Bundesministerium
für Arbeit und Soziales

DENK FABRIK DIGITALE ARBEITS-GESELLSCHAFT

Netzwerk KI in der Arbeits- und Sozialverwaltung

Selbstverpflichtende Leitlinien für den KI-Einsatz in der behördlichen Praxis der Arbeits- und Sozialverwaltung

Diese Publikation wurde im Netzwerk KI in der Arbeits- und Sozialverwaltung erarbeitet. Mitgliedsorganisationen im Netzwerk sind:

Das Netzwerk ist ein Projekt der Abteilung Denkfabrik Digitale Arbeitsgesellschaft des Bundesministeriums für Arbeit und Soziales, wird von ihr finanziell gefördert und koordiniert mit Unterstützung des Think Tank iRights.Lab.

Inhaltsverzeichnis

1. Einleitung

Künstliche Intelligenz (KI) wird in mehr und mehr Bereichen des täglichen Lebens eingesetzt. Auch für die öffentliche Verwaltung sind die Potenziale von KI-Anwendungen enorm. Jedes Jahr werden Millionen Anträge auf Rente, Sozialhilfe oder Arbeitslosengeld gestellt und Informationsanfragen an die Behörden gerichtet. KI-Systeme können die Beschäftigten der Arbeits- und Sozialverwaltung bei ihren Aufgaben unterstützen, Prozesse effizienter machen und Bearbeitungszeiten verkürzen. Nicht zuletzt die Pandemie hat gezeigt, wie wichtig eine moderne, effektive Verwaltung ist. Doch schon heute fehlt vielen Verwaltungen des Bundes und der Länder Personal. Der demografische Wandel wird diese Situation weiter zuspitzen: Wenn die sogenannten Babyboomer aus dem Berufsleben ausscheiden, werden etwa zehn Millionen Menschen weitere Rentenanträge stellen, die bearbeitet werden müssen. Gleichzeitig werden den Behörden dadurch Mitarbeiter*innen fehlen. Das stellt den Sozialstaat vor große Herausforderungen, bei deren Bewältigung KI-gestützte Innovationen helfen können.

Gleichzeitig haben die Arbeits- und Sozialverwaltungen beim KI-Einsatz eine besondere Verantwortung. Die Behörden verarbeiten sehr sensible Daten, und ihre Entscheidungen und Leistungen haben direkte Auswirkungen auf die Bürger*innen, oft in besonders herausfordernden Lebenssituationen. Bereits jetzt kommen KI-Anwendungen punktuell in der Arbeits- und Sozialverwaltung zum Einsatz, etwa beim automatischen Erkennen von Studienbescheinigungen in der Bundesagentur für Arbeit oder bei der Identifikation von aussichtsreichen Regressfällen in der Berufsgenossenschaft Energie Textil Elektro Medienerzeugnisse (BG ETEM). Die Entwicklung und Anwendung von KI in der Arbeits- und Sozialverwaltung steht damit zwar noch am Anfang, gleichwohl ist es von großer Bedeutung, sich bereits jetzt gemeinsam auf grundlegende Rechte und Pflichten, Werte und Prinzipien für den Einsatz von KI in der Arbeits- und Sozialverwaltung zu verständigen und praxisbezogene Leitlinien zu entwickeln.

Der gesetzeskonforme, wertebasierte und wertegebundene Einsatz von KI ist die Grundlage, um ihr Potenzial im Sinne der Gesellschaft und für eine moderne Verwaltung nutzen zu können. Künftig wird mit der EU-Verordnung über Künstliche Intelligenz, die derzeit als Entwurf (COM(2021) 206) vorliegt und verhandelt wird, ein Regelwerk speziell für den Einsatz von KI bestehen. Bereits heute treffen etwa die Datenschutzgrundverordnung oder das Sozial- und Verwaltungsrecht verbindliche Regelungen zu einzelnen Aspekten. Empfehlungen und Prinzipien wie etwa die „Ethik-Leitlinien für eine vertrauenswürdige KI“ der High-Level Expert Group für KI der Europäischen Kommission, das Gutachten der Datenethikkommission, der Abschlussbericht der Enquete-Kommission zu KI des Deutschen Bundestages, die KI-Strategie der Bundesregierung, die „Hambacher Erklärung zur Künstlichen Intelligenz“ der Datenschutzkonferenz oder die Empfehlungen der OECD zu KI beeinflussen die gegenwärtige Diskussion um einen Gestaltungsrahmen für KI. Konkrete Beiträge zu diesem Thema kommen auch aus der Zivilgesellschaft, wie beispielsweise die Algo.Rules von der Bertelsmann Stiftung in Zusammenarbeit mit dem iRights.Lab, das Impact-Assessment-Tool für automatisierte Entscheidungssysteme von AlgorithmWatch im Auftrag des Kantons Zürich oder das Konzeptpapier des Deutschen Gewerkschaftsbundes zum Einsatz von Künstlicher Intelligenz (KI) in der Arbeitswelt.

Das „Netzwerk KI in der Arbeits- und Sozialverwaltung“ hat diese „selbstverpflichtenden Leitlinien für den KI-Einsatz in der behördlichen Praxis“ erarbeitet, um den bestehenden Rahmen zu ergänzen und Orientierung sowie Handhabe für die Anwendung in der behördlichen Praxis im Einklang mit den gesetzlichen Vorgaben zu bieten. Angesichts des zunehmenden Einsatzes von KI auch in der Arbeits- und Sozialverwaltung ist es wichtig, schon vor der Verabschiedung eines regulatorischen Rahmens auf EU-Ebene den Praxiseinsatz zu begleiten.

Die Leitlinien bieten den Behörden eine praxisbezogene Orientierung. Sie geben eine knappe Einführung in die Themen und bieten Handlungsempfehlungen, Orientierungsfragen und Checklisten. Damit werden etwa Projektverantwortliche bei Konzeption, Entwicklung und Betrieb von KI-Systemen mit dem Ziel einer menschenzentrierten Prozessgestaltung unterstützt. Zugleich können sich auch Entscheider*innen, Personalräte, Anwender*innen und Entwickler*innen zu den Grundlagen wertebasierter KI-Gestaltung informieren und ihre jeweiligen Rollen besser ausfüllen. Zudem richten sich die Leitlinien auch an die Öffentlichkeit und damit auch an potenziell Betroffene von KI-basierten Entscheidungen. Auch ihnen gegenüber wird transparent gemacht, welche Werte, Prinzipien und Empfehlungen dem KI-Einsatz zugrunde liegen, und so die Grundlage für Vertrauen und Akzeptanz geschaffen.

Den ersten Teil der Leitlinien bildet eine Wertegrundlage für den KI-Einsatz, auf die sich das Netzwerk verständigt hat und die aus sieben Wertepaaren „Menschenzentrierung & Gemeinwohl“, „Fairness & Nichtdiskriminierung“, „Erklärbarkeit & Transparenz“, „Privatsphäre & Persönlichkeitsschutz“, „Sicherheit & Robustheit“, „Intervenierbarkeit & Verantwortung“ und „ökologische Nachhaltigkeit & Ressourcenschonung“ besteht. Außerdem werden in den weiteren Kapiteln vier zentrale Bereiche der Gestaltung von KI-Systemen in der Arbeits- und Sozialverwaltung näher betrachtet. Die Leitlinien sollen so helfen,

- die Einführungsprozesse menschenzentriert zu gestalten und gemeinsam mit den Stakeholdern Ziele für den KI-Einsatz zu definieren (vgl. Kapitel 3),
- Folgen des geplanten KI-Einsatzes frühzeitig abzuschätzen und mögliche Risiken für unterschiedliche Personengruppen, aber auch die Gesellschaft als Ganzes systematisch zu bewerten (vgl. Kapitel 4),
- eine gute Datenqualität sicherzustellen und Bias zu vermeiden (vgl. Kapitel 5) sowie
- Transparenz über den Einsatz, die Ziele und die Funktionsweisen der KI-Anwendungen zu schaffen und Erklärbarkeit herzustellen (vgl. Kapitel 6).

Damit soll die Einführung KI-basierter Innovationen in der Arbeits- und Sozialverwaltung erleichtert werden. Gleichzeitig soll gewährleistet werden, dass diese den in der Wertegrundlage und den übrigen Kapiteln beschriebenen Werten, Prinzipien und Qualitätsanforderungen entsprechen. Die Leitlinien wurden dabei in erster Linie mit Blick auf Systeme maschinellen Lernens verfasst. Sie sollten jedoch auch bei nichtlernenden Systemen und somit bei allen algorithmischen Entscheidungssystemen (ADM-Systemen) angewendet werden. Im Zentrum stehen dabei stets die soziotechnische Einbettung der Technologie und ihre Auswirkungen auf Bürger*innen, Beschäftigte sowie auf die Gesellschaft.

Entstehung und Weiterentwicklung der Leitlinien: partizipativ, behördenübergreifend und praxisorientiert

*Die Leitlinien wurden vom Netzwerk „KI in der Arbeits- und Sozialverwaltung“ kollaborativ erarbeitet. Das Netzwerk ist ein Projekt der Denkfabrik Digitale Arbeitsgesellschaft des Bundesministeriums für Arbeit und Soziales und wird von ihr finanziell gefördert und koordiniert mit Unterstützung des Think Tanks iRights.Lab. Die Denkfabrik hat im Frühjahr 2021 alle Behörden[1] des Geschäftsbereichs des BMAS in das Netzwerk eingeladen. Aus zwanzig Behörden sind durch Expert*innen und Mitarbeiter*innen im Netzwerk vertreten, die sich aktiv an der Erarbeitung der Leitlinien beteiligt haben. Zahlreiche Expert*innen haben in Workshops, den sogenannten KI-Labs, vielfältige Perspektiven und Fachwissen beigetragen.*

Das Netzwerk wird sich auch weiterhin zum Thema KI-Einsatz in der Verwaltung austauschen. Dabei werden die Leitlinien regelmäßig überprüft und an neue technologische Entwicklungen, sich verändernde gesellschaftliche Anforderungen und rechtliche Vorgaben angepasst sowie um Lernerfahrungen aus der Anwendung in der Verwaltungspraxis ergänzt. Ein wesentlicher Meilenstein bei der kontinuierlichen Weiterentwicklung wird es sein, die Leitlinien nach der Verabschiedung der europäischen KI-Verordnung zu überprüfen und ggf. an den dann geltenden Rechtsrahmen anzupassen. Diese Weiterentwicklung der Leitlinien bietet die Chance, sie den jeweiligen gesellschaftlichen, rechtlichen und verwaltungstechnischen Anforderungen entsprechend in einem bewährten partizipativen Prozess aktuell zu halten. Rückmeldungen zu den Leitlinien sind herzlich willkommen und können bei der weiteren Arbeit berücksichtigt werden.

*In dieser ersten Version der Leitlinien fokussierte sich das Netzwerk zunächst auf die vier zentralen Schwerpunktthemen: menschenzentrierte Einführung, Risiko- und Folgenabschätzung, Datenqualität und Bias sowie Transparenz und Erklärbarkeit. Weitere wichtige Aspekte konnten dabei bisher noch nicht detailliert einfließen. Dazu gehören etwa notwendige Kompetenzen bei Verwaltungsmitarbeiter*innen und ihre Vermittlung in Aus- und Weiterbildung, die grundsätzliche Frage nach der personellen Aufstellung der Behörden, vertiefende Auseinandersetzungen mit Datenschutzaspekten sowie der Zusammenhang der Leitlinien mit anderen Erwägungen wie der Wirtschaftlichkeitsbetrachtung. Der Austausch zu diesen Themen soll in der weiteren Arbeit des Netzwerks vertieft werden.*

Sie wollen sich mit dem Netzwerk austauschen oder haben Anregungen? Schreiben Sie uns gerne eine E-Mail: ki-in-der-verwaltung@bmas.bund.de

1 Dies schließt die Sozialversicherungsträger ein.

2. Wertegrundlage

Das Netzwerk „KI in der Arbeits- und Sozialverwaltung“ hat sich gemeinsam auf grundlegende Rechte, Werte und Prinzipien für den Einsatz von KI verständigt. Diese Werte basieren dabei auf den Ethik-Leitlinien für eine vertrauenswürdige KI der High-Level Expert Group für KI der Europäischen Kommission, dem Gutachten der Datenethikkommission, dem Abschlussbericht der Enquete-Kommission zu KI des Deutschen Bundestages, der KI-Strategie der Bundesregierung, der „Hambacher Erklärung zur Künstlichen Intelligenz“ der Datenschutzkonferenz und den Empfehlungen des Council on Artificial Intelligence der OECD. Die erarbeiteten sieben Wertepaare sind:

- „Menschenzentrierung & Gemeinwohl“,
- „Fairness & Nichtdiskriminierung“,
- „Erklärbarkeit & Transparenz“,
- „Privatsphäre & Persönlichkeitsschutz“,
- „Sicherheit & Robustheit“,
- „Intervenierbarkeit & Verantwortung“ und
- „ökologische Nachhaltigkeit & Ressourcenschonung“

Menschenzentrierung & Gemeinwohlorientierung

Menschenzentrierung bedeutet, dass der Mensch und sein Wohlergehen Ausgangspunkt und Ziel des Einsatzes von KI sind. KI ist für den Menschen da und nicht umgekehrt. Eine menschenzentrierte Entwicklung und Anwendung von KI bedeutet, dass von den Menschen und ihren Bedürfnissen her gedacht wird, um Vertrauen und Akzeptanz zu schaffen und die Rechte der Bürger*innen und Mitarbeitenden zu wahren. Der Einsatz von KI bietet die Möglichkeit, das Zusammenspiel zwischen Technik, Mensch und Umwelt neu zu gestalten und zu verbessern. Entscheidend dabei ist, dass KI-Systeme ganzheitlich betrachtet und im jeweiligen Nutzungskontext verstanden werden. Hierfür gilt es, alle Akteure, deren Nutzungsanforderungen, Bedürfnisse und Werte in die Entwicklung und Implementierung einzubeziehen. Gemeinwohlorientierung betont den Blick auf das Gemeinwesen. Das bedeutet, dass KI möglichst allen in einer Gesellschaft zugutekommen soll und dass mögliche soziale Folgen des KI-Einsatzes und denkbare Auswirkungen auf gesellschaftliche Grundwerte wie Demokratie und Rechtsstaatlichkeit berücksichtigt werden.

Was bedeutet das für die Arbeits- und Sozialverwaltung?

*Planung, Entwicklung und Einsatz von KI sind so zu gestalten, dass diese Werte stets geachtet werden. Bei der Zielsetzung eines KI-Einsatzes ist zu reflektieren, ob die KI wirklich dem Menschen bzw. dem Gemeinwohl dient. Dabei sind sowohl die Bürger*innen als Adressat*innen des Verwaltungshandelns als auch die Mitarbeitenden in den Behörden in den Blick zu nehmen. Inwiefern wird der KI-Einsatz ihren Bedürfnissen gerecht und erfolgt in ihrem Sinne? Hier kann KI Mitarbeitende z. B. von monotonen und belastenden Routineaufgaben entlasten, Wartezeiten für die Bürger*innen verkürzen oder die Qualität von Services und Entscheidungen steigern. Ein inklusiver und barrierefreier Ansatz für Menschen mit Behinderung ist beim KI-Einsatz zu gewährleisten.*

Fairness & Nichtdiskriminierung

KI-gestützte Entscheidungen müssen fair sein. Die Antwort auf die Frage, was fair bzw. gerecht im Einzelnen bedeutet, fällt je nach moralischer, kultureller oder weltanschaulicher Grundposition mitunter recht unterschiedlich aus und ist damit ein Aushandlungsprozess. Deshalb ist es wichtig, alle beteiligten und betroffenen Gruppen bei der Entwicklung von KI einzubeziehen. Auf der Ebene des Rechts treffen die Grundrechte wesentliche Werteentscheidungen, an die staatliche Stellen in ihrem Handeln unmittelbar gebunden sind. Dazu zählt auch das Gebot der Gleichbehandlung, nach dem im Wesentlichen gleiche Sachverhalte nicht ohne eine sachliche Rechtfertigung ungleich behandelt werden dürfen. Darüber hinaus legt das Grundgesetz einen besonderen Schutz vor diskriminierenden Ungleichbehandlungen wegen bestimmter Merkmale fest. Darunter fallen die Benachteiligung oder Bevorzugung wegen einer Behinderung, aus rassistischen Gründen sowie wegen des Geschlechts, der Abstammung, der Sprache, der Herkunft, des Glaubens oder der politischen Anschauung einer Person. Solche diskriminierenden Ungleichbehandlungen können nur im Ausnahmefall aus besonders schwerwiegenden Gründen gerechtfertigt sein. Diese Anforderungen des Grundgesetzes werden beispielsweise im Allgemeinen Gleichbehandlungsgesetz (AGG) sowie in den Sozialgesetzbüchern weiter konkretisiert.

