Nachweislich eine gute Entscheidung: Qualitätssicherung für künstlich-intelligente Verfahren in der Industrie

Annelie Pentenrieder(✉), Ernst A. Hartmann, und Matthias Künzel
Institut für Innovation und Technik in der VDI/VDE Innovation + Technik GmbH, Berlin, Deutschland
{pentenrieder,hartmann,kuenzel}@iit-berlin.de

Zusammenfassung. Welche Arten von Künstlicher Intelligenz (KI) sollen in europäischen Industrieunternehmen eingeführt und genutzt werden? Wie kann es gelingen, europäisch-demokratische Werte wie Mitbestimmung, Transparenz, Widerspruchsmöglichkeit und Anpassungsfähigkeit für die Nutzung von KI-Technologien zu gewährleisten? Diese Fragen werden aktuell unter den Debatten um erklärbare KI und KI-Zertifizierung verhandelt. Der folgende Beitrag legt an einem konkreten Fallbeispiel aus der Industrie die ALTAI-Kriterien an, die von der High Level Expert Group für die Gestaltung „vertrauenswürdiger KI" formuliert wurden. Entlang der drei ausgewählten Kriterien „Menschliches Handeln und Aufsicht", „Transparenz" und „Robustheit" wird exemplarisch skizziert, wie Erklär- und Kontrollierbarkeit KI-basierter Verfahren im industriellen Arbeitsumfeld zunehmend umgesetzt und prüfbar gemacht werden können. Es zeigt sich, dass technische Benutzeroberflächen als Teamarbeit gestaltet werden müssen und dass die Zusammenarbeit unterschiedlicher Unternehmen in der Bereitstellung von Datensätzen und Algorithmen im Fokus einer Prüfung stehen muss (Software-Genese). Als mögliches Handlungsfeld werden Auditing-Verfahren vorgestellt.

Schlüsselwörter: Erklärbare KI · KI-Zertifizierung · Partizipative Technikgestaltung

1 Bedarfe und Herausforderungen für nachweislich erklär- und kontrollierbare künstlich-intelligente Systeme in der Industrie

Verfahren Künstlicher Intelligenz (KI) erfordern in ihrer aktuellen Nutzung ein hohes Maß an Vertrauen, da die Nachvollziehbarkeit ihrer technischen Prozesse eine große Herausforderung darstellt. Sowohl die adaptive (selbstlernende) Beschaffenheit von Algorithmen, deren Logiken sich nicht direkt erschließen, als auch die in zahlreichen Schritten überarbeiteten Datensätze, die die Grundlage moderner KI-Verfahren bilden, erschweren eine Nachvollziehbarkeit. In vielen kritischen Einsatzsituationen – wie etwa in der industriellen Fertigung – kann ein solches Vertrauen nicht erbracht werden. Die Nutzung KI-basierter Systeme ist damit noch immer ein Risiko, da

E. A. Hartmann (Hrsg.): *Digitalisierung souverän gestalten II,* S. 51–63, 2022.
https://doi.org/10.1007/978-3-662-64408-9_5

durch technische Fehlentscheidungen, die nicht nachvollzogen werden können, hoher Schaden für die Unternehmen entstehen kann.

Insbesondere für moderne KI-Verfahren (wie das maschinelle Lernen) sind Verfahren zur Qualitätssicherung bisher nicht ausreichend etabliert, weder für die Überprüfung durch geschultes Fachpersonal im Unternehmen selbst noch für die Kontrolle durch außenstehende Prüfstellen. Zwar sieht die europäische Datenschutz-Grundverordnung (DSGVO) bereits Transparenzpflichten für komplexe Algorithmen vor, in der Praxis müssen jedoch erst Wege für die Umsetzung dieser rechtlichen Anforderungen gefunden werden. Zunehmend werden Erklärungsmodelle für künstlich-intelligente Systeme erforscht und erprobt (vgl. Kapitel Anforderungen, Anwendungen und Lösungsansätze erklärbarer Künstlicher Intelligenz von Kraus und Ganschow in diesem Band), doch auch diese erfordern noch immer ein hohes Expertenwissen und eignen sich darum selten für eine externe Überprüfung. Bisher gibt es nur wenige Modellprojekte, die in Zukunft einer unabhängigen Zertifizierungsstelle oder auch Fachpersonal ohne Informatik-Expertise dienen könnten.[1] Für eine generelle Debatte zur erklärbaren KI bleibt die Frage zu beantworten, an welchen technischen Details und für welche Nutzerschaft eine Erklärbarkeit konkret ausgearbeitet werden soll.

Aktuell werden Ansätze zur Erklärbarkeit verstärkt in der Informatik und in den Ingenieurwissenschaften erarbeitet (siehe Bundesministerium für Wirtschaft und Energie, Kraus, T. et al. (2021)) und weniger in interdisziplinären Forschungsprojekten, die eine soziologisch und psychologisch informierte Nutzer- und Organisationsforschung einbinden. Es zeichnet sich jedoch ab, dass ein interdisziplinärer Ansatz für nachweisbare Erklär- und Kontrollierbarkeiten sehr gewinnbringend wäre, da die notwendigen Erklärungen, um Vertrauen zu schaffen, nicht allein im technischen Verständnis liegen. Auch manche sozialen und organisatorischen Bedingungen, in denen die Technologie heute entwickelt und angewandt wird, müssen für eine Nachvollziehbarkeit strukturell offengelegt werden.

