: Welche Personengruppen haben ein Interesse an dem System und wie sehen diese Interessen aus? Stakeholder sind Personen bzw. Gruppen, die ein berechtigtes Interesse an einem Produkt bzw. an einem KI-System haben. Diese müssen nicht notwendigerweise die Benutzer*innen eines Systems sein [6].

[6] J. McManus. "A stakeholder perspective within software engineering projects." 2004 IEEE International Engineering Management Conference (IEEE Cat. No. 04CH37574). Vol. 2. IEEE, 2004.

Beispiel: Die geschäftsführende Person eines Unternehmens erwartet sich durch den Einsatz eines Chatbots für Personalauskünfte zufriedenere Mitarbeitende, weil wiederkehrende Auskünfte im Personalbereich schneller und unbürokratischer beantwortet werden können. Somit ist diese Person ein Stakeholder, obwohl sie – in ihrer Rolle als Geschäftsführer*in – das KI-System nicht unmittelbar benutzt.

Anforderungen: Welche Anforderungen wurden an das zu entwickelnde bzw. einzusetzende KI-System gestellt? Jedes software-intensive System ist gekennzeichnet durch die Anforderungen, die an ein solches gestellt werden. Grundsätzlich unterscheidet man zwischen funktionalen und nicht-funktionalen Anforderungen [7]. Die Liste an Anforderungen ist die Spezifikation dessen, was an Erwartungshaltung an das KI-System gestellt wurde. Hierbei ist auch die negative Abgrenzung ebenfalls zu berücksichtigen: also auch Aspekte bzw. Funktionsweisen, für die es keine Anforderungen gibt. Der*die Auftraggeber*in beschreibt die Funktionsweise des KI-Systems mit funktionalen und nicht-funktionalen Anforderungen. In der Praxis sind diese im Lasten- bzw. Pflichtenheft definiert. Bei traditionellen Vorgehensweisen, die an Wasserfallmodellen angelehnt sind, ist das Lastenheft das zentrale Anforderungsdokument.
Analoge Artefakte gibt es jedoch auch bei agilen Vorgehensmodellen: die Anforderungen sind dort in sog. User Stories bzw. Tasks abgebildet und beschreiben in ihrer Gesamtheit die (nicht) umgesetzten Funktionen und Funktionalitäten. Bei agilen Vorgehensmodellen heißen die zentralen Artefakte und Datenbanken der Anforderungen „Product Backlog" und „Sprintbacklog".

[7] H. Balzert. Lehrbuch der softwaretechnik: Basiskonzepte und requirements engineering. Springer-Verlag, 2010.

Beispiel: Die Geschäftsführung setzt seinen Chatbot für wiederkehrende Auskünfte im Personalbereich ein. Ein*e neue*r Mitarbeiter*in bemerkt, dass der Chatbot keine Anfragen in englischer Sprache beantworten kann. Durch Analyse der Anforderungen im Lastenheft oder im Product Backlog wird festgestellt, dass es bislang noch keine expliziten Anforderungen an den Chatbot zur Mehrsprachigkeit gibt. Gemeinsam mit dem Geschäftsführer werden konkrete Anforderungen an englischsprachige Auskünfte, formuliert, die bei der nächsten Weiterentwicklung umgesetzt werden sollen.

Dokumentation: Wie sind die Funktionen und Eigenschaften des Systems bzw. der Teilsysteme beschrieben? In der Praxis eingesetzte Softwaresysteme werden in der Regel dokumentiert [8] (siehe dazu auch Anforderungen aus ISO 9001:2015). Die Dokumentation ist immer ein wichtiger Bestandteil eines Softwaresystems und gibt Auskunft über die erwartungsgemäße Funktionsweise sowie die Randbedingungen und Vorraussetzungen, die erfüllt sein müssen, damit das System korrekt funktioniert. Die Dokumentation ist insbesondere dann wichtig, wenn Subsysteme von Dritten erstellt und implementiert werden. Große Softwaresysteme sind gekennzeichnet durch einen modularen Aufbau und durch Kommunikation und Integration über Schnittstellen. Die einzelnen Module können hier von verschiedenen Parteien und Dritten (z. B. weitere Vertragspartner, Unterauftragnehmer, oder Open Source Communities) bereitgestellt werden.

[8]

Phillips, P. Jonathon, et al. "Four principles of explainable artificial intelligence." Gaithersburg, Maryland (2020). Phillips, P. Jonathon, et al. "Four principles of explainable artificial intelligence." Gaithersburg, Maryland (2020).

Beispiel: Der Chatbot ist aus unterschiedlichen Komponenten aufgebaut. Darunter auch eine Komponente, die die Antworten des Chatbots für den Benutzenden in natürlicher Sprache (Text) formuliert. Es handelt sich um die sogenannte NLG-Komponente (Natural Language Generation). In der Praxis erweist sich diese Komponente als unzuverlässig, weil sie keine geschlechtsneutralen Formulierungen unterstützt. Sie muss ersetzt werden. Die Dokumentation verrät, dass die Komponente aktuell von Cloud Anbieter X in der Version 0.9 bezogen wird. Darüber hinaus sind die Schnittstellen (Datenformate, etc.) beschrieben. Auf Basis dieser Dokumentation wird das Problem eingegrenzt und die NLG Komponente kann durch die Komponenten eines anderen Anbieters ersetzt werden.

3.3. Ebene 2: Lebenszyklus von software-intensiven KI-Systemen

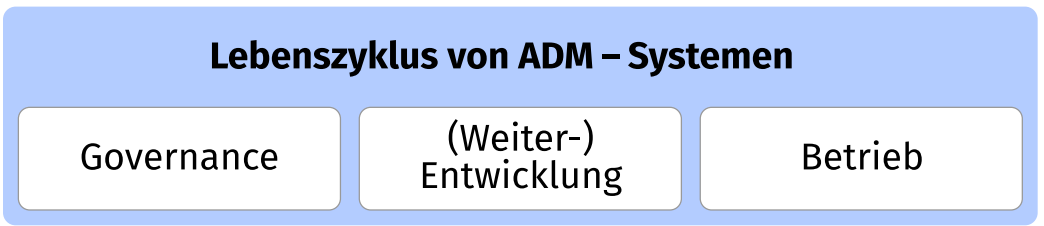


Abbildung 3: Governance, (Weiter-)Entwicklung und Betrieb als entscheidende Elemente der Ebene 2 „Lebenszyklus von ADM – Systeme"

Lebenszyklus von software-intensiven ADM Systemen: Neben grundsätzlichen Festlegungen des Einsatzes von Künstlicher Intelligenz sowie der Spezifikation der Software durch Anforderungen und Dokumentation, ist eine exakte Betrachtung des Lebenszyklus von Software von Bedeutung. Bei bisherigen Betrachtungen wird oftmals nur die vorliegende Software bzw. das vorliegende Softwaresystem untersucht. In der Praxis ist dies jedoch unzureichend [9]. Software ist eingebettet in einen komplexen Lebenszyklus. Transparenz über den Lebenszyklus ist erforderlich, wenn das Softwaresystem und sein Verhalten möglichst nachvollziehbar gemacht werden sein soll.

[9] Handelsblatt. "Kartellamt rügt Lufthansa: Solche Algorithmen werden ja nicht vom lieben Gott geschrieben", 28.12.2017.

Bisherige Betrachtungen untersuchen ein Softwaresystem nur zu einem einzigen Zeitpunkt, was für komplexe Systeme unzureichend sein kann. Gerade große Systeme und Plattformen entwickeln sich permanent weiter, sie sind gekennzeichnet durch „continuous deployment" [10]. Der Lebenszyklus umfasst dabei mehrere, unterschiedlicher Aspekte, die von zentraler Bedeutung sind:

1. Governance
2. (Weiter-)Entwicklung
3. Betrieb

[10] Vereinfacht ausgedrückt handelt es sich bei Continuous Deployment (CD) um eine Sammlung von automatisierten Methoden, die Änderungen im Softwarecode von Entwicklern sofort bewertet, testet, und auf ein Produktivsystem ausrollt. Neue Funktionen und Anpassungen können damit nahezu in Echtzeit bereitgestellt werden. Das Produktivsystem erfährt dadurch mehrere Updates am Tag. Man spricht in dem Zusammenhang von Continuous Improvement.

Governance: Im Kontext software-intensiver Systeme spielt die IT-Governance eine zentrale Rolle. Sie stellt sicher und legt fest, dass die IT und alle damit zusammenhängenden Fragen, beispielsweise das Verhältnis von In- und Out-Sourcing oder das Verhältnis von selbstentwickelten Lösungen und Kauflösungen (darunter auch Cloud und SaaS), im unternehmerischen Kontext gestellt und beantwortet werden. Die Governance umfasst dabei mindestens die Führung, Organisationsstruktur, Prozesse sowie die strategischen Ziele und grundlegenden IT-Prinzipien eines Unternehmens. [11] Im Rahmen von KI-Systemen und unter Berücksichtigung deren Lebenszyklus und Struktur müssen bestehende Prinzipien überarbeitet und neue, präziser auf die Zielstellungen angepasste Elemente geschaffen werden.

[11] Fröhlich, M., Fröhlich, M., & Glasner, K. (2007). IT Governance. Gabler.

Beispiel: Zur Umsetzung eines erklärten Unternehmensziels „Steigerung der Mitarbeiterzufriedenheit" soll ein Chatbot eingesetzt werden. Dessen Entwicklung inhouse wäre zu teuer und würde viel Zeit in Anspruch nehmen. Man entscheidet sich – im Einklang mit den strategischen Zielen der IT – dafür, die Entwicklung zu outsourcen und so viele Funktionen als Cloud Service zu konsumieren. Die IT-Leitung erkennt, dass dadurch herkömmliche Qualitätsmanagement-Methoden nicht länger ausreichen. In Zusammenarbeit mit der Geschäftsführung werden die IT-Governance-Prinzipien durch präzise Anforderungen an Prozesse und Dokumentation angepasst, sodass man weiterhin eine sehr große Transparenz über die eingekauften Leistungen hat.

(Weiter-)Entwicklung: Die Entwicklung von software-intensiven Systemen wird in Prozessen durchgeführt. Je nach Komplexität und Größe werden hierfür unterschiedliche Verfahren eingesetzt. In der jüngsten Vergangenheit haben sich Methoden, die sich mehr an agilen Vorgehensmodellen [12] orientieren, gegenüber herkömmlichen Methoden (darunter auch sog. „Wasserfall-Methoden“) durchgesetzt. Diese Vorgehensmodelle steuern nicht nur den Entwicklungsprozess, sondern definieren auch die Phasen, welche von der Anforderungserhebung bis zum Roll-out und kontinuierlichen Weiterentwicklung durchlaufen werden. Im Rahmen der Erhöhung der Transparenz von KI-Systemen spielt dieser Aspekt eine zentrale Rolle. Diverse Quellen in der Literatur weisen bereits darauf hin, dass KI-Systeme auch das Ergebnis eines sozio-technischen Vorgangs sind. [13] [14] [15] Das bedeutet, dass KI-Systeme nicht nur ein Stück Technologie sind, sondern an ihrer Entwicklung auch Menschen und Personen beteiligt sind. Die Artefakte und Zwischenergebnisse, die im Rahmen der Entwicklung produziert werden, können wichtige Hinweise über das zu erwartende Verhalten (Funktion und Fehl-Funktion) von Software enthalten. Darüber hinaus müssen in den Entwicklungsprozessen auch Phasen zur dezidierten Sicherstellung der Qualität, z. B. durch Code-Reviews, oder Softwaretests, vorhanden sein.
Das gilt nicht nur für den Entwicklungsprozess, sondern auch analog für die Weiterentwicklung, die in der Regel den Entwicklungsprozess eines Teilsystems oder einer neuen Funktion darstellt.

Beispiel: Zu Beginn des Entwicklungsprojekt Chatbot gibt es unterschiedliche Ansichten bzgl. der Projektdurchführung. Da jedoch bereits zu Beginn absehbar ist, dass neue fachliche und technische Anforderungen während des Projekts hinzukommen werden, entscheidet sich das Projektteam für ein agiles Vorgehen. Neue Anforderungen werden zu jedem Zeitpunkt erfasst, genau beschrieben („User Stories) und vom Product Owner im sogenannten „Product Backlog“ gespeichert. Das Team wählt in kurzen, aber regelmäßigen Zeitabständen („Sprints“) eine Menge von User Stories aus und setzt diese um. Damit ist für jede User Story sehr präzise beschrieben, wann und von wem diese bearbeitet und umgesetzt wurde. Außerdem behält sich das Projekt Team die Flexibilität vor auf neue Anforderungen schnell reagieren zu können.
Die Menge der User Stories gibt nicht nur Auskunft über die umgesetzten Anforderungen, sondern auch über die nicht erfassten Anforderungen, welche im Kontext von KI-Systemen ebenfalls aufschlussreich sind.

[12]
Beck, Kent, et al. “Manifesto for agile software development.” (2001): 2006. sowie Shore, James. The Art of Agile Development: Pragmatic guide to agile software development. O‘Reilly Media, 2007.

[13]
Zweig, K. A., & Krafft, T. D. (2018). Fairness und Qualität algorithmischer Entscheidungen. In R. Mohabbat Kar, B. E. P. Thapa, & P. Parycek (Hrsg.), (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft (S. 204-227). Berlin: Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT).

[14]
Baxter, Gordon, and Ian Sommerville. “Socio-technical systems: From design methods to systems engineering.” Interacting with computers 23.1 (2011): 4-17.

[15]
Holton, Robert, and Ross Boyd. ‘Where are the people? What are they doing? Why are they doing it?’(Mindell) Situating artificial intelligence within a socio-technical framework. Journal of Sociology 57.2 (2021): 179-195.

Betrieb (Operations): Der dritte große Bereich, aus dem sich Eigenschaften eines KI-Systems ergeben können, betrifft den Betrieb (sog. Operations). Der Betrieb eines software-intensiven Systems [16] umfasst auch die Systeme, Prozesse und Rollen zur Aufrechterhaltung der Funktionen, die dem erwartungsgemäßen Verhalten entsprechen. Wichtig ist beispielsweise, dass zentrale Komponenten – beispielsweise IT-Infrastruktur – überwacht werden. Ausfälle und sicherheitskritische Events müssen erkannt werden. In dem Zusammenhang zählt auch das Aktualisieren der IT-Infrastrukturen durch notwendige (Sicherheits-)Updates zu einem wichtigen Aufgabengebiet für Operations. Für Cloud Services zählt hierzu auch das Überwachen von Release- und Deployment-Zyklen, sodass Änderungen von wichtigen Funktionen sofort erkannt und bewertet werden können. Im Kontext von KI-Systemen kann hierunter auch das pro-aktive Bewerten des aktuellen Einsatzbereichs des Systems zählen. Insofern erkannt wird, dass die Qualität der Entscheidungen eines trainierten ML Modells während des Betriebs signifikant sinkt, muss dies erkannt und bewertet werden. Aus dem Bereich Operations können damit Entscheidungen zur Verbesserung eines ML-Modells vorbereitet werden.

Beispiel: Der Chatbot für Personalauskünfte wird von den Mitarbeitenden sehr gut angenommen. Mit der Benutzung steigt die Erwartungshaltung und der Chatbot wird immer öfter zu Themen angefragt, die bei der Entwicklung nicht berücksichtigt wurden, z. B. Urlaubsplanung und Kurzarbeit. Der Bereich Operations hat ein Monitoring installiert und wird darüber benachrichtigt, dass in den letzten Tagen viele Anfragen zum Thema Kurzarbeit gestellt wurden. Die Konfidenz zur Angabe der Zuverlässigkeit des ML-Modells, das im Chatbot verwendet wird, sinkt rapide ab. Operations hat hierzu anonymisierte Daten vorliegen und kann die Logdateien auswerten, um mit den Verantwortlichen die Weiterentwicklung des ML-Modell mit zusätzlichen Daten abzustimmen, sodass die neuen Anfragen richtig beantwortet werden können.

[16]

Berente, Nicholas, et al. “MANAGING ARTIFICIAL INTELLIGENCE.” MIS Quarterly 45.3 (2021).

3.4. Ebene 3: ADM-Systeme

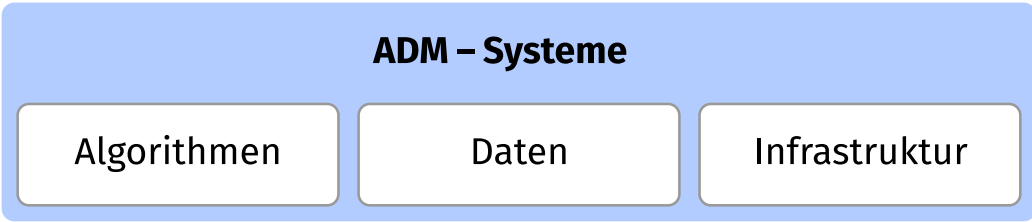


Abbildung 4: Algorithmen, Daten und Infrastruktur als entscheidende Elemente der Ebene 3 „ADM – Systeme“

ADM-Systeme: Aufbauend auf den grundsätzlichen Einsatzmöglichkeiten von Künstlicher Intelligenz für einen Anwendungsfall und dem Lebenszyklus, in den ein KI-System eingebettet ist, nimmt das eigentliche ADM-System eine zentrale Rolle in den Betrachtungen ein. [17] Dabei handelt es sich um das Softwaresystem, um das sich der Lebenszyklus aus Ebene 2 dreht. Dies ist das Ergebnis eines komplexen Entwicklungsprozesses. Bei modernen software-intensiven Systemen ist der Entwicklungsprozess nie abgeschlossen, sondern wird kontinuierlich weiterentwickelt. Um das Verhalten eines derartigen Systems transparent zu machen und entsprechend bewerten zu können, ist es notwendig, diese Dynamik zu erkennen und zu berücksichtigen.