Wer ein KI-System entwickelt oder in seiner Organisation einsetzt, muss daher unbedingt vermeiden, dass die Verwendung des KI-Systems diskriminierende Auswirkungen hat, insbesondere im Hinblick auf die durch das Grundgesetz besonders geschützten Merkmale. Staatliches Handeln muss den Interessen der Betroffenen außerdem im Sinne einer Verfahrensfairness gerecht werden: Auch in einem durch KI unterstützten Verwaltungsverfahren muss gewährleistet sein, dass die einer solchen Fairness dienenden und im Gesetz vorgeschriebenen Maßnahmen wie Anhörungen oder Beteiligungen von Interessenvertretungen beachtet werden. So können bessere Ergebnisse erzielt und damit deren Akzeptanz erhöht werden.

Was bedeutet das für die Arbeits- und Sozialverwaltung?

*Beim KI-Einsatz kann es ungewollt zu Diskriminierungen kommen, beispielsweise wenn die Datensätze, mit denen eine KI trainiert wird, verzerrt sind, etwa weil bestimmte Gruppen über- oder unterrepräsentiert sind (also ein Bias vorliegt). Dann besteht die Gefahr, dass KI-Systeme die bestehenden Ungleichgewichte aus dem Analogen reproduzieren und so verstärken. Denn ohne explizites Gegensteuern bilden die Ergebnisse von KI-Systemen die über die Trainings- und Betriebsdaten eingespeiste diskriminierende Wirklichkeit ab. Durch die Auseinandersetzung mit den Datensätzen, die aus der bisherigen Verwaltungspraxis gewonnen wurden, kann der Prozess zur Einführung von KI zugleich dazu beitragen, bestehende Diskriminierungen aufzudecken und Lösungen dafür zu finden. Um Diskriminierung durch KI-basierte Systeme zu vermeiden, sind deshalb die Qualität der Trainingsdaten und der KI-Modelle sowie eine entsprechende Kompetenz und Sensibilität der Entwickler*innen und Anwender*innen äußerst wichtig. Um Diskriminierungen durch KI-Systeme erkennen zu können, müssen die Systeme ausreichend transparent und erklärbar sein (vgl. „Erklärbarkeit & Transparenz“). Ein diverses Team kann außerdem helfen, Diskriminierung zu vermeiden oder frühzeitig zu entdecken und zu beseitigen.*

Diskriminierungsfreiheit von KI-Systemen ist auch deshalb essenziell, weil sie typischerweise nach der Einführung eine Vielzahl von behördlichen Entscheidungen beeinflussen. Sollten die Outputs eines KI-Systems diskriminierend sein, hätte dies Auswirkungen auf jede einzelne dieser Entscheidungen. Umgekehrt sind – bei entsprechender Erklärbarkeit – Fehler leichter zu erkennen und deren Abstellen führt zu einer Verbesserung in allen Anwendungsfällen des KI-Systems. Der Einsatz von KI kann durch die Konsistenz der Rechenoperationen und die Reproduzierbarkeit von Outputs zu konsistenteren Entscheidungen der Behörde beitragen.

Erklärbarkeit & Transparenz

Erklärbarkeit und Transparenz im Zusammenhang mit KI bedeuten, dass die Beteiligten und Betroffenen die Funktionsweise und die Outputs von KI-Systemen verstehen, überprüfen und hinterfragen können. Je nach Rolle (z. B. KI-Entwickler*in, Behörden und deren Mitarbeiter*innen oder Bürger*in) und Vorwissen sind hier unterschiedliche Informationen und Erklärungen gefragt. Nur so können andere Werte effektiv umgesetzt werden, etwa indem erkannt werden kann, ob Datensätze verzerrt sind oder die KI-Anwendung diskriminierende Parameter nutzt. Zugleich sind sie Grundlage für eine menschliche Kontrolle und Korrektur des KI-Systems. Erklärbarkeit und Transparenz bedeuten auch, dass die Bürger*innen immer erkennen können, dass sie es mit einem KI-System zu tun haben (Kennzeichnung etwa für Chatbots) oder ein KI-System in den Entscheidungsprozess (auch vorbereitend) involviert war.

Was bedeutet das für die Arbeits- und Sozialverwaltung?

*KI-Modelle sollten so konzipiert sein, dass erklärbar und damit nachvollziehbar ist, wie die von ihnen gemachten Vorschläge zustande kommen. Gerade im Kontext der Arbeits- und Sozialverwaltung haben perspektivisch viele unterschiedliche Personen mit einem KI-System zu tun. Die Entwickler*innen und Anwender*innen innerhalb einer Behörde, Aufsichtsführende und Beteiligte aus anderen Behörden, aber auch Bürger*innen sollten bei Bedarf verstehen können, wie das KI-System arbeitet. Dazu sollten Bürger*innen bei jeder Entscheidung informiert werden, wenn zu ihrer Vorbereitung KI-Systeme verwendet wurden. Daran anschließend sollten für Bürger*in-*

nen niedrigschwellige Möglichkeiten gewährt werden, sich über das Zustandekommen der KI-Entscheidung zu informieren.

Insgesamt geht es darum, den Menschen in der Zusammenarbeit mit KI-Systemen zu stärken, indem er notwendige Informationen erhält. Das kann bedeuten, dass die zuständigen Behördenmitarbeitenden angezeigt bekommen, welcher Fehlerwahrscheinlichkeit die Ergebnisse unterliegen, oder die Möglichkeit haben, die Entscheidung „manuell" zu überprüfen und zu korrigieren, indem sie auf die hinter etwaigen Benutzeroberflächen liegenden Daten und Dokumente zugreifen können.

Privatsphäre & Persönlichkeitsschutz

Privatsphäre ist der persönliche Bereich, in dem jede*r sich frei und individuell entfalten kann und in dem jede*r gegen öffentliche und staatliche Einsichtnahme geschützt ist. Dieser Schutz wird erweitert durch das Grundrecht auf informationelle Selbstbestimmung: Jede*r hat die Hoheit, darüber zu entscheiden, ob, wann und wie auf sie*ihn verweisende Daten verwendet werden. Auch wenn einzelne Daten für sich genommen möglicherweise nur einen geringen Informationsgehalt aufweisen, kann der Umgang mit ihnen je nach Ziel der Datenerhebung und der Verknüpfung verschiedener Daten erhebliche Auswirkungen auf die Privatsphäre und die Verhaltensfreiheit der Betroffenen haben. Soweit KI-Systeme personenbezogene Daten verarbeiten, besteht ein besonderes Risiko, da KI-basiert beispielsweise Persönlichkeits- und Verhaltensprofile aus Datensätzen abgeleitet, Bewertungen vorgenommen und zugleich für die Adressat*innen folgenreiche Entscheidungen getroffen werden können. Die Erstellung von Persönlichkeits- und Verhaltensprofilen durch staatliche Stellen ist grundsätzlich nicht erlaubt. Werden personenbezogene Daten für Entscheidungen verwendet, die voraussichtlich ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen zur Folge haben, oder fließen in die typischerweise großen Bestände von Trainingsdaten ein, sind diese Daten einer Risikoabschätzung zu unterwerfen (Datenschutzfolgenabschätzung). Allgemein ist im gesamten Lebenszyklus eines KI-Einsatzes die Einhaltung der Datenschutzbestimmungen sicherzustellen, wobei insbesondere die DSGVO vorrangig zu beachten ist.

Was bedeutet das für die Arbeits- und Sozialverwaltung?

*Die Arbeits- und Sozialverwaltung verarbeitet personenbezogene Daten der Bürger*innen, die oftmals besonders sensibel sind (z. B. Daten zu Kranken-, Ausbildungs- und Erwerbsbiografien; Informationen zur persönlichen, familiären, sozialen und ökonomischen Situation). Die Verwaltung muss deshalb ein besonderes Augenmerk auf den Schutz der Privatsphäre und die informationelle Selbstbestimmung legen sowie die diese Rechte gewährleistenden Datenschutzvorschriften einhalten. Dabei sind auch beim Einsatz von KI-Systemen die Grundsätze für die Verarbeitung personenbezogener Daten zu beachten. Dies bedeutet unter anderem, dass die Daten grundsätzlich nur für den Zweck, für den sie ursprünglich erhoben worden sind und soweit sie zu diesem erforderlich sind, verwendet werden dürfen und auch nur so lange vorgehalten werden dürfen, wie es dieser Zweck gebietet. Zudem müssen die Betroffenen in verständlicher Weise darüber informiert werden, welche und für welche Zwecke Daten zu ihrer Person mithilfe von KI-Systemen verarbeitet werden. Die Daten müssen sachlich richtig und, soweit es darauf ankommt, auch aktuell sein. Zudem sind Daten, für welche der Personenbezug nicht erforderlich ist, zu anonymisieren. Die Mitarbeitenden, die im Kontext von KI mit der Verarbeitung personenbezogener Daten betraut sind, sind besonders für die Anforderungen des Datenschutzes zu sensibilisieren. Funktionierende Datenschutzprozesse fördern gleichzeitig Vertrauen und Akzeptanz bei den Betroffenen: Die Datenschutzfreundlichkeit der KI-Systeme zahlt sich somit doppelt aus. Daneben sind auch die Beschäftigten in der Sozialverwaltung zu schützen. Deshalb ist insbesondere der Einsatz von KI-Systemen abzulehnen, die zur Überwachung der eigenen Beschäftigten genutzt werden können.*

Sicherheit & Robustheit

Sicherheit hat zwei Dimensionen: Im Sinne von Security bedeutet sie, dass ein KI-System angemessen gegen Missbrauch, Angriffe und Sicherheitsverletzungen geschützt ist (z. B. gegen Hacking) und dass es angemessene Notfallpläne für auftretende Sicherheitsrisiken gibt. Safety zielt auf den Schutz der Menschen, die mit dem System interagieren, ab.

Robustheit bedeutet, dass die von KI-Systemen erzeugten Ergebnisse unter allen Bedingungen korrekt reproduzierbar und zuverlässig sind und dass ein KI-System Sachverhalte richtig beurteilt (Präzision). Dies hat eine äußerst hohe Bedeutung bei Anwendungen, die in der Arbeits- und Sozialverwaltung verwendet werden, um zu beurteilen, ob die Voraussetzungen für bestimmte Leistungen vorliegen.

Was bedeutet das für die Arbeits- und Sozialverwaltung?

In der Praxis bedeutet Sicherheit, dass in Abstimmung mit den zuständigen Stellen (z. B. IT-Sicherheit) eine Risikoabschätzung vorgenommen und ein Schutzsystem aufgesetzt wird. KI-spezifische Risiken sind beispielsweise sog. adversarial attacks, d. h. die Manipulation von Trainings- oder Betriebsdaten, um die Ergebnisse zu verzerren.

*Für die Bereiche, in denen der Ausfall eines KI-Systems starke Auswirkungen hätte, erhält die Sicherheit eine noch größere Bedeutung. Dies gilt insbesondere – aber bei Weitem nicht nur – für kritische Infrastrukturen. Hier sind deshalb für den Betrieb von IT-Systemen übliche Vorkehrungen zu prüfen und wo erforderlich zu treffen, wie das Vorhalten von Back-up-Systemen, die Einhaltung des aktuellen Standes von Wissenschaft und Technik, Schulungen der System-Nutzer*innen sowie das Entwickeln von Notfallplänen.*

Intervenierbarkeit & Verantwortung

KI-Systeme müssen während des Einsatzes anpassbar und abschaltbar sein. Zudem sind Verantwortungsbereiche bei der Planung, Entwicklung und beim Einsatz von KI klar festzuschreiben und zuzuteilen, sodass zu jeder Zeit klar ist, wer verantwortlich ist, und sich diese Person auch verantwortlich fühlt und handelt. Besonders wichtig – auch mit Blick auf das zentrale Ziel der menschenzentrierten KI – ist, dass die Letztentscheidung immer bei einem Menschen liegt. Ein verbindliches Recht auf eine menschliche Letztentscheidung wird Betroffenen bereits durch die DSGVO gewährt. Daneben hat das Prinzip der menschlichen Aufsicht beispielsweise auch in das deutsche Verwaltungsrecht sowie den Kommissionsentwurf zur KI-Verordnung Eingang gefunden.

Intervenierbarkeit aus Sicht der Bürger*innen bedeutet, dass die ihnen zustehenden förmlichen Beschwerdemöglichkeiten (Widerspruch und Klage) durch den KI-Einsatz nicht eingeschränkt werden dürfen. Die dafür notwendige Erklärbarkeit des KI-Systems und die Zuweisung von klaren Verantwortlichkeiten müssen deshalb gewährleistet sein. Zudem kann eine zusätzliche Möglichkeit zur Information und Beschwerde über den KI-Einsatz einer Behörde nötig oder sinnvoll sein, um Bürger*innen jenseits förmlicher Verfahren Interventions- und Beschwerdemöglichkeiten zu geben, etwa zu eingesetzten Chatbots.

Was bedeutet das für die Arbeits- und Sozialverwaltung?

Damit die betreibenden Behörden die Kontrolle über das eingesetzte KI-System haben, braucht es über die jederzeitige technische Anpassbarkeit und Abschaltbarkeit hinaus ein entsprechendes Fachwissen bei den zuständigen Stellen. Sobald ein Interventionsbedarf besteht, muss eine Stelle fähig und befugt sein, die notwendigen Anpassungen vorzunehmen und/oder das System einstweilen außer Betrieb zu nehmen. Dazu bedarf es beim Betrieb, aber auch schon bei der Entwicklung von KI-Systemen klarer Rollenbeschreibungen, Verantwortungsbereiche und Entscheidungsbefugnisse. Ferner können Notfallpläne, Back-up-Systeme u. Ä. für solche Situationen angezeigt sein (vgl. „Sicherheit & Robustheit"). Das nötige Wissen ist möglichst breit in den Behörden der Arbeits- und Sozialverwaltung aufzubauen, damit die Beschäftigten informiert mit den KI-Systemen interagieren, Fehler erkennen und melden können. Gleichzeitig muss bei der Entwicklung und beim Einsatz von KI dem „Automatisierungsbias" Rechnung getragen werden. Danach kann ein Mensch, der auf Grundlage einer KI-basierten Empfehlung oder

*Information eine Entscheidung trifft, zu einem übermäßigen Vertrauen in das durch die KI ermittelte Ergebnis neigen. Auch deshalb müssen menschliche Entscheidungsträger*innen das Ergebnis der KI-Anwendung nachvollziehen und informiert einschätzen können.*

Ökologische Nachhaltigkeit & Ressourcenschonung

Ökologische Nachhaltigkeit und Ressourcenschonung beschreiben den vorausschauenden und rücksichtsvollen Umgang mit natürlichen Ressourcen und die Verpflichtung, die Lebensbedingungen zukünftiger Generationen zu sichern und zu erhalten. Ökologisch nachhaltig zu handeln, bedeutet Umweltverschmutzung zu vermeiden, biologische Vielfalt zu erhalten und den Klimawandel zu bekämpfen.

Die Entwicklung und der Einsatz von KI verbrauchen Ressourcen und setzen Treibhausgase frei. Je öfter KI zum Einsatz kommt und je höher der Rechenaufwand einzelner KI-Anwendungen wird, desto wichtiger wird es deshalb, auch hier Nachhaltigkeitsaspekte, Ressourcen- und Energieeffizienz mitzudenken. Nachhaltige KI (Sustainable AI) umfasst sowohl den Einsatz von KI für mehr Nachhaltigkeit als auch die Nachhaltigkeit von KI selbst. Dabei geht es etwa um den Aufbau und die Nutzung energieeffizienter Rechenzentren oder die Entwicklung und Implementierung von maschinellen Lernmodellen und KI-Modellen, die weniger stromintensiv und möglichst langfristig einsetzbar sind.

Die Forschung an „grüner KI“ (Green AI) hat das Ziel, KI-Methoden zu entwickeln, die den Rechenaufwand von KI-Systemen reduzieren, den Energiebedarf senken und so den nachhaltigen Einsatz von KI ermöglichen.

Was bedeutet das für die Arbeits- und Sozialverwaltung?

Bei nachhaltiger KI kommt der Verwaltung vor allem bei der Planung, Entwicklung und Beschaffung von KI große Verantwortung zu. Im Rahmen des technisch Möglichen sollte daher auf möglichst nachhaltige KI-Systeme gesetzt werden. So können staatliche Behörden durch ihre Marktmacht bei der Beschaffung die Nachfrage nach nachhaltiger KI erhöhen und so einen aktiven Beitrag zum Klimaschutz leisten. Darüber hinaus kann auch in der Verwaltung KI zum Klimaschutz genutzt werden, etwa wenn KI in der Gebäudesteuerung zum Einsatz kommt und so hilft, Energie einzusparen.

