Menschenzentrierte industrielle Künstliche Intelligenz:

Ansätze zur Gestaltung akzeptierter und vertrauenswürdiger KI-basierter Services in der Produktion

---

Vom Fachbereich Sozialwissenschaften der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau zur Verleihung des akademischen Grades Doctor rerum naturalium (Dr. rer. nat.) genehmigte

kumulative Dissertation

vorgelegt von
Janika Kutz

D 386
2024

Dekan: Prof. Dr. Michael Fröhlich

Vorsitz Prüfungskommission: Prof. Dr. Marcus Höreth

Erstbegutachtung: Prof. Dr. Thomas Lachmann

Zweitbegutachtung: Prof. Dr. Katharina Hölzle

Datum der Einreichung: 15.02.2024
Datum der Disputation: 15.05.2024

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Danksagung

Mit der Abgabe meiner Dissertation geht ein wichtiger Abschnitt meiner wissenschaftlichen Ausbildung zu Ende. An dieser Stelle möchte ich mich bei allen bedanken, die mich auf diesem Weg begleitet und unterstützt haben.

Mein Dank gilt insbesondere meinem Doktorvater Prof. Dr. Thomas Lachmann. Vielen Dank für die Zusammenarbeit in den vergangenen Jahren, deinen methodischen und fachlichen Input und alles, was ich durch die Zusammenarbeit lernen durfte. Bedanken möchte ich mich auch bei Jan Spilski, der mir von Anfang an als Ansprechpartner zur Verfügung stand, den Prozess meiner Promotion begleitet und mich stets unterstützt hat.

Danken möchte ich auch meinen lieben Kolleginnen und Kollegen des Forschungs- und Innovationszentrums für Kognitive Dienstleistungssysteme (KODIS) des Fraunhofer IAO. Vielen Dank für die gute Zusammenarbeit, die bereichernden Diskussionen und die fachlichen Anregungen während des gesamten Promotionsprozesses. Besonders hervorheben möchte ich an dieser Stelle Dr. Bernd Bienzeisler und Dr. Jens Neuhüttler, die mir die Arbeit an dieser Dissertation überhaupt erst ermöglicht haben. Sie haben mich von Anfang an sowohl fachlich als auch persönlich mit motivierenden Worten unterstützt. An dieser Stelle möchte ich auch Frau Prof. Dr. Katharina Hölzle hervorheben, die mir mit ihrer langjährigen Erfahrung beratend zur Seite stand und mir wertvolles Feedback zu meiner Arbeit gegeben hat. Danke auch an mein Team „Public Service Innovation“ für das Verständnis in der Endphase der Fertigstellung der Arbeit.

Besonderen Dank möchte ich meinem Mann und besten Freund Patrick aussprechen, der immer an mich geglaubt hat und mein emotionaler Rückhalt auf diesem Weg war. Vielen Dank für deine Geduld, die nötigen Freiräume und dass du mich auf diesem Weg immer entlastet hast. Danken möchte ich auch meiner Familie und meinen Freunden, die mich während des gesamten Prozesses unterstützt und ermutigt haben.

Nicht zuletzt möchte ich mich bei allen Kooperationspartnern und Studienteilnehmern bedanken, die mir ihre Zeit und ihr Wissen zur Verfügung gestellt haben.

Durch die Hilfe und Unterstützung aller genannten Personen wurde diese Doktorarbeit erst möglich. Herzlichen Dank!

Zusammenfassung

Das Ziel der vorliegenden Forschungsarbeit besteht darin, einen Beitrag zur erfolgreichen Gestaltung auf Künstlicher Intelligenz (KI) basierender Services in Produktionsumgebungen zu leisten, der über eine rein technisch-getriebene Entwicklungsperspektive hinausgeht. Dafür wird eine ganzheitliche Betrachtungsperspektive entlang der Ebenen Mensch, Technik und Organisation eingenommen. In einer ersten Analyse wurden Herausforderungen und Erfolgsfaktoren im Zusammenhang mit der Entwicklung, Einführung und dem Betrieb von industriellen KI-basierten Services auf allen Ebenen eines sozio-technischen Systems identifiziert. In einem zweiten Schritt wurden Möglichkeiten zur Stärkung ausgewählter Erfolgsfaktoren oder Lösung spezifischer Herausforderungen definiert, die sowohl menschenzentrierte als auch organisatorische Bereiche betreffen. In anwendungsnahen Forschungsprojekten wurden unter Berücksichtigung der Erfahrungen von Expertinnen und Experten für industrielle KI-Anwendungen zwei Lösungsbereiche ermittelt, die zu einer Optimierung im industriellen Umfeld führen können. Zum einen wurde der Bedarf eines Rollenmodells zur Entwicklung interner KI-basierter Services identifiziert, zum anderen wurde erkannt, dass Ansätze aus dem Bereich Human-Centered Artificial Intelligence (HCAI) ein hohes Potenzial haben, zentrale Herausforderungen der KI-Entwicklung und KI-Nutzung im industriellen Kontext zu überwinden. Insbesondere die Anwendung von HCAI-Design-Prinzipien wird als wertvoll für die industrielle Praxis angesehen. Deutlich wurde aber auch, dass es im breiten Forschungsfeld HCAI einer Kontextualisierung bestehender Ansätze bedarf und eine an die produktionsspezifischen Rahmenbedingungen angepasste Konzeption von Methoden notwendig ist. Auf dieser Erkenntnis aufbauend wurden zwei Modelle konzipiert, die als methodische Unterstützung für primär technisch-versierte Entwicklerinnen und Entwickler KI-basierter Services in der Produktion dienen. Dies ist zum einen das „Generische Rollenmodell zur systematischen Entwicklung interner KI-basierter Services in der Produktion“ und zum anderen das „Vorgehensmodell zur Nutzung von Design-Prinzipien in der ko-kreativen Gestaltung menschenzentrierter KI-basierter Services“. Durch die Anwendung der Modelle können zentrale Herausforderungen der KI-Entwicklung überwunden und ausgewählte Erfolgsfaktoren gestärkt werden. Das Rollenmodell beschreibt entlang der Entwicklungsphasen eines KI-basierten Services idealtypisch, welche Rollen an der Entwicklung beteiligt werden sollen, welche zentralen Aufgaben die einzelnen Rollen übernehmen und wie intensiv diese in den Entwicklungsprozess eingebunden werden sollen. Das Vorgehensmodell fokussiert die menschenzentrierte Entwicklung KI-basierter Services, indem es durch einen ko-kreativen Ansatz die Anwendung von Design-Prinzipien bei der Gestaltung industrieller KI-basierter Services unterstützt. Beide Modelle fördern die stärkere Einbindung verschiedener Mitarbeitenden, insbesondere auch von Endanwenderinnen und Endanwendern, in den Entwicklungsprozess. Insgesamt trägt die Arbeit dazu bei, industrielle KI-basierte Services aus einer ganzheitlichen, menschenzentrierten Perspektive zu betrachten und durch die Anwendung der Modelle die Akzeptanz von KI-basierten Services zu stärken und die Vertrauenswürdigkeit zu erhöhen.

Abstract

The aim of this dissertation is to provide a contribution to the successful design of Artificial Intelligence (AI)-based services in production environments, going beyond a purely technology-driven development perspective. A holistic view of people, technology, and organization is taken to achieve this. A first analysis identified challenges and success factors related to the development, implementation, and operation of industrial AI-based services at all levels of a socio-technical system. In a second step, possibilities for strengthening selected success factors or addressing specific challenges were identified, covering both human-centered and organizational issues. In application-oriented research projects, considering the experience of experts in industrial AI applications, two solution areas have been identified that could lead to optimization in an industrial environment. On the one hand, the need for a role model for the development of internal AI-based services was identified. On the other hand, it was recognized that approaches from the field of Human-Centered Artificial Intelligence (HCAI) have a high potential to overcome the central challenges of AI development and AI application in an industrial context. In particular, the application of HCAI design principles is seen as valuable for industrial practice. However, it also became clear that in the broad field of HCAI research, existing approaches need to be contextualized and methods adapted to the production-specific framework. Building on this insight, two models were designed to provide methodological support for developers of AI-based services in production who are primarily technically experienced. The first is the "Generic role model for the systematic development of internal AI-based services in production". The second is the "Process model for applying design principles in the co-creative design of human-centered IAI-based services". By applying the models, key challenges of AI development can be overcome, and selected success factors can be strengthened. Along the development phases of an AI-based service, the Role Model ideally describes which roles should be involved in developing the service, what key tasks each role should take on, and how intensively they should be involved in the development process. The process model focuses on the human-centered development of AI-based services by using a co-creative approach to support the application of design principles in the design of industrial AI-based services. Both models promote a greater involvement of different stakeholders, especially end users, in the development process. Overall, the work contributes to a holistic, human-centered view of industrial AI-based services and, through the application of the models, increases the acceptance and trustworthiness of AI-based services.

Inhaltsverzeichnis

Abbildungs- und Tabellenverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

BMBF	Bundesministerium für Bildung und Forschung
DL	Deep Learning
HCAI	Human-Centered Artificial Intelligence
KI	Künstliche Intelligenz
ML	Machine Learning
OECD	Organisation für wirtschaftliche Zusammenarbeit und Entwicklung
PAIR	People + AI Research
PoC	Proof of Concept

1 Einleitung und Motivation

Künstliche Intelligenz (KI) ist eine der Schlüsseltechnologien der heutigen Zeit und somit ein wesentlicher Faktor zum Erhalt der Wettbewerbsfähigkeit Deutschlands und Europas (Bundesministerium für Bildung und Forschung, Referat Künstliche Intelligenz, 2018). Dies spiegelt sich auch im Bereich der industriellen Fertigung wider. In Anbetracht multipler Krisen der vergangenen Jahre und damit aufkommender Herausforderungen, wie gestörter Lieferketten und der Sorge um eine gesicherte Energieversorgung, ist der Aufbau resilienter und intelligenter Fertigungen sowie Logistikprozesse zunehmend bedeutend, um weiterhin wettbewerbsfähige sowie hochwertige Produkte fertigen zu können (Bienzeisler et al., 2023). Die Digitalisierung und insbesondere die Nutzung von KI versprechen hier große Potenziale zur Optimierung von Fertigungsprozessen entlang der gesamten Wertschöpfungskette (Diemer et al., 2020). Insgesamt können Produktionsprozesse flexibler, effizienter sowie nachhaltiger gestaltet und neue Geschäftsmodelle erschlossen werden (Mockenhaupt, 2021). Doch auch wenn die Potenziale gemeinhin als hoch eingeschätzt werden, ist der operative Einsatz industrieller KI-Anwendungen nicht weitverbreitet (Pokorni et al., 2021). Erfolgreich umgesetzte Pilotprojekte sind häufig vorzufinden, doch die serienreife Anwendung sowie skalierungsfähige KI-Systeme sind derweil kaum umgesetzt (Lee et al., 2019). Die Gründe hierfür sind vielschichtig, u. a. bedingt durch technische Herausforderungen, betriebswirtschaftliche Unsicherheiten sowie Kompetenzdefizite im Bereich des KI-Einsatzes (ifaa - Institut für angewandte Arbeitswissenschaft e. V., o. J.; Lundborg & Gull, 2021). Auch in der Zusammenarbeit mit Industrieunternehmen spiegelt sich dieses Bild wider. Allgemein wird der Bedarf zur Nutzung von KI als hoch eingeschätzt, jedoch ist die Umsetzung dessen eine große Herausforderung. Viele Unternehmen befinden sich noch in einem Experimentierstadium und erproben erste Anwendungen von KI (Kämpf & Langes, 2023). Dies wird insbesondere durch den grundlegenden Wandel der Arbeits- und Beschäftigungsbedingungen bedingt, der mit der Einführung von KI-Anwendungen einhergeht (Adler et al., 2022). Mitarbeitende werden mit neuen Qualifikations- und Kompetenzanforderungen konfrontiert, neue Jobrollen werden geschaffen und die Mensch-Technik-Interaktion gewinnt zunehmend an Bedeutung (Abel et al., 2019; Plattform Lernende Systeme, 2019). Diese tiefgreifenden Veränderungen können zu einer mangelnden Akzeptanz gegenüber KI-Technologien führen. Mitarbeitende haben Sorge, ihren Arbeitsplatz zu verlieren, Angst vor Leistungskontrollen und ein mangelndes Vertrauen in KI-Anwendungen (Abel et al., 2019). Um auf diesen einschneidenden Wandel akzeptanzförderlich zu reagieren, ist es wichtig, KI-Anwendungen als Teil eines sozio-technischen Systems zu verstehen und den bislang primär technisch-getriebenen Entwicklungsprozess um organisationale und mitarbeitendenzentrierte Aspekte zu erweitern (Abel et al., 2019). Um dies zu erreichen, gewinnen menschenzentrierte Gestaltungsaspekte industrieller Arbeitsumgebungen zunehmend an Bedeutung. Verdeutlicht wird dies durch die Einführung des Konzeptes der ‚Industrie 5.0', das darauf abzielt, Menschen in den Mittelpunkt der Produktion zu stellen, und eine menschenzentrierte Gestaltung von Produktionsumgebungen sowie Maschinen fokussiert (Breque et al., 2021). In diesem Zusammenhang sind auch die menschenzentrierte Gestaltung und Einführung industrieller KI-basierter Services sowie eine nachhaltige Gestaltung sich dadurch verändernder Arbeitsprozesse unerlässlich. Doch insbesondere dies ist für Unternehmen herausfordernd (Bundesministerium für Bildung und Forschung, Referat Künstliche Intelligenz, 2023), wie auch eigene Erfahrungen in Kooperationsprojekten bestätigen. Die Entwicklung und Umsetzung von industriellen KI-Use-Cases werden meist von IT-Abteilungen in Zusammenarbeit mit der Produktion verantwortet. Die Expertise dieser Abteilungen liegt insbesondere in der technischen Ausgestaltung und dem Aufbau

funktionsfähiger KI-Systeme, wodurch menschenzentrierte Aspekte häufig nicht Gegenstand der Betrachtung sind. Es bedarf hier einer Unterstützung der verantwortlichen Entwicklerinnen und Entwickler, um künftig eine ganzheitliche Perspektive einnehmen zu können, die den Menschen bzw. Mitarbeitenden in den Mittelpunkt stellt.

Ein Forschungsfeld, das sich mit der menschenzentrierten Gestaltung von KI beschäftigt, wird als ‚Menschenzentrierte KI' (engl. Human-Centered AI (HCAI)) bezeichnet. In den letzten Jahren hat HCAI zunehmend an Aufmerksamkeit gewonnen (Capel & Brereton, 2023). Insbesondere die potenziell negativen Konsequenzen wie Diskriminierung, Verletzung der Datensicherheit und Privatsphäre, die durch KI-Systeme entstehen können, haben Wissenschaftlerinnen und Wissenschaftler dazu motiviert, die Entwicklung von HCAI voranzutreiben (Xu & Gao, 2023). Das Ziel von HCAI besteht darin, KI-Systeme zu gestalten, die den Menschen unterstützen, seine Fähigkeiten verbessern, anstatt ihm Schaden zuzufügen oder ihn zu ersetzen (Xu et al., 2023). Bei der Entwicklung menschenzentrierter KI-Systeme liegt der Fokus darauf, die menschlichen Bedürfnisse und Interessen zu berücksichtigen und vertrauenswürdige, zuverlässige, ethische und sichere KI-Systeme bereitzustellen (Auernhammer, 2020; Capel & Brereton, 2023; Ozmen Garibay et al., 2023; Pokorni et al., 2021). Auch wenn der Bedarf für HCAI von Wissenschaft, Industrie und Politik erkannt wurde, sind Entwicklungsprozesse von KI-Systemen weiterhin primär technisch getrieben (Cremers et al., 2019; Lee, 2020). Um dies zu verändern, braucht es Methoden und Vorgehensmodelle, die Orientierung und Hilfestellung zur Umsetzung von HCAI entlang des KI-Lebenszyklus bieten und zur Anwendung in spezifischen Einsatzbereichen konzipiert sind. Insbesondere diesen Bedarf adressiert die vorliegende Arbeit im Kontext industrieller KI-basierter Services.

Die vorliegende Arbeit befasst sich mit zentralen Herausforderungen industrieller KI-basierter Services sowie Erfolgsfaktoren zur erfolgreichen Entwicklung und Implementierung dieser. Folgende Forschungsfrage wird in diesem Zusammenhang beantwortet: „Welche Herausforderungen und Erfolgsfaktoren gibt es im Zusammenhang mit der Entwicklung, Einführung und dem Betrieb industrieller KI-basierter Services entlang der Ebenen Mensch, Technik und Organisation?". Aufbauend auf den identifizierten Erfolgsfaktoren und Herausforderungen wird im Rahmen der zweiten Forschungsfrage untersucht, welche spezifischen Methoden und Modelle zur Lösung ausgewählter Herausforderungen sowie Stärkung ausgewählter Erfolgsfaktoren in Bezug auf organisations- und menschenzentrierte Aspekte beitragen können. Das übergeordnete Ziel besteht darin, ein menschenzentriertes KI-Design in produktiven Arbeitsumgebungen zu fördern und damit bestehende Herausforderungen bei der Entwicklung und Nutzung industrieller KI-Systeme zu überwinden, bzw. einen Beitrag zur erfolgreichen Implementierung dieser zu leisten. Insbesondere soll dies durch die Bereitstellung anwendungsnah konzipierter Methoden und Modelle geschehen, die den Anforderungen und Bedürfnissen der Industrie gerecht werden und die gleichzeitig den Menschen in den Mittelpunkt stellen. Daraus ergibt sich die dritte zu beantwortende Forschungsfrage: „Wie können Methoden und Modelle zur Förderung eines menschenzentrierten KI-Designs im industriellen Setting gestaltet sein?". Die vorliegende Arbeit konzentriert sich dabei nicht auf die technischen Herausforderungen bei der Gestaltung menschenzentrierter KI-Anwendungen, sondern widmet sich den Gestaltungsaspekten, die entweder auf organisatorischer Ebene oder auf Mitarbeitendenebene anzusiedeln sind.

Nachfolgend wird in Kapitel 2 der theoretische Hintergrund der vorliegenden Arbeit aufbereitet, indem der Stand der Forschung im Bereich industrielle KI sowie im Bereich HCAI dargestellt wird. In Kapitel 3 werden zum einen die zentralen Forschungsfragen der Arbeit sowie der Zusammenhang der integrierten Publikationen dargestellt. Zum anderen sind die

Publikationen in Originalfassung integriert. Anschließend werden in Kapitel 4 die Ergebnisse der einzelnen Publikationen in einen Gesamtzusammenhang gestellt, auf theoretische und praktische Implikationen eingegangen sowie ein Ausblick auf weitere Forschungsbedarfe gegeben. Abschließend wird in Kapitel 5 ein kurzes Fazit der Arbeit dargestellt.

2 Hintergrund und theoretischer Rahmen

In den folgenden Kapiteln wird der theoretische Rahmen beschrieben, der der vorliegenden Arbeit zu Grunde liegt. Dabei wird zunächst auf das Thema der industriellen KI eingegangen (Kapitel 2.1) und anschließend wird das Thema der menschzentrierten KI beleuchtet (Kapitel 2.2). Durch die Betrachtung beider Themenfelder - industrielle KI und menschenzentrierte KI - wird ein umfassendes Verständnis für den Hintergrund der vorliegenden Arbeit sowie für die zentralen Herausforderungen im Kontext beider Themenfelder geschaffen.

Nach einer Einführung in das Thema werden verschiedene Ansätze und Methoden beschrieben, die eine menschenzentrierte Gestaltung KI-basierter Services ermöglichen.

2.1 Künstliche Intelligenz in der Produktion

Kernthema dieser Arbeit ist der Einsatz von KI im Produktionsumfeld. Daher wird dieses Thema in den folgenden Abschnitten näher beleuchtet. Neben einer begrifflichen Einordnung werden verschiedene Anwendungsfelder industrieller KI beschrieben sowie Potenziale und Hindernisse des Einsatzes aufgezeigt. Abschließend werden, abgeleitet aus den zentralen Erkenntnissen der Literatur, drei zentrale Herausforderungen der industriellen KI zusammengefasst, die im Kontext der vorliegenden Arbeit besondere Berücksichtigung finden.

2.1.1 Industrielle KI: Einführung und Begriffsverständnis

Die vierte industrielle Revolution, in Deutschland unter dem Begriff ‚Industrie 4.0' bekannt, beschreibt die zunehmende Digitalisierung produzierender Unternehmen, getrieben durch die Einführung cyber-physischer Systeme, eine Zunahme von Datenströmen und Big Data, Cloud-Technologien sowie additiven Fertigungsverfahren (Bundesministerium für Wirtschaft und Energie [BMWi], 2015). In diesem Zusammenhang erfahren auch KI-Technologien zunehmend Aufmerksamkeit im industriellen Umfeld und kommen vermehrt zum Einsatz (Jan et al., 2023). Grundsätzlich gilt KI als Teilgebiet der Informatik, mit dem Ziel kognitive Fähigkeiten wie Problemlösen, Lernen und Planen mithilfe von Computersystemen zu verwirklichen (Plattform Lernende Systeme, 2019, S. 4). Unterschieden wird dabei zwischen „schwacher" und „starker" KI. Schwache KI-Systeme werden für die menschenähnliche Lösung spezifischer Aufgaben und Probleme eingesetzt, wohingegen die starke KI danach strebt, eine Intelligenz zu entwickeln, die die gleichen kognitiven Fähigkeiten wie ein Mensch besitzt (vgl. Mockenhaupt, 2021; Scheuer, 2020; VDMA et al., 2020). In der Unternehmenspraxis zum Einsatz kommen bislang schwache KI-Systeme (Pokorni et al., 2021), beispielsweise in Form von Bilderkennung, Mustererkennung und Sprachverarbeitung.

Eine einheitliche Definition von KI existiert in Wissenschaft und Praxis bislang nicht. Eine allgemeine Orientierung bietet jedoch eine gängige Klassifizierung entlang der Begriffe: ‚Künstliche Intelligenz' (KI), ‚Maschinelles Lernen' (ML) und ‚Tiefes Lernen' (engl. Deep Learning (DL)) (vgl. Abbildung 1).