Diese Bedingungen stehen im Fokus des folgenden Beitrags: Um KI-Systeme konstruktiv erklären und kontrollieren zu können, bedarf es einer Untersuchung, wie Entwicklung, Training, Konfiguration und Anwendung KI-basierter technischer Verfahren heute organisiert sind und welche interdisziplinären Analysen notwendig sind, um zentrale Stellen im Entwicklungsprozess ausfindig zu machen, die diejenigen kennen sollten die die Software nicht gestaltet haben, aber diese nutzen, prüfen oder kontrollieren müssen.

[1] Das Projekt des Fraunhofer IPA „Slem“, https://www.slem-projekt.de/, zuletzt abgerufen am 3. August 2021, der Sonderforschungsbereich/Transregio der Universität Paderborn „Constructing Explainability“, https://trr318.uni-paderborn.de/, zuletzt abgerufen am 3. August 2021, und das Projekt des Fraunhofer IIS „TraMeExCo“, https://www.iis.fraunhofer.de/de/ff/sse/machine-learning/transparent-medical-expert-companion.html, zuletzt abgerufen am 3. August 2021, erforschen „erklärbare KI“ interdisziplinär und nutzerzentriert. Insbesondere in Paderborn ist diese Interdisziplinarität deutlich im Fokus. Es sind die Fächer Linguistik, Psychologie, Medienwissenschaft, Soziologie, Wirtschaftswissenschaft und Informatik beteiligt.

2 Kategorien für die Prüfbarkeit von KI-Systemen

Damit KI durch unabhängige Dritte, z. B. im Rahmen von Zertifizierungen geprüft werden kann, muss sie erklär- und kontrollierbar gemacht werden. Erklärbarkeit bedeutet in diesem Kontext, inwieweit das System in der Lage ist, für die (typischen) Nutzer:innen (vgl. dazu auch das Kapitel „Humanzentrierte künstliche Intelligenz" von Wirth et al. in diesem Band) verständliche Erklärungen bestimmter Ergebnisse seiner informationsverarbeitenden, algorithmischen Prozesse bereitzustellen bzw. auch die grundlegende Logik dieser Prozesse verständlich zu machen (das Wie und Warum). Die Erklärbarkeit künstlich-intelligenter Systeme wird bereits seit Längerem in der Informatik und verwandten Disziplinen intensiv diskutiert (Confalonieri et al. 2021).

Die Kontrollierbarkeit hingegen geht einen Schritt weiter und bezieht sich auf die Möglichkeiten der Nutzer:innen, in einem von künstlich-intelligenten Systemen geprägten Arbeitsumfeld qualitativ unterschiedliche Handlungsziele, -wege und -schritte zu wählen, wobei die einzelnen Wege möglichst sicher zu den jeweiligen Zielen führen sollen. Eine Änderung der Prozessschritte zur Erreichung bestimmter Ziele muss dem Nutzenden durch das Technikdesign ermöglicht werden. Die zugrunde liegenden Konstrukte der Handlungsregulation und der Kontrolle stammen aus der Arbeits- und Ingenieurpsychologie (Oesterreich 1981; Hartmann 2020). Beide Aspekte werden im folgenden Beitrag am konkreten Einsatz einer KI-Technologie in der Industrie auf ihre Ermöglichung hin geprüft (vgl. dazu auch den Beitrag von Hartmann in diesem Band).

Zur Prüfung der Umsetzbarkeit der Erklär- und Kontrollierbarkeit (ggf. auch für eine zukünftige Zertifizierung) wird für die Analyse des Praxisbeispiels die Bewertungsliste ALTAI (The Assessment List for Trustworthy Artificial Intelligence) herangezogen. Die High-Level Expert Group on Artificial Intelligence der Europäischen Kommission hat darin sieben Kategorien für die Vertrauenswürdigkeit von künstlich-intelligenten Systemen aufgestellt (Stix 2020). Unternehmen können anhand der Liste ihre KI-Produkte selbst auf Vertrauenswürdigkeit hin überprüfen. Die Bewertungskriterien fordern unter anderem, dass menschliches Handeln und Aufsicht in die Prozesse integrierbar sein müssen, dass die technische Robustheit und Sicherheit von KI-Systemen gewährleistet werden und dass Transparenz auf verschiedenen Ebenen angeboten wird (Stix 2020). Diese drei Kategorien „Robustheit", „Transparenz" und „Menschliches Handeln und Aufsicht" werden im Folgenden auf das Praxisbeispiel angewendet und dabei gleichzeitig weiterentwickelt.

Ein weiterer Kriterienkatalog, der für KI-Systeme in der Industrie relevant sein könnte, ist der Kriterienkatalog, den Kraus et al. (2021) auf Basis von Arrieta et al. 2019 zu acht übergeordneten Zielen, die eine Erklärbarkeit ermöglichen sollen, zusammenfasst. Eine nachweisbare Erklärbarkeit soll dabei unter anderem „Kausalitätsbeziehungen plausibilisieren", „Informationsgewinn erhöhen", „Konfidenz bestimmen", „Interaktionsmöglichkeiten verbessern" und „Verantwortlichkeiten klären". (Bundesministerium für Wirtschaft und Energie, Kraus et al. 2021: 16) Wenngleich Erklärbarkeit je nach Anwendungsfall eines KI-Verfahrens immer wieder neu gestaltet werden muss, bieten solche Kriterien die Möglichkeit, Übertragbarkeiten zu testen (Bundesministerium für Wirtschaft und Energie, Kraus et al.