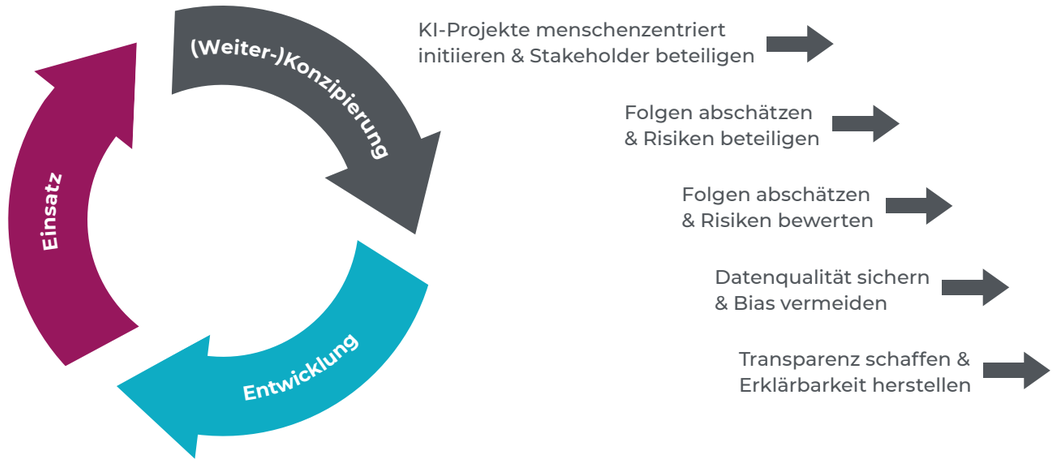
3. Einführungsprozesse menschenzentriert gestalten & Ziele definieren

3.1 Einführung

Die Einführung von IT-Systemen – und damit auch von KI-Anwendungen – beginnt mit der Klärung zentraler Fragen: Was soll erreicht, verbessert oder gelöst werden? Wie sollen diese Ziele erreicht werden? Wen wird das wie betreffen? Diese Fragen sollten ergebnisoffen beantwortet und eine Vorfestlegung auf eine bestimmte Technologie vermieden werden. Sollte sich dabei ein lernendes System als passendes Instrument herausstellen, werden Weichen nicht nur für den Projektaufbau, sondern auch für die Ausgestaltung des KI-Systems bereits in der Einführungsphase gestellt. Deshalb ist es zentral, bereits in dieser initialen Phase von KI-Projekten Menschenzentrierung „by design“ einzubauen, um sie im weiteren Gestaltungsprozess zu beachten. Auch der Entwurf zur KI-Verordnung sieht vor, dass Risikomanagementmaßnahmen und Vorkehrungen zur Sicherstellung menschlicher Aufsicht bereits in der Konzeptionsphase vorzusehen sind.

Die Gestaltung von KI-Systemen folgt, vereinfacht dargestellt, drei Phasen: der Konzeption des Systems, bei welcher etwa Ziele und Umfang festgelegt werden, der Entwicklung, bei welcher das System technisch gestaltet wird, sowie dem Einsatz. Die Phasen lassen sich in viele kleinere Schritte aufteilen, in welchen jeweils Maßnahmen zu treffen sind, um die Menschenzentrierung und Wertebindung des zu entwickelnden Systems sicherzustellen.[2] Die Reihenfolge der Schritte ist dabei, wie auch die Gestaltung von KI-Systemen insgesamt, stark vom Einzelfall abhängig. In einer bedarfszentrierten, agilen[3] Arbeitsweise werden diese drei Phasen etwa nicht abschließend nacheinander durchschritten, sondern können durch Iterationsstufen immer wieder wiederholt und verwoben werden. Außerdem sind KI-Systeme mit ihrer Einführung nicht zu Ende gestaltet. Der Einsatz und die Reaktionen auf das System sowie dessen Auswirkungen führen zu weiteren Überlegungen, Verbesserungen und der Weiterkonzeption. KI-Systeme und ihre Anwendungskontexte können sich während des Einsatzes verändern. Deshalb ist sicherzustellen, dass sie entsprechend ihrem Anwendungskontext und ihrer Risikobewertung regelmäßig überprüft und wenn nötig angepasst werden.

In den weiteren Kapiteln der Leitlinien werden zentrale Aspekte bei der wertegebundenen Einführung von KI-Systemen näher beleuchtet: Folgen abschätzen und Risiken bewerten, Datenqualität sichern und Bias vermeiden sowie Transparenz schaffen und Erklärbarkeit herstellen. All diese Aspekte ziehen sich dabei durch alle drei Phasen der Gestaltung von KI-Systemen, und die Auseinandersetzung mit ihnen beginnt bereits im Rahmen der Konzeptionsphase.



2 Für eine ausführliche Darstellung mit Praxishilfen siehe Algo.Rules „Handreichung für die digitale Verwaltung“. Abrufbar unter https://algorules.org/fileadmin/files/alg/Handreichung_fuer_die_digitale_Verwaltung_Algo.Rules_12_2020.pdf

3 Das bedeutet nicht, dass strikt nach agilen Methoden gearbeitet werden muss, sondern soll eine Problemzentrierung und Offenheit sicherstellen. Die Praxis zeigt zahlreiche Mischformen von „agilen“ und „Wasserfall“-Methoden.

Im Folgenden liegt der Fokus auf der initialen Phase, die den Beginn der Konzeption darstellt. Hier werden wesentliche Erfolgsfaktoren für KI-Projekte gesetzt bzw. beeinflusst – auch wenn diese im Rahmen einer agilen Entwicklung laufend angepasst und iteriert werden. So werden insbesondere:

- Projektziele definiert, vor allem übergeordnete Ziele und Leitbilder sowie später mess- und überprüfbare Unterziele;
- Stakeholder beteiligt, um ihre Perspektiven und ihr Wissen einzubinden sowie Akzeptanz zu schaffen;
- der Gesamtprozess skizziert, indem ein passendes Prozess- und Beteiligungsdesign gewählt wird.

Abschließend findet sich eine Übersicht über Stakeholder und ihre Perspektiven, die in ein Projekt und insbesondere in die initiale Phase eingebunden werden sollten.

3.2 Initiale Phase: KI-Projekte menschenzentriert planen

KI-Projekte können unterschiedliche Startpunkte haben – denn die Idee, ein KI-Projekt zu beginnen, kann auf unterschiedliche Art und Weise entstanden sein: So kann etwa der Wunsch, KI als Technologie zu erproben, im Vordergrund stehen, um als Behörde Erfahrungen zu sammeln. Ebenso kann es sein, dass die Beschäftigten einen konkreten Bedarf nach einem KI-System formuliert haben oder ein zu lösendes Problem aufgekommen ist, für welches der passende Ansatz noch nicht feststeht, KI aber ein mögliches Werkzeug darstellt.

In der initialen Phase sind zahlreiche Aspekte zu klären, Stakeholder einzubeziehen und richtungsweisende Entscheidungen zu treffen. Dazu wurden allgemeine Empfehlungen und Orientierungsfragen im Netzwerk erarbeitet.[4] Wie wichtig die einzelnen Orientierungsfragen sind und in welcher Reihenfolge sie bearbeitet werden sollten, hängt von der Ausgangslage des jeweiligen KI-Projektes ab: Je nachdem etwa, von wem der Impuls kommt oder wie die bestehenden Prozesse bei der Einführung von KI-Systemen ablaufen, sind Stakeholder und Leitfragen unterschiedlich anzugehen.

3.3 Allgemeine Empfehlungen

- **Beteiligung so offen wie möglich gestalten:** Partizipation macht KI-Projekte erfolgreicher. Sind relevante Stakeholder angemessen beteiligt worden, können ihre Perspektiven bei der Entwicklung berücksichtigt werden. Co-kreative Einführungsprozesse führen zu einem besseren Produkt, etwa indem Fehler früher erkannt und Bedarfe besser erfasst werden.[5] Das trägt zudem dazu bei, dass die Akzeptanz des Systems gesteigert und das System schneller und besser genutzt wird.[6] Hierbei ist dann die Beteiligung der zukünftigen Anwender*innen besonders zentral. Allgemein sollte sichergestellt werden, dass alle Stakeholder sich – entsprechend ihrer Rolle – effektiv beteiligen können. So kann es erforderlich sein, auf Bedürfnisse wie spezifische zeitliche Verfügbarkeiten oder fehlende wirtschaftliche Ressourcen insbesondere marginalisierter oder zivilgesellschaftlicher Gruppen Rücksicht zu nehmen. Dabei können Aufwandsentschädigungen in bestimmten Situationen ein taugliches Instrument sein.
- **Arbeitsprozesse analysieren, modellieren und optimieren:** Eine genaue Kenntnis der Prozesse ist Voraussetzung für eine Optimierung oder (Teil-)Automatisierung mithilfe von KI-Systemen. Bei dieser Analyse sollte auf das Prozesswissen aller Beteiligten, v. a. der Beschäftigten der betroffenen Bereiche, zurückgegriffen werden. Bevor aber mit einer Automatisierung begonnen wird, ist zu prüfen, ob der Prozess bereits eine angemessene Qualität aufweist, damit nicht aus einem schlechten Prozess ein schlechter automatisierter Prozess wird.
- **Entwicklung ergebnisoffen und bedarfszentriert anlegen:** Selbst wenn in der initialen Phase viele grundlegende Fragen besprochen werden – sie dürfen nie als abschließend beantwortet betrachtet werden. Erste Vermutungen zu Bedarfen können sich als falsch herausstellen, erste Lösungsansätze mit KI wenig effektiv sein und andere Lösungsmöglichkeiten können sich auftun. Im Zentrum muss immer das zu lösende Problem stehen, z. B. definierte Bedarfe von Stakeholdern. Umgekehrt sollte nicht von einem konkreten Ansatz oder einer Technologie her gedacht werden. Letztere können und sollen stets dem zu lösenden Problem beziehungsweise Ziel angepasst werden. Beispielsweise sollte das Ziel bei zu langen Wartezeiten für Verwaltungsleistungen nicht sein, dass die Sachbearbeiter*innen schneller arbeiten (können), sondern dass die Bürger*innen ihre Leistung schneller erhalten. Dafür sind die Prozesse von Antragseingang bis Bescheid zu analysieren und Optimierungsansätze zu erarbeiten.
- **Folgen abschätzen und mögliche Risiken bewerten:** Mögliche soziale Folgen des KI-Einsatzes für die Betroffenen und denkbare Auswirkungen auf gesellschaftliche Grundwerte wie Demokratie und Rechtsstaat müssen frühzeitig berücksichtigt und mögliche Risiken im Sinne einer Risikoklassenanalyse bewertet werden (vgl. ausführlich Kapitel Folgen abschätzen & Risiken bewerten).
- **Diversität in allen Rollen schaffen:** Diversität in den Entwicklungs- und Projektteams trägt dazu bei, dass KI-Systeme fehlerresistenter, fairer und damit besser gestaltet eingesetzt werden. So können diverse Teams etwa mögliche Quellen von Diskriminierung und Bias früher erkennen und entsprechend gegensteuern.[7] Diversität umfasst dabei verschiedene Dimensionen, insbesondere den fachlichen und den persönlichen Hintergrund. So wird empfohlen, dass Systeme nicht nur von Informatiker*innen gestaltet werden, sondern je nach Projekt z. B. auch von Sozialwissenschaftler*innen oder von Arbeitspsycholog*innen. Zudem sollte das Team Diversität etwa in Bezug auf soziale Herkunft, Migrationsgeschichte und geschlechtliche Identität aufweisen.[8] Ziel ist es, die Vielfalt der Gesellschaft und insbesondere der Anwender*innen und Betroffenen widerzuspiegeln. Diversität einzufordern ist dabei gerade für die öffentliche Hand wichtig, um so auch eine Nachfrage nach einer entsprechenden Teamzusammensetzung bei privatwirtschaftlichen Partnern oder Auftragnehmern zu schaffen und gesamtgesellschaftlich Vorbild zu sein.
- **Transparente und regelmäßige Kommunikation sicherstellen:** Um Stakeholder und ggf. die breite Öffentlichkeit ausreichend über die Entwicklung des KI-Systems zu informieren, sollte proaktiv und offen zum Projekt kommuniziert werden. So können Stakeholder die Gestaltung des Systems eigenständig begleiten und sich selbstständig einbringen.
- **Voneinander lernen:** Viele Behörden in der Arbeits- und Sozialverwaltung, aber auch darüber hinaus, sammeln gerade Erfahrungen zur Gestaltung und Einführung von KI-Systemen. Deshalb kann es hilfreich sein, sich früh im Prozess umzuschauen und nach Projekten zu suchen, die ähnlich angesetzt sind oder ähnliche Ziele verfolgen. Ein Austausch[9] mit den jeweiligen Projektbeteiligten kann helfen, Lernerfahrungen auf das eigene Projekt zu übertragen. Umgekehrt können andere vom eigenen Erfahrungswissen profitieren, wenn dieses geteilt wird.

3.4 Checkliste

1. **Zu lösendes Problem und Ziele bestimmen**
   Was soll mit dem KI-System erreicht werden?
2. **Stakeholder identifizieren und beteiligen**
   Welche Stakeholder bringen welche Interessen ein?
   Wie sollten sie beteiligt werden?
3. **Projektaufbau gestalten**
   Wie kann ein agiles, offenes und menschenzentriertes Projektmanagement sichergestellt werden?

Zu den einzelnen Schritten:

3.4.1 Zu lösendes Problem und Ziele bestimmen
Was soll mit dem KI-System erreicht werden?

Orientierungsfragen zur Bestimmung und Diskussion der Projektziele:

- Welches Problem soll gelöst werden und welche Ziele sollen erreicht werden?
  - Welches Problem war Ausgangspunkt der Überlegungen?

4 Ausgangspunkt dafür war auch die Algo.Rules „Handreichung für die digitale Verwaltung“. Abrufbar unter https://algorules.org/fileadmin/files/alg/Handreichung_fuer_die_digitale_Verwaltung_Algo.Rules_12_2020.pdf

5 Krüger, Lischka (2018): Damit Maschinen den Menschen dienen. Abrufbar unter https://algorithmenethik.de/wp-content/uploads/sites/10/2018/05/Algorithmenethik_L%C3%B6sungspanorama_final_online.pdf

6 Na et al. (2022): Acceptance Model of Artificial Intelligence (AI)-Based Technologies in Construction Firms. Abrufbar unter https://www.mdpi.com/2075-5309/12/2/90/html

7 U. a. https://pollicy.org/resource/inclusion-not-just-an-add-on/, Livingston (2020): Preventing Racial Bias in Federal AI. Abrufbar unter https://doi.org/10.38126/JSPG160205

8 Weitere Dimensionen der Diversität finden sich in: Diversitätsorientierte Organisationsentwicklung. Abrufbar unter http://raa-berlin.de/wp-content/uploads/2018/12/RAA-BERLIN-DO-GRUNDSAETZE.pdf

9 In den sogenannten KI-Labs des Netzwerks „KI in der Arbeits- und Sozialverwaltung“ stellten die teilnehmenden Behördenvertreter*innen eigene KI-Systeme vor und teilten ihre Erfahrungen bei der Entwicklung und dem Betrieb. Solche oder ähnliche Veranstaltungsformate können helfen, Wissen behördenübergreifend zu teilen und sich zu vernetzen.

- Für wen soll dieses Problem gelöst werden?
- Zu welchem übergeordneten Ziel soll damit beigetragen werden?
- Welche weiteren, etwa ökonomischen, haushälterischen oder finanziellen Ziele sind zu erreichen?
- Welche Arbeitsprozesse hängen mit dem zu lösenden Problem zusammen? Welche Arbeitsprozesse in der Behörde oder bei Anwender*innen sollen sich verändern? Wie soll sich der Prozess verbessern? Wie kann er neu gedacht werden? Wie soll sich der Prozess aus Sicht der Mitarbeitenden verändern? Z. B. welche Aufgaben möchten sie selbst bearbeiten, welche sollen aus ihrer Sicht automatisiert werden?
- Welche Rolle kann ein KI-System in diesen bestehenden Abläufen spielen? Ist KI überhaupt ein passendes Mittel zur Lösung des Problems?

- Welche Auswirkungen auf die Gesellschaft oder die Grundrechte von Mitarbeiter*innen oder Betroffenen kann der KI-Einsatz haben? (vgl. Kapitel Folgen abschätzen & Risiken bewerten) Welche Folgerungen für die Ziele sind daraus zu ziehen?

- Wie kann sichergestellt werden, dass das Projekt die Ziele erreicht?
  - Wann ist das Projekt ein Erfolg? Wie lässt sich dieser messen?
  - Was sind Voraussetzungen für diesen Erfolg? Wie kann deren Erfüllung sichergestellt werden?

3.4.2 Stakeholder identifizieren und beteiligen

Welche Stakeholder bringen welche Interessen ein? Wie sollten sie beteiligt werden?

Orientierungsfragen zur Bestimmung der Stakeholder und ihrer Perspektiven:

- Wer sind relevante Stakeholder (Bestimmung anhand der Liste weiter unten)?

- Wie sollen die Stakeholder beteiligt werden?
  - Welches Wissen und welche Perspektiven bringen die Stakeholder ein, von denen das Projekt profitieren kann?
  - Wie sieht eine gelungene Beteiligung aus? Welche Formate sind zielführend, etwa weil sie zum Projekt, zu den Stakeholdern und zur Arbeits- und Mitwirkungskultur der Behörde passen?
  - Wer sollte bereits initial mitsprechen? Wer sollte erst später in der Konzeptions- oder Entwicklungsphase beteiligt werden?
  - Sollen sie dauerhaft, ereignis- oder bedarfsorientiert oder in regelmäßigen Abständen beteiligt werden?

- Welche Erwartungen und Interessen haben die Stakeholder gegenüber dem Projekt?
  - Wie wird das Projekt initial von den Stakeholdern wahrgenommen? Wird es per se als problematisch aufgefasst? Was bedeutet das für die Beteiligungsprozesse?