[17] Waltl, B., Vogl, R. Increasing Transparency in Algorithmic- Decision-Making with Explainable AI. Datenschutz Datensich 42, 613–617 (2018).

Man kann die Betrachtungen des Lebenszyklus jedoch einschränken und sich auf drei entscheidende Komponenten fokussieren, wenn es um die Analyse des Verhaltens eines ADM-Systems geht:

1. Algorithmen
2. Daten
3. Infrastruktur

Algorithmen: Die algorithmischen Entscheidungsfindung (algorithmic decision-making) ist auf den Einsatz von Algorithmen angewiesen. Algorithmen haben eine lange Tradition in der Informatik. [18] Es handelt sich bei Algorithmen um eindeutige Handlungsvorschriften, die ein bestimmtes, vorgegebenes Problem lösen. [19] Diese Handlungsvorschriften können auch von Maschinen und in Computerprogrammen ausgeführt werden, was ihren Einsatz besonders attraktiv macht. Mittlerweile existiert eine Vielzahl von verschiedenen Algorithmen, die jeweils unterschiedliche Eigenschaften haben und sich für die Lösung eines gegebenen Problems mehr oder weniger gut eignen. Im Kontext von KI sind Algorithmen aus den drei wichtigen Bereichen besonders relevant: [20]

[18] Algorithmen gab es schon, bevor es die Informatik als Begriff und als Wissenschaft gab. Sie sind also keine genuine Erfindung der Computerwissenschaft. Gleichwohl hat die Forschung enorm zu unserem heutigen Verständnis von und unser Wissen über Algorithmen beigetragen.

[19] S. Russel, P. Norvig. “Artificial intelligence: a modern approach.” (2002).

[20] Waltl, B., Vogl, R. Increasing Transparency in Algorithmic- Decision-Making with Explainable AI. Datenschutz Datensich 42, 613–617 (2018).

- Regel-basierte Methoden
- Statistische Methoden
- Neuronale Netze

Sehr beliebt sind mittlerweile auch sogenannte Ensemble Methoden [21] , also Algorithmen, die mehrere unterschiedliche Algorithmen einsetzen, die sich durch eine gezielte Permutation von Parametern unterscheiden. Durch das gezielte Einsetzen von mehreren Methoden und der Aggregation der Teilergebnisse zu einem Endergebnis können die Nachteile von einzelnen Verfahren besser kompensiert werden. Damit kann man zum Beispiel sogenanntes Overfitting [22] für bestimmte Verfahren besser kontrollieren.

Bisherige Analysen [23] zeigen, dass sich Algorithmen – je nach zugrundeliegender Methode – unterschiedlich gut eignen, um die Entscheidungsabläufe transparent, erklär-, und nachvollziehbar zu machen. Mit der Auswahl des zu verwendenden Algorithmus werden mehrere Entscheidungen, möglicherweise implizit, getroffen. Diese Entscheidung bestimmt maßgeblich mehrere Eigenschaften des KI-Systems mit. So werden beispielsweise große Bereiche einer möglichen Erklärbarkeit signifikant durch die Wahl des Algorithmus bestimmt (siehe Abschnitt 3.5). Eine Gegenüberstellung der Algorithmen sowie deren Transparenzeigenschaften wurde in der Literatur bereits im Ansatz durchgeführt. [24] Wichtig für den Bereich des Testings und der Auditierung ist, dass ein besonderer Fokus auf den Algorithmus (und seine Parameter) innerhalb eines ADM-Systems gelegt wird.

Beispiel: Aufmerksame Mitarbeiter*innen stellen fest, dass der Chatbot bei Personalauskünften in bestimmten Situationen falsche Auskünfte gibt. Der Chatbot versteht die Anfragen zwar richtig, verweist jedoch auf das falsche Dokument in der Wissensdatenbank innerhalb des Intranets. Diese Situationen lassen sich reproduzieren. Das bedeutet, dass auf die gleichen Anfragen die gleichen, falschen Antworten gegeben werden. Die Mitarbeitenden wenden sich mit der Beobachtung an die IT-Abteilung und leiten die Anfragen mit der Bitte um eine Überprüfung der falschen Auskünften weiter. Die IT-Abteilung sichtet die zugrundeliegenden Algorithmen: 1) Ein neuronales Netz zur Extraktion der Information aus den Anfragen (die sog. NLU- Komponente [25]) und 2) ein Entscheidungsbaum zur Zuordnung der Information aus den Anfragen. Die IT-Abteilung stellt fest, dass die NLU-Komponente richtig funktioniert. Für die Zuordnung einer Anfrage zu einer Antwort wird jedoch ein regel-basierter Entscheidungsbaum verwendet. Der Entscheidungsbaum kann sehr schnell und einfach von den Personen in der IT-Abteilung

[21]

Dietterich T.G. (2000) Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

[22]

Dietterich, T.G. “Overfitting and undercomputing in machine learning.” ACM computing surveys (CSUR) 27.3 (1995): 326-327.

[23]

Raschka, Sebastian. “Model evaluation, model selection, and algorithm selection in machine learning.” arXiv preprint arXiv:1811.12808 (2018).

[24]

Waltl, B., Vogl, R. Increasing Transparency in Algorithmic- Decision-Making with Explainable AI. Datenschutz Datensich 42, 613–617 (2018).

[25]

NLU bedeutet Natural Language Understanding

überprüft werden. Es zeigt sich, dass der Entscheidungsbaum unvollständig ist. Die IT-Abteilung kann die entsprechenden Regeln innerhalb des Entscheidungsbaums rasch ergänzen und die Anfragen werden zukünftig richtig beantwortet.

Daten: Die Relevanz von Daten für KI wurde in der Literatur schon sehr häufig hervorgehoben [26] und wird im Rahmen dieser Arbeit nochmals unterstrichen. In Kombination mit Algorithmen stellen die Daten das Kernstück eines ADM-Systems dar. Die Algorithmen setzen mehr oder weniger komplexe mathematische Verfahren [27] ein, um höherwertige Information aus Daten zu gewinnen. Diese Information kann das Erkennen von Mustern (beispielsweise bei der Verarbeitung natürlicher Sprache oder Gesichtserkennung und Videoauswertung) oder das Vorhersagen von Trends und Bewertungen (beispielsweise bei der Bewertung der Kreditwürdigkeit) umfassen.

Den Daten kommt daher eine besondere Rolle zu. In den Daten sind die Muster und Grundstrukturen dessen angelegt, was durch maschinelle Lernverfahren erkannt werden kann. Die vorhandenen Muster können sich durch den Einsatz von maschinellem Lernen sogar noch verstärken. Dies kann besonders dann problematisch werden, wenn die Daten nicht repräsentativ [28] sind oder Bias (Verzerrungen, z. B. durch Bevorzugung oder Benachteiligung bestimmter Personengruppen) enthalten. Bryce Goodman bringt dies auf den Punkt: „Machine learning depends upon data that has been collected from society, and to the extent that society contains inequality, exclusion or other traces of discrimination, so too will the data." [29]

Die Überprüfung der verwendeten Daten ist somit ein wichtiger Bestandteil, um nachvollziehen zu können welche Entscheidungen ein KI-System getroffen hat.

Beispiel: Aufmerksame Mitarbeiter*innen stellen fest, dass der Chatbot zum Thema Zeiterfassung keine Auskünfte gibt und bei derartigen Anfragen Information zu einem anderen Thema anbietet. Die NLU-Komponente der Software, die für die Informationsextraktion aus den Anfragen zuständig ist, wird von der IT-Abteilung untersucht. Es zeigt sich, dass die NLU-Komponente die Begriffe und Terminologie zur Zeiterfassung nicht erkennt. Die Fachleute untersuchen die Trainingsdaten, die zum Training der NLU-Komponente verwendet wurden und stellen fest, dass in den Trainingsdaten (d. h. den beispielhaften Anfragen, die bei der Entwicklung des Systems verwendet wurden) keine Datensätze aus dem Themenbereich Zeiterfassung enthalten sind. Für das nächste Update werden geeignete Datensätze dafür erstellt und beim Training der NLU-Komponente einbezogen.

[26]
O'Leary, Daniel E. "Artificial intelligence and big data." IEEE intelligent systems 28.2 (2013): 96-99.

[27]
Ein zentrales Prinzip des Maschinellen Lernens ist das Approximieren von Funktionen, welche eine mathematische Beziehung zwischen Ein- und Ausgabedaten darstellt. Hierfür sind nicht notwendigerweise mathematisch aufwändige Funktionen notwendig.

[28]
Ntoutsi, Eirini, et al. "Bias in data‐driven artificial intelligence systems—An introductory survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.3 (2020): e1356.

[29]
B. Goodman, and S. Flaxman. "European Union regulations on algorithmic decision-making and a 'right to explanation'." AI magazine 38.3 (2017): 50-57. And B. Goodman. Computer Says No: Economic Models of (Algorithmic) Discrimination. (2016) Unpublished paper, Oxford, UK.

Infrastruktur: Ein weiterer Bereich, der für die Entwicklung und den Betrieb eines ADM-Systems von Bedeutung ist, ist die zugrundeliegende Infrastruktur. Gerade für die Entwicklung bedarf es leistungsfähiger Infrastrukturen, welche die zum Teil sehr rechenintensiven Trainingsoperationen durchführen. Die Entwicklung der Softwaresysteme ist zumeist sehr datenintensiv, weshalb viele der Operationen auf einer speziell dafür vorgesehenen Hardware durchgeführt wird. [30]

[30] Dai, Wei, and Daniel Berleant. "Benchmarking contemporary deep learning hardware and frameworks: A survey of qualitative metrics." 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI). IEEE, 2019.

Aus Effizienzgründen wird Hardware auch mehr und mehr in zentralen Bereichen gebündelt, z. B. in großen Rechenzentren, oder über sog. Cloud Services konsumiert. Die Hardware wird dann nicht mehr gekauft, sondern für den erforderlichen Zeitraum gemietet. Dies ermöglicht es auch Unternehmen mit kleineren oder weniger spezialisierten IT-Abteilungen innovative Projekte anzugehen und umzusetzen. Die Konfiguration der IT-Infrastruktur hat Auswirkungen an den Datenfluss und den Speicherort der Daten, darunter fallen auch trainierte ML-Modelle. Das Entscheidungsverhalten des KI-Systems wird dadurch nicht direkt beeinflusst. Möglicherweise ist man durch die flexible Skalierung der IT-Infrastruktur befähigt, noch leistungsfähigere Algorithmen zu verwenden oder mit einem vergleichbaren Ressourceneinsatz noch aufwändigere Trainingsverfahren für ML-Modelle einzusetzen, was sich mittelbar auf das Entscheidungsverhalten auswirkt.
Bei der ganzheitlichen Betrachtung eines KI-Systems und insbesondere beim Auditieren muss auch die Infrastruktur berücksichtigt werden. Hieran sind aufgrund besonders schützenswerter Daten möglicherweise spezielle Anforderung an die Informationssicherheit zu stellen. [31]

[31] Siehe auch BSI Publikationen und Arbeitsbereiche https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/it-grundschutz_node.html, https://www.bsi.bund.de/DE/Home/home_node.html

Beispiel: Ein Unternehmen setzt einen Chatbot zur Beantwortung wiederkehrender Personalanforderungen ein. Der Chatbot stößt bei den Mitarbeitenden auf große Akzeptanz und Beliebtheit. Die Mitarbeitenden wissen es zu schätzen, dass sie sich mit einfachen Fragen an eine KI wenden können und damit die Kolleg*innen in der Personalabteilung entlasten. Die große Nachfrage führt dazu, dass der Server, auf dem der Chatbot und seine Komponenten bereitgestellt werden, aufgerüstet werden muss. Im Auftrag des Geschäftsführers und nach Abstimmung mit dem Leiter der IT-Abteilung wird geprüft, ob für dieses Projekt nicht eine Cloud-Infrastruktur gemietet werden kann. Vor dem Hintergrund, dass Personalanfragen sensible Mitarbeiterdaten umfassen können die Daten außerhalb des Unternehmens übermittelt und verarbeitet werden, stellt die IT-Abteilung entsprechend höhere Ansprüche und Kosten an die Informationssicherheit und kann die Applikation deshalb auch via Cloud-Infrastruktur nutzen.

3.5. Ebene 4: Entscheidungsstrukturen

Abbildung 5: Performance, Interpretierbarkeit und Fehleranfälligkeit als entscheidende Elemente der Ebene 4 „Entscheidungsstrukturen“

Entscheidungsstrukturen: Eingebettet in die ADM-Systeme der Ebene 3 sind die sogenannten Entscheidungsstrukturen. Dabei handelt es sich um Bereiche, die konkrete Entscheidungen (siehe Ebene 5) ermöglichen. Die Entscheidungsstrukturen hängen oft mit den zur Entscheidungsfindung verwendeten Algorithmen [32] (siehe Ebene 4) zusammen. Sie stellen die Ergebnisse der Umsetzung bzw. der Implementierung der Algorithmen und insofern konkrete Ausprägungen der Algorithmen dar.

Die Entscheidungsstrukturen ergeben sich somit aus den eingesetzten Algorithmen und den zur Verfügung stehenden Daten. Sie stellen die „trainierten Modelle“ im Bereich Maschinellem Lernen dar. Wie in Abschnitt 3.4 kurz dargestellt gibt es unterschiedliche Klassen von Algorithmen. Ein trainiertes Modell kann je nach eingesetztem Algorithmus unterschiedliche Formen annehmen: regelbasiert, Entscheidungsbäume, Ensemble von Entscheidungsbäume, statistische Verfahren und Wahrscheinlichkeitsverteilungen, neuronale Netze, etc. Innerhalb der regelbasierten Verfahren ist der Einsatz und die Verwendung von Entscheidungsbäumen sehr etabliert. Entscheidungsbäume müssen nicht notwendigerweise von Menschen, sondern können auch automatisiert von Algorithmen erstellt werden. Die Ergebnisse von komplexeren Algorithmen, wie beispielsweise das Ergebnis des Trainings neuronaler Netze, können hingegen nicht mehr von Menschen durchgeführt werden. Ihre Erstellung ist ausschließlich Softwaresystemen überlassen. Das Ergebnis des Erstellvorgangs, die Entscheidungsstruktur, ist ein wesentlicher Bereich eines KI-Systems und wird durch den Forschungsbereich Explainable AI (XAI) untersucht. [33] Eine zentrale Erkenntnis ist, dass sich nicht alle Entscheidungsstrukturen gleichermaßen für eine Interpretation durch Menschen eignen. Es gibt jedoch zusätzlichen Methoden und Verfahren, um zugrundliegende Entscheidungsstrukturen erklärbar sowie test- und auditierbar und zu machen. [34]

[32]
Nguyen, Giang, et al. “Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey.” Artificial Intelligence Review 52.1 (2019): 77-124.

[33]
Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey.” 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, 2018.

[34]
Molnar, Christoph. “Interpretable machine learning. A Guide for Making Black Box Models Explainable”, 2019.

Bei Entscheidungsstrukturen spielen drei Bereiche eine zentrale Rolle, die im Rahmen von Auditierung und Testen fokussiert werden. Diese werden auch in der Assessment List for Trustworthy AI von der HLEG AI der Europäischen Kommission vorgeschlagen : [35]

1. Fehlerrate (Accuracy) und Performance
2. Interpretierbarkeit (Interpretability and Explainability)
3. Fehleranfälligkeit (Robustness)

Fehlerrate und Performance: Die zentralen Maße, welche zur Bewertung (Evaluation [36]) einer Entscheidungsstruktur herangezogen werden, sind Metriken [37], die Auskunft über die Fehlerraten geben. Je nach eingesetzten Verfahren und Aufgabe zählen hierzu beispielswiese Precision und Recall (insb. bei Klassifikationsverfahren), Purity und Dunn Index (insb. bei Clusterverfahren), Bleu Score (bei maschineller Übersetzung oder Mean Absolute Error (insb. bei Regressionsverfahren). Für jede Entscheidungsstruktur und Anwendung gibt es verschiedene Metriken, welche je nach Anwendungsfall und Einsatzgebiet (siehe Ebene 1) unterschiedlich eingesetzt werden können. Metriken erlauben eine intersubjetive Bewertung der Fehlerraten und der Performance. Dies Auswahl der Metriken ist jedoch nicht trivial, da eine Vielzahl von (Qualitäts-)Metriken existieren und diese zu unterschiedlichen Ergebnissen führen können.
Grundlage der Bewertungen durch Metriken sind Datensätze, deren Qualität im Vorfeld sichergestellt wurde (sog. Goldstandard oder Test-Data). Unter Verwendung dieser Datensätze wird im Rahmen der Evaluierungsphase die Entscheidungsstruktur bzw. das ML-Modell bewertet. Diese Bewertung ist reproduzierbar und kann mit geringem Aufwand, zum Teil auch automatisiert (z. B. im Rahmen eines kontinuierlichen Monitorings) durchgeführt werden.

Neben der Fehlerrate spielt die Performance eine wichtige Rolle in der Praxis. Hierbei sind insbesondere zwei Bereiche relevant: einerseits die sogenannte Inferenzzeit, die benötigt wird, damit ein ML-Modell eine Entscheidung trifft und andererseits die Trainingszeit innerhalb derer ein ML-Modell trainiert wird. [38] Die Performance unterschiedlicher ML-Modelle unterscheidet sich zum Teil erheblich voneinander. Das betrifft sowohl die Trainingsphase, aber auch die Inferenzzeit. Damit sind – je nach Einsatzgebiet – manche ML-Modelle für die Praxis (noch) nicht einsetzbar. Insbesondere in zeitkritischen Anwendungsbereichen (z. B. Autonomes Fahren) spielt dies eine Rolle.