2021: 16) und zwischen den empirischen Befunden der Einzelfälle zu vermitteln. Als Initiative in dieser Sache ist zudem die KI-Normungsroadmap des Deutschen Instituts für Normung zu nennen sowie die Initiative des Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme IAIS und des Bundesamtes für Sicherheit in der Informationstechnik, die aktuell Zertifizierungsansätze für KI-Systeme entlang konkreter Use Cases auf ihre Praxistauglichkeit testen (Deutsches Institut für Normung 2020; Fraunhofer IAIS 2021).

3 KI-Szenario in der Industrie: Fehlererkennung an Metallteilen

In diesem Abschnitt wird die Verwendung eines KI-Verfahrens in der Industrie vorgestellt und im Anschluss anhand der ALTAI-Kriterien „Robustheit", „Transparenz" und „Menschliches Handeln und Aufsicht" geprüft. Es handelt sich um ein fiktives Szenario, das aus Einblicken in die aktuelle Praxis unterschiedlicher Unternehmen zusammengestellt wurde. Die Grundlage der Szenario-Beschreibung und seiner anschließenden Analyse bilden in 2019 durchgeführte Arbeitsplatzstudien bei fünf Unternehmen im metallverarbeitenden Gewerbe und einer begleitenden Literaturrecherche. Ein Teil der Unternehmen erprobt bereits prototypische KI-Verfahren in einzelnen Prozessen, der andere Teil der Unternehmen sieht zwar das Potenzial von KI-Verfahren für einzelne Prozessschritte, benennt aber ebenso auch die Herausforderungen aktueller KI-Technologien für die Umsetzung in der Praxis. Diese Erfahrungsberichte machen die organisatorischen und technischen Neuheiten von KI als Technologie in der Praxis deutlich. Durch die Synthese der Erfahrungen zu einem Fall kann geprüft werden, an welchen Stellen Erklär- und Kontrollierbarkeiten mitgedacht werden müssen und wo eine Zertifizierung Transparenz – und damit Qualitätssicherung – in KI-Verfahren einbringen könnte.

Das mittelständische Metallverarbeitungsunternehmen „Round" liefert Metallteile für die Luft- und Raumfahrt. Aufgrund seiner hohen Risiken ist dieser Industriezweig stark reglementiert und zertifiziert – jeder Bearbeitungsschritt muss nachvollziehbar und transparent sein, um eine lückenlose Qualitätssicherung zu ermöglichen. Das Unternehmen plant den Einsatz eines maschinellen Lernverfahrens zur Fehlererkennung. An den Metallteilen sollen dünne Risse, Einkerbungen oder Dellen erkannt werden, die mit bloßem Auge kaum zu sehen sind. Eine Herausforderung für das Unternehmen ist, dass auch beim Einsatz dieses neuen Verfahrens externe Prüfer:innen eine Qualitätssicherung durchführen können sollen und dass auch die eigenen Facharbeiter:innen, die das neue Tool nutzen, die KI-basierte Fehlererkennung mit ihrer eigenen Erfahrung überprüfen können sollen. Für das Unternehmen ist klar, dass die KI-Technologie in diesem kritischen Verfahrensschritt nur eine Ergänzung und kein Ersatz menschlicher Überprüfung sein kann. Eine Erklärbarkeit, die sich direkt an die Maschinennutzer:innen richtet, ist darum essenziell. Eine vollautomatisierte Weiterleitung anscheinend fehlerfreier Metallteile wird auch in Zukunft beim Unternehmen „Round" nicht erfolgen.

Es arbeitet darum an der Erklärbarkeit dieses KI-basierten Systems. An den maschinellen Lernverfahren ist neu, dass Algorithmen auf Basis einer selbsttätigen

Mustererkennung aus vorhandenen Datensätzen hervorgehen, anstatt als regelbasierte Rezepte von Entwickler:innen selbst geschrieben zu werden, wie es in vorherigen KI-Generationen der Fall war. Dadurch verändern sich die Berechnungslogiken stetig und sind nicht ohne Weiteres nachzuvollziehen (vgl. Pentenrieder 2020: 32). Die Verfahren (v. a. neuronale Netze) sind zudem „vielschichtig und verflochten. […] Modelle reproduzieren Muster und Strukturen aus einer limitierten Datengrundlage", deren Logiken die darauf aufbauenden Entwickler:innen nicht kennen (Mangelsdorf et al. 2021: 1).

Aus Sicht der Produktionsleitung bietet das maschinelle Lernverfahren jedoch den Vorteil, „Toleranz" in die Fehleranalyse einzubringen, was bedeutet, dass die KI Fehler erkennen könnte, die alte Bildverarbeitungen nicht erkannt haben, und ggf. auch auf neue Fehlersystematiken hinweisen kann, von denen auch langjährig erfahrene Facharbeiter:innen lernen könnten. Aktuell erkennt das Trainingsmodul vier verschiedene Geometrien und acht verschiedene Sorten von Metallteilen. Mit diesem neuen Verfahren könnten dann auch Metallteile für unterschiedliche Zwecke gleichzeitig in einer Verarbeitungsschiene bearbeitet werden.

4 Das KI-Beispiel entlang der ALTAI-Kriterien

In diesem Abschnitt werden die ALTAI-Kriterien „Robustheit", „Transparenz" und „Menschliches Handeln und Aufsicht" skizzenhaft auf das Praxisbeispiel der KI-basierten Fehlererkennung angewendet. Aus dem Transfer der EU-Kriterien in die Praxis werden anschließend weitere Handlungsbedarfe abgeleitet.