- Welche Auswirkungen auf die Stakeholder lassen sich abschätzen?
  - Anhand welcher Kenngrößen (z. B. Zufriedenheit von Beschäftigten und Bürger*innen, verkürzte Bearbeitungszeit, bearbeitete Anträge pro Tag) lassen sich Auswirkungen messen?
  - Wie kann dadurch die Zieldefinition konkretisiert werden?
  - Wie muss der Projektaufbau gestaltet sein?

3.4.3 Projektaufbau gestalten

Wie kann ein agiles, offenes und menschenzentriertes Projektmanagement sichergestellt werden?

Orientierungsfragen zur Beteiligung von Stakeholdern im Projekt:

- Welcher grundsätzliche Projektaufbau ist zu wählen?
  - Wie sind die bestehenden Strukturen in der Behörde aufgebaut? Was bedeutet dies für den Projektaufbau?
  - Wie kann agiles Arbeiten sichergestellt werden?

- Wie ist es möglich, die Kenngrößen während des Projekts zu erfassen?

- Welche Räume und Formate sind für die Beteiligung zu schaffen?
  - Müssen Arbeitskreise, projektbegleitende Gruppen o. Ä. geschaffen werden? Welche Aufgaben haben diese und wer ist mit welcher Perspektive beteiligt?

- Wie ist eine ausreichend breite Beteiligung sicherzustellen?
  - Welche Zielgröße ist hinsichtlich der Beteiligung und der Diversität von Stakeholdern zu erreichen? Welche Gruppen sind beispielsweise unbedingt zu beteiligen und in welchem Maße, welche Gruppen können optional beteiligt werden?

3.5 Übersicht Perspektiven und Stakeholder: Was bringen sie in der initialen Phase ein?

Unterschiedliche Stakeholder sollten bei der Gestaltung des späteren KI-Systems mitbestimmen und in den Entwicklungsprozess involviert sein. Sie bringen

jeweils wertvolle Perspektiven ein, die zu einem besseren System beitragen können.

Im Folgenden sind alle Perspektiven und potenziell zugehörige Stakeholder, die in der initialen Phase von Bedeutung sind, im Überblick dargestellt. Dieser basiert auf der Algo.Rules „Handreichung für die digitale Verwaltung“ und wurde im Netzwerk ergänzt und an den Arbeitsbereich der Arbeits- und Sozialverwaltung angepasst.[10] Eine Perspektive kann dabei von einer Person oder Organisation(seinheit) eingenommen werden, ebenso kann eine Person oder Organisation mehrere Perspektiven gleichzeitig haben. Die Auflistung kann helfen, Perspektiven und Stakeholder für das eigene KI-Projekt zu identifizieren und ihre strukturierte Beteiligung in der Konzeptionsphase und darüber hinaus sicherzustellen. Dabei sind nicht alle Stakeholder durchgehend oder gleich zu beteiligen, sondern je nach Phase und Fragestellung einzubinden.

[Alphabetische Darstellung]

Anwendungsperspektive

- Sie betrachtet die Interaktion der Anwender*innen mit dem KI-System bei seinem Betrieb.

- Beispiele: Beschäftigte in der Behörde wie Sachbearbeiter*innen, die die KI-Anwendung in ihrem Geschäftsbereich bedienen oder nutzen, Bürger*innen, die z. B. Chatbots oder Formularausfüllhilfen nutzen (können dann gleichzeitig auch Betroffene sein)

- Initial können Anwender*innen als Expert*innen für die zu optimierenden Prozesse und für mögliche Veränderungen ihrer Arbeitswelt eingebunden werden.

Betroffenenperspektive

- Sie betrachtet, wie sich der Einsatz des KI-Systems auf Betroffene – v. a. ihre Interessen und/oder Grundrechte – auswirkt.

- Beispiele: Bürger*innen, Arbeitssuchende, Beschäftigte, Mitgliedsorganisationen. Können auch durch Intermediäre, z. B. Interessenvertreter*innen, vertreten sein.

- Initial können Betroffene als Expert*innen für die Auswirkung des KI-Einsatzes auf ihre Lebenswelt eingebunden werden.

Datenperspektive

- Sie betrachtet die Arbeit mit und die Verwaltung von Datenbeständen, die Grundlage für das Training und/oder den Einsatz des KI-Systems sein können.

- Beispiele: Fachverantwortliche, ggf. eigene Data-Science-Abteilungen in Behörden, Datenanalyst*innen, Dateneigner*innen

- Initial können diese Stakeholder als Expert*innen einen Überblick über die verfügbaren Daten und ihre Qualität, Aufwände für eine KI-fähige Aufbereitung o.Ä. geben.

Datenschutzperspektive

- Sie stellt sicher, dass die Vorgaben zum Datenschutz eingehalten werden, und berät in Fragen der Datensicherheit und Privatsphäre.

- Beispiele: Datenschutzbeauftragte

- Initial können Datenschutzbeauftragte einzuschätzen helfen, ob und inwieweit personenbezogene Daten durch das geplante KI-System verarbeitet werden und ob dies zulässig ist.

Entscheidungsperspektive

- Sie stellt sicher, dass die Leitungsebene abgebildet wird. Darunter fallen die Zuteilung von Ressourcen (z. B. Zeit, Geld und Personal), die Definition übergreifender Vorgaben, die Einbettung in den übergeordneten politischen Rahmen sowie die Übernahme der Gesamtverantwortung.

- Beispiele: Team- oder Abteilungsleitung, Behördenleitung

- Initial kann ihre Unterstützung notwendige Schritte erleichtern bzw. ermöglichen. Neben Fragen der Wirtschaftlichkeit spielen beispielsweise die öffentliche Wahrnehmung des KI-Projekts, die Sicherstellung des „Erfolges“ sowie die Einbettung des KI-Projektes in übergeordnete politische und behördliche Strategien eine Rolle.

Einsatz- und Implementierungsperspektive

- Sie betrachtet die organisatorische und technische Implementierung der KI-Systeme in bestehende Prozesse. Das beinhaltet die Verknüpfung des Systems mit bestehenden Daten und die Einbettung in eine Anwendungsumgebung. Sie überwacht zudem den fachgemäßen Einsatz und stellt die Software für Anwender*innen bereit.

10 Ausgangspunkt war die Algo.Rules „Handreichung für die digitale Verwaltung“, S. 9–10. Abrufbar unter https://algorules.org/fileadmin/files/alg/Handreichung_fuer_die_digitale_Verwaltung_Algo.Rules_12_2020.pdf

- Beispiele: IT-Abteilungen der Behörden, technische Administrator*innen/Betreiber*innen sowie ggf. externe IT-Dienstleister

- Initial können diese Stakeholder einen Einblick in die organisatorischen Abläufe und die technische Infrastruktur geben, in die das KI-System eingebettet wird. Ggü. den Anwender*innen bringen sie einen Gesamtblick auf die Prozesse ein. Sie können zudem wesentliche Aspekte der technischen und – in Bezug auf die Betriebskosten – wirtschaftlichen Machbarkeit einschätzen.

Entwicklungsperspektive
- Sie betrachtet die Entwicklung des KI-Systems, inklusive Modell und Rahmensoftware in technologischer Hinsicht. Dabei werden Vorgaben von Koordinator*innen umgesetzt.

- Beispiele: externe private IT-Dienstleister, im Fall von Eigenentwicklungen Personen aus IT-Abteilungen der Behörde, Product-Owner.

- Auch wenn Entwickler*innen erst später das System entwickeln, sind sie auch initial einzubinden, um die technische Machbarkeit, Einschätzungen über Aufwand und technische Anforderungen zu bestimmen.

IT- und Informationssicherheitsperspektive
- Sie sorgt für die Gewährleistung der Sicherheit von IT-Systemen v. a. gegen Angriffe von außen bzw. für die Gewährleistung der Vertraulichkeit, Verfügbarkeit und Integrität technischer Systeme. Dazu gehört auch die Beratung der Behördenleitung und Begleitung der KI-Entwicklung. Ohne ihre Zustimmung kann ein KI-System nicht in Einsatz gebracht werden.

- Beispiele: IT-Sicherheitsbeauftragte, Informationssicherheitsbeauftragte

- Initial können diese Stakeholder über die Perspektive der IT-Sicherheit Aspekte der (technischen) Machbarkeit abschätzen sowie einschlägige Sicherheitsstandards identifizieren.

Koordinationsperspektive
- Sie ist federführende Schnittstelle für Planung und Entwicklung und die Interaktion zwischen den Entwickler*innen, Projektträgern und Implementierer*innen. Sie übersetzt zum einen Bedarfe und Ziele in Vorgaben und Prozessschritte und ist verantwortlich dafür, dass die anderen Beteiligten die Vorgaben technisch und praktisch umsetzen. Sie verantwortet zudem eine gelingende Kommunikation des Projektes in der Behörde.

- Beispiele: Referent*innen, Projektmanager*innen

- Initial laufen bei Koordinierer*innen bereits die Fäden zusammen und sie sind maßgeblich an der frühen Konzeption des KI-Systems beteiligt.

Personalrats- und Beschäftigtenperspektive
- Sie vertritt die Interessen des Personals gegenüber der Behördenleitung.

- Beispiele: Personalrat

- Initial kann der Personalrat die Beschäftigtenperspektive einbringen (bzw. muss der Personalrat in bestimmten Fällen einbezogen werden) und dabei eine gelingende Beteiligung der und Kommunikation ggü. den Beschäftigten erleichtern. Bedenken der Beschäftigten wie eine Verhaltens- und Leistungskontrolle, Beschäftigtendatenschutz, Arbeitsplatzverluste oder (zu schneller) Wandel der Arbeitswelt können so frühzeitig erkannt werden. Begleitende Schulungskonzepte für spätere Anwender*innen können z. B. vereinbart werden. Ihre Beteiligung kann maßgeblich zur Akzeptanz beitragen.

Planungsperspektive
- Sie stellt in einzelnen Institutionen den Bedarf an einem KI-Systemen fest und formuliert diesen (z. B. in Form von Anforderungsskizzen, Ausschreibungen oder Aufträgen). Sie gestaltet zudem die Ziele für die Software und plant den späteren Einsatz. Sie muss auch die Bedarfe der Zielgruppen im Blick haben.

- Beispiele: Referatsleitung. Kann identisch mit der Einsatzperspektive sein.

- Initial vertreten Planer*innen einerseits die Perspektive des grundsätzlichen Bedarfs an KI-Systemen und formulieren andererseits – gemeinsam mit anderen Rollen, v. a. den Koordinierer*innen – erste Anforderungsskizzen.

Prüfungs- und Qualitätssicherungsperspektive
- Sie gewährleistet eine unabhängige Prüfung bzw. Qualitätssicherung. Sie erfolgt – je nach Kontext, Projekt und Projektstand – unter verschiedensten Gesichtspunkten (z. B. Anwendungsfreundlichkeit, Technik, Freiheit von Bias). Es kann eigene Stellen geben und/oder durch andere Rollen mitabgedeckt werden, wobei die Unabhängigkeit etwa von der Entwicklungseinheit sichergestellt werden muss.

- Beispiele: Tester*innen, Qualitätssicherungsbeauftragte

- Initial können diese Stakeholder rechtzeitig die Weichen dafür stellen, dass ein hinreichendes Testing bzw. Qualitätssicherungsmechanismen in das Projekt eingebaut werden.

Vertretung für Gleichstellung/Frauen/Diversity
- Sie widmet sich der Förderung und Durchsetzung der Gleichstellung bzw. Diversity in der Behörde. Sie vertritt damit oft Interessen benachteiligter Gruppen.

- Beispiele: Gleichstellungsbeauftragte*r, Frauenbeauftragte*r, Beauftragte*r für Chancengleichheit

- Initial spielen diese Vertretungen gerade dann eine besondere Rolle, wenn das angedachte KI-System die Interessen der vertretenen Gruppen besonders betrifft. Um eine solche Betroffenheit zu erkennen, sollte ihre Expertise eingeholt werden.

Vertretung von Menschen mit Behinderung
- Sie vertritt die Belange von Menschen mit Behinderung und sichert ihre Inklusion und Gleichstellung in der Behörde. Sie arbeitet zudem auf Barrierefreiheit hin.

- Beispiele: Schwerbehindertenbeauftragte*r, Beauftragte*r für die Belange von Menschen mit Behinderung

- Initial spielt diese Vertretung gerade dann eine besondere Rolle, wenn das angedachte KI-System die Interessen der Menschen mit Behinderung besonders betrifft. Um eine solche Betroffenheit zu erkennen, sollte ihre Expertise eingeholt werden.

4. Folgen abschätzen & Risiken bewerten

4.1 Einführung

Gerade in der öffentlichen Verwaltung ist es beim Einsatz von KI-Anwendungen im Rahmen der jeweils geltenden rechtlichen Bestimmungen wichtig, frühzeitig erwartbare Folgen abzuschätzen und mögliche Risiken zu bewerten.

Diese praktische, auf konkrete Anwendungen und ihr jeweiliges Einsatzgebiet bezogene Abschätzung von Folgen und die Bewertung von Risiken sind ein zentraler Schritt bei der Planung des Einsatzes von KI-Systemen. Werden mit Unterstützung eines KI-Systems beispielsweise Anträge auf Sozialleistungen abschließend bearbeitet, kann eine Fehlentscheidung für Betroffene sehr weitreichende Folgen haben. Gleichzeitig gibt es auch Einsatzmöglichkeiten für KI-Systeme, bei denen eine Fehlentscheidung deutlich geringere negative Auswirkungen hat. Beispielsweise wenn ein Chatbot nur unverbindliche Auskünfte und Informationen einer Behördenwebsite in anderer Form bereitstellt.

Die möglichen Risiken sind zwingend zu ermitteln, um sich daraus ergebende Anforderungen an den Implementierungsprozess, die technische Ausgestaltung des Systems, dessen Einbettung in bestehende Prozesse und Abläufe in einer Behörde formulieren sowie entsprechende Maßnahmen bestimmen zu können.

Eine erste Einschätzung sollte daher schon in der Planungsphase stattfinden. Aber auch während des Einsatzes sollten die Folgen erneut geprüft werden, beispielsweise wenn es zu Beschwerden, Ungenauigkeiten oder Fehlern kommt oder wenn es Veränderungen am KI-System oder am Anwendungskontext gibt. Bei kritisch eingeschätzten KI-Systemen ist zudem sicherzustellen, dass sie auch ohne konkreten Anlass regelmäßig überprüft und wenn nötig angepasst werden.

Für die Einschätzung möglicher Folgen und Risiken des Einsatzes von KI-Systemen in Behörden kann etwa ein Verfahren in Anlehnung an die „Kritikalitätsmatrix“ von Tobias Krafft und Katharina Zweig[11] genutzt werden. Dabei werden anhand von zwei Dimensionen die potenziellen Folgen bzw. Risiken von KI-Systemen eingeschätzt und bewertet. Die beiden Dimensionen sind der potenzielle „Schaden“ im Zusammenhang mit den Auswirkungen des Einsatzes des KI-Systems sowie die Abhängigkeit der Betroffenen vom eingesetzten KI-System. Dieser Bewertungsansatz bezieht sich somit auf die Auswirkungen des KI-Einsatzes und erlaubt so eine Risikoeinschätzung zu der jeweiligen konkreten KI-Anwendung unter Berücksichtigung ihres Einsatzgebietes. Auch die Empfehlungen der Datenethikkommission der Bundesregierung[12] sowie der Entwurf der EU-Kommission für eine Verordnung über Künstliche Intelligenz (COM(2021) 206) beinhalten risikobasierte

11 Krafft, Tobias & Zweig, Katharina (2019): Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse. Ein Regulierungsvorschlag aus sozioinformatischer Perspektive. Online verfügbar unter https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf. Die Ausrichtung der Matrix wurde später von Krafft & Zweig aktualisiert, sodass oben rechts die „kritischsten“ Systeme verortet werden, vgl. http://aalab.cs.uni-kl.de/resources/img/RM_KI.png

Ansätze als zentrale Elemente. Mit ihrem Inkrafttreten wird die EU-Verordnung verbindliche Regelungen für eine Risikoeinstufung nach allgemeinen, abstrakten Kriterien treffen, an die sich festgeschriebene Anforderungen knüpfen. Die im Folgenden dargestellte konkrete, praktische Risikoeinschätzung hat sich dabei innerhalb des von der EU-Verordnung gesetzten allgemeinen Rahmens zu bewegen und darf insbesondere nicht zu einer Unterschreitung dieses Rahmens führen. Das im Netzwerk erarbeitete Verfahren zur Risikobewertung auf Basis einer Kritikalitätsmatrix wurde im Vorgriff auf die KI-Verordnung entwickelt und wird unmittelbar nach Verabschiedung der EU-Verordnung erneut überprüft. Das hier vorgestellte Verfahren muss sich in bestehende Rechtsrahmen für den Einsatz von KI, wie sie z. B. Art. 22 DSGVO und § 31a SGB X bei automatisierten Verfahren vorsehen, einfügen.

Das Schädigungspotenzial wird anhand der Oberfrage bestimmt: Welchen potenziellen individuellen und gesellschaftlichen Schaden kann das KI-System erzeugen? Die Abhängigkeit wird anhand der Oberfrage bestimmt: Wie hoch ist die Abhängigkeit von der KI-gestützten Entscheidung und welche Möglichkeiten der Re-Evaluierung gibt es? Die Operationalisierung dieser Fragen erfolgt durch Prüffragen (siehe → Checkliste).