Beispiel: Ein Unternehmen verwendet ein KI-System, um Lebensläufe vorab zu analysieren und zu bewerten. Das KI-System setzt hierfür eine Natural Language Understanding Komponente auf Basis von neuronalen Netzen ein. Das Unternehmen

[35]
European Commission, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, 2021.

[36]
Hossin, Mohammad, and Md Nasir Sulaiman. “A review on evaluation metrics for data classification evaluations.” International journal of data mining & knowledge management process 5.2 (2015): 1.

[37]
Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. „Machine learning interpretability: A survey on methods and metrics.“ Electronics 8.8 (2019): 832.

[38]
Lim, Tjen-Sien, Wei-Yin Loh, and Yu-Shan Shih. „A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms.“ Machine learning 40.3 (2000): 203-228.

expandiert und ab sofort werden Lebensläufe in einer neuen, bislang noch nicht relevanten Sprache eingereicht. Die Geschäftsführung möchte wissen, ob die neu eingereichten Lebensläufe genauso fehlerfrei analysiert werden, wie die bisher eingereichten Lebensläufe. Die Mitarbeitenden bewerten viele der neu eingereichten Lebensläufe manuell (sie erstellen damit einen Testdatensatz bzw. Goldstandard) und vergleichen das manuelle Ergebnis mit dem Ergebnis des KI-Systems. Die Auswertung der beiden Standardmetriken, nämlich Precision und Recall, zeigt, dass das KI-System für die neue Sprache völlig ungeeignet ist. Gemeinsam mit den Mitarbeitenden legt die Geschäftsführung fest, das KI-System bis auf Weiteres nicht für Lebensläufe der neuen Sprache anzuwenden.

Interpretierbarkeit: Entscheidungsstrukturen und ML-Modelle von KI-Systemen sind in softwareintensive Systeme eingebettet. Sie sind damit in erster Linie für die Anwendung und Ausführung durch Software entwickelt worden. Software hat – im Vergleich zum Menschen – ganz andere Möglichkeiten, die Information in den ML-Modellen zu verarbeiten und zur Entscheidungsfindung zu nutzen. Eine große Anzahl an Parametern kann verarbeitet werden und viele, möglicherweise nicht-lineare, Rechenoperationen können parallel bzw. sequenziell ausgeführt werden, um eine Entscheidung zu treffen. Für Menschen sind diese Entscheidungsstrukturen, also die Kombination aus den anzuwenden Rechenoperationen und den dazugehörigen Parametern, nicht mehr nachvollziehbar. [39] Dies gilt insbesondere bei sehr fortgeschrittenen Entscheidungsstrukturen wie beispielsweise neuronale Netzen. Einfachere Entscheidungsstrukturen, etwa Entscheidungsbäume mit einer überschaubaren Anzahl an Parametern, können hingegen – möglicherweise – noch von Menschen interpretiert werden. Die Wahl des Algorithmus spielt also eine zentrale Rolle, wenn eine unmittelbare Interpretierbarkeit durch den Menschen gewährleistet werden soll. Andernfalls bleibt nur eine mittelbare Interpretation durch den Einsatz von spezialisierten Verfahren, die entweder Auskunft über den Einfluss von bestimmten Eingangsvariablen auf das Ergebnis geben, oder die Entscheidungsstrukturen stark vereinfacht darstellen, sodass Menschen diese interpretieren können. [40]

Beispiel: Die Geschäftsführung ist überrascht über die Entscheidung eines KI-Systems, einen Bewerber aufgrund seines Lebenslaufs nicht zu einem Vorstellungsgespräch einzuladen. Die Geschäftsführung sagt gegenüber ihren Mitarbeiter*innen, sie hätte ein „gutes Bauchgefühl“ bei dem Bewerbenden, da dieser ein sehr ungewöhnliches, aber spannendes Profil habe. Sie bittet die Mitarbeitenden herauszufinden, auf Basis welcher Information das KI-System entschieden hat. Die Mitarbeitenden nehmen das KI-System und insbesondere das ML-Modell unter die

[39] Lipton, Zachary C. „The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.“ Queue 16.3 (2018): 31-57.

[40] Molnar, Christoph. “Interpretable machine learning. A Guide for Making Black Box Models Explainable”, 2019.

Lupe. Dem ML-Modell liegt ein neuronales Netz zugrunde, sodass die Mitarbeitenden nicht unmittelbar erklären können, wie die Entscheidung zustande kam. Mit Hilfe von statistischen Methoden, z. B. LIME [41], können sie jedoch herausfinden, dass das neuronale Netz die Information, dass der Bewerbende in einem sehr ungewöhnlichen Fach an einer kleinen Universität im Ausland promoviert hat, mit sehr negativem Einfluss auf die Entscheidung gewichtet hat. Die Geschäftsführung ist darüber zwar nicht erfreut, beschließt jedoch das KI-System weiterhin einzusetzen. Zusätzlich veranlasst sie, dass die drei einflussreichsten Kriterien (Features) bzgl. der Entscheidung des KI-Systems immer mit ausgegeben werden, um die Entscheidungsfindung grob einschätzen zu können.

[41] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

Fehleranfälligkeit / Robustheit: In der Praxis gibt es oftmals große Variabilität in der Datenqualität. Daten sind unvollständig, fehlerhaft und anfällig für Manipulation. Dies kann dazu führen, dass kleinste Abweichungen von Daten zu großen Fehlern und gravierenden Fehlentscheidungen führen. Auch bzw. insbesondere sehr leistungsfähige ML-Modelle sind davor nicht geschützt. Werden solche Manipulationen gezielt vorgenommen, spricht man von sog. Adversarial Attacks [42]. Die Manipulation ist von Menschen oftmals nicht erkennbar. Gerade im Bereich Computer Vision, zum Beispiel bei der Objektklassifizierung durch Algorithmen, werden Bildern gezielt mit einem Rauschen manipuliert, das den Bildinhalt nicht verändert. Die Algorithmen hingegen, die dieses Rauschsignal verarbeiten, werden derartig gestört, dass sie ein falsches Ergebnis erzielen. Analog gilt dies zum Beispiel auch beim Einsatz von KI bei der Analyse von Sprache in Audioaufnahmen. [43]

[42] Xu, H., Ma, Y., Liu, HC. et al. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. Int. J. Autom. Comput. 17, 151–178 (2020).

[43] Sehr anschauliche Beispiele aus dem Bereich Speech-to-text unter https://adversarial-attacks.net/

Um diese Probleme zu lösen und robuste, zuverlässige Algorithmen zu entwerfen, hat sich in den letzten Jahren der Forschungsbereich Robust Machine Learning entwickelt. Mit zunehmender Forschung im Bereich Maschinelles Lernen werden auch in diesem Bereich wichtige Forschungsergebnisse erwartet. Insbesondere in sicherheitskritischen und sensiblen Bereichen ist es unbedingt notwendig, robuste Algorithmen zu verwenden.

Beispiel: Ein Unternehmen möchte das KI-System, das zur Analyse des Lebenslaufs eingesetzt wird, um eine Funktion erweitern, sodass auch das Portraitbild eines Bewerbenden sowie Bilder der jeweiligen Person in den Sozialen Medien (z. B. LinkedIn, Instagram, Facebook) ausgewertet werden. Ziel der Auswertung soll ein Profiling und eine Bewertung sein, ob der Bewerbende aufgrund seiner Freizeitgestaltung zum Unternehmen passt oder nicht. Ein Tech Start-up verspricht, hierzu einen sehr leistungsfähigen Softwarebaustein über einen Cloud Service bereitstellen zu können. Nachdem der erste Prototyp fertiggestellt ist und die interne IT-Abteilung die zusätzliche Funktion ausgiebig testet, zeigt sich, dass die Bilderkennung sehr fehle-

ranfällig und wenig robust ist. Bei fiktiven Bewerbenden wurden absichtlich manipulierte Bilder hinterlegt und die Software konnte sehr leicht getäuscht werden. Die Ergebnisse waren nicht zuverlässig. Die Geschäftsführerung stoppt das Projekt nach Rücksprache mit der IT-Abteilung.

3.6. Ebene 5: ADM Entscheidungen

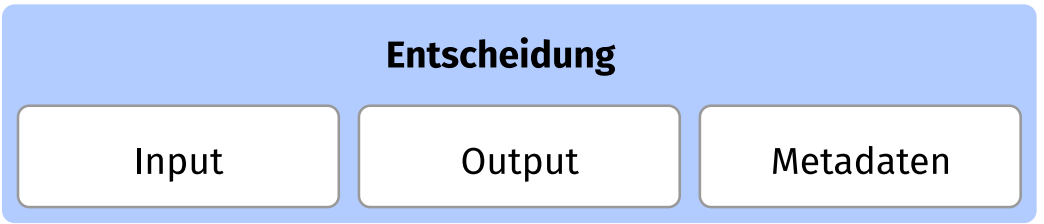


Abbildung 6: Input, Output und Metadaten sind die entscheidenden Elemente der Ebene 5 „Entscheidungen“

Entscheidung: Aus den Entscheidungsstrukturen, die wiederum in mehrere übergeordnete Ebenen eingebettet sind, ergeben sich die Entscheidungen eines KI-Systems. Bei den Entscheidungen handelt es sich um die Ergebnisse eines KI-Systems, die an den Benutzenden zurückgemeldet werden. Die Vorgänge auf den Ebenen 1-4 sind den Anwendenden meistens nicht zugänglich. Diese sind während des Erstellprozesses und während der Weiterentwicklung wichtig und notwendig. Die Entscheidung hingegen wird an den Endanwender zurückgespielt. Diese muss nicht notwendigerweise unmittelbar durch Algorithmen und den entwickelten Entscheidungsstrukturen mitgeteilt werden. Entscheidungen können auch mittelbar, z. B. durch eine dritte Person, mitgeteilt werden. Ein bedeutender Vorteil des dezidierten Trennens der Entscheidungsstruktur von der Entscheidung im Kontext von Auditierung und Testing liegt in der verbesserten und angemesseneren Analyse der zugrundeliegenden Vorgänge. Dies zeigt sich vor allem dadurch, dass es zahlreiche Methoden gibt, die Entscheidungsstrukturen unabhängig von einer konkreten Entscheidung zu analysieren bzw. transparent zu machen.

Eine Entscheidung ist der Vorgang, in dem eine Entscheidungsstruktur auf Basis von Eingabedaten (Input) eine Ausgabe (Output) hervorbringen. Dabei können noch diverse Metadaten berücksichtigt bzw. erzeugt werden. Diese sind aus zwei Gesichtspunkten relevant: einerseits können Metadaten auf der Inputseite berücksichtigt und

damit in den Entscheidungsvorgang einbezogen werden. Andererseits können Metadaten zusätzlich zur Ausgabe einer Entscheidung entstehen. Dies ist zum Beispiel der Fall, wenn zusätzlich zu einer Entscheidung auch noch die Konfidenz berechnet wird, mit der eine Entscheidung getroffen wurde.

Im Kontext der Entscheidungen spielen drei Bereiche eine zentrale Rolle, die im Rahmen von Auditierung und Testen fokussiert werden:

1. Eingabedaten (Input)
2. Ausgabedaten (Output)
3. Metadaten [44]

[44]

Schelter, Sebastian, et al. "Automatically tracking metadata and provenance of machine learning experiments." Machine Learning Systems Workshop at NIPS. 2017.

Eingabedaten (Input): Die der Entscheidungsstruktur zur Entscheidungsfindung zur Verfügung stehenden Eingabedaten (Input) sind ein wesentlicher Bestandteil, um eine Entscheidung bzw. das Verhalten eines KI-Systems zu bewerten. Die Eingabedaten sind von den Datenbeständen in Ebene 3 abzugrenzen. Während die Datenbestände in Ebene 3 sich auf den Trainings- und Entwicklungsprozess beziehen, werden die Daten, die auf Ebene 5 genutzt werden auf die eigentliche Entscheidungsfindung gerichtet.
Es gibt KI-Systeme, in denen die Eingabedaten der Ebene 5 auch zum Training verwendet werden (z. B. bei selbst-lernenden Systemen), dann sind jedoch die Prüf- und Analysekriterien aus der entsprechenden Ebene zu verwenden. Relevante Aspekte der Ebene 5 zielen insbesondere auf die Bereitstellung und auf die Richtigkeit der Eingabedaten ab. Eine zentrale Frage hierbei ist, ob Daten von Benutzern direkt bereitgestellt, also eingegeben werden, oder ob Daten vorher noch verarbeitet und mit zusätzlichen Daten angereichert werden. Die durchgeführten Verarbeitungsschritte müssen im Zweifelsfall protokolliert werden, damit diese nachvollzogen werden können. Da für die Nachvollziehbarkeit einer Entscheidung die Eingabedaten relevant sind, muss im Kontext der Auditierung auch geprüft bzw. festgelegt werden, ob Eingabedaten gespeichert werden müssen. Das bedeutet, dass die Eingabedaten protokolliert werden müssen, z.B. in einer Datenbank. Technisch kann dies einen nicht unerheblichen Mehraufwand bedeuten. Rechtlich muss eine Anforderung an das Speichern noch gesondert bewertet werden. Dies kann auch an weitere Anforderungen geknüpft werden, z.B. KI-Systeme in besonders kritischen Bereichen.

Beispiel: Der Geschäftsführer eines Unternehmens möchte den beliebten Chatbot für Personalanfragen weiterentwickeln lassen. Mehrere Möglichkeiten stehen hierfür zur Verfügung. Er bittet seine Mitarbeiter eine Auswertung zu erstellen, zu welchen Themen der Chatbot besonders häufig angefragt wird. Der Mitarbeiter wertet

die Eingabedaten (Input) aus, die sehr umfangreich in einer verschlüsselten Datenbank gespeichert werden. Bei der Auswertung stellt der Mitarbeiter fest, dass dem Chatbot viele Daten bereitgestellt werden, die einerseits personenbezogen sind, wie beispielsweise Name und E-Mail-Adresse eines Anfragestellers, und die andererseits nicht für die Entscheidungsfindung relevant sind. Nach Rücksprache mit dem Geschäftsführer und dem Betriebsrat einigt man sich darauf, dass die Daten sparsamer erhoben und für zukünftige Anfragen keine personenbezogenen Daten mehr verwendet werden sollen.

Ausgabedaten (Output): Die Ausgabedaten (Output) stehen oft im Zentrum des Diskurses von KI-Systemen und insbesondere der algorithmischen Entscheidungsfindung. Dies ist nachvollziehbar, weil das Ergebnis einer Entscheidung in manchen Fällen nicht unstrittig und für den Betroffenen unangenehm sein kann. Das Ergebnis ist dasjenige, was oftmals in Frage gestellt wird und im Kontext dieses Artikels wird klar, dass sich das Ergebnis aus der Vielzahl von größeren und kleineren Entscheidungen entlang des Entwicklungsprozesses ergibt. Auch wenn am Ende für den Betroffenen oft nur das Ergebnis einer Entscheidung zählt. Analog zu den Eingabedaten können auch die Ausgabedaten protokolliert werden (Logging). Die strukturierte Gegenüberstellung von Eingabedaten und Ausgabedaten (Input und Output) wird in vielen Testverfahren angewandt. [45] Eingabedaten lassen sich synthetisch erzeugen, sodass das Testen in dem Fall einem Art Experiment gleicht, welche die Entscheidungsstruktur zwingt zu einem bestimmten Input ein Output zu generieren. Dieses lässt sich dann mit einem – nicht notwendigerweise vorab – festgelegten Ergebnis vergleichen, um auf ein mögliches Fehlverhalten rückschließen zu können. Entscheidungsstrukturen, z.B. die Ausgabedaten eines ML-Modells sind in der Regel reproduzierbar. [46] [47] Das heißt, dass ein ML Modell, das nicht verändert wurde, auf einen gleichen Input den gleichen Output produziert. Diese Annahme liegt vielen Testverfahren, die geeignet sind, um das Verhalten eines KI-Systems zu analysieren, zugrunde. Es mag der Einwand aufkommen, dass dies gerade bei selbst-lernenden System nicht mehr der Fall ist, hierbei ist jedoch zu beachten, dass sich bei selbst-lernenden Systemen das zugrundeliegende ML Modell ständig weiterentwickelt. Die Entscheidungsstruktur verändert sich also. Damit ist die Annahme, dass das ML Modell unverändert bleibt, nicht mehr gültig. Der Zustand, in dem sich eine Entscheidungsstruktur befindet, ist also ein entscheidender Faktor, der beim Vergleich von Eingabe- und Ausgabedaten nicht vernachlässigt werden darf.
Um ein KI-System zu verstehen bzw. um eine Entscheidung nachvollziehen zu können ist die Analyse, die sich ausschließlich auf den Output konzentriert, weder sinnvoll noch zielführend. Die Kontextinformation auf den Ebenen 1 – 5 muss berücksichtigt werden.

[45]

Segura, Sergio, et al. “A survey on metamorphic testing.” IEEE Transactions on software engineering 42.9 (2016): 805-824.

[46]

Tatman, Rachael, Jake VanderPlas, and Sohier Dane. “A practical taxonomy of reproducibility for machine learning research.” (2018).