4.1 Robustheit: Training des neuronalen Netzes im Team

Das Unternehmensteam, das das maschinelle Lernverfahren implementiert und betreut, besteht aus dem Instandhaltungsleiter, einer Bachelor-Studentin und einem Roboterprogrammierer. Sie haben Bilder von Fehlern in Metallteilen gesammelt und für das Training aufbereitet. Mithilfe der eigenen Arbeitserfahrung haben sie die Bilder der Metallteile manuell mit Labeln wie „Das ist ein Fehler" und „Das ist kein Fehler" versehen („Annotierung"). Über die Label werden Fehlermerkmale festgeschrieben, die bestimmte Abhängigkeiten im neuronalen Netz erzeugen. Das Team „trainiert" damit die Mustererkennung im neuronalen Netz. Ziel ist es, dadurch auch neue Fehlertypen, die das System bisher noch nicht gesehen hat, identifizieren zu können.

Außerdem testet das Team entlang verschiedener Parameter, ob das Netz adäquat funktioniert und betreibt damit „Netzpflege". Dazu verändern sie etwa die Beleuchtung, um zu prüfen, wie robust die Fehlererkennung auch bei dunklen Abbildungen funktioniert. Zudem verwenden sie Bilder mit unterschiedlicher Bildqualität, um bewusst Störungen in den Testdaten zu erzeugen und anschließend die Effekte der Störungen überprüfen zu können. Auch grenzen sie den Suchbereich, in dem die Fehler identifiziert werden sollen, immer wieder neu ein, um zu analysieren, wo ein bestimmter Fehler besonders häufig vorkommt. Zwar gibt es zum Thema technische Robustheit auch informatische Verfahren, die die Robustheit der Algorithmen gewährleisten. Diese Handarbeit in der Annotierung und Netzpflege ist

jedoch ein Beispiel, wie auch mittels organisationaler Maßnahmen die Robustheit der Technik gewährleistet wird – in diesem Fall damit, dass die Annotierung vom eigenen Team mit gutem Domänenwissen durchgeführt wird. Das System wird hinsichtlich seiner Belastbarkeit überprüft, insbesondere wie robust es gegenüber Veränderungen reagiert (ALTAI 2021: 9). Dies wird im ALTAI-Katalog als Kriterium zur „Robustheit des technischen Systems“ gefordert.

> „Technische Robustheit erfordert,“ so die Übersetzung aus dem ALTAI-Programm, „dass KI-Systeme mit einem präventiven Ansatz für Risiken entwickelt werden und dass sie sich zuverlässig und wie vorgesehen verhalten, während unbeabsichtigte und unerwartete Schäden minimiert und nach Möglichkeit verhindert werden. Dies sollte auch im Falle möglicher Veränderungen in ihrer Betriebsumgebung oder der Anwesenheit anderer Agenten (menschlich oder künstlich) gelten.“ (ALTAI 2021: 9).

Will man technische Robustheit nicht nur technisch sondern auch organisatorisch gewährleisten, dann muss diese Technik auch funktionieren, wenn andere Kolleg:innen des Teams das System nutzen und die Ergebnisse der Fehleranalyse ggf. anders interpretieren. Das Team, das die KI im Unternehmen implementiert, hat darum die Aufgabe, über das eigene Wissen um die KI hinaus auch bei den Kolleg:innen ein Bewusstsein für die Hintergrundarbeiten am neuronalen System zu erzeugen. Für die Benutzeroberfläche wurde darum ein Design gewählt, das die manuellen Annotierungen (Erstellung der Label) des KI-Teams von den Labeln unterscheidet, die über das neuronale Netz erstellt wurden. Grün werden Schäden markiert, die von Hand identifiziert und markiert wurden. In Orange bzw. Gelb werden Schäden markiert, die entsprechend der Berechnungen des neuronalen Netzes mit einer Wahrscheinlichkeit von 96 bzw. 90 % ein Schaden sein könnten. Durch diese farbliche Unterscheidung kann eine Fachkraft, die die erkannten Schäden später auf ihre Relevanz überprüft, plausibilisieren, ob die Empfehlung auf einer menschlichen oder auf einer berechneten Vorleistung basiert. Diese farblichen Markierungen tragen nicht nur zur Robustheit bei, sondern zugleich zur Transparenz und Nachvollziehbarkeit, weil deutlich wird, welche Urteile woher kommen (Mensch oder Algorithmus).

An diesem Aspekt zeigt sich, dass die selbstlernende Adaptivität, die das neuronale Netz bei der Fehlererkennung hat, neue Gestaltungsformen erfordert, um Robustheit, Transparenz und Nachvollziehbarkeit bei einem KI-basierten Automatisierungsschritt an Mensch-Maschine-Schnittstellen zu gewährleisten. Da die Prozesse dynamisch (selbstlernend) sind, ist auf eine ganz neue Weise zu klären, was informierte Maschinenbediener:innen über die zugrunde liegenden Mechanismen in der Software (organisatorisch, technisch etc.) wissen müssen, um im Fehlerfall („War es nur Schmutz oder doch ein Riss?“) die Empfehlung des neuronalen Netzes kontrollieren und anpassen zu können. Gerade bei KI-Empfehlungen ist es eine große Herausforderung, dass mit größerer Korrektheit (das System liegt zu 96 % und 90 % richtig) Fehler immer schwieriger nachzuvollziehen sind und damit auch der beherzte Eingriff sowie die instantane Problemlösung der Fachkraft immer schwieriger werden. Um diese anspruchsvolle Handlungsfähigkeit von KI-Nutzenden zu gewährleisten, müssen organisationale Strukturen für die Endnutzer:innen offengelegt werden. Die Abb. 1 zeigt die Organisationsstruktur, die einer KI zugrunde liegen kann. Im Zentrum steht die „Interpretation“ der Fachkraft, die mit Blick auf das „statistische Modell“ eine

Entscheidung trifft (z. B. die Auswahl eines fehlerfreien Metallteils). Wie bei anderen komplexen Techniksystemen sind auch bei KI-Verfahren zahlreiche Vorarbeiten zur Bereitstellung von Daten und Algorithmen beteiligt (siehe orangene Karten).