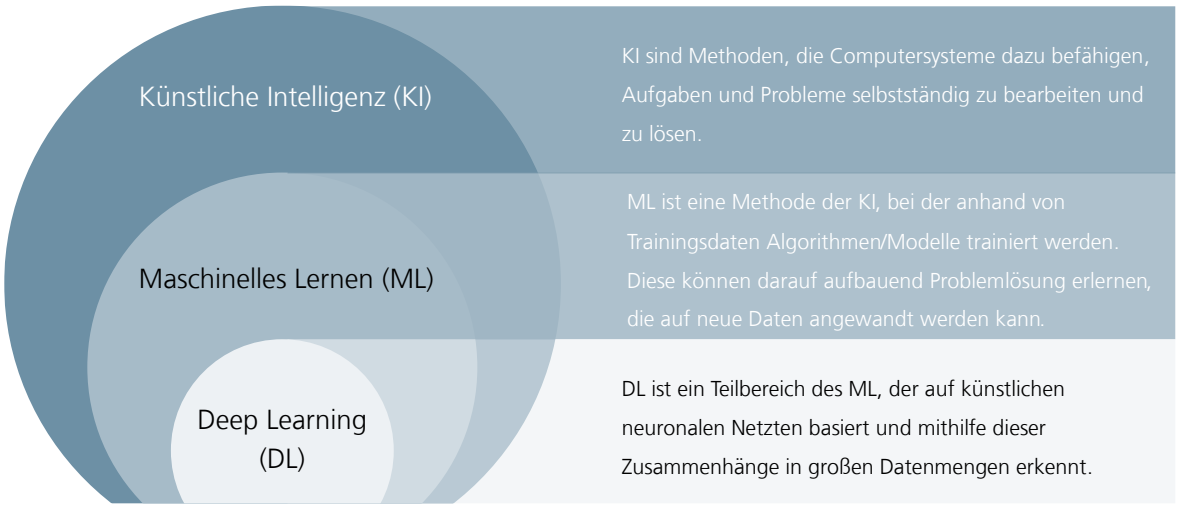


Abbildung 1. Begriffsdefinitionen KI. Eigene Darstellung angelehnt an die Ausführungen von Plattform Lernende Systeme, o. J., 2019; VDMA et al., 2020.

Ein zentrales Teilgebiet der KI ist ML. Beim ML werden KI-Modelle anhand vorhandener Daten trainiert, um darin Muster und Informationen zu erkennen und daraus Vorhersagen oder Entscheidungen abzuleiten. Diese antrainierten Modelle können dann wiederum auf unbekannte Datensätze angewandt werden (Fischer et al., 2022; Pokorni et al., 2021). Innerhalb des ML wird zwischen verschiedenen Lernverfahren unterschieden: überwachtes Lernen, unüberwachtes Lernen und verstärkendes Lernen.

- **Überwachtes Lernen (engl. supervised learning):** Beim überwachten Lernen sind für den Lernprozess neben den Eingabedaten auch die zugehörigen gewünschten Ausgabedaten bekannt. Das KI-Modell lernt, aus gelabelten (beschrifteten/gekennzeichneten) Input-Daten einen definierten Output abzuleiten. Diese gelernte Verbindung kann anschließend auf neue Daten übertragen werden. (François-Lavet et al., 2018; Mockenhaupt, 2021)
- **Unüberwachtes Lernen (engl. unsupervised learning):** Beim unüberwachten Lernen werden ungelabelte Daten verwendet, bei denen dem System keine vorgegebene Lösung bekannt ist. Das KI-Modell erkennt unbekannte Muster in den Daten und lernt ohne vorgegebene Zielwerte. (François-Lavet et al., 2018; Mockenhaupt, 2021)
- **Bestärkendes Lernen (engl. reinforcement learning):** Beim bestärkenden Lernen lernt ein KI-Modell durch das Prinzip von Versuch-und-Irrtum und basierend auf Belohnungen. Auch hier kommen ungelabelte Daten zum Einsatz, und die Zielwerte sind unbekannt. In diesem Szenario wird vorausgesetzt, dass Resultate des Modells bewertet werden können. Je besser das Ergebnis, umso höher die Belohnung für das KI-Modell. Dadurch erlernt das KI-Modell, schrittweise Aktionen zu wählen, die zu einer hohen Belohnung führen, und verändert sich somit über die Zeit im Lernprozess. (François-Lavet et al., 2018; Mockenhaupt, 2021)

Im industriellen Kontext wird KI definiert als „systematic discipline focusing on the development, validation, deployment and maintenance of AI solutions (in their varied forms) for industrial applications with sustainable performance“ (Peres et al., 2020, S. 220122). Die vorliegende Arbeit orientiert sich an diesem Verständnis von KI, das die oben beschriebenen Varianten und Lernformen einschließt.

Überdies können industrielle KI-Anwendungen als digitale Dienstleistungen verstanden werden, die entweder im Unternehmen eingesetzt oder externen Kunden zur Verfügung gestellt werden. Dies charakterisiert sich durch drei Merkmale von Dienstleistungen, welche auch im Kontext von KI-Anwendungen zum Tragen kommen. Zum einen die Integration des externen Faktors: KI-basierte Services können nur durch Beteiligung externer Akteure (z. B.

Mitarbeitende, Kunden) erfolgreich entwickelt und eingesetzt werden (Bienzeisler et al., 2023). Zum anderen durch einen hohen Anteil immaterieller Leistungsbestandteile KI-basierter Services, die zu Veränderungen von Objekten, Prozessen oder Informationen führen (Tombeil et al., 2020), und zuletzt durch die Heterogenität der Leistung, bedingt durch sich verändernde Umgebungsbedingungen, variierende Input-Daten sowie die selbstständige Weiterentwicklung der KI-Modelle (Bienzeisler et al., 2023). Vor diesem Hintergrund wird KI in dieser Arbeit als Teil einer Dienstleistung verstanden und dementsprechend als KI-basierter Service bezeichnet.

2.1.2 Industrielle KI: Nutzung und Anwendungsfelder

KI-basierte Services können entlang der gesamten industriellen Wertschöpfungskette eingesetzt werden. Typische Anwendungsfelder für KI sind die Produktionssteuerung, das Qualitätsmanagement, die Instandhaltung, Robotik und digitale Assistenzsysteme (Peres et al., 2020; Pokorni et al., 2021). Aktuelle Studien zeigen, dass KI-basierte Services bereits angewandt werden, jedoch bislang nicht flächendeckend zur Anwendung kommen. Einer aktuellen Studie des ifo Instituts (2023) zufolge nutzt jedes dritte Industrieunternehmen in Deutschland bereits KI (17,3 %) bzw. plant den Einsatz (12,9 %) (Schaller et al., 2023). In der 2022 durchgeführten Studie „Künstliche Intelligenz in produzierenden Unternehmen" des Instituts für angewandte Arbeitswissenschaft wurden 459 Mitarbeitende unterschiedlicher produzierender Unternehmen befragt. Der Studie nach gaben 36 % der Befragten an, KI werde bereits im Unternehmen eingesetzt, 37 % gaben an, die Nutzung sei in Planung und wiederum 27 % gaben an, KI würde weder genutzt werden noch sei eine Nutzung in Planung (Harlacher et al., 2023). In der folgenden Abbildung 2 ist dargestellt, welche KI-Technologien am häufigsten von produzierenden Unternehmen eingesetzt werden.

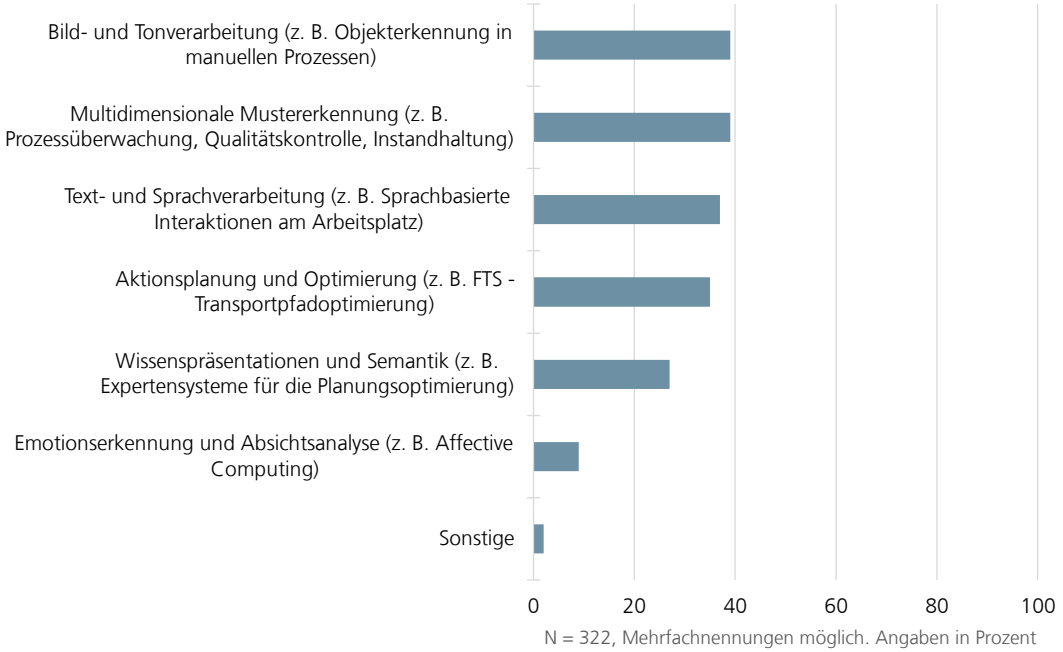


Abbildung 2. Übersicht eingesetzter oder geplanter KI-Technologien bei produzierenden Unternehmen. Eigene Darstellung in Anlehnung an ifaa – Institut für angewandte Arbeitswissenschaft e. V., o. J.

Zu unterscheiden ist außerdem zwischen industriellen KI-basierten Services, a) die von Unternehmen bei externen Anbietern eingekauft werden (z. B. prädiktive Fernwartungen von Maschinen als Dienstleistungen der Maschinenhersteller) und b) innerhalb eines Unternehmens entwickelter und angebotener KI-Anwendungen (z. B. Eigenentwicklungen zur Erkennung von Qualitätsmängeln an Produkten). Die vorliegende Arbeit fokussiert insbesondere auf letztere KI-Anwendungen, die innerhalb eines Unternehmens entwickelt, implementiert und betrieben werden.

Drei typische Beispiele für KI-basierte Services zur Anwendung in der Produktion sind folgend aufgelistet:

- **Automatisierte Qualitätskontrolle:** KI-Modelle werden eingesetzt, um Abweichungen oder Defekte in hergestellten Produkten zu erkennen. Dies kann z. B. die Prüfung von Bauteilen auf Risse, Kratzer, Schweißspritzer o. Ä. auf Basis von Bilddaten oder die Prüfung von Schweißnähten und -punkten auf Basis der Auswertung von Sensordaten sein. (Mockenhaupt, 2021; Pokorni et al., 2021)
- **Digitale Assistenzsysteme:** KI-basierte digitale Assistenzsysteme können Mitarbeitende bei ihren Arbeitsaufgaben entlasten, indem sie kontext- und zeitspezifische Informationen bereitstellen, die den Arbeitsprozess unterstützen (Pokorni et al., 2021).
- **Prozessoptimierung:** KI-Modelle können eingesetzt werden, um Fertigungsprozesse zu analysieren und zu optimieren, basierend auf Daten verschiedener Quellen. So kann der gesamte Produktionsablauf transparent abgebildet werden und auf Verbesserungsmöglichkeiten hin geprüft werden. (Mockenhaupt, 2021; Pokorni et al., 2021)

2.1.3 Industrielle KI: Anforderungen und Hindernisse des Einsatzes

Die begrenzte flächendeckende Anwendung von industriellen KI-basierten Services resultiert aus verschiedenen Herausforderungen, die mit der Entwicklung und dem Betrieb dieser Dienstleistungen einhergehen. Lee et al. (2019) beschreiben fünf zentrale Hindernisse, die den operativen Einsatz KI-basierter Services behindern:

1) Beweise, die den Erfolg industrieller KI nachweisen, sind nicht ausreichend vorhanden.
2) Fehlen eines systematischen Ansatzes zur Nutzung von KI.
3) Maschinendaten sind nicht standardisiert und strukturiert, da sie in unterschiedlichen Formaten erfasst und protokolliert werden.
4) Mangel an Ausfalldaten/Fehlerdaten
5) Die dynamischen Anwendungskontexte erfordern möglicherweise menschliches Eingreifen, um die Ergebnisse zu überprüfen und zu validieren.

Insbesondere für produzierende Unternehmen, die charakterisiert sind durch standardisierte, effiziente und hochkomplexe Fertigungsprozesse, ist es bedeutend, dass KI-Systeme skalierbar, reliabel und sicher sind (Lee, 2020). Eine große Herausforderung von Industrieunternehmen ist die Schaffung skalierbarer KI-Lösungen. Skalierbarkeit wird verstanden als „die Fähigkeit eines Systems, eines Netzwerks oder eines Prozesses zur bedarfsgerechten Größenanpassung“ (Mockenhaupt, 2021, S. 235). Um dies zu erreichen, müssen jedoch Herausforderungen in Bezug auf eine mangelnde Datenqualität und -verfügbarkeit überwunden werden, und es muss sichergestellt werden, dass die Systeme robust sind und reproduzierbare Ergebnisse liefern (Lee, 2020; Peres et al., 2020). Weitere Hindernisse sind die Gewährleistung des Datenschutzes (Harlacher et al., 2023) sowie die Integration KI-basierter Services in bestehende Produktionssysteme, die bedingt durch historisch gewachsene Prozesslandschaften von

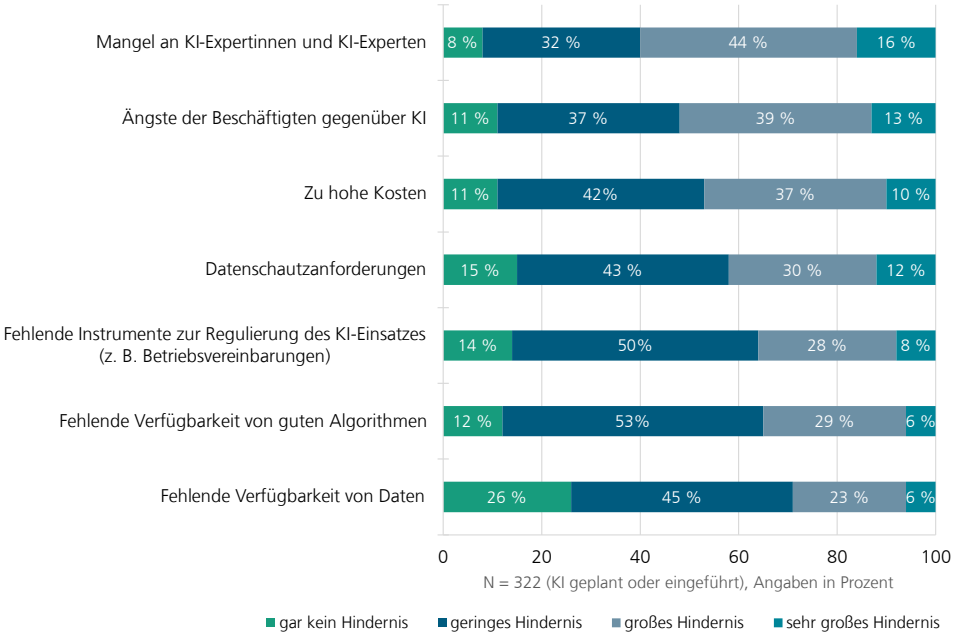


Abbildung 3. Hindernisse bei der Einführung von KI-Systemen in produzierenden Unternehmen (N=332). Eigene Darstellung nach Harlacher et al., 2023.

Heterogenität geprägt sind (Mockenhaupt, 2021). Abbildung 3 gibt einen Überblick genannter Hindernisse bei der Einführung von KI-Systemen in produzierenden Unternehmen.

Die Einführung von KI kann zu disruptiven Veränderungen in Produktionsumgebungen führen. Dies bedeutet, dass Prozesse und Strukturen neu ausgerichtet werden müssen (Mockenhaupt, 2021). Doch haben Unternehmen bislang kein standardisiertes Vorgehen zur Integration KI-basierter Services in ihre bestehenden Produktionsprozesse und Geschäftsmodelle festgelegt (Peres et al., 2020). Entwicklungsprozesse KI-basierter Services sind aktuell primär technisch getrieben (Cremers et al., 2019; Lee et al., 2019; Xu et al., 2021). Organisationale sowie menschenzentrierte Entwicklungsaspekte treten dagegen in den Hintergrund. Dabei ist eine ganzheitliche Betrachtung entscheidend für die erfolgreiche Einführung KI-basierter Services. Dies verdeutlicht eine Arbeit von Hoffmann et al. (2021), die 16 Anforderungen an die Entwicklung und Implementierung industrieller KI definiert haben (vgl. Tabelle 1).

Tabelle 1. Anforderungen an die Entwicklung und Implementierung industrieller KI nach Hoffmann et al., 2021

Kategorie	Anforderung
Anpassung der KI	1. Schrittweise Einführung 2. Human-in-the-loop 3. Datenverfügbarkeit
Entwicklung der KI	4. Virtuelles Lernen der Systeme (z. B. Einsatz digitaler Zwillinge) 5. Anpassung an veränderte Umgebungsbedingungen 6. Einfaches Design und Nutzung (Komplexität verbergen)
Integration in bestehende Produktionssysteme	7. KI-Systeme sollten keine Schlussfolgerungen aus KI-generierten Daten ziehen 8. Grenzen des Vertrauens in KI festlegen 9. KI-Systeme sollten voneinander lernen
Sicherheit (Safety/Security)	10. Sicherheit

Kategorie	Anforderung
	11. Robustheit gegenüber unerwünschten Eingaben
Vertrauen in Funktionsfähigkeit	12. Rückverfolgbarkeit und Transparenz von Entscheidungen 13. Frei von Bias 14. Vertraulichkeits-/Zuverlässigkeitsmaße ausgeben 15. Vertrauen/Qualitätsklassifizierung definieren 16. Nachweis der Fähigkeiten

Insbesondere die ganzheitliche Betrachtung KI-basierter Services über die technische Entwicklung hinaus stellt Unternehmen vor eine große Herausforderung (Pokorni et al., 2021). Doch sind es neben den technischen Hürden jene auf organisationaler und menschlicher Ebene, die überwunden werden müssen, um letztlich das volle Potenzial KI-basierter Services heben zu können. Einen Beitrag zur Überwindung organisationaler und mitarbeitendenzentrierter Hürden industrieller KI-basierter Services zu leisten, ist Kern der vorliegenden Arbeit.

2.1.4 Industrielle KI: Bestandteil eines sozio-technischen Systems

Die ganzheitliche Betrachtung KI-basierter Services über die Ebenen Mensch, Technik und Organisation hinweg geht einher mit dem Verständnis dieser als Teil sozio-technischer Arbeitssysteme. Kern einer sozio-technischen Systemperspektive ist es, soziale und technische Komponenten eines Systems in ihrer Wechselwirkung zu betrachten. Ein zentrales Modell zur Betrachtung sozio-technischer Arbeitssysteme ist das MTO-Konzept nach Ulrich (2013), welches davon ausgeht, „dass Mensch, Technik und Organisation in ihrer gegenseitigen Abhängigkeit und ihrem Zusammenwirken zu reflektieren sind" (Ulich, 2013, S. 5). Bisherige Forschungsarbeiten zu ‚Industrie 4.0' konzentrieren sich jedoch stark auf technologische Aspekte, menschliche Faktoren sind dagegen kaum Gegenstand der Betrachtungen (Abel et al., 2019; Winkelhaus et al., 2021). Dabei kommt dem Menschen auch in digitalisierten, automatisierten und intelligenten Produktionsumgebungen eine zentrale Rolle zu, denn trotz des vermehrten Einsatzes von KI-Technologien wird es auch weiterhin eine menschliche Aufsicht zur Steuerung und Kontrolle in Produktionsumgebungen brauchen (Mockenhaupt, 2021, S. 230). Industrielle KI-Anwendungen sind mehr als Algorithmen; sie verbinden Menschen und Objekte/Dinge durch Systeme miteinander, mit dem Ziel, evidenzbasiert Optimierungen im Produktionssystem zu erreichen (Lee, 2020, S. 19). Der Mensch ist somit zentraler Bestandteil industrieller KI-basierter Services, denn nur wenn das Zusammenspiel von Mensch und Technik optimal gestaltet ist, können Produktivität und Effektivität des Produktionssystems gesteigert werden (Winkelhaus et al., 2021). Aus diesem Grund sollten KI-basierte Services immer unter Berücksichtigung der Mitarbeitenden sowie der bestehenden Organisationsstrukturen gestaltet werden (Gabriel et al., 2022). Mit dem Konzept der ‚Industrie 5.0' rückt auch die Europäische Kommission den Menschen in den Fokus industrieller Entwicklungen. Neben Nachhaltigkeit und Resilienz ist die Menschenzentrierung, welche menschliche Bedürfnisse und Interessen in den Mittelpunkt von Produktionsprozessen stellt, ein Kernelement von ‚Industrie 5.0' (Breque et al., 2021).