Hinweis:

*Der Vorschlag der Europäischen Kommission (EU-Kommission) für eine EU-Verordnung über Künstliche Intelligenz (COM(2021) 206) schafft neben der Kategorie der verbotenen KI-Systeme (etwa für einige Formen des „Social Scorings" durch Behörden) auch die Kategorie der „Hochrisiko-KI-Systeme". Um die notwendige Rechtssicherheit zu gewährleisten, sieht der Verordnungsentwurf keine Risikobewertung jedes Systems anhand einer konkreten Prüfung im Einzelfall vor. Vielmehr stuft der Verordnungsentwurf bereits (abstrakt) alle KI-Systeme, die in einigen aufgezählten Anwendungsbereichen genutzt werden sollen, als Hochrisiko-KI-Systeme ein. Im Bereich der (Sozial-)Verwaltung werden etwa KI-Systeme, „die bestimmungsgemäß von Behörden oder im Namen von Behörden verwendet werden sollen, um zu beurteilen, ob natürliche Personen Anspruch auf öffentliche Unterstützungsleistungen und -dienste haben und ob solche Leistungen und Dienste zu gewähren, einzuschränken, zu widerrufen oder zurückzufordern sind", als Hochrisiko-KI-Systeme eingeordnet. Der Verordnungsentwurf sieht vor, dass die EU-Kommission auch nach Inkrafttreten der Verordnung im Wege sogenannter delegierter Rechtsakte weitere Anwendungsbereiche für KI-Systeme dem Hochrisiko-Bereich zuordnen kann; hierfür muss sie eine differenzierte Abschätzung der Folgen und Risiken des KI-Einsatzes für den jeweiligen Anwendungszweck vornehmen. Eine solche Bewertung lag auch der ursprünglichen Auswahl der Anwendungszwecke für den Hochrisiko-Bereich zugrunde. Nach dem Verordnungsentwurf sind für alle Hochrisiko-KI-Systeme konkrete Anforderungen und Verpflichtungen an die Anbieter*innen und Nutzer*innen vorgesehen, beispielsweise hinsichtlich Risikomanagement, Datenqualität oder Transparenz. Das vorgeschriebene Risikomanagementsystem für Systeme, die in den Hochrisiko-Bereich fallen, sieht auch eine Risikobewertung des jeweiligen einzelnen Systems vor. Neben den „Hochrisiko-KI-Systemen" sollen für Systeme mit geringerem Risiko Codes of Conduct angewendet werden. Der Verordnungsentwurf befindet sich derzeit in Abstimmung durch den Rat und das Europäische Parlament.*

12 Das Gutachten (von 2019) ist hier abrufbar: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6

4.2 Checkliste

1. **Schädigungspotenzial bestimmen:** *Welchen potenziellen individuellen und gesellschaftlichen Schaden kann das KI-System haben?*

2. **Abhängigkeit bestimmen:** *Wie hoch ist die Abhängigkeit von der KI-gestützten Entscheidung und welche Möglichkeiten der Re-Evaluierung gibt es?*

3. **Verortung auf der Kritikalitätsmatrix vornehmen:** *Wie sind die möglichen Folgen des KI-Systems einzuschätzen?*

Zu den einzelnen Schritten:

4.2.1 Schädigungspotenzial bestimmen: *Welchen potenziellen individuellen und gesellschaftlichen Schaden kann das KI-System haben?*

Um das Schädigungspotenzial eines KI-Systems zu bestimmen, werden die möglichen Folgen einer KI-basierten Fehlentscheidung betrachtet. Dabei ist von einem plausiblen Worst-Case-Szenario auszugehen; auf die Wahrscheinlichkeit des Schadenseintritts kommt es in diesem Schritt nicht an. Dafür kann es hilfreich sein, Szenarien zu möglichen Folgen zu bilden und durchzuspielen. Dies bietet sich gerade dann an, wenn die Fehler und ihre Folgen sehr unterschiedlich ausfallen können und nicht ein passender Fall für die Risikobestimmung gebildet werden kann. Zur Bestimmung des Schädigungspotenzials gibt es zwei Prüffragen mit konkretisierenden Unterfragen:

1 **Auswirkungen auf Personen:** Welche Personen sind wie und mit welcher Intensität betroffen?

- **Wer ist betroffen?**
  Hier sollte neben möglichen **Typen von Betroffenen** (z. B. Antragsteller*innen oder Anwender*innen in der Behörde) auch die **Menge der jeweils Betroffenen** abgeschätzt werden. Neben Menschen können auch juristische Personen (vor allem Vereine und Unternehmen) Betroffene sein.

- **Worin sind diese Personen betroffen? Wie stark sind legitime Interessen betroffen?**

  Hier sind insbesondere **Grund- und Menschenrechte, aber auch sonstige Ansprüche** (z. B. Ansprüche auf Sozialleistungen) in jedem Falle zu betrachten. Wichtige Grundrechte im Kontext der Arbeits- und Sozialverwaltung sind etwa die Ausbildungs- und Berufsfreiheit, das Recht auf körperliche Unversehrtheit, die Persönlichkeitsrechte, insbesondere das Recht auf informationelle Selbstbestimmung, und der Gleichbehandlungsgrundsatz neben besonderen Diskriminierungsverboten (etwa bezüglich Geschlecht, Herkunft, Alter und Glaube).

- **Wie hoch ist die individuelle Betroffenheit?**

  Die individuelle Betroffenheit ist qualitativ zu bestimmen. Dabei sind objektive Aspekte (z. B. Höhe des potenziellen finanziellen Schadens oder Bedeutung einer nichtmonetären Leistung wie etwa die Teilnahme an einer Rehamaßnahme oder an einer Weiterbildungsmaßnahme) und konkrete individuelle Auswirkungen (z. B. Abhängigkeit der Betroffenen von der finanziellen Leistung, besondere persönliche Umstände und soziale Folgen einer Leistungsverweigerung) zu berücksichtigen.

2 **Auswirkungen auf Gesellschaft und gesellschaftliche Güter oder Grundprinzipien:** Wie sehr birgt das System direkt oder indirekt, kurz- oder langfristig das Risiko, die Gesellschaft als Ganzes oder gesellschaftliche Güter zu beeinträchtigen?

- **In welchem Maße ist die Gesellschaft über die Ebene individueller Betroffenheit hinaus „als Ganzes" betroffen?**

  Dies kann etwa der Fall sein, wenn das grundsätzliche Vertrauen in die Richtigkeit staatlicher Informationen erschüttert wird oder das KI-System Auswirkungen auf größere gesellschaftliche Prozesse hat, wie beispielsweise Wahlen, Arbeitnehmer*innen-Vertretungen, den öffentlichen Diskurs, das prinzipielle Verhältnis zwischen Arbeitnehmer*innen und Arbeitgeber*innen.

- **In welchem Maße sind gesellschaftliche Güter wie Rechtsstaatlichkeit, Demokratie, Sozialstaatlichkeit oder Umwelt betroffen?**

  Digitale Technologien nehmen direkt und indirekt Einfluss auf die Gesellschaft und können dadurch auch zu Herausforderungen bei gesellschaftlichen Gütern oder Grundprinzipien führen. Beim Einsatz von KI-Systemen ist deswegen zu betrachten, welche Auswirkungen diese beispielsweise auf die Entfaltung der Demokratie oder auf die soziale Gerechtigkeit haben können.

4.2.2 Abhängigkeit bestimmen: *Wie hoch ist die Abhängigkeit von der KI-gestützten Entscheidung und welche Möglichkeiten der Re-Evaluierung gibt es?*

Die Abhängigkeit von einem KI-System wird anhand der Dimensionen Umschaltbarkeit, (menschliche) Kontrolle und Wiedergutmachung bestimmt. Dazu gibt es drei Prüffragen mit konkretisierenden Unterfragen:

1 **Umschaltbarkeit (oder Switchability):** Wie gut ist es möglich, dem KI-System bzw. dessen Entscheidung auszuweichen?

- **Besteht Umschaltbarkeit aus Sicht der Behörde?** Ist der Prozess auch ohne Unterstützung des eingesetzten KI-Systems durchführbar? Wie leicht ist es für die Anwender*innen, eine Entscheidung ohne Unterstützung des KI-Systems zu treffen? Gibt es die Möglichkeit, das KI-System gegen ein anderes auszutauschen? Ist eine Entscheidung ohne Unterstützung des eingesetzten KI-Systems nicht möglich, besteht eine höhere Abhängigkeit von diesem System. Auch die fehlende Möglichkeit, das KI-System zu wechseln, steigert die Abhängigkeit.

- **Besteht Umschaltbarkeit aus Sicht der Bürger*innen?** Können Bürger*innen dem KI-System einer Behörde durch Wechsel der Behörde ausweichen? Oder können die Bürger*innen in dem Prozess der Behörde der KI ausweichen? Wie gestalten sich alternative Zugänge beispielsweise zu einer Leistung? Gibt es nicht-KI-basierte Zugänge? Wie leicht sind diese Alternativen zugänglich?

- **Ist gewährleistet, dass die Überprüfung, die im Rahmen einer Beschwerde (z. B. eines Widerspruchs) stattfindet, ohne Verwendung des KI-Systems durchgeführt wird?**

  Insbesondere sollte bei einem Widerspruch tatsächlich re-evaluiert werden und nicht einfach derselbe Prozess ohne Änderungen wiederholt werden, sondern eine substanzielle Einzelfallprüfung durch einen Menschen stattfinden.

2 **Menschliche Kontrolle:** Inwieweit werden die Entscheidungen und Handlungen eines KI-Systems regelmäßig zusätzlich durch sinnvolle menschliche Interaktion geprüft?

- **Inwieweit wird der Output, den das KI-System generiert, im Zuge der Entscheidung geprüft? Welche Rolle spielt das KI-System dabei in dem (Entscheidungs-)Prozess, in den es eingebettet ist?**

  Je weniger die Ergebnisse eines KI-Systems im Entscheidungsprozess durch Menschen überprüft werden, also je selbstständiger die Entscheidung des Systems erfolgt, desto kritischer im Sinne der Matrix ist das System zu bewerten. Vollautomatisierte Entscheidungen wären unter diesem Gesichtspunkt maximal kritisch.[13] Bei entscheidungsunterstützender Einbettung ist auf die tatsächliche Wirksamkeit der menschlichen Kontrolle und Entscheidung abzustellen. Hierbei ist beispielsweise relevant, dass Sachbearbeiter*innen ausreichende Informationen, Zeit und Kompetenzen erhalten bzw. haben, um die Prüfung durchzuführen. Auch gilt es zu prüfen, inwieweit die Bearbeiter*innen tatsächlich noch einen Entscheidungsprozess durchführen. Um zu verhindern, dass die Ergebnisse von KI-Systemen lediglich abgenickt werden, gibt es technische Möglichkeiten (z. B. wenn das System mehrere Indikatoren liefert und Bearbeiter*innen diese aktiv nutzen müssen) oder Gestaltungsmöglichkeiten im sozialen Prozess. Hier ist auch die Frage wichtig, wie in einer Behörde damit umgegangen wird, wenn ein*e Bearbeiter*in sich gegen den Vorschlag einer KI-Anwendung entscheidet.

Vertiefung: Wie kann effektive Kontrolle bei einer Vorsortierung durch KI eingeordnet werden?

Wird von KI-Systemen eine Aufbereitung oder Vorsortierung vorgenommen, aber jeder Fall händisch vollumfänglich geprüft, besteht maximale menschliche Kontrolle. Hat umgekehrt die entscheidende Person praktisch keinerlei Zeit und/oder Informationen oder fehlen ihr Beurteilungskompetenzen, um die Vorarbeit zu prüfen, und „winkt" sie deshalb (nicht nur ausnahmsweise) als Entscheidung durch, so ist die menschliche Kontrolle als minimal einzustufen. Dazwischen liegen Fälle, in denen der Mensch nicht jeden Fall prüft, sondern sowohl die vorgeschlagenen Fälle bearbeitet und vor einer Entscheidung vollumfänglich prüft sowie alle anderen Fälle bezüglich ihrer Nichtauswahl z. B. auf informierter Basis stichprobenartig prüft.

Bei Vorsortierungen durch KI-Systeme – z. B. durch Bilden einer Rangliste – ist zu beachten, dass hier bezüglich zweier KI-Outputs die menschliche Kontrolle geprüft werden muss. Einerseits muss betrachtet werden, ob und inwieweit die ausgewählten Fälle händisch geprüft und in eine Entscheidung überführt werden. Andererseits muss geprüft werden, welche Wirkung der Vorgang der Vorauswahl hat und ob dieser auch menschlicher Kontrolle unterliegt. Denn durch die Vorauswahl wird der Kreis der Fälle, die weiterbearbeitet werden können, vorbestimmt.

13 Allerdings kommt es bei der Gesamtbewertung eines solchen Systems auf den Anwendungskontext an. Hat beispielsweise die Entscheidung kaum Relevanz, kann ein Fehler praktisch keinen Schaden verursachen, und sind die nachträglichen Korrekturmöglichkeiten (für Betroffene und Anwender*innen) sehr gut, dann kann sich ein solches System als insgesamt unkritisch darstellen.

3 **Korrigierbarkeit:** (Wie gut) Ist es möglich, gegen eine KI-basierte Entscheidung vorzugehen oder sie zu korrigieren?

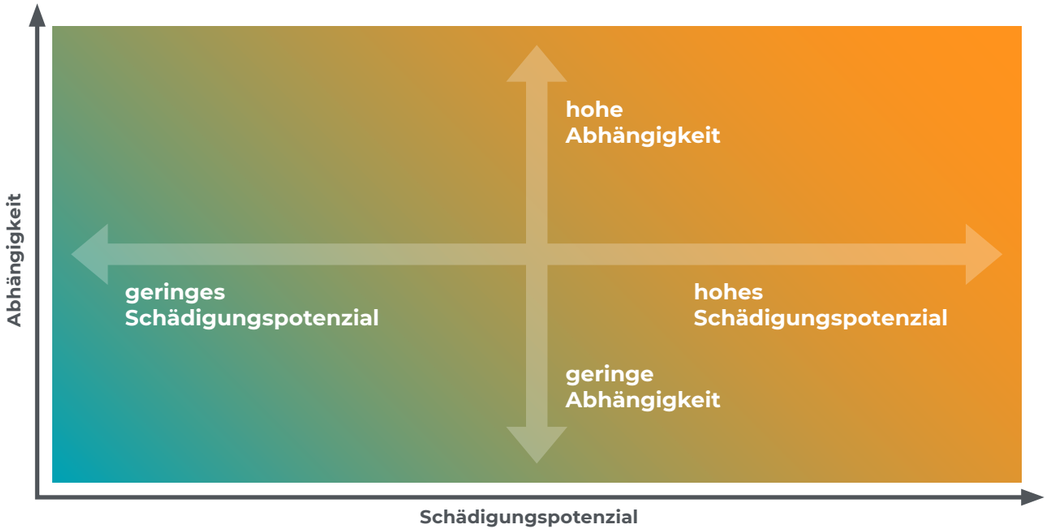
- **Welche Möglichkeiten haben die Betroffenen, gegen die Entscheidung vorzugehen? Gibt es z. B. Rechtsbehelfe oder andere Möglichkeiten? Wie zugänglich sind diese Möglichkeiten?** Ist es den konkret betroffenen Personen tatsächlich möglich, das Instrument einzusetzen? Dies ist dann nicht beziehungsweise nur eingeschränkt der Fall, wenn kein Wissen um die Möglichkeit besteht, Betroffene sich der Interaktion mit einer KI nicht bewusst sind und/oder das Verfahren zu kompliziert oder aufwendig ist.

- **Wie effektiv sind die Möglichkeiten?**

- Können z. B. gesamte Entscheidungen angefochten werden und wird bei einer Beschwerde eine umfängliche Prüfung vorgenommen oder nur eine kursorische? **Wie viel Zeit benötigt die Behörde, um eine Beschwerde angemessen zu bearbeiten? Wie ist die Situation für die Betroffenen während der Prüfung?** In der Zeit der Bearbeitung kann sich die Situation der Betroffenen verschlechtern und weiterer Schaden entstehen, etwa weil Leistungen, von denen die betroffenen Personen abhängig sind, nicht ausgezahlt werden.

4.2.3 Verortung auf der Kritikalitätsmatrix vornehmen: *Wie sind die möglichen Folgen des KI-Systems einzuschätzen?*

Entsprechend den oben genannten Prüffragen wird eine Einordnung auf den beiden Achsen vorgenommen. Dabei wird jeweils vom ungünstigsten Fall ausgegangen.



4.3 Beispielhafte Folgenabschätzungen

Um die verschiedenen Abstufungen und Übergänge innerhalb der Matrix besser einschätzen zu können, finden sich im Folgenden zwei – realitätsnahe, aber fiktive – Beispiele. Dabei soll verdeutlicht werden, dass die Kritikalität erheblich vom Kontext abhängt. In der Dimension Abhängigkeit hängt die Kritikalität v. a. von der Einbettung der KI in den Entscheidungsprozess ab.