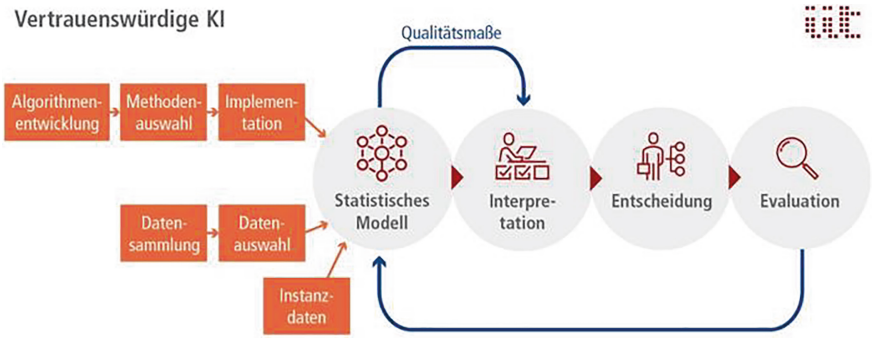

Abb. 1. Vertrauenswürdige KI. Basierend auf einem Vortrag von Katharina Zweig (2020) zur Rückverfolgung eines Unfalls durch KI beim autonomen Fahren. (Eigene Darstellung in Anlehnung an den Vortrag von Frau Zweig)

Über die in Abb. 1 dargestellten organisatorischen Prozessschritte sollte sich die Fachkraft, die eine KI nutzt, informieren können, um ihren Handlungsspielraum ausweiten zu können, z. B. flexibel bei Fehlern eingreifen zu können und Anpassungen zu veranlassen, wenn das System nicht adäquat funktioniert. Entlang der organisationalen Strukturen muss rekonstruiert werden können, wie das System dazu kommt, einen Riss oder einen Fehler zu erkennen und welche Prozessparameter eine Auswirkung auf das Ergebnis haben. Dies ist ein iterativer Prozess, in dem die wesentlichen Aspekte nur mit den Nutzenden gemeinsam identifiziert werden können.

4.2 Transparenz: Softwaregenese als entscheidender Faktor für Erklärbarkeit am Beispiel der Unternehmen „Round", „Square" und „Line"

Unter Transparenz fassen die ALTAI-Kriterien drei Aspekte: Rückverfolgbarkeit (Traceability), Erklärbarkeit (Explainability) und die offene Kommunikation über die Grenzen des KI-Systems (ALTAI 2021: 14). Das Praxisbeispiel zeigt, dass diese drei Aspekte auch eng an die Gewährleistung technischer Robustheit gekoppelt sind. Es ist eine Herausforderung, den Endnutzer:innen die in Abb. 1 dargestellten organisatorischen Zusammenhänge so zu erklären, dass sie diese entlang ihrer Erfahrung abwägen können: Wer kann die Frage beantworten, warum an einem Blech ein Fehler nicht erkannt wurde, den eine Fachkraft bei der Qualitätssicherung identifiziert hat? Wie findet die Fachkraft den wesentlichen Ansprechpartner, um Feedback geben zu können, um den Fehler in das Software-Programm zu integrieren? Wer trägt die Verantwortung im Schadensfall?

Im Beispiel des fiktiv konstruierten Unternehmensfalls erschwert ein weiterer organisatorischer Aspekt die Herstellung von Transparenz: Das Unternehmen erstellt

nicht nur selbst Trainingsdaten, sondern kauft diese auch von anderen Unternehmen ein. Erste Trainingsdaten und vortrainierte Algorithmen hat das mittelständische Metallverarbeitungsunternehmen „Round“ beim Software-Unternehmen „Line“ gekauft und in das eigene System übernommen. Aktuell wird die eingekaufte Software von einem weiteren Unternehmen namens „Square“ auf die Bedarfe von „Round“ angepasst. Dabei wird beispielsweise die Fehleranfälligkeit der Software justiert, denn im eingekauften System könnte das neuronale Netz Fehler gelernt haben, die für die Metallteile von „Round“ keine Rolle spielen.

Damit waren drei Unternehmen zu unterschiedlichen Zeiten an der Konstruktion der Software bzw. bei der Erhebung von Daten und der Erzeugung der adaptiven Algorithmen beteiligt. Das Software-Unternehmen „Line“ entwickelte das KI-Modul und lieferte damit die Grundsoftware, doch auch sie hatten bereits Teile der Software von einem Start-up zuvor aufgekauft. Das Metallverarbeitungsunternehmen „Round“ beauftragte wiederum das Unternehmen „Square“, das die Software auf die benötigten Bedarfe anpasst und die Software bei „Round“ in Betrieb nimmt. Eine solche Beteiligung unterschiedlicher Unternehmen zeigt, dass nicht nur vielfältige Bearbeitungsschritte, sondern auch unterschiedliche, an der Bearbeitung beteiligte Unternehmen in die Prüfung miteinbezogen werden müssen. Um eine Qualitätssicherung zu ermöglichen und die Fehlerursachen später konkret zurückverfolgen zu können, muss darum ein wesentlicher Fokus auf die Softwaregenese gelegt werden – also auf die Entwicklung und Entstehung von Software. Gerade im maschinellen Lernen werden die Daten, aus denen wertvolle Informationen extrahiert werden sollen, vielfach umgearbeitet, an unterschiedlichen Stellen, zu unterschiedlichen Zwecken und in unterschiedlichen Unternehmen erhoben und neu kombiniert (zum Beispiel bei Firmenübernahmen oder beim Kauf von Datensätzen). Nicht nur Daten werden dabei in unterschiedlichen Kontexten verwendet, sondern auch die Algorithmen werden auf sog. „legacy code“[2] aufgebaut und in verschiedenster Weise kombiniert.