[47]

An der Stanford University wurde 2021 ein eigenes “Center for Open and REproducible Science” (CORES) gegründet. Das Ziel „develop and nurture transparency and reproducibility in the collection, analysis, and dissemination of data across all domains of scientific activity.“

Beispiel: Aufmerksamen Mitarbeiter:innen fällt auf, dass der Chatbot eines Unternehmens, der gerade im Bereich Personalauskünfte sehr beliebt ist, für eine Anfrage ein falsches Ergebnis liefert. Sie bitten die Kollegen in der IT dies zu untersuchen. Die Kollegen können das Verhalten nachvollziehen. Sie passen die Anfrage in einem kontrollierten Experiment an, in dem sie unterschiedliche Begriffe verwenden, um dem Chatbot die leicht angepasste Anfrage immer wieder zu stellen und die Veränderungen am Output des Chatbots zu untersuchen. Durch diesen Vorgang können sie die Fehlerquelle sehr genau eingrenzen und stellen fest, dass es bestimmte Begriffe gibt, die der Chatbot offensichtlich nicht versteht. Dabei handelt es sich um sehr unternehmensspezifische Abkürzungen, welche jedoch leicht im Vokabular des Chatbots mitaufgenommen werden können. In einer Rücksprache mit dem Leiter der IT-Abteilung wird festgelegt, dass in Zukunft ein Testdatensatz angelegt werden soll. Dieser Testdatensatz soll gängige Anfragen (Input) an den Chatbot und die zu erwartenden Antworten (Output) umfassen. Damit soll eine hohe Qualität auch bei Weiterentwicklungen gewährleistet sein.

Metadaten: Eine Entscheidung eines KI-Systems ist nicht nur auf die Berechnung bzw. das Produzieren eines Ergebnisses beschränkt. Wie oben bereits festgestellt, ist das konkrete Ergebnis für das Verstehen und Nachvollziehen einer Entscheidung in vielen Fällen nur wenig hilfreich. Entscheidungsstrukturen generieren zusätzlich zu einer Entscheidung auch Metadaten, die für das nachträgliche Bewerten oftmals sehr hilfreich sind. Die konkreten Metadaten sind je nach Anwendungsfall, eingesetztem Algorithmus und sich ergebender Entscheidungsstruktur anzupassen bzw. festzulegen. Beim Einsatz von maschinellen Lernverfahren können die Metadaten beispielsweise die Konfidenz einer algorithmischen Entscheidung beinhalten. [48] Solche Metadaten können aber auch noch zusätzlich generiert werden, zum Beispiel durch den Einsatz von statistischen Methoden und Tests. Dies könnte insbesondere in Anwendungsbereichen eingesetzt werden, in denen der Einfluss von bestimmten Parametern auf eine Entscheidung automatisiert festgestellt werden muss.

[48] Towards Data Science Artikel "How to use confidence scores in machine learning models" 19.01.2021.

Metadaten können aber auch weitere Informationen zu einer Entscheidung beinhalten, die für die Qualitätssicherung relevant sein könnten. Beispielsweise welches ML-Modell oder welche Modellversion für die Berechnung einer Ausgabe verwendet wurde. Gerade in großen, komplexen Systemen in denen parallel unterschiedliche Entscheidungsstrukturen getestet werden, sogenanntes A/B Testing, bei Suchmaschinen oder im E-Commerce, können in Metadaten wertvolle Informationen enthalten sein. Darüber hinaus kann in den Metadaten auch gespeichert werden,

ob und welche Verfahren zur Vor- und Nachbearbeitung (Pre- und Post-Processing) eingesetzt werden.

Beispiel: Die Mitarbeiter eines Unternehmens machen sehr gute Erfahrung mit dem Chatbot. Das Vertrauen ist auch durch die kontinuierliche Weiterentwicklung immer weiter gestiegen. Jedoch kommt es immer wieder zu Verwirrungen, weil der Chatbot manche Anfragen falsch versteht. Bis diese neuen Anfragen verstanden werden, dauert es immer ein paar Wochen, weil die Logik von der IT-Abteilung erweitert werden muss. Der Leiter der IT-Abteilung schlägt vor, dass bei den Entscheidungen des zugrundeliegende ML-Modell, das für die Analyse der Eingabe eingesetzt wird, auch der sog. Confidence-Score ausgegeben werden soll. Der Score soll dem Mitarbeiter ein Indikator für die Zuverlässigkeit der Antwort sein. Ein niedriger Score ist indiziell für eine falsche oder unvollständige Antwort. Die Mitarbeiter halten das für eine sehr gute Idee und sie gehen noch weiter: Antworten mit einem sehr niedrigen Score sollen erst gar nicht ausgegeben werden. In dem Fall soll der Chatbot darauf hinweisen, dass er keine passende Antwort zu der Anfrage kennt. Das Auswerten und der Einsatz der Metadaten führen also zu einer erhöhten Zufriedenheit und verbesserten User Experience.

4.
Auditieren und Self Assessments von KI-Systemen

4.1. Übersicht

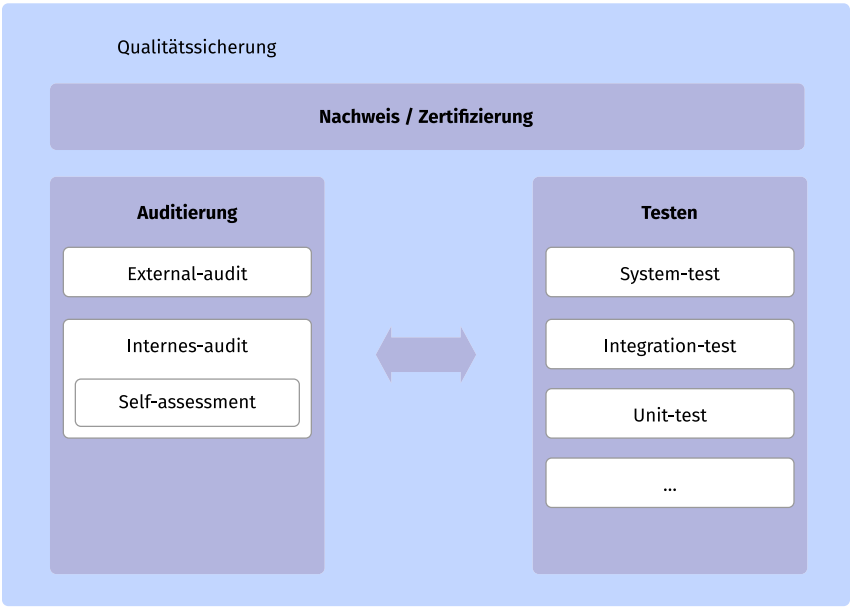


Abbildung 7: Qualitätssicherung von KI-Systemen als Zusammenspiel von Auditierung, Testen und Zertifizierung

Zur Sicherstellung von Qualität und zum Nachweis der An- bzw. Abwesenheiten von bestimmten Eigenschaften gibt es bereits etablierte Methoden, die sich mit ihren Eigenschaften ergänzen und im Zusammenspiel zu einem geeigneten Mittel des Qualitätsmanagements werden. In anderen Bereichen und Ländern, beispielsweise der Finanzwirtschaft, sind die Bereiche Auditierung, Testen sowie Zertifizierung etabliert und auch entsprechend in den Gesetzen verankert . [49]

[49] Comptroller of the Currency Administrator of National Bank “Internal and External Audits: Comptrollers Handbook“, 2003

Im Kontext der vorliegenden Arbeit soll insbesondere auf die **Auditierung** (siehe Abschnitt 4.2 bis 4.7) detaillierter eingegangen werden. Bei genauer Betrachtung lässt sich die Auditierung in zwei große Bereiche zergliedern: das externe Audit (siehe Abschnitt 4.2 und 4.3) einerseits und das interne Audit (siehe Abschnitt 4.4) andererseits. Letzteres kennt wiederum eine weitere Sonderform, die Selbsteinschätzung (engl. Self Assessment) (siehe Abschnitt 4.5).

Das **Testen** ist insbesondere bei Softwaresystemen bzw. software-intensiven Systemen ein gängiges Verfahren zur Sicherstellung definierter Anforderungen und Messung der Qualität. Dies kann auch für Software und Algorithmen im Bereich KI angewandt werden. [50]

Der übergeordnete Bereich des **Nachweises und der Zertifizierung** ist der Bestandteil des Qualitätsmanagements, der über eine erfolgreiche oder nicht erfolgreiche Durchführung eines Audits bzw. eines Testverfahrens Auskunft gibt. Unter bestimmten Umständen können Zertifikate ausgestellt werden, die eine Art Prüfsiegel darstellen, welches bestätigt, dass bestimmte Standards (Konsens) bei der Überprüfung eingehalten wurden. Darunter fällt in vielen Fällen auch, dass die überprüfende Organisationseinheit akkreditiert ist. Das bedeutet, dass sie die technischen, organisatorischen und prozessualen Fähigkeiten besitzt, den Standard bei der Überprüfung einzuhalten. Diese Akkreditierung wird von Akkreditierungsstellen ausgestellt bzw. bescheinigt. Im Bereich KI in der Arbeitswelt gibt es bereits einige Standards von nationalen und internationalen Unternehmen und Behörden. [51]

[50]

Jöckel, Lisa, et al. „Towards a Common Testing Terminology for Software Engineering and Artificial Intelligence Experts.“ arXiv preprint arXiv:2108.13837 (2021).

[51]

Bundesministerium für Arbeit und Soziales „KI in der Arbeitswelt: Potenziale erkennen, Transparenz schaffen “, Zugriff am 13.10.2021.

4.2. Audit und Audit Programm

Definition: Nach dem IEEE Standard 1028-2008 handelt es sich bei (Software) Audits¬ um einen „systematischen, unabhängigen und dokumentierten Prozess, um objektive Nachweise zu erlangen und diese objektiv zu bewerten, um festzustellen, inwiefern die Prüfkriterien erfüllt sind“. [52]

[52]

Aus dem engl. Original. “an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures.“ “IEEE Standard for Software Reviews and Audits,” in IEEE Std 1028-2008 , pp.1-53, 15 Aug. 2008, doi: 10.1109/IEEESTD.2008.4601584.

Das Audit nach IEEE zielt in seiner Definition auf drei zentrale Aspekte ab:

1. Der systematische, unabhängige und dokumentierte Prozess: Das zentrale Element der Durchführung eines Audits ist die Arbeit und Vorgehensweise nach einer vorab festgelegten Systematik, die klar und eindeutig ist, sodass der Interpretationsspielraum so groß wie notwendig, aber klein wie möglich ist. Das Audit soll unabhängig sein. Es soll sich um einen eigenständigen Prozess, der nicht in den Entwicklungs- und Implementierungsprozess eingebunden ist, handeln. Der Grund hierfür ist, das Interessens- und Zielkonflikte minimiert werden. Die Dokumentation des Prozesses ist die Grundlage für die Durchführung und ermöglicht die Entkopplung des Audits und der handelnden Personen. Sie leistet einen entscheidenden Beitrag zu Unabhängigkeit eines Audits.

2. Der objektive Nachweis: Nach dem IEEE-Standard soll das Audit vor allem dazu dienen Fakten, Evidenzen und Beweismittel zu erheben, die bei der objektiven Bewertung einer (nicht) Konformität hinsichtlich der Prüfkriterien helfen.
3. Die objektive Bewertung und Feststellung der Erfüllung von Prüfkriterien: Auf Basis der objektiven Nachweise soll das Audit auch eine objektive Bewertung auf Basis von ex ante definierten Prüfkriterien abgeben. Dabei sollen Fakten anhand von zwei Arten klassifiziert werden: bedeutend (major) und unbedeutend (minor). Bedeutende Fakten sind dadurch gekennzeichnet, dass sie einen signifikanten Einfluss auf die Qualität haben. Bedeutende Fakten sind sogenannte „Findings".

Werden Audits in einem größeren und strategischen Zusammenhang durchgeführt ist es notwendig, die Audits mit ihren unterschiedlichen Zielsetzungen an einem gemeinsamen und übergeordneten Ziel auszurichten. Zu diesem Zweck existieren **Auditprogramme**. Das Auditprogramm ist vom Auditplan zu unterscheiden. Während der Auditplan eine Beschreibung der durchzuführenden Tätigkeiten und Abläufe innerhalb eines Audits darstellt ist das Auditprogramm die übergeordnete Einordnung, Koordination und Planung einer Menge bzw. einer Serie von Audits. Das Auditprogramm ist von drei Kernaspekten geprägt:

- Strategische Einordnung und Ausrichtung: Hierunter fällt die Berücksichtigung von Unternehmenszielen und Anforderungen, die über das zu auditierende KI-System hinausgehen. Unter den Einflussfaktoren, die hierbei berücksichtigt werden können, sind z.B. die Digitalisierungs- und Datenstrategie eines Unternehmens, IT-Infrastrukturziele (Cloud und On-Prem) und übergeordnete IT-Projekte (Cloud Data Hubs, Lakes und Warehouses).
- Abstimmung und Koordination: Hierunter fällt die Berücksichtigung aller geplanten Audits und insbesondere die Abstimmung der jeweils verfolgten Ziele und Kernbereiche der jeweiligen Audits. Hierbei können Schwerpunkte, z.B. Daten und Datenbestände, Interpretierbarkeit von ML Modellen, etc., gesetzt und berücksichtigt werden.
- Planung und Vorbereitung: Hierunter fällt die Berücksichtigung der konkreten für die Durchführung relevanten Aspekte wie beispielsweise die Triggerevents für Audits und die für die Durchführung und Vorbereitung notwendigen Ressourcen. Ob die Audits in regelmäßigen Zeitabständen oder bei Bedarf durchgeführt werden und welche Ressourcen, insb. Personal und Infrastruktur, benötigt und zu welchem Zeitpunkt diese bereitgestellt werden.

4.3. KI-Audit

Definition: Ein KI-Audit beschreibt die Durch- bzw. Ausführung eines Audits zur ganzheitlichen Überprüfung eines KI-Systems, eines Subsystems davon (hierunter fallen auch Datenbestände) oder interner Kontrollmechanismen unter Verwendung messbarer Prüfkriterien. Das KI-Audit ist der Spezialfall eines Audits, welcher den Aufbau, die Komponenten und die Charakteristiken von KI-Systeme berücksichtigt.

Das KI-Audit stellt die Anwendung von Audits mit den bewährten, erprobten und akzeptierten Verfahren und Vorgehensweise auf KI-Systeme dar. Dabei werden die Charakteristiken von KI-Systemen berücksichtigt. Dies betrifft insbesondere den Aufbau von KI-Systemen (siehe 5 Ebenen von KI-Systemen in Abschnitt 3) sowie das komplexe Zusammenspiel von verschiedenen Komponenten und Subsystemen, die in Kombination mehr ergeben als sich durch die Analyse der Subsysteme zeigt. Das prominenteste Beispiel ist das Zusammenspiel von Daten und Algorithmen: weder durch die isolierte Betrachtung und Analyse des Datensatzes noch durch den vollständig beschriebenen Algorithmus ist das Verhalten einer Entscheidungsstruktur bzw. eines trainierten ML-Modells vorhersehbar. Erst die Kombination von Algorithmus und Daten legt die im Algorithmus angelegten Freiheitsgrade (Variablen) fest. Das Verhalten der Entscheidungsstruktur ist die Emergenz die das Zusammenspiel von Subsystemen ergibt.

Gegenstand eines KI-Audits ist deshalb notwendigerweise das KI-System in seiner Gesamtheit. Die Charakteristiken des KI-Audits betreffen darüber hinaus die Aus- und Durchführung des Audits unter Einbeziehung der fünf Ebenen. Die einzelnen Ebenen und die Komponenten darin können je nach Zielsetzung unterschiedlich stark berücksichtigt werden. Eine Vernachlässigung führt jedoch zu einer unvollständigen Analyse.

Ein KI-Audit bedarf der entsprechenden Ressourcen und methodischen Kompetenzen. Als zusätzliche Herausforderung zur Auditierung eines herkömmlichen Softwaresystems ist es vor allem die Datenzentriertheit von KI-Systemen, die das KI-Audit besonders komplex macht. Die Analyse großer Datenbestände muss möglich sein und erfordert deshalb auch entsprechende Werkzeuge (Big Data) und Expertisen (Data Literacy, Statistik, Data Analytics, Data Science, usw.).

Das KI-Audit kann weiter unterteilt werden in ein externes und internes KI-Audit (siehe Abschnitt 4.4 und 4.5) sowie in ein Self Assessment (siehe Abschnitt 4.6). Die konkrete Zusammenführung von Aspekten und Kriterien des Audits mit den fünf Ebenen aus Abschnitt 3 wird in Abschnitt 5 durchgeführt.

4.4. Externes KI-Audit

Ein externes KI Audit beschreibt die Durch- bzw. Ausführung eines Audits zur ganzheitlichen und vollständigen Überprüfung eines KI-Systems, eines Subsystems davon (hierunter fallen auch Datenbestände) oder interner Kontrollmechanismen unter Verwendung messbarer Prüfkriterien durch eine Organisationseinheit, die sich außerhalb der auditierten Organisation befindet. Ausgelagerte oder mitvergebene interne Prüfungsaktivitäten werden nicht als Teil eines externen Audits betrachtet.

Das externe KI-Audit ist dadurch gekennzeichnet, dass sich die überprüfende Organisationseinheit außerhalb, das bedeutet organisatorisch getrennt und ohne Einfluss von der auditierten Organisation, befindet. Bei der Erstellung des Auditplans ist dies zu berücksichtigen, weil dies Auswirkungen auf die Bereitstellung von Information, darunter auch Daten, hat.

4.5. Internes KI-Audit

Ein internes KI Audit beschreibt die unabhängige und objektive Prüfung und Bewertung eines Systems oder eines Subsystems davon (hierunter fallen auch Datenbestände) unter Verwendung messbarer Prüfkriterien durch eine Organisationseinheit, die sich nicht notwendigerweise außerhalb der zu auditierenden Organisation befindet (interne Revision).

Das interne KI-Audit ist dadurch gekennzeichnet, dass sich die überprüfende Organisationseinheit nicht außerhalb von der auditierten Organisation befindet. Das interne KI-Audit kann deshalb auch von der internen Revision oder einer anderen akkreditierten Organisationseinheit durchgeführt werden.