Beispiel 1 für Abhängigkeit: Vollständigkeitsprüfung von Anträgen

Ein KI-System wird eingesetzt, um digital eingegangene (Leistungs-)Anträge auf Vollständigkeit zu prüfen. Ist ein Antrag unvollständig, liefert es eine Warnung. Falls dieses System direkt bei der Antragstellung eingesetzt wird und die Warnung an die*den eintragende*n Bürger*in herausgegeben wird, sind sowohl Schädigungspotenzial als auch Abhängigkeit niedrig einzuschätzen. Bürger*innen erkennen den Einsatz

dieses Prüfsystems sofort, können ihre Fehler korrigieren und den Antrag dann einreichen. Gibt das KI-System nicht nur eine Warnung aus, sondern muss das Prüfsystem für die Anträge grünes Licht geben, steigt die Abhängigkeit geringfügig an, da Bürger*innen dem System nun nicht mehr ausweichen können. Wird das Prüfsystem nicht als Unterstützung bei der Antragstellung, sondern bei der Prüfung in der Behörde angewendet, steigt die Abhängigkeit weiter, weil das System nun eine (wenn auch geringfügige) Rolle im Entscheidungsprozess zum Antrag spielt. Wird nur eine Warnung ausgegeben, die zwingend von der Behörde geprüft werden muss, bevor eine Aufforderung verschickt wird, fehlende Informationen oder Dokumente nachzureichen, ist die Abhängigkeit geringer, als wenn das Prüfsystem unvollständige Anträge automatisch aussortiert und deren Prüfung faktisch kaum stattfindet, bevor ein automatisch erstelltes Schreiben erstellt wird, das fehlende Dokumente anfordert. Die Abhängigkeit steigt auch dann, wenn die Behördenmitarbeitenden nicht die Ressourcen oder Kompetenzen haben, um das Ergebnis des KI-Systems gründlich zu prüfen, bevor der Antrag abgelehnt wird. Ebenso problematisch ist es, wenn eine gründliche Prüfung für die Sachbearbeiter*innen mit hohen Hürden (z. B. Begründungsaufwand gegenüber höheren Stellen) verbunden ist. Die Abhängigkeit steigt erheblich, wenn das System vollautomatisiert über die Vollständigkeit von Anträgen entscheidet und Bürger*innen automatisch eine Ablehnung erhalten und zudem nicht ohne Weiteres einen neuen Antrag stellen können.

Beispiel 2 für Schädigungspotenzial: Auszahlung von (Sozial-)Leistungen

[Im Folgenden wird ausschließlich die Dimension Schädigungspotenzial und nur der individuelle Schaden betrachtet]

KI-Systeme, die bei der Auszahlung von Leistungen verwendet werden, sind desto kritischer, je höher die in Rede stehende Summe ist, je stärker die Abhängigkeit der Betroffenen von der Leistung ist und je mehr Personen betroffen sind. Geht es um 1.000 € anstatt um 100 €, sind eher ärmere Menschen oder 500 statt 50 Personen betroffen, stellt sich das System als kritischer dar. Geht es um Leistungen zur Existenzsicherung (auch der Angehörigen) liegt tendenziell eine höhere Kritikalität als bei sonstigen Leistungen vor.

Die praktische Herausforderung liegt in der belastbaren Abschätzung, wer worin wie stark von einer Fehlentscheidung betroffen ist.

4.4 Maßnahmen zum Umgang mit hoher Kritikalität treffen: Welche Folgerungen ergeben sich aus der Kritikalitätsbewertung?

Mit höheren Risiken steigen tendenziell die Anforderungen an den Betrieb eines KI-Systems. Eine abschließende und vollständige Darstellung der Zusammenhänge zwischen KI-System, seinem Anwendungskontext und nötigen Maßnahmen gibt es derzeit noch nicht. Der Entwurf der KI-Verordnung der EU-Kommission sieht beispielsweise für Hochrisiko-KI-Systeme umfangreiche Mindestanforderungen vor und verbietet bestimmte, besonders kritische Anwendungen. Innerhalb des von der KI-Verordnung gesetzten Rahmens ist die Anwendung der Kritikalitätsmatrix als unterstützendes Instrument bei der Risikobewertung für die behördliche Praxis zu sehen, das bei der Entscheidung über den Einsatz eines KI-Systems und zur Entwicklung von Sicherungsmaßnahmen eingesetzt werden kann. Weitere Vorgaben (behördenintern oder aus bestimmten Rechtsvorschriften), die den Einsatz von KI-Systemen regeln, sind damit aber nicht ausgeschlossen, sondern müssen vollumfänglich beachtet werden. Maßnahmen und Orientierungsfragen zu einzelnen Bereichen wie Datenqualität oder Erklärbarkeit und Transparenz finden sich in den folgenden Kapiteln. Nach Inkrafttreten der KI-Verordnung werden die selbstverpflichtenden Leitlinien für die Praxis auch mit Blick auf die Folgenabschätzung überprüft und angepasst.

5. Datenqualität sicherstellen & Bias vermeiden

5.1 Einführung

Eine hohe Datenqualität ist von essenzieller Bedeutung für jedes datenbasierte Verwaltungshandeln. Bezogen auf KI gilt: Für alle KI-Anwendungen bedarf es einer guten Datengrundlage, also hinreichend vieler aktueller, aussagekräftiger, repräsentativer und fehlerfreier Daten. Die konkreten Anforderungen an die Datengrundlage sind im Einzelfall je nach Kontext und KI-Modell zu bestimmen. Umgekehrt führen schlechte Daten bzw. eine geringe Datenqualität oft dazu, dass ein Training für den gewünschten Einsatz nicht möglich ist oder dass die Outputs des trainierten KI-Systems ungenauer und weniger verlässlich sind sowie mehr Testläufe erforderlich sind. Außerdem erhöhen schlechte Daten die Wahrscheinlichkeit, dass die Wirklichkeit verzerrt dargestellt wird, also ein Bias vorliegt. Daher stellt der Entwurf der EU-Kommission für eine KI-Verordnung auch verbindliche Qualitätskriterien für Trainings-, Validierungs- und Testdatensätze auf, die zur Entwicklung eines Hochrisiko-KI-Systems zu verwenden sind.

Eine hohe Datenqualität bei der Entwicklung von KI-Systemen sicherzustellen, zahlt nicht nur auf eine spezifische KI-Anwendung ein, sondern kann das Gesamtniveau der Daten innerhalb einer Verwaltungseinheit heben und damit weitere datenbasierte Anwendungen ermöglichen (wie z. B. Business-Intelligence-Anwendungen, Dashboards etc.).[14] Sobald personenbezogene Daten betroffen sind, müssen die sich aus dem Datenschutzrecht ergebenden Anforderungen, wie Rechtmäßigkeit, Zweckbindung, Datenminimierung und Vertraulichkeit, aber auch die Prinzipien Datenrichtigkeit, -aktualität und -vollständigkeit, die zugleich Datenqualitätsanforderungen sind, erfüllt werden. Ein Fokus dieses Kapitels liegt auf dem Zusammenhang von Datenqualität und Bias. KI-Systeme übernehmen, wenn nicht aktiv gegengesteuert wird, die den Trainingsdaten inhärenten verzerrten Abbildungen der Wirklichkeit. Beruht z. B. die Entscheidung über Sozialleistungen maßgeblich auf Ergebnissen eines KI-Systems, muss sichergestellt werden, dass die zugrunde liegenden Daten keine Verzerrungen, etwa bzgl. Geschlecht, Ethnie, Religion oder Alter, aufweisen. Eine hohe Datenqualität verringert die Wahrscheinlichkeit von Bias und reduziert damit eine Ursache von diskriminierenden KI-Entscheidungen, gegen die das Grundgesetz einen besonderen Schutz festlegt, der beispielsweise im Allgemeinen Gleichbehandlungsgesetz (AGG) sowie in den Sozialgesetzbüchern weiter konkretisiert wird (vgl. Wert „Diskriminierungsfreiheit“).

Was ist ein Bias? Welche Bedeutung hat er für Diskriminierung?

„In der Informatik bezeichnet man mit Bias ein Fehlverhalten, das auf einer systematischen Verzerrung beruht. Da das Verhalten von KI-Systemen auf gelernten Zusammenhängen basiert, ist in der Regel die Beschaffenheit der dafür verwendeten Trainingsdaten für den Bias in KI-Systemen ursächlich.“[15] So definiert es die KI-Enquete-Kommission des Bundestags in ihrem Abschlussbericht. Bias (engl. „Verzerrung, Vorurteil“) stellen für eine datenbasierte Arbeit und dabei v. a. Systeme maschinellen Lernens eine erhebliche Herausforderung dar.[16] Denn ein KI-basiertes System ist nur so gut wie die Daten, mit denen es trainiert wurde, getreu der Faustregel: „Garbage in – garbage out!“

Ein Bias liegt beispielsweise in einem Datensatz vor, wenn dieser eine verzerrte Abbildung der Realität wiedergibt. So könnten in einem Datensatz wesentlich mehr Daten von Männern als von Frauen erfasst sein, obwohl die Grundgesamtheit ausgeglichen ist. Für die Prüfung ist es daher wichtig, die Struktur der Grundgesamtheit zu kennen, auf die sich die KI-Anwendung bezieht (beispielsweise die Anteile von Frauen und Männern unter allen Beschäftigten). Ein Bias in den Daten führt zu Diskriminierung, „wenn die Datenauswahl ein systematisches Fehlverhalten

14 Dies setzt i. d. R. ein Zurückspielen der bereinigten Daten in die Ausgangsdatenquelle, z. B. die Fachverfahren, voraus.

15 Abschlussbericht der KI-Enquete des Bundestages auf BT-Drs. 19/23700, S. 60.

16 Zugleich können Bias bei Lernvorgängen von KI eine wichtige Rolle spielen, um das Generalisieren von Zusammenhängen zu ermöglichen, vgl. Abschlussbericht der KI-Enquete des Bundestages auf BT-Drs. 19/23700, S. 60 m. w. N.

des KI-Systems hervorruft, sodass Menschen aufgrund von äußeren und inneren Persönlichkeitsmerkmalen ungerechtfertigt bevorteilt oder benachteiligt werden“[17]. Ein Bias liegt auch vor, wenn ein KI-System bestehende Diskriminierung reproduziert.[18] Das bedeutet, dass es in diesen Fällen selbst bei einer perfekten Abbildung der Realität in den Daten zu Diskriminierung kommt. Ganz allgemein wird ein KI-System bestehende Diskriminierungen reproduzieren, wenn nicht aktiv gegengesteuert wird.

Ursachen für Bias sind vielfältig:[19] Sie können in „verzerrten“ Datenerhebungsprozessen begründet sein, etwa weil nicht repräsentativ erhoben wird, etwa indem bestimmte Gruppen überproportional stark in den zugrunde liegenden Fällen vertreten sind oder an Umfragen teilnehmen[20] oder weil Sensoren zur Datenerhebung nicht repräsentativ verteilt sind. Im Fall von KI kann zudem eine Verzerrung auftreten über die Auswahl der Trainingsdaten und auch im laufenden Einsatz über die Auswahl der Betriebsdaten. Erfolgt eine Auswahl nicht repräsentativ, wird auch eine ursprünglich repräsentative Datengrundlage verzerrt bzw. (bzgl. Betriebsdaten) ein verzerrter Output generiert.

Neben einem Bias in den Daten gibt es vielfältige weitere Arten von Bias, z. B. kognitive, statistische und induktive Bias,[21] die sich auch auf die Gestaltung von KI-Systemen auswirken können.

Um Bias und Diskriminierung zu erkennen und zu vermeiden bzw. zu korrigieren, müssen sowohl die Daten und die Lern- und Korrekturprozesse als auch die soziotechnische Einbettung des KI-Systems betrachtet werden.

Beispiele für Bias in den Daten

Zahlreiche Beispiele für verzerrte Daten aus den letzten Jahren sind bekannt. Etwa eine KI-Anwendung von Amazon, die einen Score für Bewerber*innen vergab und dabei Männer massiv bevorzugte, weil das System aus Daten der Vergangenheit lernte.[22] Dabei wurde das Geschlecht zwar nicht explizit in den Trainingsdaten verwendet, allerdings gab es andere, stark mit dem Geschlecht korrelierende Merkmale in den Lebensläufen, anhand derer die KI das Geschlecht „erkennen“ konnte. Wird eine solche Diskriminierung gegenüber einem bestimmten Geschlecht von den Verantwortlichen identifiziert, ist bei der Beschaffung und Auswertung weiterer Daten darauf zu achten, dass dieser Bias korrigiert wird.

Beispiel: Bias aufgrund unterschiedlicher Organisationsstrukturen bei der (betriebs-)ärztlichen Versorgung in den Vergleichsgruppen Chemie- und Lederindustrie

Es wurde seitens der Berufsgenossenschaft analysiert, wie viele (und welche) Berufskrankheiten in dem jeweiligen Industriezweig auftreten. Dazu wurden die Fälle erkannter Berufskrankheiten zusammengetragen und ins Verhältnis zur Mitarbeitendenzahl gesetzt. Ergebnis war, dass die Chemieindustrie deutlich mehr Berufskrankheiten (pro Mitarbeitende*n) aufwies als die Lederindustrie. Dieses Ergebnis beruht aber auf einer Überrepräsentation von Krankheitsfällen in der Chemieindustrie, was in der unterschiedlichen Erhebung der Fälle begründet ist. Ursache war das wesentlich dichter organisierte Gesundheitsnetz in der Chemieindustrie, das z. B. eine regelmäßige Untersuchung durch Betriebsärzte vorsah, was die Lederindustrie nicht kannte. Dadurch war in vielen Fällen überhaupt erst die Möglichkeit geschaffen worden, Berufskrankheiten zu erkennen. Ohne Korrektur kann dies zum Fehlschluss führen, dass die Chemieindustrie „gefährlicher“ als die Lederindustrie sei.

17 Abschlussbericht der KI-Enquete des Bundestages auf BT-Drs. 19/23700, S. 61.

18 Vgl. Abschlussbericht der KI-Enquete des Bundestages auf BT-Drs. 19/23700, S. 61.

19 Vgl. weitere Ursachen für Diskriminierungen, BMFSFJ, Wegweiser digitale Debatten. Teil 2: Algorithmenvermittelte Diskriminierung, S. 7. Abrufbar unter https://www.bmfsfj.de/resource/blob/186300/961021829a491933cf24e8f06ff8018f/wegweiser-digitale-debatten-teil-2-data.pdf

20 Dazu gibt es im englischsprachigen Raum den griffigen Ausdruck „WEIRD sample“ (Western, Educated, Industrialized, Rich und Democratic Society).

21 Vgl. eine Übersicht bei Bias in algorithmischen Systemen – Erläuterungen, Beispiele und Thesen der Initiative D21, abrufbar unter https://initiative21.de/app/uploads/2019/03/algomon_denkimpuls_bias_190318.pdf

22 Vgl. The Guardian (2018): Amazon ditched AI recruiting tool that favored men for technical jobs. Online verfügbar unter https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

5.2 Checkliste: Welche Schritte sind zur Sicherung einer guten Datenqualität nötig?

Die folgenden Schritte helfen dabei, sich mit dem Thema gute Datenqualität einleitend auseinanderzusetzen. Die Schritte sind nicht zwingend linear abzuarbeiten und bilden einen Ausgangspunkt für die Begleitung von KI-Prozessen aus Sicht der Datenqualität.

1. **Ziele des KI-Einsatzes, (Daten-)Bedarfe und anwendungsbezogene Datenqualitätskriterien definieren:** *Welches Problem soll gelöst werden und welche Daten (-quantität) werden dafür in welcher Qualität benötigt?*
2. **Bestand verfügbarer Daten ermitteln und Datenqualität prüfen:** *Erfüllen die verfügbaren Daten die definierten Qualitätsanforderungen?*
3. **Datenaufbereitung und -bereinigung durchführen:** *Wie können die verfügbaren Daten so aufbereitet werden, dass sie die erforderliche Qualität aufweisen?*
4. **Bias finden und verhindern:** *Wie können die Mitarbeitenden (v. a. Datenverantwortliche, aber auch Anwender*innen) für Bias sensibilisiert werden? Welche (technischen) Möglichkeiten zur Vermeidung von Bias gibt es? Wie können die diskriminierenden Auswirkungen eines Bias verhindert werden?*

Schritte im Detail

5.2.1 Ziele des KI-Einsatzes, (Daten-)Bedarfe und anwendungsbezogene Datenqualitätskriterien definieren

Welches Problem soll gelöst werden und welche Daten werden dafür in welcher Qualität benötigt?

Ziele des KI-Einsatzes und Datenbedarfe definieren
Die Ziele des KI-Einsatzes und der geplante Anwendungskontext bestimmen in entscheidender Weise Art, Umfang, Menge und Qualität der erforderlichen Daten. Das eingesetzte KI-Modell und dessen Art, aus den Daten zu lernen, bestimmt ebenfalls die Anforderungen an die Datengrundlage. Derzeit benötigen viele Modelle maschinellen Lernens viele Daten mit je nach Einsatzzweck sehr unterschiedlichen Eigenschaften. Beim verstärkenden Lernen kann mit relativ wenigen Daten gestartet werden, dafür ist mehr Input während des Betriebs nötig.