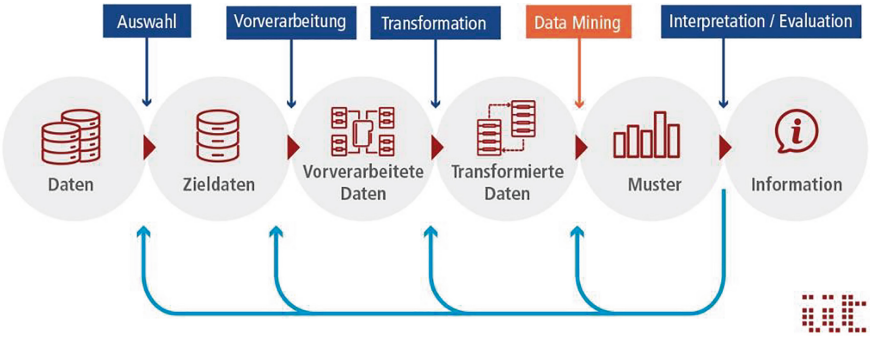

Abb. 2. Interpretierbarkeit von KI. (Eigene Darstellung basierend auf Cynthia Rudin 2019)

[2] Programmcode, der in der Regel älter und historisch gewachsen ist anstatt geplant erstellt. Der Code erfüllt seine Funktion, kann aber häufig nicht überprüft werden, weil Kommentare oder Spezifikationen von den vorherigen Programmierer:innen fehlen.

Die Grafik in Anlehnung an Cynthia Rudin zeigt die vielschrittigen Umarbeitungsprozesse, die aus Daten schlussendlich „Informationen“ machen.[3] Zunächst werden aus Datensätzen jene „Zieldaten“ ausgewählt, die für eine konkrete Aufgabe Relevanz haben und einen Mehrwert bieten könnten. Diese werden anschließend „vorverarbeitet“, „transformiert“ bzw. bereinigt (siehe Abb. 2), sodass „Muster“ abgeleitet werden können, die für die konkreten Aufgaben einen Informationsgehalt bieten. Im Beispiel der Fehleranalyse von Metallteilen könnte eine solche Rückverfolgung notwendig sein, wenn Fehler auf Basis gelernter Daten entstehen, weil die Spezifika des Blechteils, das analysiert werden soll, andere Spezifika sind als die des Materials, auf denen die Lernverfahren trainiert wurden.

4.3 Menschliches Handeln und Aufsicht: Auditing für KI-Software

Eine nutzerzentrierte Gestaltung der (wie in Abschn. 4.1. dargestellt) ermöglicht, dass Fachkräfte kritische, reflektierte und unabhängige Entscheidungen in der Nutzung komplexer Technologien treffen können. Vorab getroffene Entscheidungen von anderen Teammitgliedern werden sichtbar und können reflektiert und diskutiert werden. Auch Haftungsfragen lassen sich im Schadensfall besser behandeln. Eine derartige Gestaltung der Benutzeroberfläche adressiert auch das ALTAI-Kriterium „Menschliches Handeln und Aufsicht“. Das Kriterium gibt vor, dass der Entscheidungsprozess menschlicher Akteure von KI-Technologien nicht übernommen, sondern lediglich unterstützt werden soll (ALTAI 2021: 7). Für die Nutzenden muss sichtbar sein, dass eine Entscheidung von einem System und nicht von einem Menschen getroffen wurde. Die farbliche Unterscheidung zwischen menschlichen und KI-erzeugten Labeln macht dies im Ansatz kenntlich. Eine weitere Herausforderung besteht jedoch auch im dargestellten Beispiel und zwar wie das übermäßige Vertrauen – „Did you put in place procedures to avoid that endusers over-rely on the AI system?“ (ebd.) – verhindert wird; dieses übermäßige Vertrauen wird hier als wesentliche Bedrohung menschlicher Autonomie verstanden. Hier sind zusätzliche Maßnahmen erforderlich.

Neben der Handlungsfähigkeit der Fachkraft fordert dieses Kriterium auch die Möglichkeit einer Auditierung KI-basierter Systeme. Eine Methode zur standardisierten Überprüfung erklärbarer und kontrollierbarer KI-Verfahren ist das „Algorithmische Auditing“ (Chiusi 2021: 13). Mit unabhängigen Audits werden Verantwortungsgefüge offengelegt, sodass trotz des komplexen vernetzten Handelns, auf dem eine KI-Empfehlung basiert, eine Qualitätssicherung möglich ist. Bis dato gibt es noch keine ausgereiften Vorschläge, so schreibt es auch der Report „Automating Society“ (Chiusi 2021, S.13). Lediglich in einer Fußnote nennen die Autor:innen des Reports erste Aspekte, die zur Entwicklung zukünftiger Auditverfahren diskutiert werden müssen (ebd.). Zunächst müssen Prüf- und Auditkriterien entwickelt werden, um daraufhin systematische, unabhängige und dokumentierbare Prozesse ausformulieren zu können. Die ALTAI-Kriterien können eine Grundlage solcher

[3] https://www.youtube.com/watch?app=desktop&v=I0yrJz8uc5Q&ucbcb=1, zuletzt abgerufen am 7. Juni 2021.