4.6. KI-Self-Assessment als KI-Audit

Ein KI-Self-Assessment beschreibt die systematische Erfassung von Eigenschaften und Verhalten eines Systems (oder Subsystems) durch messbare Prüfkriterien sowie die Bewertung dieser Eigenschaften durch die Organisationseinheit, die auch die Konzeption und (Weiter-)Entwicklung des KI-Systems verantwortet.

Das KI-Self-Assessment (dt. Selbsteinschätzung) ist eine Sonderform des internen KI-Audits. Die Bewertung läuft ebenfalls nach einer ex ante festgelegten Systematik ab und die Prüfkriterien müssen weiterhin messbar und feststellbar sein. Die Prüfkriterien sind wiederum so auszulegen, dass das KI-System ganzheitlich und vollständig bewertet wird. Das schließt also auch Subsysteme (darunter fallen auch Datenbestände) ein. Die Prüfkriterien müssen so gewählt sein, dass auch ein Rückschluss auf die emergenten Eigenschaften möglich ist.

Im Gegensatz zum externen oder herkömmlichen internen Audit liegt die Verantwortung der Durchführung der Bewertung bei der Organisationseinheit, die auch für die Entwicklung des KI-Systems zuständig ist. Dies hat den Nachteil, dass ein Interessenskonflikt der handelnden Personen nicht ausgeschlossen werden kann. Daher genießt das KI-Self-Assessment nicht notwendigerweise die gleiche Glaubwürdigkeit, wie ein externes KI Audit oder das herkömmliche interne Audit. Die Organisationseinheit, die die Konzeption und (Weiter-)Entwicklung verantwortet, kann das KI-Self-Assessment (oder Teile davon) auch delegieren und z.B. an einen Zentralbereich übergeben. Je nachdem Ausgestaltung der Übergabe und bleibt es jedoch ein Self Assessment, auch wenn die Bewertung der Prüfkriterien delegiert wurde. Entscheidend ist die Planung und Vorbereitung sowie die Letztverantwortung, die im Rahmen des Self Assessment nicht delegiert oder aufgeteilt werden kann. Die Kerneigenschaft des Self Assessment ist, dass die Verantwortung bei der Organisationseinheit liegt, die auch die Konzeption und (Weiter-)Entwicklung des KI-Systems verantwortet.

Ein bedeutender Vorteil ist jedoch der verhältnismäßig geringe organisatorische Aufwand (Overhead), mit dem ein KI-Self-Assessment durchgeführt werden kann. Ein weiterer Vorteil ist auch, dass die Prüfkriterien den Projektmitarbeitern bekannt sind und sie dadurch eine Sensibilisierung für derartige Fragestellungen im Rahmen der Projektarbeit erfahren. Darüber hinaus können die Prüfkriterien von dem Projektteam angepasst und fortgeschrieben werden, falls bei der Bewertung auffällt, dass die Kriterien nicht adäquat sind. Der Kriterienkatalog, z.B. Checkliste, wird somit zu einem impliziten Bestandteil der Entwicklung des KI-Systems. Ausufernde und schwerfällige Self-Assessment-Verfahren können hingegen wieder schnell als zusätzlicher Overhead wahrgenommen werden. Es ist auf die Angemessenheit, also das richtige Verhältnis zwischen Nutzen und Aufwand, zu achten.

Tabelle 1: Fünf Vor- und Nachteile eines Self Assessments im Kontext von KI-Systemen

Vorteile	Nachteile
Geschäfts- und Betriebsgeheimnisse müssen nicht offengelegt werden	Interessenskonflikte bei der Bewertung der Prüfkriterien
Sensibilisierung des Projektteams für kritische Fragestellungen und Prüfkriterien vor und während der Projektdurchführung	KI-Self-Assessment kann nicht den Stellenwert (Glaubwürdigkeit) wie externes KI-Audit haben
Prüfkriterien können vom Projektteam angepasst und für das vorliegende Projekt präzisiert werden	Zusätzlicher Aufwand für das Projektteam (Overhead)
Flexible Einbindung des Assessments in den Projektverlauf	Zusätzliche Fähigkeiten müssen aufgebaut und bereitgehalten werden
Assessmentergebnisse stehen externen Dritten nicht zur Verfügung	Aktualisierung der Prüfkriterien erfordert Aufwand und Ressourcen

4.7. Fallstudie: Assessment List for Trustworthy Artificial Intelligence

Am 17. Juli 2020 stellte die High-Level Expert Group on Artificial Intelligence (HLEG AI) eine Self-Assessment-Liste für Trustworthy AI vor. [53] Die Liste ist das Ergebnis einer Studie, die vom 26. Juni 2019 bis zum 1. Dezember 2019 durchgeführt wurde und an der sich über 350 Teilnehmer beteiligten. Auf Basis des Feedbacks wurde die Assessment Liste überarbeitet und in ein Web-Tool überführt. Das Web-Tool stellt einen interaktiven Fragebogen dar, der die die Zielgruppen „AI Developers“ und „AI Deployers“ adressiert. Das Tool ist online verfügbar und nach Registrierung ohne zusätzliche Kosten benutzbar. [54]

[53] European Commission, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, 2021.

[54] Stand September 2021

4.7.1. Aufbau und Struktur

Das Self Assessment orientiert sich an den „Ethics Guidelines for Trustworthy AI“ der HLEG AI und insbesondere an den sieben Kernanforderungen einer vertrauenswürdigen KI: [55]

1. Vorrang menschlichen Handelns und menschliche Aufsicht
2. Technische Robustheit und Sicherheit

[55] HLEG AI: High-Level Expert Group AI, „Ethik-Leitlinien für eine vertrauenswürdige KI“, 2018.

3. Datenschutz und Datenqualitäts-management
4. Transparenz
5. Vielfalt, Nichtdiskriminierung und Fairness
6. Gesellschaftliches und ökologisches Wohlergehen
7. Rechenschaftspflicht

Die Liste für das Self Assessment ist nach an den sieben Kernanforderungen in Abschnitte strukturiert. Innerhalb dieser Abschnitte befindet sich eine kurze Einleitung zu dem jeweiligen Abschnitt, eine Einordnung des Abschnitts in den Gesamtkontext, eine Gliederung in Unterabschnitte, sowie Fragestellungen, deren Beantwortung über den Stand des KI-Systems in der Dimension Auskunft geben.

Tabelle 2: Abschnitte und Unterabschnitte des Self Assessments der HLEG AI.

Abschnitt	Unterabschnitte (Verwendung englischer Originalbegriffe)
Vorrang menschlichen Handelns und menschliche Aufsicht	1. Human Agency and Autonomy 2. Human Oversight
Technische Robustheit und Sicherheit	1. Resilience to Attack and Security 2. General Safety 3. Accuracy 4. Reliability, Fall-back plans, and Reproducibility
Datenschutz und Datenqualitäts-management	1. Privacy 2. Data Governance
Transparenz	1. Traceability 2. Explainability 3. Communication
Vielfalt, Nichtdiskriminierung und Fairness	1. Avoidance of Unfair Bias 2. Accessibility and Universal Design 3. Stakeholder Participation
Gesellschaftliches und ökologisches Wohlergehen	1. Environmental Well-being 2. Impact on Work and Skills 3. Impact on Society at large or democracy
Rechenschaftspflicht	1. Auditability 2. Risk Management

4.7.2. Inhalte

Die Hauptinhalte der sieben Abschnitte sind die Fragen, die bewusst als offene Fragen mit Interpretationsspielraum formuliert sind. Beispielfragen sind in Tabelle 1 dargestellt. Die Fragenliste des Self Assessment ist derart aufgebaut, dass Fragen auch Unterfragen, die zur Präzisierung oder zur Lenkung der Antwort in eine bestimmte Richtung beitragen.

Tabelle 3: Beispielfragen zu den sieben Abschnitten des Self Assessment der HLEG AI (Quelle: HLEG AI)

Abschnitt	Ausgewählte Fragen zur Illustration (Quelle: HLEG AI [56])
Vorrang menschlichen Handelns und menschliche Aufsicht	• Could the AI system generate confusion for some of all end-users or subjects on whether they are interacting with a human or AI system? • Is the AI system designed to interact, guide, or take decisions by human end-users that affect humans or society? • Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way?
Technische Robustheit und Sicherheit	• Did you put measures in place to ensure the integrity, robustness, and overall security of the AI system against potential attacks over its lifecycle? • Could the AI system have adversarial, critical or damaging effects (e.g., to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use? • Did you put in place a series of steps to monitor, and document the AI system's accuracy?
Datenschutz und Datenqualitäts-management	• Is your AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)? • Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data? • Did you align the AI system with relevant standards (e.g., ISO, IEEE) or widely adopted protocols for (daily) data management and governance?

[56] Fragen wurden aus dem offiziellen Dokument (doi: 10.2759/791819) entnommen und werden Originalsprache dargestellt.

Abschnitt	Ausgewählte Fragen zur Illustration (Quelle: HLEG AI [56])
Transparenz	• Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system? • Can you trace back which data was used by the AI system to make a certain decision(s) or recommendation(s)? • Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?
Vielfalt, Nichtdiskriminierung und Fairness	• Did you put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system? • Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI system? • Where relevant, did you consider diversity and representativeness of end-users and or subjects in the data?
Gesellschaftliches und ökologisches Wohlergehen	• Could the AI system have a negative impact on society at large or democracy? • Does the AI system impact human work and work arrangements? • Could the AI system create the risk of de-skilling of the workforce?
Rechenschaftspflicht	• Did you establish mechanisms that facilitate the AI system's auditability (e.g., traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)? • Did you ensure that the AI system can be audited by independent third parties? • Did you consider establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas?

Insgesamt umfasst das Self-Assessment-Tool eine Vielzahl von Fragestellungen für die Abschnitte und Unterabschnitte. Die Antwortmöglichkeiten sind

1. Yes
2. To some extent
3. No
4. Don't know

Einige der Fragen bieten, je nach Antwort auch noch ein Freitextfeld an, das zu Beschreibung bzw. zur Begründung der gewählten Antwort genutzt werden kann.

4.7.3. Umsetzung und Auswertung

Das Self Assessment ist in zwei verschiedenen Formaten verfügbar:

1. Frei verfügbares PDF-Dokument [57]
2. Web-basierte Applikation mit interaktiven Elementen [58]

Auf der offiziellen Webseite des Self Assessement wird ein PDF-Dokument zum freien Download angeboten. Dieses Dokument beinhaltet alle Fragen, sowie eine kurze Einführung zu dem Thema und der Zielsetzung. Zur Klarstellung der Terminologie wird auch ein Glossar bereitgestellt, das wesentliche Begriffe und Phrasen erläutert. Die Antwortmöglichkeiten sind in dem PDF-Dokument jedoch nicht vorgegeben.

Die web-basierte Umsetzung des Self Assessment stellt neben dem PDF-Dokument die zweite Variante zur Anwendung dar. Die web-basierte Variante kann nach Registrierung ohne zusätzliche Kosten verwendet werden (Stand September 2021). Inhaltlich sind beide Formate nicht komplett deckungsgleich. Bei der web-basierten Version erscheinen, je nach abgegebener Antwort, zusätzliche weitere Fragen, die zur Präzisierung der Antwort beitragen. Die web-basierte Version bietet darüber hinaus interaktive Elemente an, die bei dem Ausfüllen des Fragebogens hilfreich sein können. So sind zum Beispiel einige Fragen mit interaktiven Schaltflächen versehen, über die weitere Hilfestellungen und Erläuterungen zu einer Frage erscheinen.
Als Ergebnis der web-basierten Applikation erhält man als Benutzer eine Übersicht über die abgegebenen Einschätzungen und Antworten auf die Fragen. Eine grafische Visualisierung als Netzdiagram bereitet die viel-dimensionale Information übersichtlich auf und zeigt Stärken und Schwächen des bewerteten KI-Systems auf.

[57] European Commission, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, 2021.

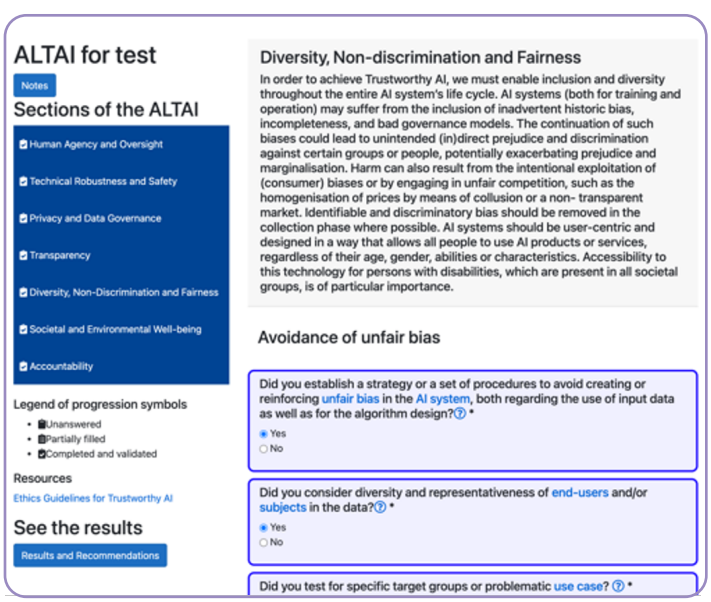
[58] Assessment List for Trustworthy Artificial Intelligence, Online Tool verfübar unter https://altai.insight-centre.org, Zugriff am 13.10.2021

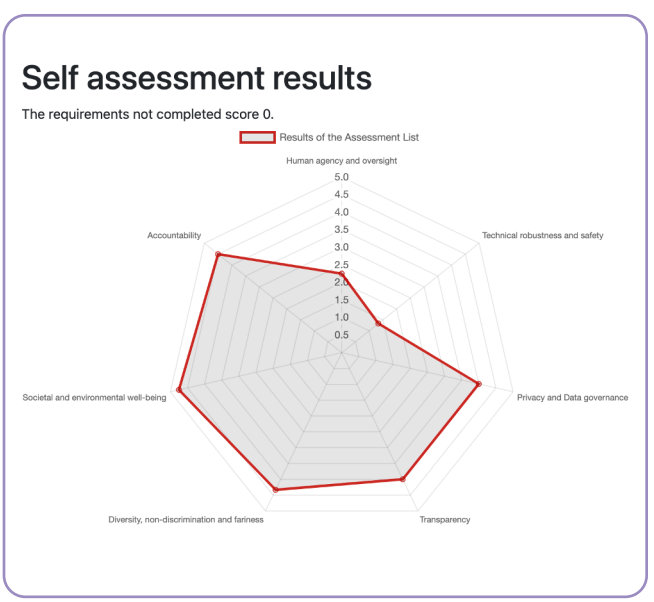
4.7.4. Bewertung und Resümee

Die ausgearbeiteten Fragen und deren freie Verfügbarkeit als PDF-Dokument sowie als interaktive Web-Applikation ist ein gelungener Beitrag zur Auditierung insbesondere zum Self Assessment. Man ist dem Anspruch ein generisches Tool, das für alle KI-Systeme verwendet werden kann, gerecht geworden. In der Allgemeinheit der Anwendung liegt jedoch auch eine Schwäche des Ansatzes: bei genauerer Betrachtung zeigt sich rasch, dass die Fragen nicht auf bestimmte Domänen angepasst sind und damit bestimmte Eigenschaften bzw. Charakteristika nicht abgefragt und überprüft werden. Im Kontext von ExamAI wäre eine Anpassung auf den Anwendungsbereich Personal- und Talentmanagement notwendig.

Die zweite große Herausforderung, die sich ebenfalls aus dem Anspruch ein allgemeingültiges System zu entwickeln ergibt, ist die unzureichende Berücksichtigung technischer Eigenschaften von KI- bzw. ADM Systemen. Die Fragen sind nach den sieben Key Requirements der HLEG AI ausgerichtet. Sie gehen nicht bzw. nur selten auf technische Rahmen- und Umsetzungsbedingungen ein. Eine Zusammenführung des in Abschnitt 3 vorgestellten Frameworks über die verschiedenen Ebenen eines KI-Systems, sowie der sieben Key Requirements wäre deshalb wünschenswert. Insbesondere, weil dies den Adressatenkreis des Tools bzw. der einzelnen Fragen besser definieren würde. Aktuell ist unklar an welcher Stelle des Lebenszyklus eines KI-Systems das Tool eingesetzt werden soll und an wen sich das Tool genau richtet. Es bleibt abzuwarten, ob sich das Tool in der Praxis bewährt.

Tabelle 4: Beispielhafte Auszüge als Screenshots aus dem Self Assessment Tool (links: Fragen, sowie deren Antwortmöglichkeiten; rechts: die Ergebnisdarstellung eines fiktiven Beispiels als Netzdiagram (Quelle: B. Waltl bzw. Online)).





5.
Ebenen des KI-Audits

5.1. Auditierung und Testen von KI-Systemen

Das in Abschnitt 3 vorgestellte ganzheitliche Modell eines KI-Systems stellt eine Grundlage für die Anwendung von Auditierungs- und Testverfahren dar. Wie oben bereits erläutert, handelt es sich bei einem KI-System um ein komplexes und software-intensives System. Unter anderem hat dies zur Folge, dass die Auswirkungen von Entscheidungen und Änderungen an einer Stelle des Gesamtsystems an einer anderen Stelle bemerkbar werden. So hat beispielsweise die Wahl des Algorithmus in Ebene 3 Auswirkungen auf die in Ebene 5 zur Verfügung stehenden Metadaten. Wird der Algorithmus während einer Weiterentwicklung geändert, weil dieser durch einen neuartigen Algorithmus ersetzt wird, so kann dies dazu führen, dass erforderliche Metadaten in Zukunft nicht mehr bereitgestellt werden können. Dies liegt daran, dass den Algorithmen zum Teil gänzlich verschiedene Prinzipien zugrunde liegen. So werden beispielsweise in Entscheidungsbäumen, bayesschen Netzen und Neuronalen Netzwerken andere mathematische Regeln umgesetzt, was zur Folge hat, dass die Ausgaben der Algorithmen anders interpretiert und verarbeitet werden müssen. Dieses Beispiel soll zeigen, dass ein ganzheitlicher Ansatz, der auf die Eigenheiten der Subsysteme eingehen kann, unerlässlich ist, wenn es um das Verstehen eines KI-Systems geht.