Durch die Einführung KI-basierter Services ist jedoch die Gestaltung sozio-technischer Systeme komplexer geworden. Aufgrund der besonderen Gegebenheiten von KI-Technologien müssen weitere Komponenten berücksichtigt werden, z. B. die Autonomie der Systeme, Transparenzpflichten, Erklärbarkeitsprobleme oder andere ethische Aspekte (Xu & Gao, 2024). Die sozio-technische Gestaltung von KI-basierten Services ist in organisationale Strukturen eingebettet, die beachtet und angepasst werden müssen. Diese Integration organisationaler

Aspekte in die Gestaltung intelligenter sozio-technischer Systeme hilft dabei, das Potenzial KI-basierter Services zu heben und gleichzeitig Mitarbeitende und deren Interessen zu berücksichtigen (Xu & Gao, 2024). Außerdem sind organisationale intelligente sozio-technische Systeme eingebettet in ein Ökosystem, welches u. a. Standards zur Verwendung von KI über Organisationen hinweg schafft (Xu & Gao, 2024). Im Kontext industrieller KI-Anwendungen können hier Initiativen wie *Catena X* genannt werden, die durch die kooperative Bereitstellung von Daten die Entwicklung KI-basierter Systeme in der Automobilbranche fördern (Göbels et al., im Druck). Diese Initiativen können ebenfalls dazu beitragen, die Förderung von KI-Anwendungen voranzutreiben, die den Menschen in den Mittelpunkt stellen und Standards für einen vertrauenswürdigen Umgang im sozio-technischen System schaffen. Die äußere Ebene des sozio-technischen Systems stellt die soziale Umwelt dar, die u. a. durch ethische Grundsätze, Gesetze sowie politische Leitlinien geprägt wird und die Grundlage für die Entwicklung und Bereitstellung von KI-basierten Services in unserer Gesellschaft bildet, die dem Menschen nützen und diesem nicht schaden. Abbildung 4 spiegelt die Komplexität sozio-technischer Systeme KI-basierter Services wider. Im Zentrum des intelligenten sozio-technischen Systems steht das menschenzentrierte KI-Design (Xu & Gao, 2024), das die Perspektiven des sozialen Subsystems und die Anforderungen des technischen Subsystems miteinander verbindet.

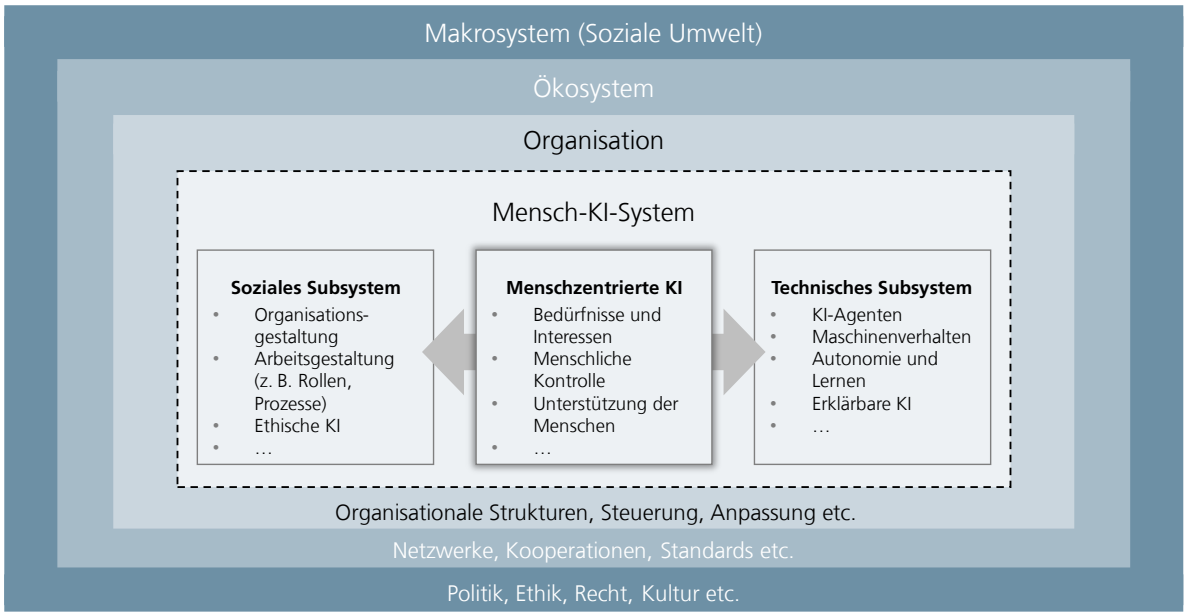


Abbildung 4. Darstellung eines sozio-technischen Systems KI-basierter Services in Anlehnung an Xu & Gao, 2024

2.1.5 Industrielle KI: Zusammenfassung der zentralen Herausforderungen

Der folgende Abschnitt fasst die drei zentralen Herausforderungen zusammen, die sich aus den vorangegangenen Kapiteln zu KI-basierten Services in der Produktion ergeben. Die Zusammenfassung fokussiert dabei insbesondere auf jene Herausforderungen, die auf Organisationaler- sowie Mitarbeitendenebene zu verorten sind.

- Herausforderung 1: Tiefgreifende Veränderungen durch KI-Systeme können bei den Mitarbeitenden Ängste auslösen und zu Vertrauens- und Akzeptanzproblemen führen.
- Herausforderung 2: Es bedarf einer Anpassung organisatorischer Prozesse und Strukturen an die Bedürfnisse, die sich aus den Veränderungen durch die Einführung KI-basierter Services im Unternehmen ergeben.

- Herausforderung 3: KI-basierte Services sind komplexe sozio-technische Systeme, die in ihrer Ganzheitlichkeit bei der Entwicklung, der Implementierung und dem Betrieb berücksichtigt werden müssen.

Einen Beitrag zur Bewältigung der aufgeführten Herausforderungen im Zusammenhang mit industriellen KI-basierten Services zu leisten, ist das zentrale Anliegen der vorliegenden Arbeit.

2.2 Menschenzentrierte Künstliche Intelligenz

Im Folgenden wird das Forschungsfeld ‚Menschzentrierte KI' ausführlich betrachtet. Ausgehend von einer Darstellung der aktuellen Entwicklungen im Themenfeld, über die Beschreibung der Zielsetzung von HCAI, der Vorstellung ausgewählter Ansätze zur Gestaltung von HCAI, werden abschließend die aus der Literatur abgeleiteten zentralen Herausforderungen dargestellt.

2.2.1 Menschenzentrierte KI: Aktuelle Entwicklungen

Das Thema der Menschenzentrierung gewinnt auch im Kontext von KI zunehmend an Aufmerksamkeit. Sowohl in der Wissenschaft, der Politik als auch seitens der Industrie wird das Thema seit einigen Jahren verstärkt diskutiert. So ist die Anzahl an wissenschaftlichen Publikationen, die sich mit dem Thema HCAI befassen, seit 2017 kontinuierlich gestiegen (vgl. Abbildung 5).

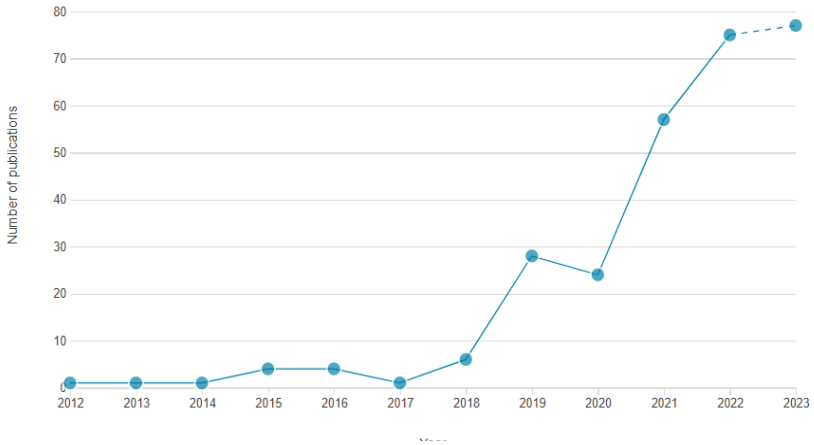


Abbildung 5. Jährliche Anzahl der Publikationen zum Thema „Human-Centered Artificial Intelligence". Stand 11.01.2024. Quelle: KATI developed by Fraunhofer INT

Auch in der internationalen Politik rückt das Thema eines verantwortungsvollen, vertrauensvollen und menschenzentrierten KI-Designs zunehmend in den Fokus. Um den Herausforderungen im Umgang mit KI-Technologien zu begegnen, haben einige Regierungen nationale KI-Strategien erarbeitet. In Deutschland wurde 2018 die „Nationale Strategie für Künstliche Intelligenz" veröffentlicht. Eines der darin festgelegten Ziele ist es, eine „verantwortungsvolle und gemeinwohlorientierte Entwicklung und Nutzung von KI" zu etablieren (Bundesministerium für Bildung und Forschung, Referat Künstliche Intelligenz, 2018). Auch andere Länder haben nationale KI-Strategien erarbeitet. So hat sich beispielsweise Dänemark zum Ziel gesetzt, eine Vorreiterrolle bei der Entwicklung verantwortungsvoller KI einzunehmen und die Entwicklung einer Grundlage für ethische und menschenzentrierte KI als Ziel definiert (Bundesministerium für Bildung und Forschung, Referat Künstliche Intelligenz, o. J.). Eine Übersicht verschiedener KI-Strategien wird vom Bundesministerium für Bildung und

Forschung (BMBF) bereitgestellt (Bundesministerium für Bildung und Forschung, Referat Künstliche Intelligenz, o. J.).

Eine Vorreiterrolle im Hinblick auf den gesamtgesellschaftlichen Umgang mit KI nimmt die Europäische Union ein. Im Jahr 2019 hat sie die „Ethik-Leitlinien für eine Vertrauenswürdige KI" veröffentlicht, die vier Prinzipien und sieben Anforderungen an ein ethisches und menschenzentriertes KI-Design definiert (High-Level Expert Group on Artificial Intelligence, 2019; Hochrangige Expertengruppe für künstliche Intelligenz, 2019). Ende 2023 wurde außerdem der EU AI Act verabschiedet, ein europäisches Gesetz zur Regulierung von KI-Technologien. In diesem werden basierend auf einem risikobasierten Ansatz Anforderungen an KI-Systeme u. a. im Hinblick auf das Risikomanagement, Data-Governance, technische Dokumentation, Transparenz- und Aufzeichnungspflicht sowie die menschliche Aufsicht festgelegt (European Commission, 2021; Kutz, Göbels, et al., 2023). Prinzipien für die Gestaltung eines menschenzentrierten Ansatzes für vertrauenswürdige KI-Systeme wurden auch seitens der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) festgelegt. Die Leitlinie beinhaltet zum einen wertebasierte Prinzipien und zum anderen Empfehlungen für politische Entscheidungsträger (OECD, 2023). Neben Wissenschaft und Politik beschäftigen sich auch Unternehmen, insbesondere große Technologiekonzerne, zunehmend mit der Gestaltung vertrauenswürdiger und menschenzentrierter KI-Systeme. Einige Konzerne definieren eigene Leitlinien für den Umgang mit KI, wie beispielsweise Bosch (Robert Bosch GmbH, o. J.), IBM (IBM Deutschland GmbH, o. J.) oder BMW (BMW Group, 2020). Unternehmen wie Microsoft und Google haben Design-Prinzipien für ein menschenzentriertes KI-Design entwickelt und als Hilfestellung für KI-Entwicklerinnen und KI-Entwickler veröffentlicht (Google PAIR, 2019; Microsoft, o. J.). Ausgewählte Leitlinien werden in den folgenden Kapiteln detaillierter beschrieben.

2.2.2 Menschenzentrierte KI: Verständnis und Zielsetzung

Menschenzentrierte KI wird allgemein verstanden als Ansatz, der Menschen in den Fokus KI-basierter Services entlang des gesamten KI-Lebenszyklus stellt (Xu et al., 2023). Bisherige Ansätze zur Entwicklung KI-basierter Services sind primär technisch getrieben (Abel et al., 2019). HCAI zielt darauf ab, diese Ansätze durch die Integration menschenzentrierter Entwicklungsaspekte zu vervollständigen (Xu et al., 2023). Ziel ist es, KI-basierte Services zu designen, die als reliabel, sicher und vertrauenswürdig wahrgenommen werden, den Menschen in seinen Fähigkeiten unterstützen und von diesem akzeptiert werden (Shneiderman, 2020b; Xu, 2019). Dabei soll nicht der Grad an Automatisierung durch eine Erhöhung der menschlichen Kontrolle reduziert werden, sondern beides in Einklang gebracht werden. Dieses Verständnis verdeutlicht Shneiderman in seinem 2020 veröffentlichten zweidimensionalen Framework für HCAI, das das Ziel unterstreicht, durch gutes Design einen hohen Grad an Automatisierung und menschlicher Kontrolle zu erreichen (Shneiderman, 2020b).

Eine allgemein gültige Definition von HCAI existiert bislang nicht, stattdessen wird HCAI als Überbegriff verwendet, welcher verschiedene Ansätze und Forschungsarbeiten einschließt. Dies verdeutlicht ein Literaturreview von Capel und Brereton (2023), die vier Hauptforschungsbereiche im Kontext HCAI identifiziert haben: (1) Erklärbare und interpretierbare KI, (2) Menschenzentrierte Ansätze für Design und Evaluation von KI, (3) Mensch-KI-Teaming und (4) Ethische KI. Unter Berücksichtigung der Arbeiten in diesen vier Forschungsbereichen definieren die Autoren HCAI wie folgt: „Human-Centered Artifcial Intelligence utilizes data to empower and enable its human users, while revealing its underlying

values, biases, limitations, and the ethics of its data gathering and algorithms to foster ethical, interactive, and contestable use" (Capel & Brereton, 2023).

In der vorliegenden Arbeit wird HCAI als Ansatz verstanden, den Anforderungen an ein ethisches und vertrauenswürdiges KI-Design gerecht zu werden, das den Menschen und seine Bedürfnisse in den Mittelpunkt stellt. Dieses Verständnis geht einher mit dem von Xu et al. (2021) veröffentlichten Framework, das drei Aspekte von HCAI (1. Technik, 2. Mensch und 3. Ethik) und deren Zusammenspiel beschreibt. Ziel des Frameworks ist es, den Menschen und seine Bedürfnisse zum Ausgangspunkt für die Entwicklung von KI-basierten Services zu machen, die Technologie zur Unterstützung menschlicher Fähigkeiten einzusetzen und erklärbare KI-Systeme zu kreieren, die einen Nutzen bieten. Gleichzeitig müssen ethische Standards bei der Entwicklung berücksichtigt werden, um verantwortungsvolle und von Menschen kontrollierbare KI-Systeme zu schaffen (Xu et al., 2021; Xu et al., 2022). Xu et al. (2023) definieren sieben Ziele des menschenzentrierten KI-Designs (vgl. Tabelle 2).

Tabelle 2. Sieben Ziele eines menschenzentrierten KI-Designs nach Xu et al., 2023.

Ziel	Definition
Vertrauenswürdige KI	Vertrauenswürdige KI-Systeme sind intelligent und zuverlässig, sodass Menschen ihnen vertrauen können. Diese Systeme treffen transparente und nachvollziehbare Entscheidungen, die fair und frei von Diskriminierung sind. Sie arbeiten vorhersehbar und sicher, auch in unterschiedlichen Situationen. Zudem schützen sie die Daten und Rechte der Nutzenden.
Skalierbare KI	Skalierbare KI-Systeme bieten einen langfristigen Nutzen in Bezug auf Daten, Modelle, Infrastruktur etc. Durch Einbindung menschlicher Rollen werden Mängel wie geringe Robustheit und mangelnde Erklärbarkeit überwunden.
Nützliche KI	Nützliche KI-Systeme werden basierend auf relevanten Anwendungsszenarien entwickelt. Sie sind bedürfnis- oder problemorientiert und bieten einen Mehrwert. Diese Systeme unterstützen die Nutzenden bei der Erfüllung ihrer Aufgaben und passen sich an veränderte Umgebungsbedingungen an.
Benutzerfreundliche KI	Benutzerfreundliche KI-Systeme bieten ein positives Nutzungserlebnis. Die Systemausgaben sind verständlich und können effektiv und effizient zum Erreichen individueller Aufgabenziele eingesetzt werden. Zudem lassen sich die Systeme problemlos in den Alltag oder die Arbeitsabläufe der Nutzenden integrieren.
Befähigung	Die KI-Systeme nutzen sowohl maschinelle als auch menschliche Intelligenz, um die Fähigkeiten des Menschen zu verbessern, zu ergänzen oder zu erweitern. Dies wird durch eine effektive Interaktion und Zusammenarbeit zwischen Menschen und KI erreicht, anstatt menschliche Fähigkeiten zu ersetzen oder zu reduzieren.
Verantwortungsvolle KI	Verantwortungsvolle KI-Systeme werden unter ethischen Gesichtspunkten entwickelt, eingesetzt und verwaltet, um für alle Beteiligten sicherzustellen, dass sie zuverlässig und sicher sind und der Menschheit nützen und nicht schaden. Sie verfügen über Mechanismen, mit denen die Menschen für die Handlungen der Systeme zur Verantwortung zu ziehen.
Menschlich kontrollierbare KI	KI-Systeme, die effektiv und sicher von Menschen gesteuert, vorhergesagt, verstanden und verwaltet werden können. Dadurch werden unbeabsichtigte Folgen wie die Schädigung von Menschen vermieden und sichergestellt, dass

Ziel	Definition
	KI immer innerhalb von Entwicklerinnen und Entwicklern oder Nutzenden definierter Grenzen handelt.

Die Gestaltung menschenzentrierter KI-Anwendungen wird als Möglichkeit verstanden, bekannte Hürden der KI-Entwicklung und Implementierung zu überwinden, sodass diese ihr volles Potenzial entfalten können (Xu & Gao, 2023). In der Literatur bereits viel diskutiert sind Konzepte und Frameworks für HCAI, die zur Aufbereitung des Problemverständnisses und als Rahmen gut geeignet sind. Hingegen dienen sie nicht als Hilfestellung zur Umsetzung von HCAI (Xu et al., 2023). Um die Ziele einer menschenzentrierten KI zu erreichen, braucht es Leitlinien und Methoden für Design, Implementierung, Betrieb, Wartung und Stilllegung sowie die Evaluierung und Verwaltung KI-basierter Services. Gegenwärtig fehlt es außerdem in Unternehmen und Politik an Führungs- und Verwaltungsstrukturen, die die Einführung einer menschenzentrierten Gestaltung von KI unterstützen und deren Umsetzung sicherstellen (Ozmen Garibay et al., 2023).

Shneiderman (2020a) definiert drei Ebenen, die Einfluss auf die Entwicklung reliabler, sicherer und vertrauenswürdiger KI-Systeme haben: 1. Team-Ebene, 2. Organisations-Ebene und 3. Industrie-Ebene. Die Entwicklungsteams sind dafür verantwortlich, reliable Systeme zu entwickeln und bereitzustellen, die die Prinzipien eines menschenzentrierten KI-Designs berücksichtigen. Die Organisation wiederum gibt den Rahmen für die Umsetzung vor, wobei Faktoren wie Führungsverhalten, Organisationskultur und unternehmensinterne Sicherheitsstandards die Entwicklung auf Teamebene beeinflussen. Team- und Organisations-Ebene sind wiederum in industrieübergreifende Strukturen eingebettet, die durch externe Kontrollen, Regulierungen und Standardsetzung die Entwicklung vertrauenswürdiger KI-basierter Services in Organisationen fördern. Alle drei Ebenen können dazu beitragen, dass die Ziele einer auf den Menschen ausgerichteten Gestaltung der KI in der Praxis berücksichtigt werden.

2.2.3 Menschenzentrierte KI: Leitlinien und Design-Prinzipien

Definierte Leitlinien sollen dazu beitragen, dass menschliche Unterschiede und individuelle Bedürfnisse bei der Technikgestaltung berücksichtigt werden, dass KI mit menschlichen Werten vereinbar ist und dass die Sicherheit und die Einflussmöglichkeiten des Menschen gewährleistet sind (Ozmen Garibay et al., 2023, S. 393). Es gibt bereits von verschiedenen Institutionen und Organisationen formulierte Leitlinien, an denen man sich im Entwicklungsprozess orientieren kann. Dabei kann zwischen allgemeingültigen Grundsätzen, die als Orientierung für die Gestaltung dienen, und handlungsorientierten Design-Prinzipien unterschieden werden.

Ein prominentes Beispiel für eine Leitlinie, die **allgemeine Grundsätze** formuliert, ist die „**Ethik-Leitlinie für eine vertrauenswürdige KI**“, die 2019 von der EU veröffentlicht wurde (Hochrangige Expertengruppe für künstliche Intelligenz, 2019). Gemäß der Leitlinie ist eine vertrauenswürdige KI dadurch gekennzeichnet, dass diese rechtmäßig, ethisch und (technisch und sozial) robust entlang des gesamten Lebenszyklus ist. Mit der Leitlinie wird ein Rahmen zur Entwicklung vertrauenswürdiger KI geboten, der insbesondere die Facetten ethische und robuste KI adressiert. In der Leitlinie werden vier ethische Grundsätze einer vertrauenswürdigen KI angegeben: 1. Achtung der menschlichen Autonomie, 2. Schadensverhütung, 3. Fairness und 4. Erklärbarkeit. Diese sollten bei der Entwicklung, der Implementierung und dem Betrieb KI-basierter Services befolgt werden. Basierend auf diesen Grundsätzen werden sieben

Anforderungen an eine vertrauenswürdige KI beschrieben, die bei der Umsetzung dieser unterstützen sollen (vgl. Tabelle 3).

Tabelle 3. Anforderungen an eine vertrauenswürdige KI nach der europäischen Ethik-Leitlinie. Definitionen basierend auf High-Level Expert Group on Artificial Intelligence, 2019.