Auditkriterien sein. Zu klären ist weiterhin die Frage, zu welchem Zeitpunkt der Softwareentwicklung ein Auditverfahren ansetzen sollte: ex ante in der Design- und Entwicklungsphase oder ex post in der Betriebsphase. In jedem Fall sollten Auditanforderungen bereits frühzeitig in der Entwicklung algorithmischer Systeme Berücksichtigung finden, um ein Technikdesign zu gewährleisten, das überhaupt prüffähig ist (ebd.).

Jedenfalls müssen die unmittelbaren Nutzer:innen über die Ergebnisse solcher Auditverfahren informiert sein, idealerweise werden sie beim Auditverfahren beteiligt. Auf diese Weise haben sie die Möglichkeit, die Fähigkeiten des KI-Systems und deren Grenzen realistisch einzuschätzen und übermäßiges Vertrauen zu vermeiden.

Am Beispiel der KI-basierten Fehlererkennung ist zu sehen, dass gerade für KI-Verfahren Daten in unterschiedlichen Unternehmen erhoben, vorverarbeitet und bereinigt werden. Kritische Umarbeitungen (z. B. von Datensätzen) können so bereits zu einem recht frühen Zeitpunkt geschehen, die ggf. zu einem späteren Zeitpunkt nicht mehr prüfbar sind, wenn Daten mehrfach transformiert wurden (siehe auch Abb. 2). Gerade wenn ein neuronales Netz jedoch aus vorherigen Fehldiagnosen lernt, könnte eine solche Nachvollziehbarkeit bei der Prüfung wesentlich sein. Es gilt darum zu diskutieren, wie sichergestellt werden kann, dass eine Prüfstelle auch später noch Umarbeitungen prüfen kann. Beispielsweise über „Logbücher“[4] könnte zumindest festgehalten werden, welche Unternehmen an der Erstellung von Algorithmen und Trainingsdatensätzen beteiligt waren.

5 Fazit

Ein Zertifikat beglaubigt etwas, wertet etwas auf – welche Werte sollten einer europäisch geprägten KI für die Nutzung bescheinigt werden? Aus dem ALTAI-Prüfkatalog zur vertrauenswürdigen KI wurden drei Kriterien ausgewählt, die auch in der Industrie die digitale Souveränität der Facharbeiter:innen unterstützen: „Technische Robustheit“, „Transparenz“ und „menschliches Handeln und Aufsicht“. An diesen Kriterien wurde ein Praxisbeispiel aus dem metallverarbeitenden Gewerbe geprüft. Daraus gehen sowohl für die interne Prüfbarkeit durch Fachkräfte als auch für die externe Qualitätssicherung folgende Handlungsbedarfe hervor.

Robustheit und Erklärbarkeit am Arbeitsplatz müssen in Teamarbeit erzeugt werden
Robustheit und Erklärbarkeit der algorithmischen Systeme am Arbeitsplatz sind Querschnittsthemen, die nicht allein von technischen Exper:innen verantwortet werden können. Die Herausforderung ist groß, eine KI so zu konstruieren und in den Arbeitsalltag einzubetten, dass eine Fachkraft überhaupt die Möglichkeit erhält, entgegen

[4] „Welche Informationen müssen verfügbar sein, damit ein Audit effektiv und verlässlich ist (z. B. Quellcode, Trainingsdaten, Dokumentation)“? […] Brauchen Prüfende physischen Zugang zu den Systemen, während sie im Betrieb sind, um einen effektiven Audit durchzuführen? (Chiusi 2021, S.13).

eines KI-Vorschlags sagen zu können, dass es sich hier um einen Fehler handelt, der zur Aussortierung eines Metallteils führen muss – anders als es die KI vorgegeben hätte. Um diese Handlungsfähigkeit der Endnutzer:innen zu gewährleisten, müssen die ungefähren Abläufe, die zur Erstellung dieser KI-Empfehlung geführt haben, für die Fachkraft transparent sein (siehe Abb. 1). Um diese Transparenz passgenau und bedarfsgerecht gestalten zu können, sollten Erklärungsansätze aus der Informatik mit der Perspektive informierter Nutzer:innen in Form von dezidierter Nutzerforschung ergänzt werden. Die Erklärungen zum KI-System müssen sich an jene richten, die die Software nicht entwickelt haben und womöglich auch kaum Kompetenzen in der Informatik und IT-Technik haben. Diese Art von nutzerzentrierter Softwareentwicklung wird über einen partizipativen Austausch zwischen Domänenwissen, KI-Kenntnissen, organisatorischem Wissen und sozialen Effekten ermöglicht. In den aktuellen Debatten zur erklärbaren KI ist ein solches nutzerzentriertes Denken, das von interdisziplinären Perspektiven gestützt wird, noch immer unterrepräsentiert.[5] Ein Konfidenzwert etwa, wie er in zahlreichen Machine-Learning-Anwendungen und auch im oben dargelegten Beispiel genutzt wird, um die Vertrauenswürdigkeit einer KI-Empfehlung als Prozentsatz anzuzeigen, reicht als nutzerzentrierte Erklärbarkeit nicht aus. Farbliche Darstellungen wie die oben dargelegte Unterscheidung zwischen den von der KI und den vom Team identifizierten Schäden sind eine weitere Maßnahme für nutzerzentrierte Transparenz, denn sie machen weitere zugrunde liegende Prozessschritte über die Benutzeroberfläche transparent. Dies hat positive Effekte für die Teamarbeit und reduziert Risiken einer Fehlentscheidung für das Unternehmen, denn die Robustheit der Technik kann somit von einem größeren Team überprüft werden. Langfristig kann ein solches Technikdesign auch auf volkswirtschaftlicher Ebene positiven Einfluss auf die Wettbewerbsfähigkeit europäischer KI-Technologien haben, wenn die Überprüfbarkeit KI-basierter Systeme ins Technikdesign einbezogen und zu einem Alleinstellungsmerkmal europäischer KI wird.