Datenqualitätsanforderungen bestimmen
Es gibt zahlreiche Qualitätsmerkmale für Daten und Datensätze. Sie reichen von A wie aktuell bis V wie vollständig. Eine erste Übersicht über bestehende Standardanforderungen gibt folgende Grafik, die jedoch keine abschließende Darstellung bieten kann.[23]

aktuell	**fehlerfrei**	**genau**	**konsistent**	**einheitlich formatiert & strukturiert**	**maschinen-lesbar**
repräsentativ	**umfangreich & granular**	**vertrauens-würdig**	**verlässlich**	**verständlich**	**vollständig**

Der **Einsatz von KI** kann sehr unterschiedliche Anforderungen an die Datenqualität stellen. So können in einem Fall viele historische Daten als Trainingsdaten erforderlich sein, etwa um die Entwicklung über Jahre nachzuzeichnen und Muster zu erkennen. Hier wären etwa die Merkmale „Aktualität“ und „Vollständigkeit“ entsprechend unterschiedlich wichtig. In einem anderen Fall kommt es auf die unmittelbare und möglichst fehlerfreie maschinelle Lesbarkeit von Akten (die im Wesentlichen Texte enthalten) an, wodurch dem Merkmal „Maschinenlesbarkeit“ eine besondere Bedeutung zukommt.

5.2.2 Bestand verfügbarer Daten ermitteln und Datenqualität prüfen: *Erfüllen die verfügbaren Daten die definierten Qualitätsanforderungen?*

23 Eine Definition der Begriffe findet sich im Glossar. Vgl. ferner den „Leitfaden für qualitativ hochwertige Daten und Metadaten“ von Fraunhofer FOKUS, S. 14 ff., abrufbar unter https://cdn0.scrvt.com/fokus/e472f1bf447f370f/32c99a36d8b3/NQDM_Leitfaden-f-r-qualitativ-hochwertige-Daten-und-Metadaten_2019.pdf
Das FAIR-Prinzip versucht die wesentlichen Kriterien zusammenzubringen im Kontext einer Öffnung der Daten (und sind damit im engeren Sinne keine Aspekte von Datenqualität): Findable (auffindbar), Accessible (zugänglich), Interoperable (interoperabel) und Re-usable (nachnutzbar).

Verfügbare Daten identifizieren & deren Qualität prüfen
Die im ersten Schritt bestimmten erforderlichen Daten und Qualitätsanforderungen werden jetzt mit verfügbaren Datenquellen und Daten abgeglichen:

- Welche Datenquellen mit welchen Daten stehen für die konkrete Entwicklung zur Verfügung? Dabei können vielfältige Datenquellen relevant werden, von denen eigene Fachverfahren der Behörden eine typische Quelle darstellen. Neben eigenen Daten können auch solche von Partnerorganisationen übermittelt, aus öffentlichen Quellen (z. B. aus amtlichen Statistiken oder Open-Data-Portalen) bezogen oder von Datenanbietern erworben werden. In diesem Fall muss geklärt werden, unter welchen Bedingungen der Zugang zu den Daten bzw. die Nutzung der benötigten Daten möglich ist.

- Weisen diese Daten die erforderliche Qualität auf? Diese Einschätzung ist essenziell und muss desto gründlicher erfolgen, je weniger bekannt die Daten bzw. ihre Quelle sind. Dies ist insbesondere bei externen Datenquellen wichtig, wobei eine hinreichende Vertrauenswürdigkeit des Datenzulieferers sicherzustellen ist. Die verfügbaren Daten sind dann entsprechend dem entwickelten Anforderungskatalog zu prüfen. Hier schließt die Datenaufbereitung und -bereinigung an.

5.2.3 Datenaufbereitung und -bereinigung durchführen:

Wie können die verfügbaren Daten so aufbereitet werden, dass sie die erforderliche Qualität aufweisen?

Eine gute Datenaufbereitung stellt einen essenziellen und zumeist aufwendigen[24] Prozess der Bereinigung und Vorbereitung von Daten für die weitere Verwendung dar.

Was bedeutet Datenaufbereitung? Und was ist der Unterschied zu Datenbereinigung?

Datenaufbereitung beinhaltet die Bereinigung von Daten. Die Daten werden in ein einheitliches, maschinenlesbares Format gebracht, Datenfehler werden beseitigt, indem z. B. Duplikate entfernt, Fehler korrigiert oder (mithilfe von Data-Augmentation) fehlende Daten hinzugefügt werden. Ergänzt wird dies um die Vorbereitung der Daten für die weitere Verwendung (z. B. zur Analyse oder als Trainingsdaten für maschinelles Lernen). Solche Vorbereitungen können in unterschiedlichster Form erfolgen, etwa durch Reduktion (Aggregierung oder Generalisierung) oder Strukturierung der Daten und Ablage in performativen Datenbanksystemen.

Schritte der Datenaufbereitung
Folgende Schritte sind wesentlich bei der Datenaufbereitung:[25]

1. Standards formulieren: die für den Anwendungsfall relevanten Qualitätsanforderungen bestimmen, etwaige Zielkonflikte erkennen und in operable Standards zur Prüfung der verfügbaren Daten übersetzen

2. Daten integrieren: Die Daten werden aus den Quellen transformiert und zusammengeführt.

3. Daten prüfen und validieren: Hier werden Fehler und Diskrepanzen der Daten zum formulierten Standard identifiziert.

4. Daten plausibilisieren und imputieren: Bei falschen, fehlenden, unzuverlässigen, veralteten o. ä. Daten werden – wenn möglich – korrekte Werte eingefügt oder falsche Daten entfernt. Die Plausibilität der korrigierten bzw. imputierten Daten sollte erneut geprüft werden.

24 Vgl. dazu Cleaning Data: Most Time-Consuming, Least Enjoyable Data Science Task, Gil Press, Forbes, March 23rd, 2016, http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says. Dies wird im Vortrag von Prof. Naumann in KI-Lab #5 dargestellt, vgl. Video-Vortrag, abrufbar unter https://www.denkfabrik-bmas.de/projekte/ki-in-der-verwaltung/ki-labs-zu-datenqualitaet-und-datenreinigung-fuer-ki-anwendungen sowie auf Folie 5 der entsprechenden Präsentation.

25 Vgl. dazu Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder, 2021, S. 102 ff. (abrufbar unter https://www.destatis.de/DE/Methoden/Qualitaet/qualitaetshandbuch.html) und den Vortrag von Prof. Neumann in KI-Lab #5, vgl. auch Folie 6 der Präsentation, abrufbar hier sowie im Vortrag, Video hier: KI-Labs zu Datenqualität und Datenreinigung für KI-Anwendungen – Denkfabrik Digitale Arbeitsgesellschaft (denkfabrik-bmas.de).

Die Schritte können (müssen) je nach Anwendungskontext in einer anderen Reihenfolge und zudem mehrfach durchgeführt werden.

Ursachen für eine niedrige Datenqualität
Es gibt zahlreiche Fehlerursachen, exemplarisch:

(1) fehlerhafte Datenerfassung oder -eingabe

(2) unterschiedliche Formate verschiedener Datenquellen

(3) unvollständige Datensätze (leere Zellen oder ganze Spalten bzw. Zeilen)

(4) inkonsistente Datentypen (Zahlen vs. Buchstaben)

(5) unterschiedliche Skalen/Einheiten

(6) uneinheitliche Bezeichnungen

(7) mehrere Tabellen in einer Datei

(8) Fehler in der Datenaufbereitung, z. B. bei der Aggregierung

(9) divergierende Strukturen bei CSV-Daten (etwa durch eine abweichende Anzahl von Zeilen oder Spalten)

(10) falsche Zuordnungen von Werten zu Variablen (z. B. bei der Übersetzung eines ausländischen Dokuments)[26]

Verfahren zur Datenaufbereitung
Es stehen verschiedene Verfahren zur Datenaufbereitung zur Verfügung. Das gewählte Verfahren muss zu den Datentypen und Dateiformaten, zum KI-Modell sowie zu den Fehlerursachen bzw. zum Ziel der Aufbereitung passen.

Eine Möglichkeit stellt die Aggregierung von Daten dar. Hier werden nicht Fehler behoben, sondern durch die Aggregierung die für den jeweiligen Anwendungskontext passende Granularitätsebenen erzeugt. Eine andere Möglichkeit ist Data-Augmentation, also eine Methode, mit der fehlende Daten durch künstliche Daten aufgefüllt werden. Fehlen z. B. Daten zu einer Gruppe oder Objektklasse (Frauen, nichtweiße Personen etc.), kann mit dieser Methode durch Schaffung synthetischer und zugleich „originalgetreuer" Daten eine ausgeglichene oder hinreichend große Grundgesamtheit – z. B. für Trainingsdaten – sichergestellt werden. Ferner finden Prozesse der Datenbereinigung statt (z. B. einheitliches maschinenlesbares Format sicherstellen, Duplikate entfernen).

Bei der Datenaufbereitung sind die **Grenzen** des jeweiligen Verfahrens zu beachten. Werden etwa mittels Data-Augmentation synthetische Daten erzeugt, muss sichergestellt werden, dass der Datensatz insgesamt weiterhin repräsentativ ist. Dies setzt eine hohe Kenntnis der Ausgangsdaten sowie ein Verständnis der Methode und entsprechende Evaluationsprozesse voraus.

5.3 Bias finden und verhindern: Wie können Mitarbeitende für mögliche Bias sensibilisiert werden? Welche (technischen) Möglichkeiten zur Vermeidung von Bias gibt es?

Bias können im gesamten Lebenszyklus von KI auftreten (Planung, Entwicklung, Einführung und Betrieb),[27] wobei bei Systemen maschinellen Lernens die Trainingsphase den Moment der Übersetzung des Datenbias in geltende Entscheidungsregeln des KI-Systems darstellt. Auch beim Betrieb muss auf repräsentative Betriebsdaten geachtet werden. Einerseits, um diskriminierende Outputs des trainierten Algorithmus zu verhindern, andererseits gerade auch dann, wenn die Betriebsdaten zum Weitertrainieren der KI genutzt werden.[28] Nur eine fortlaufende Evaluation des KI-Systems stellt sicher, dass später auftretende Risiken erkannt und in Bezug auf Bias notwendige Anpassungen am KI-System vorgenommen werden können. Für die behördliche Praxis ist die Vermeidung von Bias entscheidend, denn je nach Anwendungskontext des KI-Systems kann ein Bias in den Daten zu Diskriminierungen führen.

5.3.1 Mitarbeitende sensibilisieren und Feedback-Schleifen vorsehen

Im Kontext von KI muss das Ziel sein, die zugrunde liegenden Daten kritisch zu prüfen und Verzerrungen zu vermeiden bzw. zu beheben. Dazu sind die Mitarbeitenden entsprechend ihren Rollen zu schulen. Im Zuge

26 Vgl. dazu auch die vielfältigen Beispiele im Vortrag von Prof. Neumann in KI-Lab #5, im Video unter https://www.denkfabrik-bmas.de/projekte/ki-in-der-verwaltung/ki-labs-zu-datenqualitaet-und-datenreinigung-fuer-ki-anwendungen sowie Folien 7 ff. der entsprechenden Präsentation.

27 Vgl. dazu Kapitel „Einführungsprozesse menschenzentriert gestalten & Ziele definieren".

28 Etwa wenn die Daten für spätere Anpassungen genutzt werden sollten oder es sich um eine selbstlernende KI handelt.

dessen sollten sie ebenso bzgl. der grundsätzlichen Fehlbarkeit von KI-Systemen sensibilisiert werden, um einen kritischen Umgang zu ermöglichen und so beispielsweise einem Automatisierungsbias[29] vorzubeugen. Zudem kann eine diverse Aufstellung des Teams dabei helfen, Verzerrungen zu erkennen, zu vermeiden oder zu beheben. Dabei kann die frühe Beteiligung der Stakeholder schon bei Planung und Entwicklung sowie im Kontext der Feststellung und Vermeidung von Diskriminierung eine Stelle für Gleichstellung, Antidiskriminierung o. Ä. beteiligt werden.

Außerdem sollten Prüf- und Feedback-Schleifen entsprechend den Bias- (und Diskriminierungs-)Risiken bei Entwicklung, Einführung und Betrieb des KI-Systems mitgedacht und eingebaut werden. Denn häufig ist bei Beginn des Trainings die Grundgesamtheit der Daten und damit die Bezugsgröße, um eine Verzerrung in den Daten erkennen zu können, nicht oder nicht genau bekannt. Zugleich können über diese Schleifen die Auswirkungen des Systems – v. a. auf Betroffene – erfasst und zurückgemeldet werden, was als Ausgangspunkt für Anpassungen genommen werden kann.

5.3.2 Verfahren zur Erkennung und zum Umgang mit Bias

Derzeit sind automatisierte **Verfahren zur Erkennung** von Bias in den Daten noch in der Erprobungsphase.[30] Diese werden künftig sicherlich ein wichtiges Instrument für die Erkennung darstellen. Allerdings setzt das Erkennen eines Bias in bestimmten Anwendungsfällen voraus, dass die Struktur der Grundgesamtheit bekannt ist, was in der Realität nicht immer der Fall ist.

Eine Möglichkeit zur **Verhinderung von Bias** ist die Beschränkung auf die tatsächlich erforderlichen Daten beim Training der KI. Denn für das jeweilige Ziel irrelevante Merkmale können z. B. bei maschinellem Lernen einen Bias hervorrufen, weil das KI-System (auch) auf Basis dieser irrelevanten Merkmale Muster entwickelt. Kommt es bei einer Anwendung beispielsweise nicht auf das Geschlecht an, dann sollte dieses Merkmal nicht in den Trainingsdatensatz aufgenommen werden, um eine Bezugnahme darauf zu vermeiden. Gleichzeitig muss darauf geachtet werden, dass das Geschlecht nicht über andere Merkmale indirekt erfasst wird. Die Entscheidung darüber, welche Merkmale erforderlich sind, setzt neben dem Verständnis über Lernvorgänge von KI auch ein tiefgreifendes Verständnis des Anwendungskontextes (d. h. insbesondere der Aufgaben und Prozesse) und der verfügbaren Daten voraus.

Eine nachträgliche **Behebung eines Bias, der im Entwicklungsprozess erkannt wurde,** kann über die Löschung von den Daten(-Merkmalen), die den Bias bzw. die Diskriminierung produziert haben, erfolgen. Eine Erfolgsgarantie bietet das aber nicht, wenn diese Merkmale z. B. mit anderen korrelieren. Auch hier ist also eine gute Kenntnis des Datensatzes sowie der Interdependenzen der Daten erforderlich. Die Folgen eines Bias in den Daten und/oder diskriminierender Outputs von KI-Systemen für bereits erfolgte Entscheidungen, etwa erlassene Verwaltungsakte, sind im Einzelfall zu prüfen: etwa ob der Verwaltungsakt dadurch rechtswidrig oder gar unwirksam ist und ob er bei bestehender Rechtswidrigkeit insbesondere heilbar, zurückzunehmen oder aufzuheben ist.

29 Dieser beschreibt die Neigung des Menschen, Vorschläge von automatisierten Entscheidungsfindungssystemen zu bevorzugen und etwaige widersprüchliche Informationen weniger ernst zu nehmen.

30 Entwickler*innen von IBM haben einen Open-Source-Werkzeugkasten bereitgestellt, abrufbar etwa über https://aif360.mybluemix.net/ oder via Github, https://github.com/Trusted-AI/AIF360

6. Transparenz schaffen & Erklärbarkeit herstellen

6.1 Einführung

Wie ein KI-System funktioniert und wie ein bestimmtes Ergebnis zustande kommt, muss für unterschiedliche Zielgruppen (z. B. Anwender*innen in den Behörden, betroffene Bürger*innen, Personalräte oder zivilgesellschaftliche Akteure) so verständlich und nachvollziehbar sein, dass diese – entsprechend ihrer Rolle – das System richtig anwenden, die Ergebnisse richtig verstehen und weiterverwenden sowie das System hinterfragen und überprüfen können.[31] Für die öffentliche Verwaltung ist dies besonders relevant, da Verwaltungshandeln für Bürger*innen transparent, erklär- und begründbar sein muss. Transparenz und Erklärbarkeit von KI-Systemen schaffen Vertrauen und Akzeptanz bei Bürger*innen gegenüber Handlungen der Verwaltung und bei Mitarbeiter*innen gegenüber dem KI-Einsatz in der Behörde.

Je nach Anwendung sind verschiedene Maßnahmen nötig, damit die Funktionsweise und Entscheidungen von KI-Systemen verständlich und nachvollziehbar sind.[32] Verwenden KI-Systeme regelbasierte Modelle mit festen Kriterien wie z. B. einfache Entscheidungsbäume mit wenigen Verzweigungen und geringer Tiefe und sind diese bekannt bzw. offengelegt, sind Entscheidungen eines solchen Systems für Anwender*innen, Betroffene und andere Zielgruppen verhältnismäßig einfach nachvollziehbar. Das setzt voraus, dass die verwendeten Daten und die Funktionsweise transparent und adressatengerecht dargestellt sind. Solche auf nachvollziehbaren Eingangsgrößen basierenden Modelle werden White-Box-Modelle genannt. Im Unterschied dazu sind Black-Box-Modelle, wie z. B. neuronale Netze, für Menschen nicht intuitiv nachvollziehbar, selbst wenn Daten und die Funktionsweise transparent sind. Bei KI-Systemen, die auf Black-Box-Modellen basieren, ist es nötig, die vom System getroffenen Entscheidungen zusätzlich erklärbar zu machen. Dafür gibt es verschiedene Erklärungsmethoden, um beispielsweise die Faktoren, welche die Ergebnisse des KI-Systems maßgeblich beeinflussen, in einer für Menschen verständlichen Weise auszudrücken. Sie werden oft unter dem Begriff Explainable AI zusammengefasst.