Anforderung	Definition
Vorrang menschlichen Handelns und menschliche Aufsicht	KI-Systeme sollen Menschen unterstützen, fundierte Entscheidungen zu treffen und ihre Grundrechte zu fördern. Gleichzeitig müssen Kontrollmechanismen, wie Human-in-the-Loop, Human-on-the-Loop oder Human-in-Command, gewährleistet werden.
Technische Robustheit und Sicherheit	KI-Systeme müssen widerstandsfähig, sicher sowie genau, reliabel und reproduzierbar sein. Für den Fall von Ausfällen/Schäden muss es einen Notfallplan geben. So kann sichergestellt werden, dass unbeabsichtigte Schäden minimiert und verhindert werden können.
Schutz der Privatsphäre und Datenqualitätsmanagement	Zusätzlich zur Gewährleistung der Privatsphäre und des Datenschutzes müssen geeignete Mechanismen für die Datenverwaltung geschaffen werden. Diese Mechanismen sollten die Qualität und Integrität der Daten berücksichtigen und einen rechtmäßigen Zugang zu den Daten ermöglichen.
Transparenz	Daten, Systeme und KI-Geschäftsmodelle sollten transparent sein. Rückverfolgbarkeitsmechanismen können dabei helfen. KI-Systeme und ihre Entscheidungen sollten verständlich für betroffene Interessensgruppen erklärt werden. Nutzende müssen sich der Interaktion mit einem KI-System bewusst sein und über dessen Fähigkeiten und Grenzen informiert werden.
Vielfalt, Nichtdiskriminierung und Fairness	Die Vermeidung von unfairen Biases ist wichtig, da sie negative Auswirkungen haben können, z. B. die Ausgrenzung gefährdeter Gruppen und die Verstärkung von Vorurteilen und Diskriminierung. Um Vielfalt zu fördern, sollten KI-Systeme für alle zugänglich sein, unabhängig jeglicher Beeinträchtigungen, und die relevanten Interessengruppen während ihres gesamten Lebenszyklus einbeziehen.
Gesellschaftliches und ökologisches Wohlergehen	KI-Systeme sollten allen Menschen, auch zukünftigen Generationen, zugutekommen. Sie sollten nachhaltig und umweltfreundlich sein und die Umwelt sowie andere Lebewesen berücksichtigen. Die sozialen und gesellschaftlichen Auswirkungen sollten sorgfältig geprüft werden.
Rechenschaftspflicht	Es sollten Mechanismen eingeführt werden, um Verantwortung und Rechenschaftspflicht für KI-Systeme und ihre Ergebnisse sicherzustellen. Die Überprüfbarkeit von Algorithmen, Daten und Entwurfsprozessen spielt dabei eine Schlüsselrolle, insbesondere bei kritischen Anwendungen. Zudem sollte ein angemessener und zugänglicher Rechtsschutz gewährleistet sein.

In der Leitlinie werden außerdem technische und nicht-technische Methoden zur Umsetzung der Anforderungen eingeführt. Diese Ausführungen geben einen ersten Überblick zu generellen Möglichkeiten der Anwendung, eine genaue Anweisung zur Umsetzung in Form detailliert beschriebener Methoden und Vorgehensmodelle bietet die Leitlinie jedoch nicht.

Die exemplarisch beschriebene sowie weitere allgemeine Leitlinien müssen künftig über eine reine Definition und Beschreibung allgemeiner Prinzipien wie Fairness, Sicherheit, Transparenz und Vertrauenswürdigkeit hinauskommen, um tatsächlich Wirkung bei der Umsetzung KI-basierter Services zu entfalten (Shneiderman, 2020a). Zudem repräsentieren allgemeine

Leitlinien nicht die Komplexität der realen Welt und sind nicht kontextspezifisch zugeschnitten (Auernhammer, 2020).

Neben den allgemeingültigen Grundsätzen gibt es auch **handlungsorientierte Design-Prinzipien**, an denen sich Entwicklerinnen und Entwickler bei der Gestaltung menschenzentrierter KI-basierter Services orientieren können. Design-Prinzipien dienen der Systematisierung von Wissen im Kontext eines Sachverhalts mit dem Ziel, die Entwicklung von spezifischen Lösungen zu unterstützen (Hermann et al., 2016).

Von international tätigen Technologiekonzernen und ihren Forschungsabteilungen wurden umfangreiche Arbeiten bereitgestellt, die Gestaltungsprinzipien für die praktische Anwendung anbieten. Hervorzuheben sind hier insbesondere die Arbeiten von Google (Google PAIR, 2019) und Microsoft (Microsoft, o. J.). Das *People + AI Guidebook* des People and AI Research Centers (PAIR) von Google inkludiert Methoden, Best-Practice-Ansätze und Beispiele für menschenzentriertes KI-Design. Es enthält zudem 23 Design Patterns zur Unterstützung bei der Gestaltung von HCAI. Die Patterns orientieren sich an spezifischen Fragen, die bei der Entwicklung KI-basierter Services auftreten (z. B. „How do I get started with human-centered AI?" oder „How do I explain my AI system to users?"). Eine detaillierte Beschreibung aller Design-Prinzipien wird online von Google bereitgestellt. In der folgenden Liste werden einzelne Design-Patterns des Guidebooks exemplarisch aufgeführt (Google PAIR, 2019):

- „Determine if AI adds value."
- „Set the right expectation."
- „Make it safe to explore."
- „Explain for understanding, not completeness."
- „Let users supervise automation."

Auch Microsoft hat in einem Online-Tool (HAX-Toolkit) Design Guidelines und Prinzipien für KI bereitgestellt. Diese fokussieren auf die Mensch-KI-Interaktion sowie ein positives Nutzungserlebnis. Basierend auf einem Literaturreview wurden 18 Prinzipien identifiziert (Amershi et al., 2019), welche um Design Patterns erweitert wurden. Die Guidelines orientieren sich am Ablauf der KI-User-Interaktion vom Erstkontakt, über die Interaktion bis hin zu eventuell auftretenden Fehlern sowie über die Zeit hinweg. Einige Design Guidelines sind folgend exemplarisch aufgelistet (Microsoft, o. J.):

- „Make clear what the system can do."
- „Match relevant social norms."
- „Make clear why the system did what it did."
- „Encourage granular feedback."
- „Learn from user behavior."

Beide Arbeiten können den Entwicklungsprozess aktiv unterstützen. Detaillierte Ausführungen zur Art und Weise der Integration und Berücksichtigung der Gestaltungsprinzipien im Entwicklungsprozess sind jedoch nicht beschrieben. Zukünftig werden Methoden und Vorgehensmodelle benötigt, die die konkrete Anwendung der Prinzipien unterstützen. Darüber hinaus ist eine Evaluierung erforderlich, ob die Anwendung der Prinzipien zu den gewünschten Effekten führt und ob sie tatsächlich einen Beitrag zur Gestaltung einer menschenzentrierten KI leisten. Insgesamt fehlt es aktuell an praktischen Ansätzen und Methoden zur Umsetzung eines menschenzentrierten KI-Designs (Hartikainen et al., 2022), wodurch die Umsetzung von HCAI in der Praxis herausfordernd ist (Bingley et al., 2023; Capel & Brereton, 2023).

2.2.4 Menschenzentrierte KI: Partizipation und Ko-Kreation

Ein weiterer Ansatz zur Unterstützung des Designs menschenzentrierter KI ist die Integration diverser Perspektiven in den Entwicklungsprozess. Im sozialen Diskurs sollen entlang des KI-Lebenszyklus unterschiedliche Interessensträger (z. B. Mitarbeitende, Endanwenderinnen und Endanwender, Ethikexpertinnen und -experten) sowie deren Bedürfnisse und Ziele berücksichtigt werden (Hochrangige Expertengruppe für künstliche Intelligenz, 2019). Solch interdisziplinäre Kollaborationen können die Entwicklung menschenzentrierter KI-basierter Services vorantreiben (Xu et al., 2023). Im Kontext der Dienstleistungsforschung wird diese Einbeziehung unterschiedlicher Stakeholder in den Entwicklungsprozess als Ko-Kreation bezeichnet (Hieber et al., 2023; Russo-Spena & Mele, 2012). Im Kontext der KI-Gestaltung wird in diesem Zusammenhang häufig von partizipativem Design oder partizipativer KI gesprochen (Adler et al., 2022; Berditchevskaia et al., 2021; Birhane et al., 2022; Capel & Brereton, 2023). Partizipatives KI-Design ist gekennzeichnet durch die Einbindung verschiedener Akteure in den Entwicklungsprozess (Berditchevskaia et al., 2021). Es ermöglicht die ko-kreative Gestaltung KI-basierter Services und somit die Integration verschiedener Perspektiven in die Ausgestaltung der Systeme (Auernhammer, 2020). Ziel von Ko-Kreation ist es, die Bedürfnisse und Perspektiven aller beteiligten Akteure zu berücksichtigen, um so Lösungen zu entwickeln, die den Anforderungen aller Beteiligten entsprechen (Hieber et al., 2023). Ein solch partizipatives Vorgehen fördert die Gestaltung vertrauenswürdiger und menschenzentrierter KI (Morley et al., 2020). Entsprechend kann Partizipation als Werkzeug für die Entwicklung von HCAI verstanden werden (Birhane et al., 2022). Berditchevskaia et al. (2021) unterscheiden vier Stufen partizipativer KI: Beratung, Mitwirkung, Zusammenarbeit und Ko-Kreation (vgl. Abbildung 6).

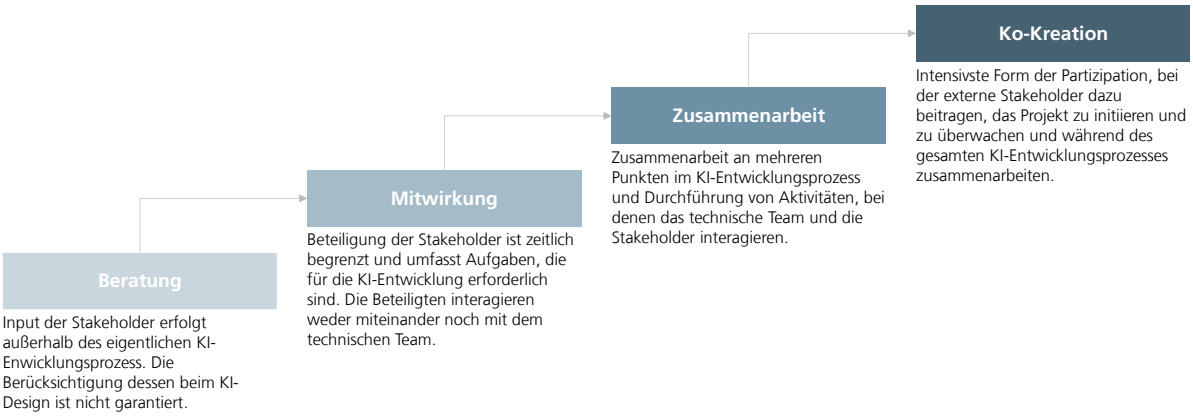


Abbildung 6. Vier Stufen partizipativer KI nach Berditchevskaia et al., 2021.

Die Einbindung diverser Stakeholder in den Gestaltungsprozess KI-basierter Services wird häufig als Erfolgsfaktor für die Gestaltung und Einführung menschenzentrierter KI genannt (Auernhammer, 2020; Berditchevskaia et al., 2021; Plattform Lernende Systeme, 2019; Xu et al., 2023). Insbesondere im Kontext des Einsatzes KI-basierter Services in Arbeitsumgebungen wird die Beteiligung von Mitarbeitenden als akzeptanzförderlicher Faktor hervorgehoben (Abel et al., 2019; Plattform Lernende Systeme, o. J.; Pokorni et al., 2021). Dennoch werden bis dato Endanwenderinnen und Endanwender selten von Beginn an in den Entwicklungsprozess eingebunden (Subramonyam et al., 2022). Generell ist eine Beteiligung der Stakeholder entlang des gesamten KI-Lebenszyklus – von der Ideenfindung, über die eigentliche Entwicklung bis hin zum Betrieb – möglich (Russo-Spena & Mele, 2012). Optimalerweise sollte die Einbindung von Mitarbeitenden in den Gestaltungs- und Einführungsprozess KI-basierter Systeme kontinuierlich und frühzeitig erfolgen, um deren

Erfahrungen und Kenntnisse integrieren zu können (Abel et al., 2019). Darüber hinaus ermöglichen partizipative Designansätze einen Perspektivwechsel von der Evaluation der Leistung eingesetzter Algorithmen hin zur Evaluation der menschlichen Leistungssteigerung bedingt durch KI-basierte Unterstützung und deren Zufriedenheit mit dem eingesetzten KI-Tool (Shneiderman, 2020a).

Für die erfolgreiche Umsetzung partizipativer Entwicklungsprozesse ist es wichtig zu verstehen, wie in der industriellen Praxis technisch-versierte und nicht-technisch-versierte Rollen miteinander kooperieren können (Subramonyam et al., 2022), und zu definieren, welche Stakeholder Teil eines ko-kreativen Entwicklungsansatzes sein sollten (Hieber et al., 2023). Für beide Aspekte bedarf es weiterer Forschung, um HCAI durch Partizipation künftig in der Praxis fördern zu können.

2.2.5 Menschenzentrierte KI: Zusammenfassung der zentralen Herausforderungen

Obwohl in den letzten Jahren die Forschungsarbeiten zu HCAI stetig zugenommen haben, befindet sich das Konzept in der praktischen Umsetzung noch in einem frühen Stadium (Hartikainen et al., 2022). Der folgende Abschnitt fasst die drei wichtigsten Herausforderungen zusammen, die sich aus den vorangegangenen Kapiteln zur menschenzentrierten KI ergeben, insbesondere im Hinblick auf die Gestaltung und Umsetzung in der Praxis.

- Herausforderung 1: Es fehlt an Strukturen in Unternehmen und Politik sowie an Methoden und Vorgehensmodellen, um die tatsächliche Umsetzung von HCAI in der Praxis zu unterstützen.
- Herausforderung 2: Die bestehenden Leitlinien und Prinzipien sind nicht kontextspezifisch und es mangelt an genauen Anleitungen zur praktischen Umsetzung.
- Herausforderung 3: Die interdisziplinäre Gestaltung von KI-basierten Services, insbesondere unter Einbeziehung der Mitarbeitenden, stellt eine Herausforderung für Unternehmen dar.

Diese zentralen Herausforderungen sowie die daraus resultierenden Forschungslücken wurden von verschiedenen Wissenschaftlerinnen und Wissenschaftlern erkannt und als eine zentrale Aufgabe für die weitere Entwicklung genannt (Ahmad et al., 2023; Hartikainen et al., 2022; Mazarakis et al., 2023; Ozmen Garibay et al., 2023). Mit dieser Arbeit soll ein Beitrag geleistet werden, diese Lücken im Kontext industrieller KI-Anwendungen zu schließen.

3 Forschungsfragen und Publikationen

Abgeleitet aus den zentralen Herausforderungen im Kontext industrieller KI-basierter Services (vgl. Kap. 2.1.5) und der Gestaltung menschenzentrierter KI (vgl. Kap. 2.2.5) sowie dem Bedarf von kooperierenden Unternehmen in Forschungsprojekten des Fraunhofer IAO werden in dieser Arbeit Möglichkeiten untersucht, wie der Einsatz industrieller KI-Anwendungen erfolgreich gestaltet werden kann. Der Fokus liegt dabei auf organisationalen und menschenzentrierten Aspekten des KI-Designs, die aufgrund eines primär technologiegetriebenen Entwicklungsprozesses in bisherigen Forschungs- und Entwicklungsarbeiten vernachlässigt wurden. Ausgehend von einer explorativen Analyse bestehender Herausforderungen und Erfolgsfaktoren des KI-Einsatzes werden potenzielle Lösungen identifiziert und Umsetzungskonzepte entwickelt.

Folgende drei Forschungsfragen bilden dabei den Rahmen der vorliegenden Arbeit:

1. Welche Herausforderungen und Erfolgsfaktoren gibt es im Zusammenhang mit der Entwicklung, Einführung und dem Betrieb industrieller KI-basierter Services entlang der Ebenen Mensch, Technik und Organisation?
2. Welche spezifischen Methoden und Modelle können zur Lösung ausgewählter Herausforderungen sowie Stärkung ausgewählter Erfolgsfaktoren in Bezug auf organisations- und menschenzentrierte Aspekte beitragen?
3. Wie können Methoden und Modelle zur Förderung eines menschenzentrierten KI-Designs im industriellen Setting gestaltet sein?

Folgende Abbildung 7 enthält einen Überblick zum Aufbau der Forschungsarbeit sowie den integrierten Publikationen.

Abbildung 7. Überblick zum Aufbau der Forschungsarbeit und der integrierten Publikationen

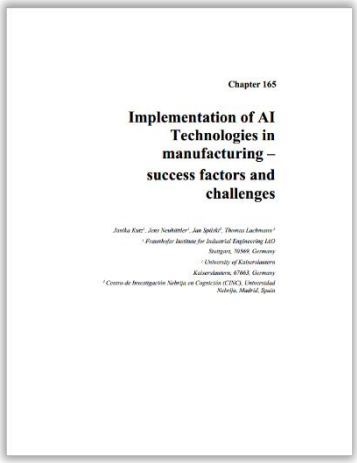
Zur Beantwortung der ersten Fragestellung wurde eine qualitativ-explorative Studie durchgeführt, in der insgesamt 31 Herausforderungen und 20 Erfolgsfaktoren auf den Ebenen Mensch, Technik und Organisation identifiziert wurden (Publikation 1). Aufbauend auf diesen Ergebnissen wurden zur Beantwortung der zweiten Forschungsfrage Methoden und Modelle zur Lösung spezifischer Herausforderungen bzw. zur Stärkung spezifischer Erfolgsfaktoren identifiziert und analysiert. In Zusammenarbeit mit einem Praxispartner und unter Berücksichtigung einschlägiger Literatur wurden zwei zentrale Handlungsstränge festgelegt, die hohes Potenzial zur Verbesserung der betrieblichen Praxis bei der Entwicklung und dem Betrieb von KI-Anwendungen versprechen. Dies sind zum einen die Entwicklung eines

Rollenmodells für die Umsetzung unternehmensinterner KI-Projekte (Publikation 2) und zum anderen die Fokussierung auf das Thema des menschenzentrierten Einsatzes KI-basierter Services im industriellen Umfeld (Publikation 3 und Publikation 4). Zur Beantwortung der dritten Forschungsfrage wurde für beide Themenfelder ein Konzept für die Umsetzung in der industriellen Praxis entwickelt (Publikation 2 und Publikation 5). Letzteres wurde vor der Konzeptentwicklung hinsichtlich der konkreten Anwendbarkeit im industriellen Setting analysiert und eingeordnet.

Die Publikationen sind in der folgenden Liste aufgeführt. Nachfolgend sind die einzelnen Publikationen nach einer jeweiligen Kurzzusammenfassung integriert.

- Publikation 1: Kutz, J., Neuhüttler, J., Spilski, J., & Lachmann, T. (2022). Implementation of AI Technologies in manufacturing - success factors and challenges. In *AHFE International, The Human Side of Service Engineering.* AHFE International. https://doi.org/10.54941/ahfe1002565
- Publikation 2: Kutz, J., Neuhüttler, J., Schaefer, K., Spilski, J., & Lachmann, T. (2023). Generic Role Model for the Systematic Development of Internal AI-based Services in Manufacturing. In T. X. Bui (Chair), *Proceedings of the 56th Annual Hawaii International Conference on System Sciences: January 3-6, 2023*. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/ea841716-73a6-4ea8-9a4e-858e8f498d6d/content
- Publikation 3: Kutz, J., Neuhüttler, J., Spilski, J., & Lachmann, T. (2023). AI-based Services - Design Principles to Meet the Requirements of a Trustworthy AI. In C. Leitner, J. Neuhüttler, C. Bassano, & D. Satterfield (Eds.), *AHFE International, The Human Side of Service Engineering.* AHFE International. https://doi.org/10.54941/ahfe1003107
- Publikation 4: Kutz, J., Neuhüttler, J., Bienzeisler, B., Spilski, J., & Lachmann, T. (2023). Human-Centered AI for Manufacturing – Design Principles for Industrial AI-Based Services. In H. Degen & S. Ntoa (Eds.), *Lecture Notes in Computer Science, Artificial Intelligence in HCI* (pp. 115–130). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35891-3_8
- Publikation 5: Kutz, J., Neuhüttler, Gladilov, N., J., Spilski, J., Hölzle, K., Lachmann, T. (Manuskript zur Vorlage). Developing Human-Centred AI in Industrial Settings: Process model for applying design principles in the co-creative design of human-centred IAI-based services

3.1 Publikation 1: Implementation of AI Technologies in manufacturing – success factors and challenges (Kutz et al., 2022)


Chapter 165

Implementation of AI Technologies in manufacturing – success factors and challenges

Janika Kutz[1], Jens Neuhüttler[2], Jan Spilski[2], Thomas Lachmann[3]
[1] Fraunhofer Institute for Industrial Engineering IAO
Stuttgart, 70569, Germany
[2] University of Kaiserslautern
Kaiserslautern, 67663, Germany
[3] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, Spain

Titel: Implementation of AI Technologies in manufacturing – success factors and challenges

Autoren: Janika Kutz, Jens Neuhüttler, Jan Spilski, Thomas Lachmann

Status: veröffentlicht

Jahr der Veröffentlichung: 2022

Zitation: Kutz, J., Neuhüttler, J., Spilski, J., & Lachmann, T. (2022). Implementation of AI Technologies in manufacturing - success factors and challenges. In AHFE International, The Human Side of Service Engineering. AHFE International. https://doi.org/10.54941/ahfe1002565

Inhalt:

Basierend auf einem qualitativen Forschungsdesign wurden Herausforderungen und Erfolgsfaktoren der Entwicklung, Implementierung und des Betriebs industrieller KI-Anwendungen identifiziert. Die identifizierten Faktoren wurden entlang der Ebenen eines soziotechnischen Systems (Mensch, Technik und Organisation) eingeordnet. Als Erfolgsfaktoren genannt wurden u. a. Partizipation und nutzerzentrierte Gestaltung, Aufbau eines geeigneten Datensatzes sowie eine ganzheitliche Betrachtung von Digitalisierungsprojekten. Beschriebene Herausforderungen sind u. a. die Berücksichtigung aller Stakeholder im Entwicklungsprozess, die Kontextsensitivität von KI-Anwendungen sowie Unklarheiten in Bezug auf Rollen und Verantwortlichkeiten bei KI-Projekten.