Förderung von neuen, offenen Experimenten partizipativer Technikgestaltung
Aus dem Bedarf an nutzerzentrierter KI-Technologie leitet sich der Bedarf nach interdisziplinären Methoden für die Technikgestaltung ab. Die Partizipation der Nutzenden sollte von Anbeginn einer Technikentwicklung auch finanziell und zeitlich miteingeplant werden. Es bedarf einer ausgeprägteren Förderung KI-bezogener Nutzerforschung, bei der gerade die fordernden und nur schwer zufriedenzustellenden Nutzenden ins Zentrum rücken und proaktiv in die Gestaltung von Technik eingebunden sind. Eine Erprobung partizipativer Verfahren in Unternehmen und mit wissenschaftlicher Begleitung ist dafür erfolgsversprechend. Selbst Schulungskonzepte können gewinnbringend auf dieser Vorarbeit aufbauen, womit sich die partizipative Vorarbeit im späteren Verlauf einer KI-Nutzung womöglich sogar als

[5] Dies wurde auch im Panel „Explainable AI: Nachvollziehbarkeit in der Anwendung“ bei den vom Bundeswirtschaftsministerium veranstalteten „Tagen der digitalen Technologien“ am 17. November 2020 wiederholt betont. Programm einsehbar unter https://www.bmwi-registrierung.de/tage-der-digitalen-technologien//Default.aspx?link=m2870, Zugegriffen: 4. August 2021.

Kostenersparnis auszahlen könnte. Die Erprobung solcher Methoden partizipativer Technikgestaltung steht im Zentrum eines Workshops des Projekts „Digitale Souveränität in der Wirtschaft“, der Anfang Dezember 2021 digital vom Institut für Innovation und Technik veranstaltet wird.

Literatur

Arrieta, A. B. et al., Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. http://arxiv.org/pdf/1910.10045v2 (2019). Zugegriffen 18. Nov. 2021

Bundesministerium für Wirtschaft und Energie, Kraus, T., et al.: https://vdivde-it.de/de/publikation/erklaerbare-ki-anforderungen-anwendungsfaelle-und-loesungen (2021). Zugegriffen: 16. Juni 2021

Chiusi, F., et al.: Algorithm Watch, Bertelsmann Stiftung. https://automatingsociety.algorithmwatch.org/wp-content/uploads/2021/01/Automating_Society_Report_2020_-_Deutsche_Ausgabe.pdf (2021). Zugegriffen: 25. Juni 2021

Confalonieri, R., Coba, L., Wagner, B., Besold, T.R.: A historical perspective of explainable artificial intelligence. WIREs Data Min. Knowl. Discov. **11,** e1391. https://doi.org/10.1002/widm.1391 (2021)

Deutsches Institut für Normung: Deutsche Kommission für Elektrotechnik. Deutsche Normungsroadmap Künstliche Intelligenz, Berlin. https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf (2020). Zugegriffen: 29. Juli 2021

Hartmann, E.A.: Digitale Souveränität in der Wirtschaft – Gegenstandsbereiche, Konzepte und Merkmale. In: Hartmann, E.A. (Hrsg.) Digitalisierung souverän gestalten. Springer Vieweg, Wiesbaden (2020)

High-Level Expert Group on Artificial Intelligence; Stix, C.: The assessment list for trustworthy artificial intelligence (ALTAI). https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (2020). Zugegriffen: 21. Juli 2021

Mangelsdorf, A., Gabriel, P., Weimer, M.: Die Zertifizierung von KI: Mehr Sicherheit für alle – oder unnötiger Ballast? iit-perspektive. https://www.iit-berlin.de/publikation/die-zertifizierung-von-ki-mehr-sicherheit-fuer-alle-oder-unnoetiger-ballast/ (2021). Zugegriffen: 29. Juli 2021

Oesterreich, R.: Handlungsregulation und Kontrolle. Urban & Schwarzenberg, München (1981)

Pentenrieder, A.: Algorithmen im Alltag. Campus, Frankfurt am Main (2020)

Poretschkin, M., et al., Fraunhofer IAIS: Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. KI-Prüfkatalog. https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (2021). Zugegriffen: 4. Aug. 2021

Rudin, C., Berkman Klein Center for Internet & Society: https://www.youtube.com/watch?app=desktop&v=I0yrJz8uc5Q&ucbcb=1. Zugegriffen: 7. Juni 2021. Basierend auf dem Artikel: Rudin, C.: Stop explaining black box machine learning models for high stake decisions and use interpretable models instead. Nat. Mach. Intell. https://www.nature.com/articles/s42256-019-0048-x (2019). Zugegriffen: 8. Juni 2021

VDI Zentrum Ressourceneffizienz. Im Auftrag des Bundesministeriums für Umwelt, Naturschutz und nukleare Sicherheit.: Potenziale der schwachen künstlichen Intelligenz für die betriebliche Ressourceneffizienz. https://www.ressource-deutschland.de/fileadmin/user_upload/downloads/studien/VDI-ZRE_Studie_KI-betriebliche-Ressourceneffizienz_Web_bf.pdf (2021). Zugegriffen: 8. Juni 2021