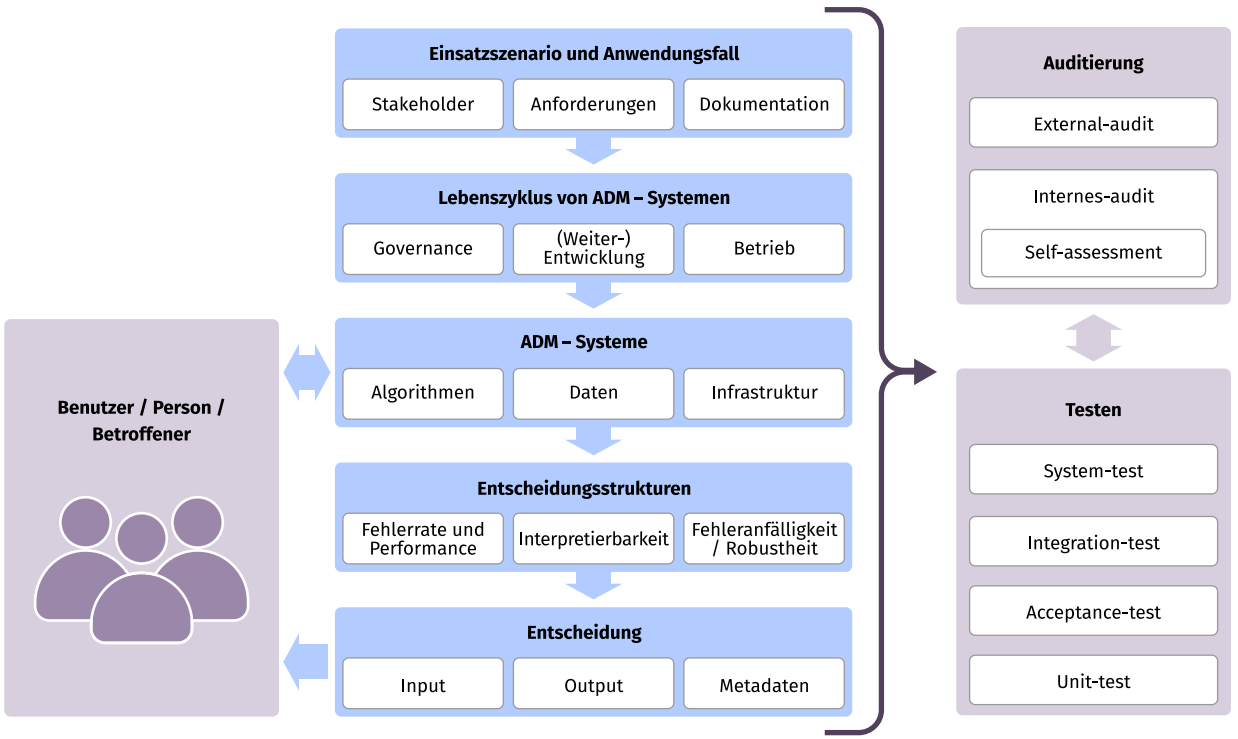


Abbildung 8: Ganzheitliche Übersicht eines KI-Systems als Grundlage für die Verortung von Auditierungs- und Testverfahren.

In Abbildung 8 wird das sich ergebende Gesamtsystem inklusive der Erweiterung von Auditierungs- und Testverfahren nochmals dargestellt. Unbestritten ist, dass sowohl Auditierungsverfahren als auch Testverfahren mit zunehmender Forschung in dem Bereich weiterentwickelt werden. Die vier Testverfahren in Abbildung 8 sind keineswegs eine vollständige Übersicht über die zur Verfügung stehenden Testverfahren von software-intensiven Systemen. Darüber hinaus ist zu erwarten, dass sich der Trend der letzten Jahre fortsetzt und es neben den bisherigen Methoden zum Testen von software-intensiven Systemen noch weitere und vermutlich auf das Testen von Algorithmen hoch-spezialisierte Testverfahren geben wird. Gerade wenn diese Testverfahren hilfreich sind, um gesetzliche Anforderungen umzusetzen bzw. Systeme leistungsfähiger und robuster zu entwickeln.

Abbildung 8 stellt darüber hinaus noch dar, dass weder Auditierungs- noch Testverfahren immer nur auf die Anwendung und Analyse einer Komponente oder einer Schicht beschränkt sind. Es ist möglich, dass die Verfahren mehrere Komponenten gleichzeitig und über mehrere Schichten hinweg analysieren. So können Testverfahren nicht nur Auskunft über die Funktionsweise des Algorithmus (Ebene 3) geben, sondern auch auf die Einblicke in die Funktionsweise der Entscheidungsstruktur (Ebene 4) gewähren und eine Analyse der Ausgabe sowie Metadaten (Ebene 5) ermöglichen. Analog gilt dies auch für Auditierungsverfahren. Wie in Abschnitt 4.1 und insbesondere in Abbildung 7 dargestellt schließen sich die Verfahren von Testen und Auditierung nicht aus, sondern ergänzen sich. Dabei kann es sinnvoll sein, sich im Rahmen eines Audits ausgewählter Testverfahren zu bedienen, um objektive Nachweise zu erhalten, die nur mit klassischen Auditmethoden (z. B. Software Code Review) nicht zu generieren wären.

Nach den Erläuterungen um das Verhältnis der Ebenen und Komponenten eine KI-Systems zu Auditierungs- und Testverfahren, soll im nächsten Abschnitt auf die Herausforderungen eingegangen werden, die bei Berücksichtigung gesetzlicher Anforderungen auftauchen.

5.2. Herausforderungen der Berücksichtigung gesetzlicher Anforderungen

5.2.1. Bestimmung und Ableitung gesetzlicher Anforderungen

Auf Basis der bisherigen Betrachtungen und Ausführungen lässt sich festhalten, dass das Auditieren eine sinnvolle und effektive Maßnahme darstellt, das Verhalten eines KI-Systems zu bewerten.

Darüber hinaus kann durch das Auditieren Information über das KI-System erlangt werden, welche das Verhalten des Systems auch erklären kann. Entscheidend dabei ist die ganzheitliche Betrachtung von KI-Systemen sowie die Berücksichtigung von spezifischen Eigenschaften, die ein KI-System von einem herkömmlichen software-intensiven System abgrenzen.

Es bleiben jedoch weitere Herausforderungen gerade in Hinblick auf die Bestimmung von gesetzlichen Anforderungen. Gesetzliche Anforderungen an KI-Systeme im HR-Bereich können aus unterschiedlichen Quellen entstehen. [59] [60] [61] Beispielhaft sei hier nur auf Vorschriften zur Vermeidung von Diskriminierung nach dem AGG, Einhaltung des Datenschutzes und der informationellen Selbstbestimmung, sowie Haftungsregime und Schadensersatzansprüche hingewiesen. [62] Daneben gibt es (in Deutschland) jedoch noch zahlreiche sektorale Bestimmungen, die nicht notwendigerweise im Kontext von Personal und Talentmanagement eine Rolle spielen, jedoch bei weiteren Überlegungen nicht außer Acht gelassen werden sollten, z.B. die Kennzeichnung algorithmischer Entscheidung im Bereich Handels- und Kapitalmarktrecht (u. a. § 80 Abs. 2 WpHG und § 16 Abs. 2. Nr. 3 BörsG). Aus bewährten und etablierten Methoden zur Regulierung des Einsatzes von algorithmischer Entscheidungsfindung können im Idealfall Ansätze übernommen und nach entsprechender Anpassung eingesetzt werden. In modernen Gesellschaften und Rechtsordnungen wird klar, dass die Bestimmung von geltenden und anwendbaren Normen schnell zu einer größeren Herausforderung wird. Hinzukommt die Überlagerung von gesetzlichen Bestimmungen und Vorschriften von nationalen und internationalen Rechtsquellen. Es bleibt abzuwarten, ob es sich bei KI-Systemen ähnlich entwickeln wird wie in anderen Industrien, wie zum Beispiel der Automobilindustrie, in denen Typzulassung und Homologation Bestandteil des Entwicklungsprozesses sind. Sodass die Ausgestaltung und der Einsatzbereich eines KI-Systems von vorneherein durch die Berücksichtigung von gesetzlichen Anforderungen mitbestimmt ist.

Tendenzen hierfür werden in dem jüngsten Vorschlag der Europäischen Kommission zur Regulierung Künstlicher Intelligenz sichtbar. Der „Artificial Intelligence Act“ wurde am 21. April 2021 von der Europäischen Kommission als Vorschlag vorgelegt und stellt einen risiko-basierten Ansatz zur Bewertung von KI-Systemen vor. Dabei werden – im Wesentlichen – Risikogruppen für KI-Systeme vorgeschlagen, darunter die folgenden: [63]

i) unannehmbares Risiko
ii) hohes Risiko
iii) geringes bzw. minimales Risiko

[59] Lewinski, Bestehende und künftige Regelungen des Einsatzes von Algorithmen im HR-Bereich, (hrsg. v. AlgorithmWatch) 2019 (zusammen mit R. de Barros Fritz und K. Biermeier)

[60] G. Borges, R. Hoffmann, A. Sesing „Kurzbegutachtung Anwendungsszenario 2: KI-Systeme im Personal- und Talentmangement“, Publikation im Rahmen von Projekt ExamAI, 2021 (Veröffentlichung ausstehend)

[61] M. Martini „Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz“ 2019, Springer, Berlin, Heidelberg,.

[62] Lewinski, Bestehende und künftige Regelungen des Einsatzes von Algorithmen im HR-Bereich, (hrsg. v. AlgorithmWatch) 2019 (zusammen mit R. de Barros Fritz und K. Biermeier), Seite 8

[63] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS Kapitel 5.2.2.

Unter Hochrisiko-KI-System fallen „solche Systemen, die erhebliche Risiken für die Gesundheit und Sicherheit oder die Grundrechte von Personen bergen." Es ist daher davon auszugehen, dass ein nicht unerheblicher Teil der Anwendungsfälle der KI-Systeme im Kontext von Personal- und Talentmanagement als Hochrisiko-KI-System eingestuft wird. Hierfür wären dann „Anforderungen an hohe Datenqualität, Dokumentation und Rückverfolgbarkeit, Transparenz, menschliche Aufsicht, Präzision und Robustheit unerlässlich, um die Risiken für die Grundrechte und die Sicherheit abzumildern, die mit diesen KI verbunden sind und die nicht durch andere bestehende Rechtsvorschriften abgedeckt werden." [64]

[64]

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS Kapitel 2.4.

Nach dem derzeitigen Stand des Vorschlags der Verordnung ist die Riskostufe eines KI-Systems auch die Grundlage für die Anforderungen, welche an das KI-System gestellt werden. Die Anforderungen werden in den Artikeln 8 – 15 präzisiert und orientieren sich von der Grundstruktur an den Rahmenbedingungen der durch die HLEG AI skizziert wurde, nämlich den sieben Key Requirements für KI, die von der HLEG AI im April 2019 vorgestellt wurden. Noch befindet sich der Entwurf im Gesetzgebungsverfahren, es ist jedoch bereits absehbar, dass die Verordnung nicht nur einzelne Wirtschaftssektoren oder einzelne KI-Systeme betreffen wird. In der aktuellen Fassung besteht bei den Anforderungen an Hochrisiko-KI-Systeme noch viel Interpretationsspielraum, sodass das Gesetz zwar eine Grundlage für die zukünftigen Generationen von KI-Systemen ist, für die konkrete Umsetzung der Anforderungen jedoch noch viel Forschungs- und Entwicklungsarbeit zu leisten sind.

Hinlänglich bekannt und vielfach untersucht ist die semantische Lücke zwischen konkreten Prüfkriterien und Gesetzesnormen. [65] Besonders deutlich zeigt sich dies, beim Versuch eindeutige Prüfkriterien aus Normen abzuleiten. Gesetzliche Normen lassen zumeist einen nicht unerheblichen Raum für Interpretation und Auslegung offen. Dies mag einerseits für das Ableiten konkreter Anforderungen an KI-Systeme unbefriedigend sein, weil nur in seltenen Fällen sichergestellt werden kann, dass Anforderungen auch die Norm vollumfänglich umsetzen. Andererseits bleibt dadurch auch Raum für Innovation und Einzelfallgerechtigkeit erhalten. Gerade in einem innovativen Umfeld wie Künstliche Intelligenz ist es unbedingt erforderlich, dass Gesetze keinen allzu engen Korridor vorgeben innerhalb dessen Anwendungen erlaubt sind. Offene und vage Formulierungen fördern die Rechtsunsicherheit einerseits, bewusst eingesetzt und klar gekennzeichnet können diese jedoch auch zu positiven Entwicklungen und insbesondere Anreize zur (Weiter-)Entwicklung von KI-Systemen sein, die den Vorgaben des Gesetzgebers entsprechen und zum Wohl der Gesellschaft, der Wirtschaft und jedes einzelnen Individuums eingesetzt werden können.

[65]

Sergot, Marek & Sadri, Fariba & Kowalski, Robert & Kriwaczek, Frank & Hammond, Peter & Cory, H.. (1986). The British Nationality Act as a Logic Program. Communications of the ACM. 29. 370-386. 10.1145/5689.5920.

5.2.2. Verortung gesetzlicher Anforderungen in KI-Systemen

Neben der Herausforderung Anforderungen aus Gesetzestexten und anderen relevanten Quellen mit normativem Charakter (siehe Abschnitt 5.2.1) abzuleiten besteht noch eine weitere Herausforderung: die Auswirkungen einer Anforderung auf die Entwicklung und den Betrieb eines KI-Systems zu verstehen und zu bewerten. Dies ist, angesichts der hohen Komplexität eines software-intensiven Systems, nicht trivial. [66] Hierbei kann eine konstruktive Differenzierung in fünf Ebenen jedoch eine große Rolle spielen. Gesetzliche Anforderungen sind nicht selten Anforderungen, die sich unmittelbar auf ein beobachtbares Verhalten beziehen. So bezieht sich beispielsweise das Allgemeines Gleichbehandlungsgesetz (AGG) mit §3 in Kombination mit §1 auf das Verhalten im Kontext der „unmittelbaren Benachteiligung“. Auf ein KI-System übertragen könnte hier abgeleitet werden, dass eine konkrete Entscheidung hinsichtlich ihres Ausgangs (Output) überprüft werden muss. Wie in Abschnitt 3.6 festgestellt kann das Verhalten eines KI-Systems auf Ebene 5 „Entscheidung“ zwar überprüft und – im Falle einer unmittelbaren Benachteiligung – gegebenenfalls auch festgestellt werden. Die Überprüfung der Ursachen kann jedoch nicht ausschließlich in Ebene 5 stattfinden. Um die Suche nach Ursachen effektiv zu gestalten, müssen auch weitere (alle) Ebenen berücksichtigt und auf den Prüfstand gestellt werden.

[66] Mens, Tom. „On the complexity of software systems.“ IEEE Computer Architecture Letters 45.08 (2012): 79-81.

Die Suche nach den Gründen für das Verhalten eines KI-Systems ist in Situationen relevant, in denen ein KI-System bereits im Einsatz ist und eine Bewertung des Verhaltens, sowie deren Ursachen durchgeführt wird (ex post). Demgegenüber stehen Situationen, in denen das Risiko für ein – nach rechtlichen Maßstäben - problematisches Verhalten eines KI-Systems vor seinem Einsatz minimiert werden soll (ex ante). Hierbei müssen Anforderungen an alle Ebenen schon während frühester Phasen und insbesondere der Konzeption und Entwicklung gestellt werden.

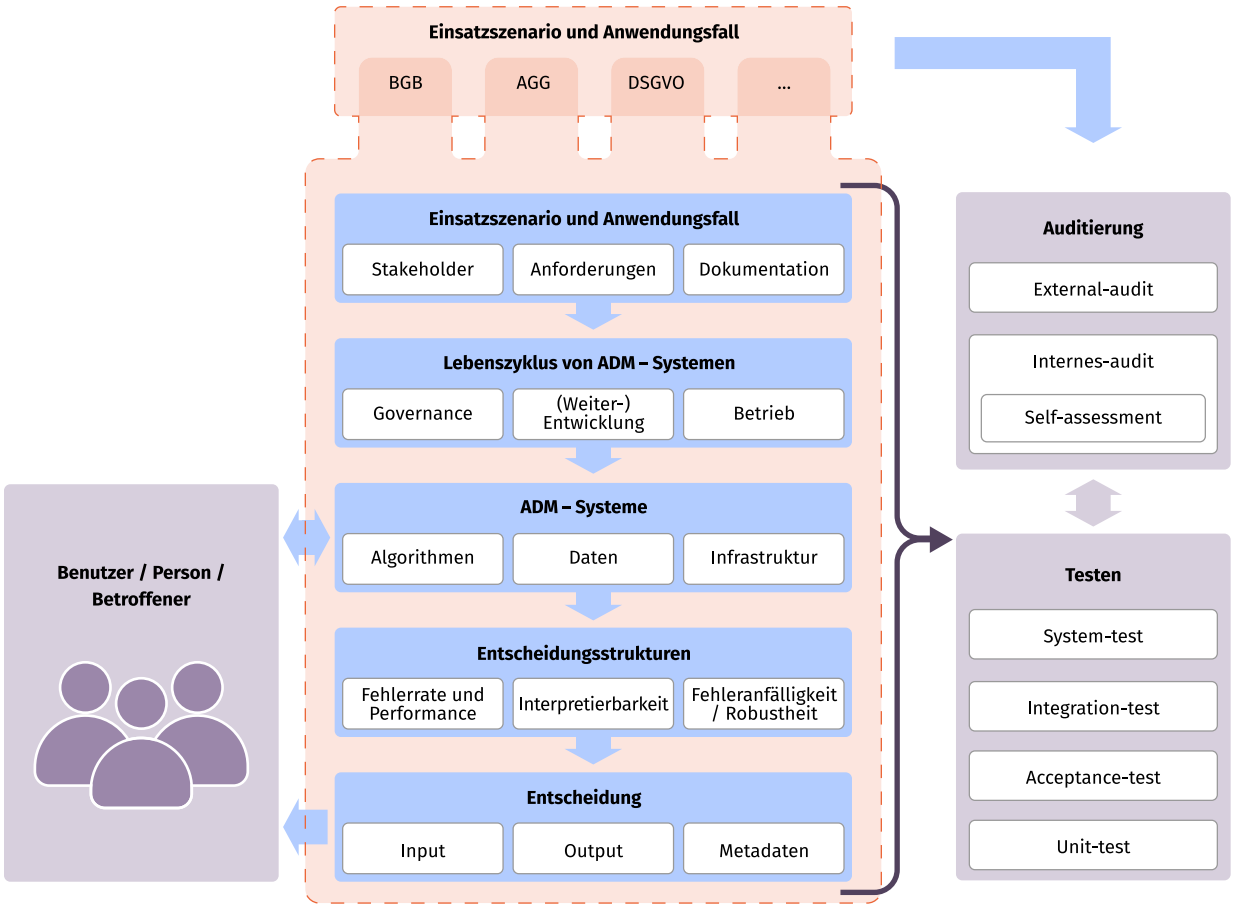


Abbildung 9: Gesetzliche Anforderungen können Auswirkungen auf allen Ebenen eines KI-Systems haben.