Einordnung in den Gesamtzusammenhang der Arbeit:

- Qualitative Analyse der Ausgangssituation, zum Aufbau eines detaillierten Verständnisses der bestehenden Herausforderungen industrieller KI
- Identifikation von Erfolgsfaktoren
- Ausgangsbasis für weitere Forschungsarbeiten

The Human Side of Service Engineering, Vol. 62, 2022, 256–261
https://doi.org/10.54941/ahfe1002565

Implementation of AI Technologies in Manufacturing – Success Factors and Challenges

Janika Kutz[1], Jens Neuhüttler[1], Jan Spilski[2], and Thomas Lachmann[2,3]

[1]Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, 70569, Germany
[2]Universitiy of Kaiserslautern, Center for Cognitive Science, 67336, Germany
[3]Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, 28015, Spain

ABSTRACT

There is a broad consensus on the potential of smart services for production and the added value their use offers. Industrial artificial intelligence (AI) has several advantages. AI technologies, for example, can strengthen resilience, support work processes, increase product quality and thus improve competitiveness. Many compa- nies have recognised these potentials and are developing AI solutions. There are many successful proof-of-concepts (PoC) and pilot projects, but AI technologies successfully implemented in the real environment are scarce. Successful implementation of smart services based on industrial AI in production operations can be understood as its repetitive use and integration into operational business, which is a prerequisite for exploiting the potentials. Currently, little is known about how to achieve successful implementation. In contrast, there is much evidence that the implementation and ope- ration of AI in manufacturing is associated with extensive challenges and barriers. The factors that positively influence the roll-out of AI technologies in manufacturing, how- ever, are little explored. Therefore, this paper focuses on the identification of success factors and barriers for the implementation and operation of AI solutions in manufa- cturing. Furthermore, it is analysed whether and how the identified success factors and barriers differ from each other in order to subsequently derive initial recommen- dations for action. The methodology is based on explorative qualitative research. First, 10 semi-structured interviews were conducted with AI experts from a German Original Equipment Manufacturer (OEM). In an expert workshop, the main findings were validated, and possible solution and support options were discussed. Our findings confirm the results found in the literature and complement them with new insights. Success factors and challenges can be found on the technical, organisational, and human side and relate most often to “data”, “development and operational processes” and “stakeholder engagement”.

Keywords: Industrial AI, Challenges, Success factors, Technology acceptance

INTRODUCTION

Artificial Intelligence (AI) services are meanwhile well known and are increasingly finding their way into everyday life. The potential of these AI

services is also considered high in the industrial context (Bérubé et al. 2021; Lundborg and Gull 2021). This paper focuses on AI services in manufactu- ring, also known as Industrial AI. In this paper Industrial AI is defined a "systematic discipline focusing on the development, validation, deployment and maintenance of AI solutions (in their varied forms) for industrial appli- cations with sustainable performance." (Peres et al. 2020). Further we under- stand Industrial AI services as socio-technical systems as well as tools which assist humans. Industrial AI has several advantages. For example, AI tech- nologies can strengthen resilience, support work processes, increase product quality and thus improve competitiveness. Despite these promising possibi- lities, few applications using Industrial AI are currently implemented in real environment. Many use cases are developed and tested in laboratory environments and rarely get beyond prototypical status (Lundborg and Gull 2021; Bérubé et al. 2021). Although these prototypes operate well, implementation is associated with many challenges. Based on our experience with industrial business partners, we have come to the assumption that the integration of AI into the real working environment is not as straightforward as it could be. An assumption also made by Bérubé et al. (2021). In addition, there needs to be a broad awareness of the challenges associated with implementing AI services, as these have an impact on successful implementation (Bérubé et al. 2021). In Literature, consensus exists that companies face particular challenges with the implementation of AI services. Commonly mentioned challenges in the literature are related to data (e.g., availability and quality of data, data governance) and the lack of necessary competencies. Further challenges mentioned include a lack of top management support and strategic vision of AI, and uncertainty in regard to the business case (e.g. Bérubé et al. 2021; Goasduff 2019; Peres et al. 2020; IDG Research Serivce 2021; Kinkel et al. 2021). Overall, however, there is little empirical research on factors influencing the successful implementation and operation of Industrial AI services. In particular, little attention is paid to success factors. Aim of this study is to identify success factors for the successful implementation and operation of AI Services in manufacturing, as well as to identify the main challenges associated with this. Therefore, the research question of this study is: *What are the success factors and challenges to the implementation of Industrial AI services in manufacturing?*

METHOD

Since there is little research in this field, the methodology is based on explorative qualitative research. First, 10 AI experts of an OEM were interviewed. Five interview partners were assigned to the IT department and five to the manufacturing department. The interview-guide included questions about daily working routines, AI projects in the company, challenges associated with the development, implementation and operation of AI services in manufacturing, and the associated success factors. Each interview was about 45 minutes. The interviews were conducted virtually via Microsoft Teams. In addition to the interviewee and the interviewer, another person attended the interview to transcribe the conversation. Success factors and challenges were

identified based on inductive categorization (Mayring 2000). Afterwards, the factors were assigned to the categories Human, Technology, and Organisation according to the Human-Technology-Organisation (HTO) concept (Ulich 2013). Identified success factors and challenges were validated in an internal expert workshop. Seven AI experts from different disciplines participated in the online workshop. First, the importance or criticalness of the factors was assessed individually in the workshop via an online questionnaire, then the factors were openly discussed, adapted, and further factors were identified. Furthermore, possible solutions and support options were discussed.

RESULTS

Success Factors and Challenges in Relation With the Implementation of AI Services in Manufacturing

Based on the experience of the total of 17 AI experts, 20 success factors and 31 challenges were identified. We used the HTO-concept to classify these factors. Out of the 20 success factors 8 were assigned to the category Human, 3 to Technology and 9 to Organisation (Table 1). Of the identified challenges, 9 were assigned to the category Human, 14 to Technology and 8 to Orga- nisation (Table 2). In the expert workshop, it became clear that depending on the AI service, challenges and success factors can have a different influe- nce on the development and implementation of AI services. Additionally, an interaction between the factors is to be expected.

Some of the listed factors are described in more detail in the following. According to the interviewees, various stakeholders are involved in the ove- rall AI service engineering process. In their opinion, the support of these stakeholders (e.g., end-users, maintenance, works council, management, Human Resources) is essential for a successful implementation. Further competencies in the field of AI, at least a basic understanding of AI in all related departments would be positive.

When setting up a suitable data set, it is important to have both high- quality data and enough data to train the AI models. The AI experts agree: the more data available for the development of the AI model, the more gene- ralizable the algorithm is and the more stable the AI model runs in productive operation. At the organisational level, rapid development cycles were menti- oned as success factor. This refers to short development cycles in which the AI service is put into productive operation as early as possible. Early testing in productive operation improves the product quality and stability of the AI service on the one hand, and on the other hand the added value of the pro- duct can be demonstrated at an early stage, thereby gaining support from stakeholders.

The lack of technology acceptance was a frequently mentioned challenge for development, implementation and operation. According to the AI experts, end-users feel threatened by AI technologies and are afraid of losing their jobs. Furthermore, the implementation leads to change and there is often a lack of willingness to embrace something new. The cooperation between

T

Table 1. Success factors assigned to Human (H), Technology (T) and Organisation (O).

	Success factor
H	Co-determination and participation of end-users in development and implementation
	Confidence in operability of the IT-System
	Managing expectations
	Qualification and competencies in the field of AI among stakeholders
	Support from stakeholders
	Trust in AI
	Usage of demonstrators User-centred development
T	Setting up a suitable data set
	Standardization of hardware, software, and AI-modules Validation of pre-defined metrics before Roll-Out
O	Added value of the AI service must be clear
	Communication strategy
	Holistic view of digitalisation project Data governance
	(Meta)-evaluation of digitalisation projects Open corporate culture
	Rapid development loops
	Synchronization of development and approval processes Using synergies between projects

IT and manufacturing was also described as challenging. For example, it is perceived as difficult to get access to manufacturing staff.

Undefined roles and responsibilities are a huge challenge coming with the introduction of a new technology into a company. It has not been conclusively clarified which stakeholders are to be included in an AI service engineering process and which responsibilities are to be assigned to which stakeholders. Further roles need to be redefined and responsibilities have to be assigned as well. This makes it particularly difficult to consider and involve all relevant stakeholders.

A pure focus on profitability by the company is perceived as a challenge for technological innovation by the AI experts. Currently, a high front-loading of resources for the development is needed, but this is according to the interviews classified differently by controlling and management.

Possible Solutions and Support Options

In the interviews as well the expert workshop possible solutions and support options were discussed. The discussed opportunities usually address more than one challenge or success factor as well across the Human, Technology and Organisation categories. According to the AI experts, promising opportunities for the successful implementation of AI services in future are interdisciplinary development of AI services, democratization of AI and innovative qualification concepts. These could be AI based on-the-job trainings, the use of digital assistance systems or training through the direct involvement

Table 2. Challenges assigned to Human (H), Technology (T) and Organisation (O).

	Challenging factor
H	Concerns and fear related to AI-Services Consider and involve all stakeholders Cooperation between IT and manufacturing Demotivation due to challenges
	Dissemination of experiential knowledge between AI-Developers False expectations of the end-user in the AI service
	Lack of competencies among stakeholders Lack of management-commitment
	Lack of technology acceptance among stakeholders
T	Complexity of AI-Services
	Context sensitivity of AI Data privacy
	Development in laboratory environments
	Development of a suitable architecture
	Identification of applicable algorithms
	Integration into existing IT-Infrastructure Lack of data quality and availability Lack of standardization
	Non-transparency of AI Onsite IT-integration Response time of AI models
	Security concerns related to cloud solutions Unbiased AI
O	Building a business-case for AI services
	Company focusses purely on profitability Ensure productive operation and support
	Established software engineering processes are insufficiently designed for AI
	Time-consuming administrative tasks and processes
	Uncertainty about development and approval processes Undefined roles and responsibilities
	Value of data is not recognized

of end users in development projects. Within a company, cooperation between development projects and a professional knowledge management for AI development should be supported. Further descriptions of best-practice use cases and guidelines for AI service development and implementation processes can be helpful for a successful implementation according to the AI experts. Suggested technical solutions include intensifying platform ecosystems for industrial AI services, promoting standardization and introducing a Machine Learning Operations (MLOps) approach model.

CONCLUSION

Using an explorative qualitative research design this study analysed success factors and challenges related to the implementation of Industrial AI services. Some challenges known from literature were also mentioned by our

T

interview partners, such as data related problems and the lack of competencies. In addition, other challenges were identified. For example, difficulties in cooperation between IT and manufacturing, the development in laboratory environments and undefined roles and responsibilities. Moreover, our study explores success factors of the implementation of Industrial AI services. Success factors mentioned by our interview partners could mainly assigned to the human and organisational side.

It is important for researchers as well as companies to understand the posi- tive and negative factors that influence the introduction of AI services. Only if these are known it is possible to create a successful implementation process. The present results are complementary to the little existing empirical know- ledge in this research area. Nevertheless, further research is needed to gain a better understanding of the influencing factors, e.g., the dependency and relative importance of the factors. Furthermore, some factors are described as success factors as well as a challenge by the interviewees. Further resea- rch is needed to analyse whether these factors are independent or represent poles of one dimension. It must also be considered that the sample of the present study is limited to AI experts. All stakeholders involved in the pro- cess of implementing AI services, especially end-users, should be interviewed in further studies to gain a comprehensive understanding. Generally, there is little experience with the successful implementation of Industrial AI services. Ethnographic research methods could provide further important insights in future.

REFERENCES

Bérubé, Mathieu/Giannelia, Tanya/Vial, Gregory (2021). Barriers to the Implementation of AI in Organizations: Findings from a Delphi Study. In: Tung Bui (Ed.). Proceedings of the 54th Hawaii International Conference on System Sciences, Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences.

Goasduff, Laurence (2019). 3 Barriers to AI Adoption. Available online at https://www.gartner.com/smarterwithgartner/3-barriers-to-ai-adoption.

IDG Research Serivce (2021). Studie Machine Learning 2021.

Kinkel, Steffen/Baumgartner, Marco/Cherubini, Enrica (2021). Prerequisites for the adoption of AI technologies in manufacturing – Evidence from a worldwide sample of manufacturing companies. Technovation, 102375. https://doi.org/10.1016/j.technovation.2021.102375.

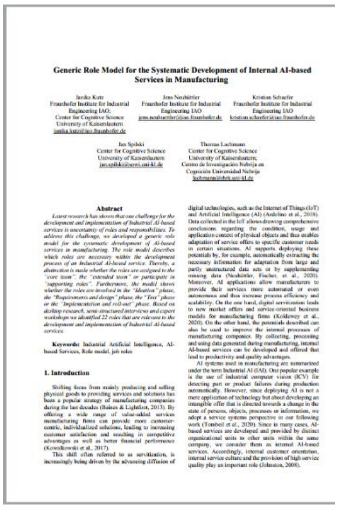
Lundborg, Martin/Gull, Isabell (2021). Künstliche Intelligenz im Mittelstand. So wird KI für kleine und mittlere Unternehmen zum Game Changer. Eine Erhe- bung der Mittelstand-Digital Begleitforschung im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz. wik consult. Bad Honnef. Available online at https://www.mittelstand-digital.de/MD/Redaktion/DE/Publikationen/ki-Studie-2021.pdf?__blob=publicationFile&v=5 (accessed 1/24/2021).

Mayring, Philipp (2000). Qualitative Content Analysis. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [On-line Journal], http://qualitative-research.net/fqs/fqs-e/2-00inhalt-e.htm 1.

Peres, Ricardo Silva/Jia, Xiaodong/Lee, Jay/Sun, Keyi/Colombo, Armando Walter/Barata, Jose (2020). Industrial Artificial Intelligence in Industry 4.0 - Systematic Review, Challenges and Outlook. IEEE Access 8, 220121–220139. https://doi.org/10.1109/ACCESS.2020.3042874.

Ulich, Eberhard (2013). Arbeitssysteme als soziotechnische Systeme–eine Erinnerung. Journal Psychologie des Alltagshandelns 6 (1), 4–12.

3.2 Publikation 2: Generic Role Model for the Systematic Development of Internal AI-based Services in Manufacturing


Generic Role Model for the Systematic Development of Internal AI-based Services in Manufacturing

Titel: Generic Role Model for the Systematic Development of Internal AI-based Services in Manufacturing

Autoren: Janika Kutz, Jens Neuhüttler, Kristian Schaefer, Jan Spilski, Thomas Lachmann

Status: veröffentlicht

Jahr der Veröffentlichung: 2023

Zitation: Kutz, J., Neuhüttler, J., Schaefer, K., Spilski, J., & Lachmann, T. (2023). Generic Role Model for the Systematic Development of Internal AI-based Services in Manufacturing. In T. X. Bui (Chair), Proceedings of the 56th Annual Hawaii International Conference on System Sciences: January 3-6, 2023. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/ea841716-73a6-4ea8-9a4e-858e8f498d6d/content

Inhalt:

Eine zentrale Herausforderung der erfolgreichen Implementierung KI-basierter Services in der Produktion sind Unklarheiten im Zusammenhang mit Rollen und Verantwortlichkeiten. Um diese Herausforderung zu überwinden, wurde basierend auf Erkenntnissen der Literatur sowie Expertinnen- und Expertenmeinungen ein generisches Rollenmodell zur systematischen Entwicklung KI-basierter Services im industriellen Umfeld entwickelt. Insgesamt konnten 22 Rollen identifiziert werden, die von der Ideenphase bis zum Roll-Out an der Entwicklung beteiligt sind. Da die Rollen in unterschiedlicher Intensität an der Entwicklung mitwirken, werden in dem Modell drei Teamebenen unterschieden: das Kernteam, das erweiterte Team sowie unterstützende Rollen.

Einordnung in den Gesamtzusammenhang der Arbeit:

- Transparenz über Rollen und Verantwortlichkeiten sowie deren Grad der Integration bei KI-Entwicklungsprojekten wurde als Möglichkeit, spezifische Herausforderungen zu adressieren, identifiziert
- Konzept eines Rollenmodells zur Anwendung im industriellen Kontext wurde entwickelt

Generic Role Model for the Systematic Development of Internal AI-based Services in Manufacturing

Janika Kutz
Fraunhofer Institute for Industrial Engineering IAO;
Center for Cognitive Science
University of Kaiserslautern
janika.kutz@iao.fraunhofer.de

Jens Neuhüttler
Fraunhofer Institute for Industrial Engineering IAO
jens.neuhuettler@iao.fraunhofer.de

Kristian Schaefer
Fraunhofer Institute for Industrial Engineering IAO
kristian.schaefer@iao.fraunhofer.de

Jan Spilski
Center for Cognitive Science
University of Kaiserslautern
jan.spilski@sowi.uni-kl.de

Thomas Lachmann
Center for Cognitive Science
University of Kaiserslautern; Centro de Investigación Nebrija en Cognición Universidad Nebrija
lachmann@rhrk.uni-kl.de

Abstract

Latest research has shown that one challenge for the development and implementation of Industrial AI-based services is uncertainty of roles and responsibilities. To address this challenge, we developed a generic role model for the systematic development of AI-based services in manufacturing. The role model describes which roles are necessary within the development process of an Industrial AI-based service. Thereby, a distinction is made whether the roles are assigned to the "core team", the "extended team" or participate in "supporting roles". Furthermore, the model shows whether the roles are involved in the "Ideation" phase, the "Requirements and design" phase, the "Test" phase or the "Implementation and roll-out" phase. Based on desktop research, semi-structured interviews and expert workshops we identified 22 roles that are relevant to the development and implementation of Industrial AI-based services.

Keywords: Industrial Artificial Intelligence, AI-based Services, Role model, job roles

1. Introduction

Shifting focus from mainly producing and selling physical goods to providing services and solutions has been a popular strategy of manufacturing companies during the last decades (Baines & Lightfoot, 2013). By offering a wide range of value-added services manufacturing firms can provide more customer-centric, individualized solutions, leading to increasing customer satisfaction and resulting in competitive advantages as well as better financial performance (Kowalkowski et al., 2017).

This shift often referred to as servitization, is increasingly being driven by the advancing diffusion of digital technologies, such as the Internet of Things (IoT) and Artificial Intelligence (AI) (Ardolino et al., 2018). Data collected in the IoT allows drawing comprehensive conclusions regarding the condition, usage and application context of physical objects and thus enables adaptation of service offers to specific customer needs in certain situations. AI supports deploying these potentials by, for example, automatically extracting the necessary information for adaptation from large and partly unstructured data sets or by supplementing missing data (Neuhüttler, Fischer, et al., 2020). Moreover, AI applications allow manufacturers to provide their services more automated or even autonomous and thus increase process efficiency and scalability. On the one hand, digital servitization leads to new market offers and service-oriented business models for manufacturing firms (Koldewey et al., 2020). On the other hand, the potentials described can also be used to improve the internal processes of manufacturing companies. By collecting, processing and using data generated during manufacturing, internal AI-based services can be developed and offered that lead to productivity and quality advantages.

AI systems used in manufacturing are summarized under the term Industrial AI (IAI). One popular example is the use of industrial computer vision (ICV) for detecting part or product failures during production automatically. However, since deploying AI is not a mere application of technology but about developing an intangible offer that is directed towards a change in the state of persons, objects, processes or information, we adopt a service systems perspective in our following work (Tombeil et al., 2020). Since in many cases, AI- based services are developed and provided by distinct organizational units to other units within the same company, we consider them as internal AI-based services. Accordingly, internal customer orientation, internal service culture and the provision of high service quality play an important role (Johnston, 2008).

URI: https://hdl.handle.net/10125/102742
978-0-9981331-6-4

Companies that want to implement their ideas for AI-based services within the company often face challenges. The development of AI-based services is a relatively new and complex field (Lim et al., 2018). Furthermore, current company structures and processes are often not designed for the efficient and quality-oriented development of new, internal services. In many cases, the difficulties start with the fact that the development processes are not clearly defined. For example, there are no clear descriptions of the tasks, the methods to be used and the necessary roles and personnel requirements (Kutz et al., 2022).

A scientific discipline that aims to support the successful development of new services is service engineering. Service engineering provides suitable processes, methods and tools, and thus enables systematic service development in companies. One of these supporting tools are generic role models. These help companies to facilitate staffing of their development teams (DIN 91364:2018). Against this background, this paper aims to support manufacturing companies in the engineering of internal AI-based services by developing a generic role model.

2. Theoretical Background

2.1 Service Engineering

Service engineering can be understood as a technical discipline that deals with the systematic development and design of services using appropriate process models, methods and tools (Bullinger et al., 2003). By providing a dedicated design methodology, the aim is to enable repeatable and thus efficient development of predominantly intangible and integrative offers (Meyer & Zinke, 2018). Particular importance is attached to the design of high service quality in a resource-efficient development process (Schuh et al., 2016).

At the center of service engineering are reference procedure models that contain detailed documentation of development processes, tasks and responsibilities and thus allow planning and monitoring of developing projects. There are different reference models for the development of new services with different focuses, which are similar in the phases and activities described (cf. Kitsios & Kamariotou, 2019; Witell et al., 2017).

In addition to approaches to developing traditional services, there are also reference models that integrate the use of data and digital elements (cf. Frank et al., 2020; Neuhüttler et al., 2020). Typically, these models include the following six generic development activities, which are combined into phases with varying levels of differentiation. Within "*Ideation*", ideas for a new service are collected and evaluated about criteria such as feasibility, economic viability, customer benefits and market potential. If an idea is selected for further pursuit, technical, organizational and personnel requirements are collected among different internal and external stakeholders during "*Requirement Analyses".* Based on the results a detailed service concept is created during the *"Design"* activities, which includes a description of the service including a system architecture as well as a process and resource model. During the following activity "*Testing*", concepts and prototypes developed so far are evaluated. This includes functional and user testing as well as price and cost simulations. Tests are followed by the "*Implementation*", in which concepts and prototypes that passed the tests are implemented, including organizational measures, regulation of responsibilities and the creation of work instructions and training measures. In the final activity *"Roll-out"*, newly developed services are scaled to all relevant areas and success is monitored closely, for example via feedback questionnaires.

Based on the activities presented, traditional approaches to service engineering often introduce role models that deal with the necessary activities and competences for development at the individual actor level (Schymanietz & Jonas, 2020)

2.2 Service Engineering for internal AI-based Services

In the past decades, service engineering has provided a large body of knowledge that is still considered to be useful and valid today (Marx et al., 2020). However, due to the changed nature of services, a need for further development of service engineering concepts and methods was expressed in scientific literature (Böhmann et al., 2018; Hunke & Schüritz, 2019). This applies in particular to the development of AI-based services, which are characterized by intensive use of data through a high degree of automation up to autonomy in decisions and actions in service delivery.