Hierbei ist zwischen Erklärungen der generellen Funktionalität des KI-Systems und Erklärungen von Einzelentscheidungen zu unterscheiden. Erklärungen der generellen Funktionalität (sogenannte Modellerklärbarkeit oder globale Erklärbarkeit) helfen dabei, Anwender*innen, Betroffenen oder anderen Zielgruppen die Funktionsweise eines KI-Modells als Ganzes nachvollziehbar zu machen, indem beispielsweise Wechselwirkungen oder Zusammenhänge von verwendeten Daten adressatengerecht dargestellt werden. Mit solchen generellen Erklärungen ist es aber in der Regel nicht möglich, individuelle Ergebnisse eines KI-Systems nachzuvollziehen, weswegen zusätzlich Erläuterungen von Einzelergebnissen (sogenannte lokale Erklärbarkeit oder Datenerklärbarkeit) bereitgestellt werden müssen. Der Datenschutz insbesondere personenbezogener Daten anderer Betroffener (z. B. Mitbewerber*innen auf eine Stelle, andere Antragsteller*innen auf staatliche Leistungen) muss dabei beachtet werden.

Eine gesetzliche Anforderung an die Erklärbarkeit von KI-Systemen ergibt sich insbesondere aus den Regelungen zu den Informationspflichten bzw. den Auskunftsrechten der DSGVO, die im Falle der Verarbeitung personenbezogener Daten „… aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person“ bei einer automatisierten Entscheidungsfindung vorschreiben (Artikel 13 Abs. 2 Buchstabe f, Artikel 15 Abs. 1 Buchstabe h). Auch wenn die DSGVO nur für die Verarbeitung personenbezogener Daten gilt, bedeutet das in diesem Zusammenhang für die Erklärbarkeit von KI-Systemen, dass für konkrete Einzelentscheidungen die wesentlichen Einflussfaktoren aufgezeigt werden müssen.

Die zitierten Regelungen in der DSGVO zur Erforderlichkeit der Erklärbarkeit gelten außerdem nur für vollständig automatisierte Entscheidungen. KI-basierte Entscheidungsunterstützungen sind von diesen Regelungen in der DSGVO nicht erfasst, die damit eine wesentliche Lücke aufweist. Da Entscheidungsempfehlungen und -unterstützungen durch KI aufgrund eines möglichen Automatisierungsbias bei menschlichen Letztentscheider*innen ähnlich große Auswirkungen auf betroffene Bürger*innen haben können wie vollständig automatisierte Entscheidungen, sollte auch in diesen Fällen Transparenz und Erklärbarkeit gewährleistet werden. Bürger*innen wie auch Mitarbeiter*innen sollten über den Einsatz von KI umfassend infor-

31 Vgl. Algo.Rules (2020): Handreichung für die digitale Verwaltung. Algorithmische Assistenzsysteme gemeinwohlorientiert gestalten.

32 Im Folgenden vgl. iit-Institut für Innovation und Technik in der VDI/VDE Innovation + Technik GmbH (2021): Erklärbare KI. Anforderungen, Anwendungsfälle und Lösungen, S. 20–21.

miert werden, damit sie erkennen können, wenn ein KI-System im Einsatz ist bzw. in eine Entscheidungsfindung einbezogen war (siehe „Erklärbarkeit & Transparenz“ in der Wertegrundlage auf S. 9–10).

Dies kann je nach Art der KI-Anwendung z. B. über eine direkte Kennzeichnung der KI-Anwendung (etwa bei Chatbots, KI-gestützten Assistenten für das Ausfüllen von Anträgen oder anderen KI-Anwendungen, mit denen die Bürger*innen direkt interagieren), über Hinweise in einem Bescheid (wenn bei der Bearbeitung vorbereitend oder unterstützend KI-Anwendungen genutzt wurden) oder auch zentral auf einer Website erfolgen (wenn in anderen Bereichen der Behörde KI-Anwendungen zum Einsatz kommen, die nicht die Antragsbearbeitung betreffen, wie etwa bei der Vorhersage von Serverüberlastungen). Auf der Website können außerdem ergänzend zu den Kennzeichnungen und Hinweisen in Bescheiden detailliertere Informationen zu den eingesetzten KI-Anwendungen gegeben werden.

Für betroffene Bürger*innen sind Erklärungen über die sie betreffende Einzelentscheidung oft ausreichend. Damit aber Entwickler*innen von KI-Systemen sowie Anwender*innen die korrekte Funktionsweise eines Systems überprüfen und gewährleisten können, sind für sie umfangreichere Erklärungen über ein KI-System zu erstellen. Dabei ist zu beachten, dass Menschen unterschiedliches technisches Hintergrundwissen mitbringen. Entsprechende Kompetenzen, die Mitarbeiter*innen für die Auseinandersetzung mit Erklärungen von KI-System benötigen, sind über Aus- und Weiterbildungen sicherzustellen. Egal wie intuitiv ein KI-Systemen ist und wie leicht verständlich die Erklärungen sind, ersetzen diese nie Schulungen. Schulungen unterstützen Mitarbeiter*innen etwa dabei, mögliche Fehler im Betrieb zu erkennen, und geben Vertrauen bei der Anwendung.

Die folgende Checkliste mit vier Schritten hilft dabei, für zentrale Zielgruppen (z. B. Entwickler*innen, Anwender*innen sowie Betroffene) passende Erklärungen zu erarbeiten. Für jeden Schritt enthält die Checkliste mehrere Orientierungsfragen, die jeweils beantwortet werden müssen. Sie kann damit auch mögliche verbindliche Regelungen konkretisieren und sinnvoll ergänzen, die in der KI-Verordnung zur Transparenz und Erklärbarkeit gegenüber Anwender*innen und Betroffenen enthalten sind.

6.2 Allgemeine Empfehlungen

- Ist für einen Anwendungsfall die Verwendung eines einfach nachvollziehbaren KI-Modells (White-Box-Modell, wie z. B. auf nachvollziehbaren Eingangsgrößen basierende Entscheidungsbäume) möglich, ist dies bei gleicher Eignung gegenüber einem weniger nachvollziehbaren KI-Modell (Black-Box-Modell, wie z. B. neuronale Netze) zu bevorzugen.

- Werden nicht nachvollziehbare KI-Modelle verwendet, können verschiedene Erklärungsmethoden eingesetzt werden. Wenn möglich, sind Prototypen oder Counterfactual Explanations[33] zu verwenden, da diese insbesondere für Anwender*innen intuitiv verständlich sind.[34]

- Ob die erarbeiteten Erklärungen passend bzw. ausreichend sind, ist mit den jeweiligen Zielgruppen vor Einführung sowie laufend zu testen und im Bedarfsfall anzupassen.

6.3 Checkliste:

1. **Zielgruppen und jeweilige Anforderungen an die Erklärungen bestimmen:** *Wer sind die zentralen Zielgruppen und was muss ihnen erklärt werden?*
2. **Erklärung der allgemeinen Funktionsweise:** *Wie kann die generelle Funktionalität eines KI-Systems der jeweiligen Zielgruppe erklärt werden?*
3. **Erklärung der konkreten Entscheidung im Einzelfall:** *Wie kann das Zustandekommen einer Einzelentscheidung eines KI-Systems der jeweiligen Zielgruppe erklärt werden?*
4. **Erklärungsstrategie bestimmen:** *Welche Hilfsmittel können zur Erklärbarkeit für verschiedene Zielgruppen verwendet werden?*

Zu den einzelnen Schritten:

6.3.1 Zielgruppen und jeweilige Anforderungen an die Erklärungen bestimmen: *Wer sind die zentralen Zielgruppen und was muss ihnen erklärt werden?*

Bei den Zielgruppen ist zu beachten, dass diese jeweils unterschiedliche Anforderungen an die Erklärungen

33 Counterfactual Explanations zeigen, welche möglichst kleinen Änderungen in den Eingabewerten nötig sind, um zu einem anderen Endergebnis zu führen.

34 Eine Orientierungshilfe über die gängigsten Erklärungsstrategien findet sich in der Studie „Erklärbare KI“ des iit-Instituts.

haben und sich außerdem in Bezug auf ihr Wissen über KI-Systeme unterscheiden. So müssen Anwender*innen in der Lage sein, eventuelle Fehler in den Ergebnissen – welche z. B. durch eine fehlerhafte Dateneingabe entstehen können – entdecken zu können. Allerdings haben Anwender*innen (häufig) wenig technisches Hintergrundwissen zu KI-Systemen, weswegen Erklärungen für sie zielgruppengerecht und leicht verständlich aufbereitet sein müssen. Im Gegensatz dazu können für Entwickler*innen Erklärungen verwendet werden, die mathematische, statistische und/oder technische Kompetenzen voraussetzen. In beiden Fällen sollten Anwender*innen und Entwickler*innen entsprechend aus- und weitergebildet werden. Dies ist nötig, damit sie die korrekte Funktionsweise eines KI-Systems gewährleisten und überprüfen können.

Anwender*innen, Entwickler*innen und betroffene Bürger*innen sollten bei der Erarbeitung von Erklärungen im Fokus stehen, aber auch weitere zentrale Zielgruppen sind zu berücksichtigen. Es kann zwar herausfordernd sein, für Zielgruppen allgemein verständliche Erklärungen zu erarbeiten, deren Hintergrundwissen zu KI-Systemen sehr heterogen sein kann wie beispielsweise bei Bürger*innen. Trotzdem ist dies für die Akzeptanz wichtig. Zusätzlich ist es hilfreich, Anwender*innen in die Lage zu versetzen, für Bürger*innen zusätzliche Erklärungen zu liefern, wenn diese Rückfragen haben.

Orientierungsfragen zur Bestimmung der Zielgruppen und ihrer Anforderungen:

- Wer sind zentrale Zielgruppen?

  Die zentralen Zielgruppen ergeben sich aus dem jeweiligen Anwendungskontext des KI-Systems. Dabei sind sowohl Stakeholder zu berücksichtigen, die am Einführungsprozess beteiligt sind (siehe Kapitel 4.5), als auch Zielgruppen, die im weiteren Entwicklungsprozess oder beim Einsatz involviert sind. Erster Ausgangspunkt zur Bestimmung der zentralen Zielgruppen sind die folgenden Gruppen:

  - **Entwickler*innen, u. a.**
    - Entwickler*innen, die das System entwerfen, konzipieren und implementieren
    - Entwickler*innen, die nach der Implementierung das System in der Behörde betreuen
    - Entwickler*innen, die das System in Bezug auf Design und Funktionalität überprüfen und testen
  - **Anwender*innen in Behörde**
  - **Bürger*innen als von einer KI-Entscheidung Betroffene oder als Anwender*innen einer KI**
    - ggf. Untergruppen innerhalb der Betroffenen, z. B. Arbeitnehmer*innen, Familien, Migrant*innen
  - **interne Stellen**
    - Entscheider*innen/Leitungsebene in Behörde
    - Personalräte
    - Gleichstellungsbeauftragte
    - Schwerbehindertenvertretung
    - Controlling/interne Prüfabteilungen
    - Datenschutzbeauftragte
    - Beauftragte für Informationssicherheit
  - **externe Stellen**
    - aufsichtsführende Stellen
    - Gerichte
    - zivilgesellschaftliche Akteure (Gewerkschaften, Verbände, NGOs etc.)

- Welche Informationen, die für Erklärungen erforderlich sind, müssen kommuniziert werden?

Mögliche zu kommunizierende Informationen können sein:

System
- Ziele des Systems
- bekannte Einschränkungen
- Designentscheidungen
- Annahmen
- Modelle
- Algorithmen
- Trainingsmethoden
- Qualitätssicherungsprozesse
- Maßnahmen zur Informationssicherheit
- Maßnahmen zum Datenschutz

Verwendete Daten
- Ort und Zeit der Datenerhebung
- Grund der Datenerhebung
- Umfang der Datenerhebung
- Methode der Datenerhebung
- Zusammensetzung des Datensatzes, Repräsentativität

Anwendung
- Applikation
- Verarbeitung
- Automatisierungsgrad und Einbettung des Systems in Entscheidungsprozess

- Welche Anforderungen an Erklärbarkeit haben die Zielgruppen? Bestimmen, was für die jeweilige Zielgruppe erklärbar sein muss:
  - ausschlaggebende Faktoren bei einer Einzelentscheidung unter Beachtung des Datenschutzes, insbesondere anderer Betroffener

  - verwendete Modelle und ihre Funktionsweise
  - verwendete Daten im Modell
  - verwendete Daten bei Einzelentscheidung
  - grundlegende Funktionsweise des Entscheidungssystems (wie kommt es zu einer Entscheidung und welche Rolle spielt dabei das KI-System?)

- Welches Wissen, welche Kompetenzen und wie viel Zeit bringen die Zielgruppen mit, um sich mit dem KI-System auseinanderzusetzen? Welche Schulungen und Weiterbildungsmaßnahmen sind ggf. sinnvoll oder notwendig?

6.3.2 Erklärung der allgemeinen Funktionsweise: *Wie kann die generelle Funktionalität eines KI-Systems der jeweiligen Zielgruppe erklärt werden?*

Jede Zielgruppe hat individuelle Anforderungen. Orientierungsfragen zur Erarbeitung einer Erklärung der allgemeinen Funktionsweise eines KI-Systems sind:

- Was sind die Ziele und der Einsatzkontext des KI-Systems?

- Welche wichtigen Kriterien spielen bei der Entscheidung des KI-Systems eine Rolle? Welches Gewicht haben diese Kriterien jeweils?

- Welches sind die Grenzen des KI-Systems? Was kann es leisten, was nicht?

- Welche Tendenzen bezüglich der Ergebnisse waren in Test- oder Betaphasen bzw. im bisherigen Einsatz erkennbar? Was waren bisherige Fehlerquoten (false positives und false negatives)?

6.3.3 Erklärung der konkreten Entscheidung im Einzelfall: *Wie kann das Zustandekommen einer Einzelentscheidung eines KI-Systems der jeweiligen Zielgruppe erklärt werden?*

Orientierungsfragen zur Erarbeitung einer Erklärung für konkrete Entscheidungen eines KI-Systems:

- Wie kann den Anwender*innen und Betroffenen kommuniziert werden, welche Faktoren für eine konkrete Ausgabe (z. B. die sie persönlich betreffende Entscheidung) relevant waren?

- Wie werden Entscheidungen dokumentiert, die im Zusammenspiel von KI-System und Mensch getroffen werden?

- Erklärbar ist nicht gleich verständlich: Wie können Informationen durch Texte und Grafiken so aufbereitet werden, dass sie leicht verständlich und interpretierbar sind?

- Wie können diese Informationen den Anwender*innen und Betroffenen niedrigschwellig zugänglich gemacht werden, z. B. als Teil der Kennzeichnung, der Ausgabe o. Ä.?

6.3.4 Erklärungsstrategie bestimmen: *Welche Hilfsmittel können zur Erklärbarkeit für verschiedene Zielgruppen verwendet werden?*

Orientierungsfragen zur Bestimmung der Erklärungsmethode:

- Mithilfe welcher technischen Maßnahmen können im Nachgang der Entscheidung die relevanten Faktoren ermittelt werden?
  - Zusätzliche Werkzeuge, welche die Ergebnisse der Software erklären und beispielsweise die für das Ergebnis ausschlaggebenden Faktoren verständlich aufbereiten.

  - Werkzeuge sind abhängig von Zielgruppe, KI-Modell, verwendeten Daten.[35]

  - Welche Grade der Erklärbarkeit gibt es? In welchen Fällen muss welcher Grad erfüllt sein, d. h., wie konkret müssen das System und seine Outputs nachvollziehbar sein?
  - Wovon hängt der erforderliche Grad der Erklärbarkeit ab? Welche Rolle spielt v. a. der Anwendungskontext und dabei die Risikoeinschätzung des Einsatzes?

  - Welche Grade braucht es für welche Zielgruppen?

- Wie kann sichergestellt werden, dass das richtige Erklärungsmodell für den jeweiligen Verwendungszweck genutzt wird? Wie kann dies verifiziert werden? Sind etwa Tests der Erklärungen mit Anwender*innen und Betroffenen oder ihren Vertretungen denkbar?

- Werden regelmäßig oder anlassbezogen (z. B. wenn am System etwas signifikant verändert wurde) Tests durchgeführt, um die Verständlichkeit der Erklärungen und die Fähigkeit von Externen, das System zu hinterfragen, sicherzustellen?

35 Eine Orientierungshilfe über die gängigsten Werkzeuge findet sich in der Studie „Erklärbare KI“ des iit-Instituts.