Grundsätzlich ist es sinnvoll, dass eine Unterscheidung zwischen einer ex post und einer ex ante Betrachtung vorgenommen wird. Die Schwierigkeit besteht in der Praxis oft darin, dass nicht jedes Verhalten eines KI-Systems ex ante bewertet und ausgeschlossen werden kann. Bekannte Herausforderungen sind, dass in der Praxis Situationen auftreten können, die während der Entwicklungsphase nicht vorhersehbar waren oder dass es einfach nicht möglich ist, ein fehlerhaftes Verhalten durch Auditierung und Testen ex ante zu erkennen. Für die Bewertung, ob es zu einem problematischen Verhalten kommt, muss das IST Verhalten mit einem SOLL Verhalten verglichen werden. Kommt es zu einer Abweichung, ist dies ein Indiz für einen Fehler im KI-System. Für komplexere Probleme, und dies ist insbesondere im Bereich HR-Systeme zu erwarten, kann jedoch kein SOLL Verhalten vorgegeben werden. Deswegen ist es nicht trivial, einen SOLL / IST Abgleich durchzuführen. In der Informatik ist dieses Problem als sogenanntes Orakel-Problem (engl. oracle problem) bekannt und wird bei dem Vorschlag herkömmliche Software-Testverfahren zu verwenden, oftmals übersehen. [67]

[67]
E. T. Barr, M. Harman, P. McMinn, M. Shahbaz and S. Yoo, “The Oracle Problem in Software Testing: A Survey,” in IEEE Transactions on Software Engineering, vol. 41, no. 5, pp. 507-525, 1 May 2015, doi: 10.1109/TSE.2014.2372785

Die Differenzierung zwischen ex post und ex ante ist insofern relevant, weil mit Verweis auf die 5 Ebenen eines KI-Systems klar wird, dass gesetzliche Anforderungen auf allen Ebenen eines KI-Systems Einfluss nehmen. So werden die Auswirkungen der Anforderungen an ein KI-System zur Vermeidung von nicht gerechtfertigter (siehe §§ 8 bis 10 und 20 AGG) unmittelbarer oder mittelbarer Benachteiligung auf allen Ebenen spürbar. Abbildung 9 stellt dies grafisch dar. Die gesetzlichen Anforderungen ziehen sich wie Querschnittsthemen über alle Ebenen hindurch. So können gesetzliche Anforderungen auch an verschiedenen Stellen eines KI-Systems Auswirkungen haben. Daraus folgt, dass es nicht notwendigerweise eine 1:1 Beziehung zwischen gesetzlichen Anforderungen und technischen Anforderungen gibt. Erschwert wird dies durch den Umstand, dass die Bestimmung von gesetzlichen Anforderungen durchaus anspruchsvoll ist und sich möglicherweise in gar keine eindeutigen technischen Anforderungen überführen lassen. Oder anders ausgedrückt: die Operationalisierung von gesetzlichen Anforderungen in eindeutige technische Anforderungen ist komplex und allein mit den Mitteln der Softwaretechnik und Informatik nicht zu beantworten. Offen ist hierbei, wie in der Praxis damit umgegangen werden soll. Es erscheint sinnvoll, entsprechende zusätzliche Entscheidungen, die zur Präzisierung getroffen werden, sorgfältig zu dokumentieren und regelmäßig zu überprüfen.

Es kann festgehalten werden, dass es nicht oder nur unzureichend möglich ist konkrete und allgemein-gültige Maßnahmen für alle KI-Systeme aus einzelnen Gesetzen abzuleiten. In Abschnitt 6 soll zumindest grob skizziert werden, wie dies exemplarisch für Anwendungsfälle aus dem Bereich HR und Talent Management aussehen kann. Grundlage wird hierbei die Differenzierung und ganzheitliche Betrachtung von KI-Systemen sein.

6. KI-Audits für Personal- und Talentmanagement

6.1. Repräsentative Anwendungsfälle nach ExamAI

Das Forschungsprojekt „ExamAI – KI Testing & Auditing“ unter der Leitung der Gesellschaft für Informatik e. V. widmet sich der Erforschung von Auditierungsverfahren für KI-Systeme im Bereich Personal- und Talentmanagement. In dem Kontext wurden sieben Anwendungsszenarien identifiziert, die repräsentativ für die den Bereich sind:

1. Automatisierte Vorschlagssysteme auf Personalplattformen: Plattformen helfen Arbeitssuchenden und Arbeitgebern bei der Zuordnung von freien Stellen zu Personen auf Basis bereitgestellter Information über Jobprofile und Fähigkeitsprofile. Hierbei kommen vermehrt KI-Systeme zum Einsatz, die bei der Auswahl und Matching Abläufe automatisieren bzw. menschliche Entscheidungen beeinflussen und vorwegnehmen. Es besteht das Risiko der Diskriminierung bzw. Ungleichbehandlung durch Algorithmen.
2. Persönlichkeitsbewertung per Lebenslauf: KI unterstützt Personen im Personalmanagement bei der Vorauswahl von Bewerber*innen durch automatisierte Checks von eingereichten Unterlagen (z.B. Lebenslauf, Anschreiben, etc.) oder auch durch Bewertung der Sprache und Gestik bei Videobewerbungen. Diese automatisierten Checks durch Algorithmen sind der Gefahr ausgesetzt, dass sie die Merkmale zur Vorauswahl weder transparent noch für Menschen nachvollziehbar sind, mit allen damit verbundenen Risiken.
3. KI-basierte Background-Checks: Background-Checks stehen vor der Herausforderung, dass vielen Daten aus unterschiedlichen Systemen zusammengetragen werden müssen. Hierbei können Algorithmen helfen, da sie großen Datenräume schnell und effizient durchsuchen können. Der Einsatz intelligenter Algorithmen liegt nahe. Doch Algorithmen können Fehler machen, weil z.B. Daten von Personen mit häufigen Namen, verwechselt, oder falsch zugeordnet werden. Dadurch kann eine Person auf Basis einer falschen Datenlage bewertet werden.

4. Chatbot der HR-Abteilung: Bereitstellung von Information für Mitarbeiter*innen ist gerade im Bereich Personalwesen eine wichtige und ressourcen-intensive Tätigkeit. Chatbots stehen Mitarbeiter*innen 24 Stunden und 7 Tage die Woche zur Verfügung und können tagesaktuelle Informationen auf Anfragen liefern. So zumindest die Theorie. Die Praxis steht vor der Herausforderung, dass der Chatbot stets auf dem neuesten Stand ist – und keine veralteten oder falschen Informationen verwendet - und Aspekte rund um Datenverarbeitung berücksichtigt.
5. Internes Jobprofil-Matching: Bei der Zuordnung von neuen und offenen Stellen und Positionen zu möglichen Kandidat*innen bzw. bei der Bewertung von Kandidat*innen kann ein KI-System helfen. Allerdings läuft ein KI-System Gefahr, dass Fachbegriffe, die auch dem Wandel der Zeit unterliegen, nicht oder falsch zugeordnet werden oder in der Entscheidungsfindung nicht berücksichtigt werden. So hat sich die offizielle Bezeichnung von Pflegefachfrau/-mann aus den Begriffen „Krankenschwester/Krankenpfleger“ (1985) über „Gesundheits- und Krankenpfleger/in“ entwickelt. Ein System, dass dies nicht berücksichtigt, läuft Gefahr einer unzulässigen Benachteiligung (z.B. nach dem Geschlecht). Matching Verfahren, deren Entscheidungen und Entscheidungsstrukturen völlig intransparent sind, bergen daher Risiken.
6. Vorhersage der Jobkündigungsbereitschaft: Die Auswertung großer Datenbestände kann Unternehmen unterstützen die Zufriedenheit, die stark mit der Kündigungsbereitschaft korreliert, vorherzusagen. Faktoren wie Stress, psychische Belastung, Über- oder Unterforderung oder Frustration sind hierfür Anhaltspunkte. Algorithmen können zu einem positiven Arbeitsklima beitragen, sofern diese transparent und nachvollziehbar sind, sodass Ergebnisse immer von Menschen bewertet werden können. Andernfalls droht die Gefahr, dass Mitarbeiter falsch bewertet und eingeschätzt werden. Die Verarbeitung von möglicherweise sensiblen Personendaten muss ebenfalls berücksichtigt werden.
7. Automatisierte Arbeitszeitzuweisung bei Gig-Workern: Algorithmen können eingesetzt werden, um den Arbeitsablauf und die Arbeitsbelastung von verschiedenen Berufsgruppen zu planen und zu bestimmen. Die Reduktion des Menschen in automatisiert und intransparent erstellte Arbeitsabläufe und Schichtpläne kann zu einer Degradierung des Menschen und einer Dehumanisierung der Arbeit führen. Transparente Entscheidungskriterien, in denen Menschen ein Mitbestimmungsrecht haben können zu gut funktionierenden Arbeitsbeziehungen zwischen Menschen und Unternehmen führen.

Bei der Betrachtung der technischen Umsetzung der sieben Anwendungsfälle ist das hier vorgestellte 5-Ebenen-Modell zum Aufbau und Zusammensetzung von KI-Systemen hilfreich. Je nach Anwendungsfall kann es selbstverständlich vorkommen, dass eine Komponente mehr oder weniger stark ausgeprägt ist. Im Prinzip jedoch sollte jede Komponente vorhanden sein. Erst der ganzheitliche Blick auf ein KI-System erlaubt es dann Aussagen zu treffen bzgl. Nachvollziehbarkeit von Entscheidungen.

Im nächsten Abschnitt werden beispielhafte Fragestellungen vorgestellt, die sich entlang der 5 Ebenen eines KI-Systems orientieren und die einen Einstieg in ein KI-Audit erlauben.

6.2. Ausgewählte Fragestellungen für ein KI-Audit

Auf Basis des vorgeschlagenen Frameworks zur Differenzierung von KI-Systemen und den Prinzipien und Möglichkeiten der Auditierung von KI-Systemen, werden in diesem Kapitel konkrete Fragen formuliert, die eine Hilfestellung einen Einstieg in die komplexe Materie des KI-Audits darstellen. Die Fragestellungen sind dabei in die 5 Ebenen und die zentralen Bestandteile eines software-intensiven KI-Systems differenziert. Daran zeigt sich der Vorteil dieses Frameworks, da ein holistisches und gleichzeitig differenziertes Vorgehen möglich ist, dass auf die Charakteristiken von KI-Systeme eingehen kann. Darüber hinaus erlaubt es spezifische Schwerpunkte auf einzelne Ebene bzw. Bestandteile zu setzen.

Die formulierten Fragen sind keineswegs vollständig. Sie können und müssen je nach Schwerpunkt und Zielsetzung eines KI-Audits variiert und angepasst werden.

6.2.1. KI-Audit für Ebene 1: Einsatzszenario und Anwendungsfall

Ebene 1: Einsatzszenario und Anwendungsfall	
Stakeholder	- Wer sind die Auftraggeber des Systems? - Wer finanziert das System und wie sieht das Geschäftsmodell aus? - Welche Benutzergruppen für das System gibt es (Administratoren, Sachbearbeiter, Anwender, etc.)? - Welche Personengruppen sind an der Konzeption, bei der Entwicklung und beim Betrieb bzw. Einsatz des Systems involviert? - Gibt es Dritte die am Einsatz bzw. an der Verwendung des Systems interessiert sind (z. B. Partnerunternehmen)?
Anforderungen	- Wie sieht der Prozess (das Vorgehen aus) in dem die Anforderungen festgelegt wurden? - Welche Anforderungen bestehen an das System (Pflichtenheft, Lastenheft, Backlog, etc.)? - Wurde eine Risikobewertung von Anforderungen während der Entwicklung durchgeführt? - Gibt es nach rechtlichen Maßstäben (z. B. DSGVO, AGG, BGB, etc.) problematische Anforderungen? - Wie wurde die Umsetzung der Anforderungen sichergestellt?
Dokumentation	- Ist eine Dokumentation für das System vorhanden und wird diese aktualisiert? - Entspricht die Dokumentation etablierten Industriestandards (z. B. ISO 9001)? - Sind die verschiedenen Aspekte des Systems dokumentiert (Betrieb und Wartung, Anwendung und Benutzung, Funktionsweise ADM System, etc.)? - Sind Schnittstellen zu Benutzern (User Interface) und anderen Systemen (Application Programmable Interface) dokumentiert? - Sind Testergebnisse sowie sicherheitskritische Bereiche dokumentiert?

6.2.2. KI-Audit für Ebene 2: Lebenszyklus von software-intensiven KI-Systemen

Ebene 2: Lebenszyklus von software-intensiven KI-Systemen	
Governance	- Wie sieht die Organisationsstruktur des Unternehmens in Bezug auf die Entwicklung, den Betrieb und die Weiterentwicklung aus? Wer ist der Owner? Wie werden Zielkonflikte erkannt, dokumentiert und aufgelöst? - Sind Prozesse zur Berücksichtigung von IT-Sicherheit, Privacy Impact, Informationsklassifizierung, Rechte/ Rollen Konzepte, etc. aufgesetzt? - Läuft das KI-System on-prem oder als Cloud Lösung? Sind die Anforderungen hierfür erfüllt und wird deren Einhaltung auch während des Betriebs sichergestellt? - Welche Prozesse sind zur Implementierung und zur Monitoring des KI-Systems aufgesetzt? - Ist die IT-Architektur beschrieben und sind SLAs, sowie Schnittstellenverträge zu IT-Systemen (Datenbanken und Datenzulieferer) aufgesetzt, zu denen es Abhängigkeiten gibt?
(Weiter-) Entwicklung	- Nach welchem Vorgehensmodell erfolgt die (Weiter-) Entwicklung des KI-Systems (z.B. Wasserfall, Agil, etc.)? - Wie werden Anforderungen erfasst und dokumentiert? Nach welchen Kriterien werden Anforderungen priorisiert? - Werden neue Anforderungen hinsichtlich deren Kritikalität und Gesetzeskonformität bewertet? - Wie wird die Umsetzung von Anforderungen überprüft und dokumentiert? - Wie werden Änderungen und Updates an die Stakeholder kommuniziert?

Ebene 2: Lebenszyklus von software-intensiven KI-Systemen	
Betrieb	- Wie wird das System während des Einsatzes überwacht (Monitoring)? Welche Komponenten sind vom Monitoring erfasst? - Wie werden Systemausfälle erkannt bzw. wie wird Fehlverhalten festgestellt (Incidentmanagement)? - In welchen Modus werden neue Updates und Releases des Systems bzw. von Subkomponenten ausgerollt? - Berücksichtigt das Betriebsmodell besondere Anforderungen, die sich aufgrund des Einsatzes von KI ergeben? - Wie sind die Mitarbeiter*innen geschult für den Betrieb und die Weiterentwicklung des KI-Systems?

6.2.3. KI-Audit für Ebene 3: ADM Systeme

Ebene 3: ADM Systeme	
Algorithmen	- Welche Algorithmen und Verfahren aus dem Bereich Maschinelles Lernen werden in dem ADM System eingesetzt? - Werden Anforderungen an Transparenz und Erklärbarkeit bei der Auswahl der Algorithmen berücksichtigt? - Wie und von wem werden die Algorithmen trainiert? - Werden die Algorithmen evaluiert und getestet? Werden die Evaluationsergebnisse dokumentiert? - Werden spezifische Verfahren angewandt um Algorithmen auf Robustheit, Fehleranfälligkeit, etc. zu prüfen?
Daten	- Welche Daten werden innerhalb des ADM Systems eingesetzt? Was ist der Ursprung der Daten? - Ist die Verwendung der Daten für den vorgesehenen Zweck legitimiert? - Wie wurde die Datenqualität und die Repräsentativität der Daten überprüft? - Welche Maßnahmen unternommen um bekannte Fehlerquellen, die auf Datenqualität zurückzuführen sind, zu erkennen bzw. auszuschließen? - Welche Maßnahmen gibt es um bei Weiterentwicklungen und sich ändernden Datensätzen (z. B. bei lernenden Systemen) Auswirkungen frühzeitig zu erkennen?