However, calls for adaptation of service engineering concepts refer not only to the new characteristics of AI- based services but also to their increasingly systemic and collaborative development. Due to the complexity of AI-based services and their impact on internal service systems, different actors with specific competences and resources need to be involved in the development (Neuhüttler, Kett, et al., 2020). On the one hand, this includes technical competences for the collection and preparation of the data basis as well as the development of algorithms and system architectures. On the other hand, due to the high degree of autonomy, there is also a need to involve other actors in the process. Concerning the acceptance of AI-based services, this includes the intensive involvement of users, but also actors who deal with security and questions of safety in the working environment. If stakeholders are not involved throughout the development process, problems arise, such as a lack of resources, commitment and customer focus as well as poor processes, which are regarded as central barriers to internal service quality (Johnston, 2008).

To coordinate and synchronize different actors, competences and tasks during interdisciplinary development projects, role models that describe the tasks and responsibilities of each actor become even more important for the successful engineering of internal AI-based services.

2.3 Definition of role and role model

Due to changes in the working environment, the use of roles and role profiles in organizations is becoming increasingly relevant. In contrast to job descriptions, job roles allow more flexibility. This means that one role can be occupied by multiple people, but also one person can occupy multiple roles. (Grote et al., 2020; Schüller & Steffen, 2021). A role can be defined as a specific function a person or organizational unit fulfills (Broy & Kuhrmann, 2013). Following the definition of social psychology a role is associated with behavioral expectations that are directed at a particular position (Maier). Also, professional roles are associated with specific behavioral expectations (Hinz, 2017). A role profile describes the tasks and responsibilities associated with a role, as well as required competences (Grote et al., 2020).

Several related roles can be combined in a role model. According to Grote et al. (2020) "role models include groups of role profiles specific to the company with clear tasks and competences as well as associated responsibilities and authorities" (p. 1).

2.4 Role models in related disciplines

Previous studies show that the relevance of suitable role models is becoming increasingly important. The development and implementation of AI-based services is still a new application field and accordingly associated with new job roles and competences. Through the growing complexity of data-based services an interdisciplinary team is needed for a successful development process (Anke et al., 2020; Tombeil et al., 2021). According to Kutzias et al. (2021) roles and competences are focused on newer data science process models. For example, the EDDA (Engineering Data Driven- Applications) process model by Hesenius et al. (2019) includes "four roles providing the necessary expertise to develop a data-driven application" (p. 38). The roles are Domain Expert, Data Scientist, Data Domain Expert and Software Engineer. Another model is the Team Data Science Process which includes six roles: Solution Architect, Project Manager, Data Engineer, Data Scientist, Application developer and Project Lead (Microsoft, 2020). There are further studies that deal with roles in data science projects. The study by Crisis et al. (2020) provides a good overview. Based on the analysis of 12 studies in the field of data science, they identified 9 roles that are needed for a data science project. In a qualitative study on smart service innovation, 17 roles were identified. The authors separate Primary roles and Secondary roles in smart service engineering (SSE) projects. Primary roles are (1) project sponsor, (2) digital innovator, (3) system integrator and (4) service operator (Anke et al., 2020). Besides the establishment of new roles, new approaches to forming project teams are required.

2.4 Summary and research questions

Internal AI-based services represent a new type of development object to which existing concepts and methods of service engineering must adapt. This also includes staffing the development projects with suitable people who have the required skills and competences. Especially in manufacturing, staffing a project team for internal AI-service development is challenging. Development projects often do not take place within the boundaries of one organizational unit, but between different units within a manufacturing company. Besides experts from the IT department, also experts from the production department are needed for successful development and implementation (Kutz et al., 2022). We were not able to find a role model that holistically addresses the development of internal IAI- based services and takes into account all necessary development objects and their specific characteristics (cf. section 2.2). To overcome this challenge the aim of our study is to develop a generic role model for the systematic development of internal AI-based services in manufacturing. Therefore, we address the following research questions:

- RQ1: What job roles are needed to successfully develop and implement internal IAI-based services?
- RQ2: What tasks are assumed by the roles in the internal IAI-based service development process?
- RQ3: During which process phases are the roles involved?

3. Research Design

Our study aims to develop a generic role model for the development of IAI-based services. The model should cover the entire service engineering process, from ideation to roll-out. A 3-stage research design was chosen to develop the role model (cf. table 1).

Table 1. Overview research design

Research Step	Method	Sample
1. Development	Interview	10 AI Expert of a European automotive OEM (Company A)
2. Evaluation	Workshop	1 Product Owner, 1 Project Manager of a European automotive OEM (Company A)

3.	Validation	Workshop	2 AI Experts of a German tool manufacturing company (Company B)

Step 1: Development of the role model: First, we conducted semi-structured interviews with 10 AI experts of a European Original Equipment Manufacturer (OEM) from the automotive sector. The interview guideline included questions about the interviewees' roles in the development of AI-based services, related tasks and challenges. Furthermore, questions were asked about additional roles respectively stakeholders which are currently involved or should be involved in future development projects. One interview was about 45 minutes and was conducted virtually via Microsoft Teams. Based on these results, the first version of the role model, consisting of 25 roles, was developed.

Step 2: Evaluation of the role model: During the second step we evaluated the role model in two workshops at the same OEM: In the first workshop the role model was discussed and reviewed with a Product Owner of an IAI-based service in the field of ICV. In the second workshop, the role model was discussed and reviewed with a Project Manager of an IAI-based service in the field of Anomaly Detection. The workshops were conducted virtually and lasted about 60 minutes. We used an online whiteboard for interactive collaboration. In particular, role titles and tasks as well as the allocation to the development phases and levels of the role model were discussed. Reflecting on these results, a second version of the role model was developed.

Step 3: Validation of the role model: Afterwards, the role model was generalized considering the literature and empirical findings. This third version, consisting of 23 roles, was then validated with two AI experts from a German tool manufacturing company. This workshop was conducted virtually along an online whiteboard. The workshop lasted about 60 minutes. Taking these workshop results into account, the final version presented in our paper was created. Based on the results of the workshop, one role was removed, so that the final version contains 22 roles.

4. Findings

4.1 Role model for the development of internal IAI- based services

The role model presented in figure 1 describes which roles are necessary within an internal IAI-based service engineering process. A total of 22 roles could be identified, which were assigned to three levels:

- Core team (5 roles)
- Extended team (8 roles)
- Supporting roles (9 roles)

The roles assigned to the core team are involved throughout the whole development process of an IAI-based service. Roles assigned to the extended team are only temporarily involved to fulfil certain tasks or to contribute their experiences and requirements. Together, these roles form the project team. Further, nine roles are categorized as supporting roles. These roles are only indirectly involved in the development of an IAI-based service. They assume control or advisory function and offer support when necessary. The role model also shows which role should be involved in which process phase of the development of IAI-based services. Process phases used within the role model are based on the process models of service engineering and the development tasks contained therein (Frank et al., 2020; Neuhüttler et al., 2019). Since this study focuses on roles rather than process steps, we have summarized the original six activities into four phases to reduce complexity. In the "Ideation" phase, ideas for solutions to existing challenges are sought and evaluated in terms of technical feasibility, economic viability, customer benefits and market potential (Waidelich et al., 2018). Consequently, it is important to involve customers and end-users from the extended team as well as management in the idea evaluation, in addition to the technical roles, such as software and AI engineers, and omain experts. Once an idea is selected for further pursuit technical, organizational and personnel requirements are collected and defined during the phase "Requirement and design". All stakeholders concerned should be involved in the requirement analysis. In addition to the roles already mentioned, the requirements of maintainers and other important supporting actors, such as IT security, works council and change management, are included and taken into account. Based on the requirements, a concept for the internal IAI-based service is developed and iteratively transferred into functional prototypes. This means the entire service process is described and specified, including a system architecture as well as a process and resource model. Accordingly, additional DevOps experts as well as data engineers and data scientists must be involved during the requirements and design phase. Subsequently, concepts and prototypes developed so far are evaluated. This third phase "Test" includes different functional and validation tests, which require the involvement of the different requirement groups. In the final phase "Implementation and roll-out concepts and prototypes that passed the tests are implemented. This includes the integration of the AI system in the production environment as well as organizational measures, regulation of responsibilities and the creation of work instructions and training measures. After further functional tests in the production environment, the service is then put into operation. At this point, specialists for deployment should be involved in the

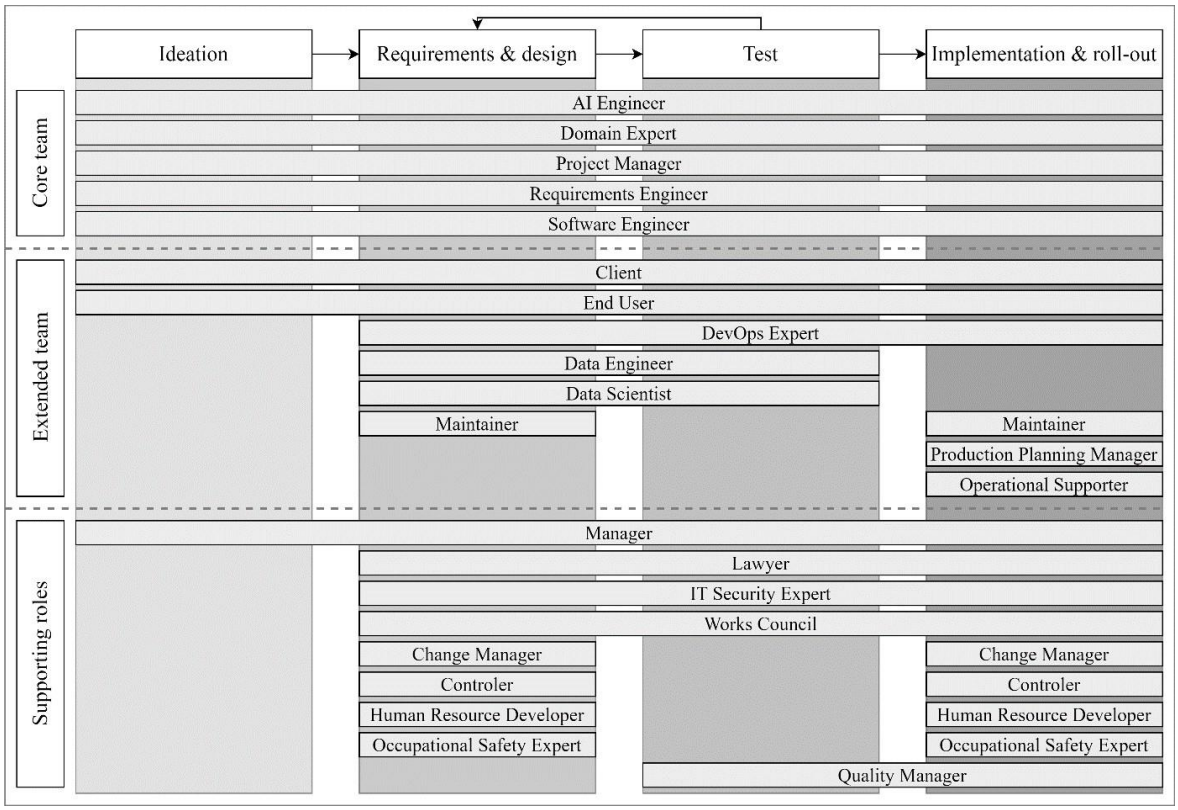


Figure 1. Role model for the systematic development of internal Industrial AI-based services

form of the roles "maintainer", "production planning manager" as well as "operational supporters". As the deployment can lead to changes in existing organizational processes and the need for competences, the roles of "change manager", "controller", "HR developer" and "security experts" should also be included at this point. As already described, the roles are partly involved in the entire development process and partly only in certain development phases. Depending on their functions and tasks, the assignment of roles to the individual process phases was worked out in the workshops (Step 2 and Step 3). Roles of the core team are actively involved in all phases, as they take on active and guiding tasks throughout the whole process. Furthermore, they must be able to keep an overview of the overall process and should be present at all project and status meetings. However, the roles of the extended team do not need to be present at all project meetings, and they only undertake certain tasks. This will be illustrated in the following using the example of end users. In the ideation phase, they can contribute ideas for new AI use cases that have arisen from practical experience. In the second phase, they contribute requirements that relate in particular to the work process, and in the third phase, they test an IAI-based service for its usability. In the last phase, "Implementation and Roll-Out", they have to be involved to acquire the necessary competences for the successful use of the new application.

4.2 Roles descriptions

The following sub-section describes the roles of the core team (table 1) and the extended team (table 2) in more detail. Role descriptions are based on the interviews and workshops, taking into account role descriptions known from literature (Anke et al., 2020; Crisan et al., 2020; Hesenius et al., 2019; Microsoft, 2020). We will refrain from a detailed description of the supporting roles at this point, as their tasks and responsibilities are quite familiar. However, this does not mean that the inclusion of these roles in the development process is less important. According to the interviewees, it is recommended to inform these roles in time about the project progress or to involve them at an early stage, as their areas of responsibility influence the successful implementation of IAI-based services.

Table 2. Role descriptions core team

Role	Description
AI Engineer	Supports the assessment of technical feasibility. Develops, adapts, trains, and validates AI models and provides AI systems for the production environment.

Domain Expert	Contributes expertise from the relevant department to the development of IAI-based services. Is responsible for the implementation on the shop floor.
Project Manager	Leads the project from ideation to roll-out. Is responsible for the project organization (e.g., planning of deadlines, documentation of project progress, release processes), project quality and available resources. Reports to the management.
Requirements Engineer	Maintains an overview of the stakeholders to be involved (e.g., production, maintenance, management) and compiles their requirements for the IAI-based service. Introduces requirements into the development process and examines whether these have been fulfilled.
Software Engineer	Develops software solutions that are necessary for the operation of an IAI-based service, such as platforms and user interfaces. Is also responsible for the connection to existing IT systems.

Table 3. Role descriptions extended team

Role	Description
Client	Requests an AI-based service to improve the production process and therefore orders the development of an IAI-based service. Defines requirements for the IAI-based service and checks whether these have been fulfilled.
DevOps Expert	Is responsible for the operation, architecture, and deployment of the required IT systems.
End User	Uses the IAI-based services in daily operations once the service is implemented. Actively supports the development process by contributing information about work processes as well as usability requirements.
Data Engineer	Collects all necessary data, prepares them, and makes data available for the following analysis. Thereby interfaces with all relevant systems are provided or created.
Data Scientist	Identifies and analyses data. Selects or develops models for data analysis. Must implement the requirements of the department and extract findings from the data.
Maintainer	Is responsible for maintenance, repairs and the operational capability of the machinery and equipment. Approves the sensors and systems used within the IAI-based service when the service is implemented.
Production Planning Manager	Reorganizes production processes to suit the IAI-based service. Enables integration into the production and workflow.
Operational Supporter	Supports users using the IAI-based service and provides advice if problems or errors occur. Needs to gain practical knowledge during the last two phases to support users after implementation.

4.3 Application of the model

The presented role model can be understood as a tool to support the systematic development of IAI-based services in manufacturing companies. Once a company decides to develop such services for internal application, the role model provides information about who should be involved in the development process. According to the study participants, a major added value of the role model developed here is the interdisciplinary team composition. Nevertheless, the cooperation between IT and manufacturing was also described as a challenge by the interviewees. For example, the AI experts explained that it is often difficult for them to gain access to production stuff. The role model presented here can help overcome this challenge by involving people from all relevant areas in development from the very beginning. At the beginning of a development process, it should be determined which person or respectively persons occupy one or more roles. The model and the detailed description of the roles help to communicate the requirements and necessary tasks for each person involved clearly.
Furthermore, it is possible to work with external partners if not all roles can be filled by internal employees. If roles cannot be filled internally, transparency is created about missing competences in the company. It is possible to train employees accordingly or to hire new people with the appropriate competences.

In addition, during the development process, the role model can be used by the core team to ensure that all relevant stakeholders are involved. Parallel roles and the tasks, responsibilities and competences associated with them should be described in more detail.

5. Discussion and Conclusion

Industrial AI-based services have several advantages for manufacturing companies. For example, they strengthen resilience, improve work processes, and increase product quality. To reach their full potential, the successful development and implementation of IAI- based services are crucial. Service engineering provides guidance to do so, in form of formalized process models, methods and tools. However, traditional service engineering approaches do not take into account specific features and circumstances of developing internal AI-based services (cf. 2.2).

Among other things, companies need support in defining roles to be included in an internal IAI-based service development process. No role model tailored to this field of application could be found in the literature. To support companies, our paper presents a generic role model developed based on qualitative research. Some of the roles known from the literature (Hesenius et al., 2019; Microsoft, 2020) for the implementation of data-driven services can also be found in our model (e.g., data engineer, data scientist or project manager). What is unique about our model is the extensive addition of roles from other disciplines. Roles of the production department, such as domain expert, end user and maintenance, have been integrated. Other important business areas are integrated through supporting roles such as human resources manager, change manager or the works council. As described by Anke et al. (2020) and Tombeil et al. (2021) before, our role model confirms that interdisciplinary teams, in our case, mainly consisting of roles from the areas of the production department and IT department, are needed for a successful development process. According to the interviewees, however, this is one of the biggest challenges that must be overcome. Measures to strengthen interdisciplinary collaboration should be addressed in future research.

The role model presented can support the development of internal IAI-based services in companies in various ways. First, it can serve as a basis for staffing internal development projects with the necessary people. Based on the role descriptions, suitable persons can be identified. In addition, they can be used to identify competence gaps from which personnel development measures can be derived.

The role model can also make an important contribution to the communication of expectations and tasks to the persons involved in development projects. The systematic preparation and presentation of activities and required components can improve transparency within the teams.

In addition, the representation of the required roles prevents necessary actors from being left out of the development process. Although not every role needs to be involved to the same extent for every project, a systematic selection helps to ensure systematic development and to make decisions transparent and traceable.

As our model is mainly based on the experiences of AI experts working in the manufacturing industry, we expect it to have high usability for practitioners. Furthermore, this role model can also be applied by human resource management and organizational development for an "AI-oriented personnel and resource management" (Ganz et al., 2021, p. 46). It should also be noted that, depending on the size of the development project, it may be useful to divide some of the defined roles into sub-roles. Moreover, it is possible to adapt the model to the company's internal requirements, for example by adding company-specific roles. However, a more detailed differentiation would have gone beyond the purpose addressed in this paper.

The results presented are subject to limitations and need to be further validated, because the generic role model was developed based on the expertise of AI experts from two companies from the automotive and manufacturing sectors. To validate our results, a larger number of companies with different types of AI-based services should be considered.

Validation should be conducted with the involvement of employees occupying one of the defined roles working in additional companies or other service development projects. In addition, the role model should be piloted in development projects and evaluated.

It also requires further research to analyze whether the role model is transferable to other industries, or the development of AI-based services manufacturers are providing to their customers.

6. Reference

Anke, J., Poeppelbuss, J., & Alt, R. (2020). It Takes More than Two to Tango: Identifying Roles and Patterns in Multi-Actor Smart Service Innovation. *Schmalenbach Business Review*, *72*(4), 599–634.

Ardolino, M., Rapaccini, M., Saccani, N., Gaiardelli, P., Crespi, G., & Ruggeri, C. (2018). The role of digital technologies for the service transformation of industrial companies. *International Journal of Production Research*, *56*(6), 2116–2132.

Baines, T., & W. Lightfoot, H. (2013). Servitization of the manufacturing firm. *International Journal of Operations & Production Management*, *34*(1), 2–35.

Böhmann, T., Leimeister, J. M., & Möslein, K. (2018). The New Fontiers of Service Systems Engineering. *Business & Information Systems Engineering*, *60*(5), 373–375.

Broy, M., & Kuhrmann, M. (2013). *Projektorganisation und Management im Software Engineering*. *Xpert.press*. Springer Berlin Heidelberg.

Bullinger, H.-J., Fähnrich, K.-P., & Meiren, T [Thomas] (2003). Service engineering—methodical development of new service products. *International Journal of Production Economics*, *85*(3), 275–287.

Crisan, A., Fiore-Gartland, B., & Tory, M. (2020). Passing the Data Baton: A Retrospective Analysis on Data Science Work and Workers. In *2020 Visualization in Data Science (VDS)*.

Deutsches Institut für Normung e.V. (2018). *91364:2018 Leitfaden für die Entwicklung neuer Dienstleistungen zur Elektromobilität.* Beuth Verlag GmbH.

Frank, M., Gausemeier, J., Hennig-Cardinal von Widdern, N., Koldewey, C., Menzefricke, J. S., & Reinhold, J. (2020). A reference process for the Smart Service business: development and practical implications. In Proceedings of the ISPIM connects. International Society for Professional Innovation Management (ISPIM).

Ganz, W., Friedrich, M., Hornung, T., Schneider, B., & Tombeil, A.-S. (2021). *Arbeiten mit Künstlicher Intelligenz: Fallbeispiele aus Produktion, Sacharbeit und Dienstleistungen*. Fraunhofer IAO.

Grote, E.-M., Pfeifer, S. A., Roltgen, D., Kuhn, A., & Dumitrescu, R. (2020). Towards Defining Role Models in Advanced Systems Engineering. In *2020 IEEE International Symposium on Systems Engineering (ISSE)* 1–7. IEEE.

Hesenius, M., Schwenzfeier, N., Meyer, O., Koop, W., & Gruhn, V. (2019). Towards a Software Engineering Process for Developing Data-Driven Applications. In *2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE)* (pp. 35–41). IEEE.

Hinz, O. (2017). *Segeln auf Sicht: Das Führungshandbuch für ungewisse Zeiten* (1. Auflage 2017). Springer Fachmedien Wiesbaden.

Hunke, F., & Schüritz, R. (2019). Smartere Produkte durch analysebasierte Dienstleistungen – Ein methodisches Werkzeug zur strukturierten Entwicklung. *HMD Praxis Der Wirtschaftsinformatik*, *56*(3), 514–529.

Johnston, R. (2008). Internal service – barriers, flows and assessment. *International Journal of Service Industry Management*, *19*(2), 210–231.

Kitsios, F., & Kamariotou, M. (2020). Mapping new service development: a review and synthesis of the literature. The Service Industries Journal, 40(9-10), 682–704. https://doi.org/10.1080/02642069.2018.1561876

Koldewey, C., Meyer, M., Stockbrügger, P., Dumitrescu, R., & Gausemeier, J. (2020). Framework and Functionality Patterns for Smart Service Innovation. *Procedia CIRP*, *91*, 851–857.

Kowalkowski, C., Gebauer, H., Kamp, B., & Parry, G. (2017). Servitization and deservitization: Overview, concepts, and definitions. *Industrial Marketing Management*, *60*, 4–10.

Kutz, J., Neuhüttler, J., Spilski, J., & Lachmann, T. (2022). Implementation of AI Technologies in manufacturing - success factors and challenges. In C. Leitner, W. Ganz, C. Bassano, & D. Satterfield (Eds.), AHFE International, The Human Side of Service Engineering. AHFE International. https://doi.org/10.54941/ahfe1002565

Lim, C., Kim, M.-J., Kim, K.-H., Kim, K.-J., & Maglio, P. P. (2018). Using data to advance service: managerial issues and theoretical implications from action research. *Journal of Service Theory and Practice*, *28*(1), 99–128.

Maier, G. *Rolle.* Springer Fachmedien Wiesbaden GmbH.

Marx, E., Pauli, T., Matzner, M., & Fielt, E. (2020). From Services to Smart Services: Can Service Engineering Methods get Smarter as well? In N. Gronau, M. Heine, K. Poustcchi, & H. Krasnova (Eds.), *WI2020 Zentrale Tracks* (pp. 1067–1083). GITO Verlag.

Meyer, K., & Zinke, C. (2018). Service Engineering – eine Standortbestimmung. In K. Meyer, S. Klingner, & C. Zinke (Eds.), *Service Engineering.* Pp. 3–17. Springer Fachmedien Wiesbaden.

Microsoft. (2020). *Team Data Science Process Documentation: What is the Team Data Science Process?*

Neuhüttler, J., Fischer, R., Ganz, W., & Urmetzer, F. (2020). Perceived Quality of Artificial Intelligence in Smart Service Systems: A Structured Approach. In M. Shepperd, F. et al. (Eds.), *Communications in Computer and Information Science. Quality of Information and Communications Technology* Vol. 1266, pp. 3–16). Springer International Publishing.

Neuhüttler, J, Ganz, W., & Spath, D. (2019). An Integrative Quality Framework for Developing Industrial Smart Services. *Service Science*, *11*(3), 157–171.

Neuhüttler, J, Kett, H., Frings, S., Falkner, J., Ganz, W., & Urmetzer, F. (2020). Artificial Intelligence as Driver for Business Model Innovation in Smart Service Systems. In J. Spohrer & C. Leitner (Eds.), *Advances in Intelligent Systems and Computing. Advances in the Human Side of Service Engineering* (Vol. 1208, pp. 212–219). Springer International Publishing.

Schymanietz, M., & Jonas, J. M. (2020). The Roles of Individual Actors in Data-driven Service Innovation – A Dynamic Capabilities Perspective to Explore its Microfoundations. In T. Bui (Ed.), Proceedings of the 53rd Annual Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2020.142

Schuh, G., Gudergan, G., & Kampker, A. (2016). *Management industrieller Dienstleistungen*. *VDI-Buch*. Springer.

Schüller, A. M., & Steffen, A. T. (2021). *Die Orbit-Organisation: In 9 Schritten zum Unternehmensmodell für die digitale Zukunft* (3., aktualisierte Auflage). GABAL.

Tombeil, A.-S., Dukino, C., Zaiser, H., & Ganz, W. (2021). *KI-Ambition als Treiber für die Realisierung von Digitalisierung: Wann ist weniger mehr? Automatisierung und Unterstützung in der Sachbearbeitung mit Künstlicher Intelligenz: Vol. 8*. Fraunhofer Verlag.

Tombeil, A.-S., Kremer, D., Neuhüttler, J, Dukino, C., & Ganz, W. (2020). Potenziale von Künstlicher Intelligenz in der Dienstleistungsarbeit. In M. Bruhn & K. Hadwich (Eds.), *Automatisierung und Personalisierung von Dienstleistungen.* 135–154) Springer Fachmedien Wiesbaden.

Waidelich, L., Richter, A., Kölmel, B., & Bulander, R. (2018). Design thinking process model review. In *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. Symposium conducted at the meeting of IEEE.

Witell, L., Gebauer, H., Jaakkola, E., Hammedi, W., Patricio, L., & Perks, H. (2017). A bricolage perspective on service innovation. Journal of Business Research, 79, 290–298. https://doi.org/10.1016/j.jbusres.2017.03.021

3.3 Publikation 3: AI-based Services – Design Principles to Meet the Requirements of a Trustworthy AI

AHFE 2023 Vol. XX, 2023 AHFE
doi: 10.54941/ahfeXXXX

AI-based Services - Design Principles to Meet the Requirements of a Trustworthy AI

Janika Kutz[1,2], Jens Neuhüttler[1], Jan Spilski[2] and Thomas Lachmann[2,3]

[1] Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, 70569, Germany
[2] Center for Cognitive Science, University of Kaiserslautern-Landau, Kaiserslautern, 67663, Germany
[3] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, Spain

ABSTRACT

The development of Human-Centered and Trustworthy AI-based services has recently attracted increased attention in politics and science. Even though that technical advances have received many of the attention lately, ethical considerations are becoming more and more important. One of the most valuable publications in this area is the "Ethics Guidelines for Trustworthy AI" of the European Commission (EC). One approach to assist developers in implementing these requirements during the development process is to provide design guidelines. The aim of this paper is to identify which action-oriented design principles can be applied to satisfy the requirements for Trustworthy AI. For this purpose, the design principles published by Major providers of commercial AI-based services were contrasted with the seven requirements of the EC. The results indicate that some design principles can be used to meet the requirements of Trustworthy AI. At the same time, however, it becomes clear that work on Ethical AI should be extended by aspects related to Human-AI Interaction and service process quality.

Keywords: AI-based services, Human-Centered AI, Trustworthy AI, Design Principles

INTRODUCTION

Artificial Intelligence (AI)-based services are increasingly used in both private and professional life. Their use offers many opportunities and benefits, but they can also cause harm (Xu and Dainoff 2021). Examples of this are stored in the *AI Incident Database* (McGregor 2021). The reasons for such failures can be biased data as well as complex and non-transparent AI systems (Kaur et al. 2023). As a result, the development and operation of AI-based services are associated with challenges and concerns. Two often mentioned challenges are a lack of technology acceptance and a lack of trust in AI-based services (Kaur et al. 2023; Kutz et al. 2022). Creating Ethically and Trustworthy AI as well as Human-Centered AI (HCAI) has therefore recently received more attention from academia and politics (Xu 2019). Globally, politicians are considering how to address the challenges caused by the advancement of AI. One of the most valuable publications in this area is the "Ethics Guidelines for Trustworthy AI" of the European Commission (EC), which defines seven requirements for Trustworthy AI (High-Level Expert Group on Artificial Intelligence 2019a). One approach to assist developers in implementing these requirements during the development process is to provide

Titel: AI-based Services – Design Principles to Meet the Requirements of a Trustworthy AI

Autoren: Janika Kutz, Jens Neuhüttler, Jan Spilski, Thomas Lachmann

Status: veröffentlicht

Jahr der Veröffentlichung: 2023

Zitation: Kutz, J., Neuhüttler, J., Spilski, J., & Lachmann, T. (2023). AI-based Services - Design Principles to Meet the Requirements of a Trustworthy AI. In C. Leitner, J. Neuhüttler, C. Bassano, & D. Satterfield (Eds.), AHFE International, The Human Side of Service Engineering. AHFE International. https://doi.org/10.54941/ahfe1003107

Inhalt:

Die Entwicklung vertrauenswürdiger und menschenzentrierter KI-Anwendungen rückt zunehmend in den Fokus der wissenschaftlichen und medialen Aufmerksamkeit. Eine zentrale Publikation in diesem Zusammenhang ist die „Ethik-Leitlinie für eine Vertrauenswürdige KI" der Europäischen Kommission, die sieben Anforderungen an vertrauenswürdige KI-Anwendungen definiert. Die Guideline gibt einen Rahmen vor, jedoch fehlen konkrete Hinweise zur Umsetzung der Anforderungen in der Praxis. Eine Möglichkeit dies zu unterstützen, ist die Anwendung von Design-Prinzipien im Entwicklungsprozess. Untersucht wurde, ob bereits publizierte Design-Prinzipien führender Technologiekonzerne angewandt werden können, um den Anforderungen einer vertrauenswürdigen KI seitens der EU gerecht zu werden. Es zeigt sich, dass einige der Anforderungen durch die Anwendung der Design-Prinzipien erfüllt werden können. Dennoch müssen über die Design-Prinzipien hinaus weitere Aspekte berücksichtigt werden.

Einordnung in den Gesamtzusammenhang der Arbeit:

- Annäherung an das Thema der menschenzentrierten KI, als mögliche Handlungsoption zur Adressierung von Herausforderungen und Erfolgsfaktoren
- Kontextunabhängige Analyse des Themas der menschenzentrierten KI-Anwendungen unter Berücksichtigung der europäischen Anforderungen
- Analyse des allgemeinen Einsatzpotenzials publizierter Design-Prinzipien

AHFE 2023: Vol. XX, 2023
doi: 10.54941/ahfeXXXX

AI-based Services - Design Principles to Meet the Requirements of a Trustworthy AI

Janika Kutz[1,2], Jens Neuhüttler[1], Jan Spilski[2] and Thomas Lachmann[2,3]

[1] Fraunhofer Institute for Industrial Engineering IAO; Stuttgart, 70569, Germany

[2] Center for Cognitive Science, University of Kaiserslautern-Landau; Kaiserslautern, 67663, Germany

[3] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija; Madrid, Spain

ABSTRACT

The development of Human-Centered and Trustworthy AI-based services has recently attracted increased attention in politics and science. Even though that technical advances have received many of the attention lately, ethical considerations are becoming more and more important. One of the most valuable publications in this area is the "Ethics Guidelines for Trustworthy AI" of the European Commission (EC). One approach to assist developers in implementing these requirements during the development process is to provide design guidelines. The aim of this paper is to identify which action-oriented design principles can be applied to satisfy the requirements for Trustworthy AI. For this purpose, the design principles published by Major providers of commercial AI-based services were contrasted with the seven requirements of the EC. The results indicate that some design principles can be used to meet the requirements of Trustworthy AI. At the same time, however, it becomes clear that work on Ethical AI should be extended by aspects related to Human-AI Interaction and service process quality.

Keywords: AI-based services, Human-Centered AI, Trustworthy AI, Design Principles

INTRODUCTION

Artificial Intelligence (AI)-based services are increasingly used in both private and professional life. Their use offers many opportunities and benefits, but they can also cause harm (Xu and Dainoff 2021). Examples of this are stored in the *AI Incident Database* (McGregor 2021). The reasons for such failures can be biased data as well as complex and non-transparent AI systems (Kaur et al. 2023). As a result, the development and operation of AI-based services are associated with challenges and concerns. Two often mentioned challenges are a lack of technology acceptance and a lack of trust in AI-based services (Kaur et al. 2023; Kutz et al. 2022). Creating Ethically and Trustworthy AI as well as Human-Centered AI (HCAI) has therefore recently received more attention from academia and politics (Xu 2019). Globally, politicians are considering how to address the challenges caused by the advancement of AI. One of the most valuable publications in this area is the "Ethics Guidelines for Trustworthy AI" of the European Commission (EC), which defines seven requirements for Trustworthy AI (High-Level Expert Group on Artificial Intelligence 2019a). One approach to assist developers in implementing these requirements during the development process is to provide

action-oriented design principles. The aim of this paper is to identify which practical design principles can be applied to satisfy the seven requirements for Trustworthy AI.

ETHICAL AND TRUSTWORTHY AI

There are an unmanageable number of guidelines and papers relevant to designing Ethical and Trustworthy AI-based systems, making it difficult for developers and researchers to draw the right conclusions. The best strategy is to limit the focus to review papers that already provide a systematic evaluation and summary of existing work (Hagendorff 2020; Jobin et al. 2019). Jobin et al., for instance, analyzed 84 ethics guidelines for AI and identified five ethical principles that are globally included (transparency, justice and fairness, non-maleficence, responsibility, and privacy). One of the most comprehensive works on the subject is provided by the EC. In 2019, the High-Level Expert Group on Artificial Intelligence published the "Ethics Guidelines for Trustworthy AI". This guideline's aim is to encourage the development of Trustworthy AI in a human-centered approach. To fulfil the four ethical principles (respect for human autonomy, prevention of harm, fairness, explainability), seven key requirements are defined (see Table 1). Nevertheless, the guidelines are set at a high level and therefore serve more as guidance and less as actual assistance for designing trustworthy AI-based services.

Table 1. Ethics Guidelines for Trustworthy AI - Key Requirements (High-Level Expert Group on Artificial Intelligence 2019b)

Requirement	Description
Human agency and oversight	"AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches"[1]
Technical robustness and safety	"AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented." [1]
Privacy and data governance	"Privacy and data governance: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data." [1]
Transparency	"Transparency: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations." [1]

Diversity, non-discrimination and fairness	"Unfair bias must be avoided, as it could could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle." [1]
Environmental and societal well-being	"AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered." [1]
Accountability	"Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured." [1]

Note. 1 = (High-Level Expert Group on Artificial Intelligence 2019a)

Since the EU is a leading entity in the field of Trustworthy AI and a first regulatory framework is expected with the EU-AI Act (European Commission 2021), we focus in this paper on the requirements formulated by the EC for Trustworthy AI.

Also considering the requirements of the EC is the AI test-guideline of Fraunhofer IAIS (2021). This guideline provides a framework for assessing trustworthiness in a structured way, and at the same time provides guidance for developers to implement these requirements. However, the focus of the guideline is on the verification and not on the provision of design principles for the development.

In 2022, Kaur et al. published a review about Trustworthy AI. They propose an overview about methods that can be used to address the requirements of the EC. Moreover, they argue that the guidelines should be added by a principle focusing on the acceptance of AI. Furthermore, they mention that "human involvement is essential in this changing era of AI..." (p. 39:28). One approach to address this is Human-Centered AI (HCAI).

HUMAN-CENTERED AI

Lately, HCAI get more popular in research. "HCAI focuses on amplifying, augmenting, and enhancing human performance in ways that make systems reliable, safe, and trustworthy." (Shneiderman 2020, 26:2) Instead of replacing people, HCAI seeks to put people at the center of AI-based services. One way of improving HCAI design is to use design principles, patterns, and guidelines throughout the development process. Of course, policymakers' guidelines provide direction for HCAI design, but only on a high level, as previously stated. Technology companies, for example, published more action-oriented guidelines. The People + AI Guidebook from Google's People + Research Center contains 23 design patterns for developing HCAI. The patterns are sorted along critical questions in the development process and more in-depth information can be found

on six thematic categories (User Needs and Defining Success, Data Collection and Evaluation, Mental Models, Explainability and Trust, Feedback and Control, Errors and Graceful Failure; Google PAIR 2019). Another comprehensive work containing action-oriented guidelines is published by Microsoft. Based on a literature review, they identified 18 design guidelines for Human-AI Interaction (Amershi et al. 2019). A web application contains detailed descriptions and practical recommendations for implementation (Microsoft 2021).

PURPOSE OF RESEARCH

Ethical and Human-Centered AI design go hand in hand, i.e., they serve the same goal of developing AI-based services that are reliable, safe, and trustworthy. While the ethical guidelines published by government organizations are at a high level, the HCAI design principles published by technology companies are more specific and provide guidance for action. To implement the seven requirements in practice, AI developers, management, and other stakeholders need action-oriented design principles or design patterns (Shneiderman 2020). The aim of this paper is to examine to what extent the application of the design principles published by Microsoft (Microsoft 2021) and Google (Google PAIR 2019) leads to a fulfillment of the requirements of Trustworthy AI set by the EC. For this purpose, the following research question is answered: Which design principles for AI-based services can be identified to fulfil the guidelines of a Trustworthy AI according to the EC? In addition, aspects are to be identified that are considered in the HCAI design but are of minor importance in the debate on ethical design.

METHOD

To answer the research question, the action-oriented design principles (18 guidelines for Human-AI Interaction by Microsoft, 23 design patterns of the People + AI Guidebook by Google) are mapped to the requirements of the EC by nine independent raters. For this, each participant got a matrix in which the seven requirements were entered in the columns and the design principles in the rows. A cross was used to indicate whether a design principle contributes to the satisfaction of the requirement. To ensure that all participants have the same understanding of the design principles and requirements, explanations were provided in the matrix. Participants required an average of 90 minutes to complete the matrix. All participants regularly deal with the development and implementation of AI-based services in their daily work or conduct research in the field of data-based services. However, participants from different disciplines were selected to fill out the matrix, e.g., AI engineers, digital developers, psychologists, or researchers in information systems. To evaluate the results, the sum was calculated for each cell. Moreover, each row and each column were summed up. A visualization of the frequencies via pie charts was chosen to present the results.

RESULTS

The evaluation of the matrices shows that the requirements "*Human agency and oversight*", "*Transparency*" and "*Technical robustness and safety*" are most frequently addressed by the application of the action-oriented design principles. The requirements "*Privacy and data governance*", "*Diversity, non-discrimination and fairness*" and "*Accountability*" are less addressed by the guidelines for Human-AI interaction (Microsoft 2022; see Figure 1). For the design patterns from the People + AI Guidebook, this is also evident for the requirements "*Diversity, non-discrimination and fairness*" and "*Accountability*" but also for "*Environmental and social well-being*" (see Figure 2).

	Requirement of the EC							
Design Principle by Microsoft	Human agency and oversight	Technical robustness and safety	Privacy and data governance	Transparency	Diversity, non-discrimination and fairness	Environmental and societal well-being	Accountability	Total (max. 63 per row)
Convey the consequences of user actions.	◕	◑	◑	◕	○	◔	◔	26
Encourage granular feedback.	◑	◔	◔	◑	◔	◔	○	21
Learn from users' behavior.	◔	○	◔	○	◔	◔	○	14
Make clear how well the system can do what it can do.	◑	◑	◔	●	○	○	◑	25
Make clear what the system can do.	◔	◑	○	●	◔	◔	◑	26
Make clear why system did what it did.	◔	◑	◔	●	◔	◔	◑	27
Match relevant social norms.	◔	○	○	○	●	◕	○	21
Mitigate social biases.	◔	◔	◔	○	●	◕	○	24
Notify users about changes.	◔	◔	○	◕	○	○	◔	16
Provide global controls.	◕	◔	◑	◕	◔	◔	○	28
Remember recent interactions.	◔	◔	◔	◔	○	◔	◔	14
Scope services when in doubt.	◕	◕	◔	◔	○	◔	○	23
Show contextually relevant information.	◑	◔	○	◑	○	◔	○	17
Support efficient correction.	●	◕	◔	◔	◑	◑	◔	28
Support efficient dismissal.	◕	◔	◔	◔	○	◔	◔	21
Support efficient invocations.	◕	◑	○	◔	○	◔	◔	19
Time services based on context.	●	◔	○	◔	○	◔	○	16
Update and adapt cautiously.	◔	◕	○	◔	◔	○	○	16
Total (max. 162 per column)	85	62	35	73	42	49	36	

Notes. The shading of the circle visualizes the number of crosses given.
● = 8 or 9 ◕ = 6 or 7 ◕ = 4 or 5 ◕ = 2 or 3 ○ = 0 or 1

Figure 1. Mapping of the Guidelines for Human-AI Interaction by Microsoft (Microsoft 2021) and the requirements for Trustworthy AI by the EC (High-Level Expert Group on Artificial Intelligence 2019b).

The results in Figure 1 also show that principles such as "*Convey the consequences of user actions*.", "*Make clear how well the system can do what it can do*." and "*Provide global controls*" address multiple requirements. Principles

like "*Learn from users' behavior.*" and "*Remember recent interactions.*" are seldom linked to the requirements of the EC.

	Requirement of the EU							
Design Principle by PAIR	Human agency and oversight	Technical robustness and safety	Privacy and data governance	Transparency	Diversity, non-discrimination and fairness	Environmental and societal well-being	Accountability	Total (max. 63 per row)
Actively maintain your dataset.	○	◕	●	○	◔	○	◔	22
Add context from human sources.	◑	◔	○	◔	○	○	○	11
Anchor to familiarity.	◑	◔	○	◔	◔	○	○	11
Automate in phases.	●	◑	◔	◑	◔	○	○	25
Automate more when risk is low.	◑	◕	○	○	◔	◔	◔	19
Be accountable for errors	○	●	◔	◔	○	○	◕	23
Be transparent about privacy and data settings.	◔	○	●	◕	◔	○	◔	27
Design for your data labelers.	○	◕	◕	○	◔	○	◔	20
Determine how to show model confidence, if at all.	◑	◕	○	●	◔	○	◔	24
Determine if AI adds value.	◔	◑	○	◑	○	◕	○	18
Embrace "noisy" data.	○	◑	◑	◔	◔	○	○	15
Explain for understanding, not completeness.	◑	○	○	◕	○	○	○	16
Explain the benefit, not the technology.	◔	◔	○	◑	○	○	○	11
Get input from domain experts as you build your dataset.	◔	◕	◑	◔	◑	◔	◑	29
Give control back to the user when automation fails.	●	◑	○	○	○	◔	○	18
Go beyond in-the-moment explanations.	◔	○	○	●	○	○	◔	16
Invest early in good data practices.	○	◕	◕	◔	◔	○	◔	23
Learn from label disagreements.	○	◑	◑	◔	◔	○	◔	20
Let users give feedback.	◑	◑	◑	◑	◑	◔	○	25
Let users supervise automation.	●	◑	○	◑	◔	○	◔	22
Make it safe to explore.	◕	◔	○	◑	◔	○	○	17
Make precision and recall tradeoffs carefully.	○	◕	◔	◔	○	◔	○	16
Set the right expectations.	○	◑	○	●	◔	◔	◑	23
Total (max. 207 per column)	78	100	61	94	44	32	42	

Notes. The shading of the circle visualizes the number of crosses given.
● = 8 or 9 ◑ = 4 or 5 ○ = 0 or 1
◕ = 6 or 7 ◔ = 2 or 3

Figure 2. Mapping of the design patterns of the People + AI Guidebook (Google PAIR 2019) and the requirements for Trustworthy AI by the EC (High-Level Expert Group on Artificial Intelligence 2019b).