Ebene 3: ADM Systeme	
Infrastruktur	- Wurde die Infrastruktur hinsichtlich regulatorischer Anforderungen bewertet? - Wurden technische Maßnahmen getroffen um die Infrastruktur abzusichern? - Wurden organisatorische Maßnahmen getroffen um die Infrastruktur zu verwalten? - Welche Infrastruktur wird zum Training eines ADM-Systems verwendet? - Welche Infrastruktur wird zum Ausführen eines ADM-Systems verwendet?

6.2.4. KI-Audit für Ebene 4: ADM Entscheidungsstrukturen

Ebene 4: ADM Entscheidungsstrukturen	
Performance und Genauigkeit	- Wie wurde die Entscheidungsstruktur (das trainierte ML Modell) evaluiert? - Welche Kriterien (Genauigkeit, Geschwindigkeit, etc.) wurden bei der Evaluation berücksichtigt? - Welche Vorhersagegenauigkeit (Accuracy, Fehler 1. Ordnung, Fehler 2. Ordnung) wurde erreicht? - Welche Maßnahmen wurden umgesetzt um einen Qualitätsabfall hinsichtlich der Kriterien während des Betriebs zu erkennen? - Welche Prozesse und Abläufe wurden aufgesetzt um die Performance und die Genauigkeit des Systems neu zu bewerten?
Interpretierbarkeit	- Wurde die Interpretierbarkeit der Entscheidungsstruktur bewertet? - Welche Maßnahmen zur Interpretation der Entscheidungsstruktur wurden getroffen? - Wurden die Maßnahmen hinsichtlich deren Eignung und Richtigkeit bewertet? - Sind die Adressaten der Information zu Interpretierbarkeit bekannt und sind entsprechende Prozesse aufgesetzt, damit diese die Information auch erhalten? - Welche Maßnahmen zur Schulung und Weiterbildung des Personals zur Umsetzung und Sicherstellung der Maßnahmen wurden getroffen?

Ebene 4: ADM Entscheidungsstrukturen	
Fehleranfälligkeit	- Wurde die Entscheidungsstruktur hinsichtlich der Fehleranfälligkeit bewertet? - Wurden absichtliche Manipulation (Missbrauch) und unabsichtliche Manipulation (fehlerhafter Gebrauch) in Betracht gezogen? - Welche technischen und organisatorischen Maßnahmen wurden umgesetzt um fehlerhaftes Verhalten der Entscheidungsstruktur zu vermeiden? - Welche technischen und organisatorischen Maßnahmen wurden umgesetzt um fehlerhaftes Verhalten der Entscheidungsstruktur zu erkennen? - Wurden die Maßnahmen mit regulatorischen Anforderungen abgeglichen?

6.2.5. KI-Audit für Ebene 5: ADM Entscheidungen

Ebene 5: ADM Entscheidungen	
Eingabe (Input)	- Welche Daten und Attribute (inkl. Datentypen) werden bei einer Entscheidung verwendet? - Werden die Daten und Attribute gespeichert bzw. gelogged, sodass eine Entscheidung nachvollziehbar ist? - Wie werden die Input Daten auf Richtigkeit überprüft? - Wie werden unvollständige Daten behandelt? - Sind regulatorische Anforderungen berücksichtigt worden? Dürfen die Daten zur Entscheidung durch das KI-System verwendet werden?
Metadaten	- Welche Metadaten werden zu einer Entscheidung erfasst (z. B. Datum, verwendetes ML-Modell, etc.)? - Welchen Stakeholdern und Benutzern sind die Informationen aus den Metadaten zugänglich? - Werden Metadaten gespeichert bzw. gelogged, sodass eine Entscheidung nachvollziehbar ist? - Können Metadaten generiert werden die auf die Zuverlässigkeit einer Entscheidung schließen lassen (z. B. Konfidenz, etc.)? - Werden Metadaten ausgewertet und zu welchen Zwecken (z.B. kontinuierliche Qualitätskontrolle)?

Ebene 5: ADM Entscheidungen	
Ausgabe (Output)	- Welche Daten und Attribute (inkl. Datentypen) werden bei einer Entscheidung generiert? - Wird der Output gespeichert bzw. gelogged? - Findet eine manuelle Nachbearbeitung bzw. Qualitätskontrolle statt oder wäre diese prinzipiell möglich (Human Oversight)? - Welche Information der Entscheidung wird an die Benutzer bzw. den Anwender übermittelt? - Sind regulatorische Anforderungen an die automatisierte Entscheidung berücksichtigt worden?

7. Fazit

7.1. Handlungsbedarfe

Die Ausführungen in dieser Arbeit zeigen, dass sich ein technischer und methodischer Rahmen zum Auditieren von KI-Systemen beschreiben und in einem integrativen Framework zusammenführen lässt. Vor allem aber zeigt sich, dass das Themenfeld komplex und vielschichtig ist. Es ergeben sich daher zahlreiche weitere Handlungsbedarfe für die Politik, Industrie, Gesellschaft und Forschung.

7.1.1. Evaluation und Praxistauglichkeit

Das KI-Audit, wie in dieser Arbeit vorgestellt, beruht auf vielen Annahmen, deren Evaluation und Bewährung in der Praxis noch ausstehen. Der Handlungsbedarf für Folgeaktivitäten besteht in der objektiven Evaluierung und der Praxistauglichkeit. Es empfiehlt sich also, dass man gemeinsam mit Partnern aus der Industrie und Praxis dieses Modell bespricht und es idealerweise auch an konkreten Szenarien anwendet. Um eine objektive Aussage über die Praxistauglichkeit treffen zu können sollen die Evaluationskriterien durch alle Beteiligten vorab festgelegt werden. Hierunter sollen auch die Rahmenbedingungen aller Beteiligten berücksichtigt werden. Unter anderem auch der Zeit- und Ressourcenbedarf. Gerade das Auditieren von komplexen Systemen kann einen hohen Ressourcenbedarf mit sich bringen.

7.1.2. Weiterentwicklung des KI-Audits

Diese Arbeit zeigt die sozio-technischen Ebenen eines Audits mit Fokus auf die Überprüfung von KI-Systemen auf (sog. KI-Audit). Ein KI-Audit ist dadurch gekennzeichnet, dass die Prüfbereiche und -kriterien die Charakteristika von KI-Systemen in den Fokus nehmen. Darunter fallen Datenbestände, Entscheidungsstrukturen auf Basis von maschinellem Lernen und Metadaten von algorithmischen Entscheidungen. Das hier vorgestellte Framework wurde auf Basis von wissenschaftlicher Literatur mit Fokus auf Anwendungsfälle im Bereich HR und Talentmanagement entwickelt. Im Bereich KI, einem hoch-dynamischen Feld mit sehr kurzen Innovationszyklen, ist eine kontinuierliche Weiterentwicklung und Anpassung auf den Stand der Technik unbedingt erforderlich.

Der Handlungsbedarf besteht darin, Strukturen zu schaffen, die den Stand der Technik bewerten und entsprechende Weiterentwicklungen des KI-Audits durchführen können.

7.1.3. Standortattraktivität und Aus- und Weiterbildungsangebote

Durch die zunehmende Relevanz von Software und software-intensiven Systemen werden Expertise und Fachkenntnisse immer wichtiger. Fehlendes Know-How wird zu einem Standortnachteil. Dies gilt im Bereich KI und insbesondere auch im Bereich Trustworthy AI und damit auch für KI-Audits. Der Handlungsbedarf besteht darin, neue Studienangebote in dem Bereich zu schaffen, bestehende Studienangebote auszubauen und die Attraktivität für Studierende zu steigern. Begleitstudien, sowie Förderprogramme und Partner aus Industrie und Wirtschaft können hier hilfreich sein. Die Angebote sollten sich jedoch nicht nur für grundständige Studien beschränken. Attraktive Bildungsangebote müssen auch für Mitarbeiter*innen in Unternehmen und Behörden geschaffen werden. Gerade Personen mit Praxiserfahrungen können von der Weiterbildung profitieren und die Attraktivität des Wirtschaftsstandorts steigern. Die Attraktivität des Wirtschaftsstandort kann bei KI nicht oft genug betont werden. Analog gilt es auch in diesem Kontext: Programme zu Förderung von Innovation, Start-ups, sowie KMUs werden in diesem hochkompetitiven und globalisierten Markt immer wichtiger.

7.1.4. Harmonisierung von Initiativen und Regulierung

Bei der Umsetzung und Weiterentwicklung des KI-Audits müssen Aktivitäten, die einen inhaltlichen Einfluss auf die Prüfkriterien und -praxis haben beachtet werden. Der Handlungsbedarf besteht darin, nationale und internationale Normungsinitiativen zu berücksichtigen. Hierfür müssen auch Strukturen, z. B. Gremien, geschaffen werden. Hierbei dürfen aktuelle nationale und internationale Gesetzesinitiativen nicht unberücksichtigt bleiben. Es ist die Aufgabe aller und insbesondere der Politik, klare und harmonisierte Vorschriften für die Praxis (welche in diesem Fall auch Aufsichtsbehörden miteinschließt) zu schaffen. Unklare Vorschriften sind kostspielig, führen zu möglicherweise teuren Konflikten und schützen die Gesellschaft nur unzureichend vor der Umsetzung fehlerhafter oder nicht-gewünschter KI.

7.1.5. Internationaler Austausch und Communities

Die konkrete Ausgestaltung von KI-Audits erfordert die Anpassung an und Berücksich-

tigung von nationalen Rahmenbedingungen. Es muss jedoch der Stand der Technik berücksichtigt werden. Dieser wird in einem erheblichen Maße von dem Fortschritt geprägt, der keine Ländergrenzen kennt. International agierende Unternehmen, länderübergreifende Communities sowie open-source Software prägen den Fortschritt im Bereich KI. Diese Umstände erfordern die internationale Zusammenarbeit und Vernetzung. Der Handlungsbedarf besteht darin, Programme aufzusetzen bzw. sich an Programmen zu beteiligen, die den internationalen Austausch von Wissen und die Zusammenarbeit fördern. Internationale Standards welche die Anforderungen von Vielen und nicht nur die Interessen Weniger berücksichtigen sind wünschenswert. Hierbei ist es wichtig sinnvolle bestehende internationale Gremien und neutrale Institutionen zu fördern.

7.1.6. Forschung

Die Bedeutung von Forschung in diesem Fachgebiet kann nicht genug hervorgehoben werden. Der Handlungsbedarf besteht darin, zusätzlich zu bestehenden Förderprogrammen, die die Weiterentwicklung von KI adressieren, Programme zur Förderung von Forschung im Bereich KI-Audits und Trustworthy AI aufzusetzen und auszubauen. Diese Forschung braucht eine starke fächerübergreifende Ausrichtung. Inter- und Transdisziplinarität sind hier von entscheidender Bedeutung. Dies schließt neben Informatik und Rechtswissenschaft zumindest auch Geistes- und Sozialwissenschaften ein. Die Forschung in dem Bereich ist von hoher gesellschaftlicher Relevanz, deswegen ist es auch sinnvoll sie als Begleitforschung zu großen Forschungsvorhaben, z. B. GAIA X, aufzusetzen und zu verankern. Die Regulierung eines dynamischen Umfelds mit sehr kurzen Innovationszyklen stellt die Gegenwart vor große Herausforderungen. Neuartige Ansätze, wie beispielsweise „Regulatory Sandboxes“, können hier von entscheidender Bedeutung werden.

7.2. Zusammenfassung

Die Ausführungen bis hierher sind ein klares Plädoyer für eine ganzheitliche Betrachtung von KI-Systemen unter Berücksichtigung ihrer Eigenschaften. Obwohl sich KI-Systeme dadurch auszeichnen, dass die zugrundeliegenden Entscheidungen und Entscheidungsstrukturen durch datenintensive Methoden beeinflusst und geprägt sind, darf nicht übersehen werden, dass KI-Systeme vor allem auch sozio-technische Systeme sind. Dies gilt insbesondere im Kontext des Personal- und Talentmanagement. Nicht nur sind die Anwender*innen und Betroffenen von Entscheidungen eines KI-Systems Menschen, sondern auch deren Auftraggeber*innen und Personen, die sich

Gedanken zur Funktionsweise und zur Umsetzung eines KI-Systems machen, spielen eine entscheidende Rolle, wenn es darum geht, ein KI-System zu verstehen bzw. Entscheidungen zu begründen und nachvollziehbar zu machen.

Im Rahmen dieser Arbeit werden die sozio-technischen Aspekte eines KI-Systems herausgearbeitet und in einem integrativen Framework zusammengebracht. Das Framework stellt 5 Ebenen eines KI-Systems vor und bringt diese zusammen:

1. Einsatzszenario und Anwendungsfall
2. Lebenszyklus von software-intensiven KI-Systemen
3. ADM Systeme
4. ADM Entscheidungsstrukturen
5. ADM Entscheidungen

Die einzelnen Ebenen erlauben eine differenzierte Betrachtung, ohne das Gesamtsystem aus dem Blick zu verlieren. Hierin liegt die große Herausforderung, die bei der Regulierung von KI nicht übersehen werden darf:
so attraktiv und naheliegend die Regulierung eines KI-Systems anhand einzelner Komponenten, z. B. Daten, Algorithmen, Transparenzkriterien, etc., sein mag, so komplex und vielschichtig ist der Aufbau und das Zusammenspiel von Komponenten eines KI-Systems. Regulierungsansätze sollten daher nie nur auf einzelne Komponenten reduziert werden.

Das Framework leistet außerdem einen Beitrag zu aktuellen Diskussionen rund um das Auditieren und Testen von KI-Systemen. Wiederum muss darauf hingewiesen werden, dass beim Auditieren das KI-System ganzheitlich betrachtet, werden muss. Für ausgewählte und sehr enge Fragestellungen kann es zielführend sein, sich ausschließlich auf einzelne Komponenten zu fokussieren, das Verhalten eines Gesamtsystems kann jedoch nur bewertet werden, wenn es alle Komponenten und seinen Lebenszyklus, sprich seiner Entwicklungshistorie, berücksichtigt.

Das Auditieren wird als sinnvolle Maßnahme zur Bewertung von KI-Systemen identifiziert und bestätigt. Die Berücksichtigung von spezifischen Eigenschaften ist daher erforderlich. Da die konkreten Inhalte und Vorgehensweisen für Audits angepasst werden können und müssen, argumentiert der Beitrag für die Entwicklung von KI-Audits. Diese können nach etablierten Vorgehensweisen (extern, intern, Self Assessment) durchgeführt werden, wobei die Vorgehensweisen verschiedene Vor- und Nachteile haben, die bei der Bewertung eine Rolle spielen können.

Zuletzt sei nochmals darauf hingewiesen, dass die Diskussionen rund um Trustworthy AI, sowie Auditierung und Testen enorm von der Bearbeitung konkreter Fragestellungen profitieren. Am Beispiel von Personal- und Talentmanagement lassen sich die Herausforderungen aufzeigen und konkrete Lösungsvorschläge entwickeln. Diese konkreten Lösungsvorschläge lassen sich dann wiederum zu einer allgemein gültigen Lösung ausbauen, sodass konstruktive und wertvolle Beiträge für ein gemeinsames Verständnis und zur Entwicklung von Trustworthy AI in der Zukunft entstehen.

8. Tabellenverzeichnis

9. Abbildungsverzeichnis

Über die Autoren

Dr. Bernhard Waltl hat an der TU München an der Fakultät für Informatik an der Schnittstelle zwischen KI und Recht promoviert. Er war 2017 Gastwissenschaftler an der Stanford Law School und hat sich dort mit Explainable AI beschäftigt. Bernhard Waltl ist im Wirtschaftsbeirat und der Fachgruppe Rechtsinformatik der Gesellschaft für Informatik e.V. (GI) aktiv. In der Vergangenheit beriet er u.a. die Bundesmonopolkommission und den SVR des Bundesjustizministeriums zum Thema Algorithmische Entscheidungsfindung und Diskriminierung durch KI. Er ist im Vorstand des Liquid Legal Institutes e.V., das er 2017 mitgegründet hat und welcher sich mit Legal Innovation und Digitalisierung des Rechts beschäftigt.

Nikolas Becker ist Verbundkoordinator des Projektes ExamAI – KI Testing & Auditing und Leiter des Teams Politik & Wissenschaft bei der Gesellschaft für Informatik e.V. (GI). Im Rahmen seiner Forschung beschäftigte er sich insbesondere mit Governance-Fragen von Technologie und Internet sowie der Regulierung von Wissen. Nach seinem M.A. in Politikwissenschaft (Universität Potsdam, UC San Diego) arbeitete er zunächst als Produktmanager für die Open-Data-Plattform CKAN.

Impressum

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

Dezember 2021

Herausgeberin

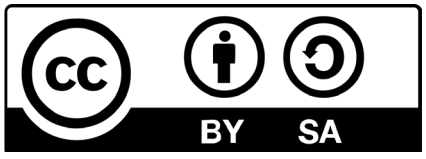
Gesellschaft für Informatik e.V. (GI)
Spreepalais am Dom
Anna-Louisa-Karsch-Straße 2
10178 Berlin

Projektleitung

Nikolas Becker
nikolas.becker@gi.de

Gestaltung

Gabriela Kapfer
http://smileinitial.plus



ExamAI – KI Testing & Auditing

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

Projektpartner*innen:

GESELLSCHAFT FÜR INFORMATIK

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

INSTITUT FÜR RECHTSINFORMATIK
UNIVERSITÄT DES SAARLANDES

Fraunhofer IESE

Stiftung Neue Verantwortung

Gefördert durch:

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

Im Rahmen des:

ki_o
Observatorium Künstliche Intelligenz in Arbeit und Gesellschaft