Design patterns of the Google + AI Guidebook that address more than one requirement are, e.g., "*Automate in phases*", "*Be transparent about privacy and data settings.*" and "*Get input from your domain experts as you build your*

dataset.". Few requirements are addressed with patterns such as "*Add context from human sources.*", "*Anchor to familiarity*." and "*Explain the benefit, not the technology.*".

DISCUSSION AND CONCLUSION

The aim of this paper is to identify action-oriented design principles that can be applied to satisfy the seven requirements for Trustworthy AI by the EC (European Commission 2021). For this purpose, the design principles published by Google (Google PAIR 2019) and Microsoft (2021) were contrasted with the seven requirements of the EC. The application of the design principles by Microsoft, as well as Google, can be used in particular to fulfil the requirements for "*Human agency and oversight*", "*Technical robustness and safety*" and "*Transparency*". The patterns from the People + AI Guidebook also address the requirement after "Privacy and data governance". To conclude, the design principles are partly suitable for designing Trustworthy AI according to the understanding of the EC. However, they are not sufficient to meet all requirements. The following requirements are less addressed by the design principles: "*Diversity, non-discrimination and fairness*", "*Environmental and societal well-being*" and "*Accountability*". Further research should consider this gap and more detailed action-oriented design principles should be formulated. This gap was identified by Kaur et al. (2022) as well: "However, there is still an implementation gap between the research and practice. So, there is a need to establish policies and standards to enforce these guidelines and existing laws into practice. " (p. 39:2). With the publication of the EU-AI Act (European Commission 2021), the implementation of the requirements for Trustworthy AI will gain additional relevance and thus also the range of practical recommendations for action.

According to the results, some of the principles can help to meet multiple requirements. Nevertheless, to classify the results, it must be considered that one principle cannot fulfil all requirements at the same time. On the one hand, this would hardly be possible due to the complexity and multifaceted requirements, and on the other hand, the design patterns would lose their specific orientation. It should also be noted that contradictions of objectives can arise when fulfilling the requirements. This is a restriction that was also made clear in the guidelines for testing Trustworthy AI systems by the Fraunhofer IAIS (2021). The same is to be expected when implementing the design principles. Consideration of how to deal with potentially conflicting goals was beyond the scope of this work. Further research needs to be done to define selection criteria, as well as methods for applying these criteria to determine the key requirements and principles for a particular AI-based service. For example, depending on the criticality of the AI-based service, as well as its field of application, the interaction strength between humans and AI, or the selected AI technology itself, differences in the relevance of the implementation of design principles may arise (High-Level Expert Group on Artificial Intelligence 2019a; Kaur et al. 2023).

HCAI aims to put people at the center of AI-based services (Shneiderman 2022). Looking at the guidelines for Human-AI Interaction formulated by Microsoft as

well as the People + AI Guidebook by Google, it becomes clear that some formulated principles do not match the requirements of the EC. This is particularly evident for those principles that focus on Human-AI interaction. As "HCAI focuses on amplifying, augmenting, and enhancing human performance…" (Shneiderman 2020, 26:2) these principles are not minor important in the discussion about AI-based services that are perceived as reliable, safe, and trustworthy. More research should be conducted to analyze how the two aspects of Human-AI interaction and Ethical AI might be considered together.

In the future, an important addition to the existing design principles could be the combination with insights and approaches from the discipline of service science and engineering. On the one hand, this discipline deals with the design of new services and takes a holistic view of the utilisation process as well as the consideration of contextual factors during utilisation. So far, such factors have only been marginally considered in the existing design principles. On the other hand, new methods are currently being researched on how the perception of quality (including the perceived trustworthiness, safety and usefulness during the usage process) can already be tested during development in order to prevent undesirable developments as early as possible (Neuhüttler et al. 2022). The approaches there do not deal with the objectively assessable technical implementation, but with the perception of users.

Our research has some limitations. The results show differences in the evaluation of matching. One possible reason for this could be that the participants might understand the requirements and design principles differently. It was not possible to verify that everyone understood the requirements and principles equally well, even though detailed descriptions for each requirement and principle attempted to ensure this. To identify reasons for the different matchings, further qualitative studies should be conducted, for example through focus group discussions. In addition, the study should be replicated with a larger sample to validate the results. Another limitation of this study is that it only included the design principles of two sources. As these works are, to our knowledge, the most comprehensive available from a practical point of view, we have nevertheless limited the scope of our study to them. An extension of the literature and desktop research, with particular emphasis on the requirements not covered by the design principles considered, could contribute to a more complete picture. In order to provide actionable principles for each requirement, a comprehensive framework should be established in the future.

REFERENCES

Amershi, Saleema/Weld, Dan/Vorvoreanu, Mihaela/Fourney, Adam/Nushi, Besmira/Collisson, Penny/Suh, Jina/Iqbal, Shamsi/Bennett, Paul N./Inkpen, Kori/Teevan, Jaime/Kikin-Gil, Ruth/Horvitz, Eric (2019). Guidelines for Human-AI Interaction. In: Stephen Brewster/Geraldine Fitzpatrick/Anna Cox et al. (Eds.). Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19: CHI Conference on Human Factors in Computing Systems, Glasgow Scotland Uk, 04 05 2019 09 05 2019. New York, NY, USA, ACM, 1–13.

European Commission (2021). Proposal for a Regulation of the European Parliament and the Council. Laying down harmonised rules on artificial intelligence. Brussels. Available online at file:///C:/Users/kutz/Downloads/regulation_ai_875509BF_-C386-0D30-2CB7E56A798BA4EA_75788.pdf (accessed 2/9/2022).

Google PAIR (2019). People + AI Guidebook. Designing human-centered AI products. Available online at https://pair.withgoogle.com/guidebook/ (accessed 04.11.22).

Hagendorff, Thilo (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines 30 (1), 99–120. https://doi.org/10.1007/s11023-020-09517-8.

High-Level Expert Group on Artificial Intelligence (2019a). Ethics guidelines for trustworthy AI. Available online at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed 2/9/2023).

High-Level Expert Group on Artificial Intelligence (2019b). Ethics Guidelines for Trustworthy AI. Brussels. Available online at file:///C:/Users/kutz/Downloads/ai_hleg_ethics_guidelines_for_trustworthy_ai-en_87F84A41-A6E8-F38C-BFF661481B40077B_60419.pdf.

Jobin, Anna/Ienca, Marcello/Vayena, Effy (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence 1 (9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Kaur, Davinder/Uslu, Suleyman/Rittichier, Kaley J./Durresi, Arjan (2023). Trustworthy Artificial Intelligence: A Review. ACM Computing Surveys 55 (2), 1–38. https://doi.org/10.1145/3491209.

Kutz, Janika/Neuhüttler, Jens/Spilski, Jan/Lachmann, Thomas (2022). Implementation of AI Technologies in manufacturing - success factors and challenges. In: The Human Side of Service Engineering, 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022), July 24-28, 2022. AHFE International.

McGregor, Sean (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. Proceedings of the AAAI Conference on Artificial Intelligence 35 (17), 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817.

Microsoft (2021). Guidelines for Human-AI Interaction. Microsoft HAX Toolkit. Available online at https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/ (accessed 2/8/2023).

Neuhüttler, Jens/Hermann, Sibylle/Ganz, Walter/Spath, Dieter/Mark, Riccarda (2022). Quality Based Testing of AI-based Smart Services: The Example of Stuttgart Airport. In: 2022 Portland International Conference on Management of Engineering and Technology (PICMET), 2022 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA, 07.08.2022 - 11.08.2022. IEEE, 1–1

Poretschkin, Maximilian/Schmitz, Anna/Akila, Maram/Adilova, Linara/Becker, Daniel/Cremers, Armin B./Hecker, Dirk/Houben, Sebastian/Mock, Michael/Rosenzweig, Julia/Sicking, Joachim/Schulz, Elena/Voß, Angelika/Wrobel, Stefan (2021). Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog). Fraunhofer IAIS. Sankt Augustin. https://doi.org/10.24406/PUBLICA-FHG-301361.

Shneiderman, Ben (2020). Bridging the Gap Between Ethics and Practice:. Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Transactions on Interactive Intelligent Systems 10 (4), 1–31. https://doi.org/10.1145/3419764.

Shneiderman, Ben (2022). Human-centered AI. Oxford, United Kingdom/New York, NY, Oxford University Press.

Xu, Wei (2019). Toward human-centered AI: A Perspective from Human-Computer-Interaction. Interactions 26 (4), 42–46. https://doi.org/10.1145/3328485.

Xu, Wei/Dainoff, Marvin (2021). Enabling human-centered AI: A new junction and shared journey between AI and HCI communities. Available online at http://arxiv.org/pdf/2111.08460v3.

3.4 Publikation 4: Human-Centered AI for Manufacturing – Design Principles for Industrial AI-Based Services

Human-Centered AI
for Manufacturing – Design Principles
for Industrial AI-Based Services

Janika Kutz[1,2(✉)], Jens Neuhüttler[1], Bernd Bienzeisler[1], Jan Spilski[2], and Thomas Lachmann[2,3]

[1] Fraunhofer Institute for Industrial Engineering IAO, 70569 Stuttgart, Germany
janika.kutz@iao.fraunhofer.de
[2] Center for Cognitive Science, University of Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
[3] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, Spain

Abstract. AI-based services are becoming more and more common in manufacturing; however, the development, implementation, and operation of these services are associated with challenges. The design of Human-Centered AI (HCAI) is one approach to address these challenges. Design guidelines and principles are provided to assist AI developers in the design of HCAI. However, these principles are currently defined for AI in general and not for specific application contexts. The aim of this work is to analyze whether existing design principles for HCAI are transferable to IAI-based services in manufacturing and how they can be integrated into the development process. In an explorative-qualitative research design, the design pattern of the People + AI Guidebook by the PAIR from Google were analyzed regarding their applicability in manufacturing environments. The finding show that a transfer of the design principles is generally possible. According to the experts, 15 of the design patterns have a direct influence on the perception of Industrial AI-based services by end-users or management and can thus increase the acceptance of them. Finally, the design patterns were assessed in terms of their application relevance and complexity in manufacturing.

Keywords: Industrial AI · Human-Centered AI · Design Principles

1 Introduction

The field of Human-Centered Artificial Intelligence (HCAI) has recently gained more importance. A number of research papers have addressed the design of HCAI applications. These studies emphasize the significance of a human-centered design approach in the development and operation of AI-based systems, as these systems have a significant impact on people's daily lives well as their working environment [1–3]. The aim of HCAI is to design AI-based systems that empower people, fulfil humans values and needs [3]. "A human-centered approach will reduce the out-of-control technologies, calm fears of robot-led unemployment, and give users the rewarding sense of mastery and accomplishment" [3].

H. Degen and S. Ntoa (Eds.): HCII 2023, LNAI 14050, pp. 115–130, 2023.
https://doi.org/10.1007/978-3-031-35891-3_8

Titel: Human-Centered AI for Manufacturing – Design Principles for Industrial AI-Based Services

Autoren: Janika Kutz, Jens Neuhüttler, Bernd Bienzeisler, Jan Spilski, Thomas Lachmann

Status: veröffentlicht

Jahr der Veröffentlichung: 2023

Zitation: Kutz, J., Neuhüttler, J., Bienzeisler, B., Spilski, J., & Lachmann, T. (2023). Human-Centered AI for Manufacturing – Design Principles for Industrial AI-Based Services. In H. Degen & S. Ntoa (Eds.), Lecture Notes in Computer Science, Artificial Intelligence in HCI (pp. 115–130). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35891-3_8, reproduced with permission from Springer Nature

Inhalt:

Die Entwicklung menschenzentrierter KI-basierter Services ist eine Möglichkeit, diverse Herausforderungen der Entwicklung und Implementierung von industrieller KI zu überwinden. Leitfäden und Design-Prinzipien können Orientierung und Unterstützung im Entwicklungsprozess geben. Jedoch sind die bisher veröffentlichten Arbeiten dazu entweder allgemein gehalten oder für KI-Anwendungen in anderen Anwendungsgebieten entwickelt worden. In einer qualitativen Studie wurde analysiert, inwiefern sich bekannte Prinzipien auf das industrielle Umfeld übertragen lassen. Untersucht wurde dies am Beispiel der Design Patterns des P*eople + AI Guidebook* des PAIR Research Centers von Google. Es hat sich gezeigt, dass eine Übertragung möglich ist, jedoch kontextspezifische Anpassungen vorgenommen werden müssen.

Einordnung in den Gesamtzusammenhang der Arbeit:

- Stärkerer Fokus auf das Thema der menschenzentrierten KI wurde als Möglichkeit zur Überwindung verschiedener Herausforderungen im industriellen Umfeld identifiziert
- Bestehende Design-Prinzipien für menschenzentrierte KI wurden auf ihre Anwendbarkeit im industriellen Umfeld hin analysiert

The following Paper: Kutz, J., Neuhüttler, J., Bienzeisler, B., Spilski, J., & Lachmann, T. (2023). Human-Centered AI for Manufacturing – Design Principles for Industrial AI-Based Services. In H. Degen & S. Ntoa (Eds.), Lecture Notes in Computer Science, Artificial Intelligence in HCI (pp. 115–130). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35891-3_8 is reproduced with permission from Springer Nature and is subject to the publisher's policy and is excluded from the selected license for the dissertation.

Human-Centered AI for Manufacturing – Design Principles for Industrial AI-Based Services

Janika Kutz[1,2(✉)], Jens Neuhüttler[1], Bernd Bienzeisler[1], Jan Spilski[2], and Thomas Lachmann[2,3]

[1] Fraunhofer Institute for Industrial Engineering IAO, 70569 Stuttgart, Germany
janika.kutz@iao.fraunhofer.de
[2] Center for Cognitive Science, University of Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
[3] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid, Spain

Abstract. AI-based services are becoming more and more common in manufacturing; however, the development, implementation, and operation of these services are associated with challenges. The design of Human-Centered AI (HCAI) is one approach to address these challenges. Design guidelines and principles are provided to assist AI developers in the design of HCAI. However, these principles are currently defined for AI in general and not for specific application contexts. The aim of this work is to analyze whether existing design principles for HCAI are transferable to IAI-based services in manufacturing and how they can be integrated into the development process. In an explorative-qualitative research design, the design pattern of the People + AI Guidebook by the PAIR from Google were analyzed regarding their applicability in manufacturing environments. The finding show that a transfer of the design principles is generally possible. According to the experts, 15 of the design patterns have a direct influence on the perception of Industrial AI-based services by end-users or management and can thus increase the acceptance of them. Finally, the design patterns were assessed in terms of their application relevance and complexity in manufacturing.

Keywords: Industrial AI, Human-Centered AI, Design Principles

1 Introduction

The field of Human-Centered Artificial Intelligence (HCAI) has recently gained more importance. A number of research papers have addressed the design of HCAI applications. These studies emphasize the significance of a human-centered design approach in the development and operation of AI-based systems, as these systems have a significant impact on people's daily lives well as their working environment [1–3]. The aim of HCAI is to design AI-based systems that empower people, fulfil humans values and needs [3]. "A human-centered approach will reduce the out-of-control technologies, calm fears of robotled unemployment, and give users the rewarding sense of mastery and accomplishment" [3].

https://doi.org/10.1007/978-3-031-35891-3_8

This also applies to the use of AI-based systems in manufacturing, referred to as Industrial AI (IAI). The use of IAI is diverse, as it can be applied along the entire value chain from logistics to assembly [4]. All these applications have one thing in common: they influence the working environment. As a result, employees at manufacturing organizations will increasingly need to collaborate with AI-based systems in the future. Even though IAI applications are becoming more common, development, implementation and roll-out are associated with numerous challenges. Challenges identified in our previous research are, e.g., "concerns and fear related to AI-Services", "false expectations of the end-user in the AI service" and "lack of technology acceptance by employees". In addition, we identified several success factors in relation to IAI-based services, e.g., "trust in AI", "confidence in operability of the IT-System", "user-centered development" and the "added value of the AI service must be clear" [5]. Designing HCAI-based services for manufacturing is a necessary step to address these challenges and to strengthen the success factors. The aim of these paper is to analyze whether known approaches to design HCAI are transferable to IAI-based services and how these can be integrated into the development process.

2 Theoretical Background

2.1 Human-Centered AI Design

Throughout the entire AI life cycle, HCAI seeks to put people back in the center of AI-based services. The aim of HCAI is not to reduce the degree of automation, but to create AI-based systems "that increase automation, while amplifying, augmenting, enhancing, and empowering people to innovatively apply systems and creatively refine them" [6]. One way to enhance HCAI design is to apply design principles, patterns, and guidelines throughout the development process. Various guidelines and principles have been published lately for the implementation of ethical and human-centered AI-based services. Principles are developed at various abstraction levels. A distinction can be made between high-level guidelines, which are mainly provided by policymakers, and concrete action-oriented principles, which are published by companies, for example [7]. The former are usually general guidelines that can provide initial guidance on HCAI design. For example, the "Ethics Guidelines for Trustworthy AI" of the High-Level Expert Group on Artificial Intelligence of the European Commission (2019) describes four principles and seven requirements for designing trustworthy AI systems [8]. However, these general guidelines contain only a few concrete development recommendations. The action-oriented principles, on the other hand, offer concrete instructions for action and can thus be understood as design principles in a narrower sense [9]. Google and Microsoft have published such specific design principles. Google's People + AI research team [10] has published 23 design patterns for HCAI in their People + AI Guidebook. The patterns are structured along common questions that arise in the development process [10]. Pattern language is used to describe the principles in the People + AI Guidebook. The patterns are "brief expressions of important ideas that suggest solutions to common design problems" [3]. Microsoft published its Guidelines for Human-AI Interaction in 2019 [1]. These can also be viewed in a web application, are described in detail, and include practical recommendations [11]. This paper focuses on design principles/design

patterns, that provide specific instructions on how to apply and implement the principles. However, more general design principles and guidelines are also crucial and can offer basic advice on HCAI design. Finally, it should be noted that the various published guidelines have one thing in common: they are generally applicable and not specifically tailored to concrete application contexts, such as manufacturing.

2.2 The Human Factor in Industrial AI-Based Services

Manufacturing companies are increasingly embracing digitalization, and along with it, the use of AI-based services in manufacturing is growing. Industrial AI applications can be used for, e.g., predictive maintenance, quality management respectively fault diagnosis or as decision support systems [12]. Furthermore, they can be understood as services, as they are not limited to technological aspects, but rather represent an immate-rial offer that aims to change the state of people, objects, processes, or information [13]. Nevertheless, AI applications are scarce in the industrial environments [4, 14]. This has several reasons, e.g., lack of skills and experience, lack of development and implementation resources and lack of access to the necessary data [4]. Moreover, the development, roll-out, and operation of IAI-based services is related to various challenges at a technical, organizational, and human level. The development of AI systems has been driven primarily by technology [15] and users at the operational level are poorly considered in the development process [16]. Thereby, IAI-based services should "connect people and things through systems..." [14], which indicates that beyond technological factors, human and organizational requirements should also be taken into account. Employees should not be viewed as objects to be persuaded, but as active participants in the transformation process, to increase the technological acceptance of Industry 4.0 solutions and thus AI-based systems [16].

Hoffmann et al. describe in their "Proposal for requirements on industrial AI solutions" a broader understanding of IAI, which includes, beyond technology, topics such as industrial applications, value creation, human-AI interaction, regulatory aspects, and ethics [12]. Along 5 main issues (Adaption, Engineering, Embedding, Safety/Security, Trust), they formulate 16 requirements in relation to IAI solutions, e.g., "Stepwise introduction", "Virtual learning" and "Proof of capabilities". This work gives AI developers in the industrial context a first orientation for the design, which goes beyond technical aspects.

2.3 Purpose of Research

As previously mentioned, there are several publications that discuss the ethical and human-centered design of AI applications and establish principles and guidelines for them. These, however, are of general application or specifically aimed at mainstream AI applications and have not been developed specifically for IAI-based services. Since IAI has some specific characteristics [12, 14], a simple application of the principles in the context of IAI is not possible. However, the need for HCAI in the industrial environment is high to overcome certain challenges. It should be the aim to develop IAI-based ser-vices that are accepted and trusted by the employees. Moreover, they should be perceived

as safe by the employees. Furthermore, semi-automated IAI-based services that combine machine learning with human experience have proven to be particularly effective. This also necessitates a human-centered design approach, where the IAI-based service assists the employees rather than replacing them [4]. Integrating HCAI applications into the industrial environment is a necessary step to successfully advance and shape digitalization and automation in industry. The purpose of this paper is to introduce the production perspective on HCAI. Moreover, practical guidance for the application of HCAI design principles in industrial environments is provided. Therefore, the following research questions are answered:

- Which design principles for HCAI known from literature are applicable to IAI-based services?
- When should the design principles be considered during the development process of IAI-based services?
- Which design principles are directly perceptible by the management and end-users?
- How relevant is their use in practice?
- How complex is their application in industrial environments?

After screening various design guidelines, principles and patterns for the design of AI-based systems, we selected the design patterns defined in Google PAIR's People + AI Guidebook [10] to answer the research questions. To analyze the applicability in the industrial environment, design patterns with complete descriptions and practical recommendations were selected. Each design pattern is thoroughly described in the Guidebook, along with helpful actionable advice. Additionally, the development of reliable, safe, and trustworthy AI-based applications should be the main objective, with a clear focus on human-centered design. This is also the case with the Guidebook, as it is according to the People + AI Research: "A friendly, practical guide that lays out some best practices for creating useful, responsible AI applications." [10].

3 Method

To answer the research questions, an explorative-qualitative research design was chosen. Online workshops were conducted with two groups, each consisting of 2 experts for Industrial AI-based services. Two 60-min sessions were held with the first group of participants, and one 120-min session was held with the second group. During the workshops, the participants worked interactively in three steps. First, the 23 design patterns from The People + AI Guidebook [10] were examined and their general applicability to the manufacturing environment was discussed. This ensured that all participants had the same level of knowledge and understanding of the design patterns. In the second step, the design patterns were located in a matrix. On the one hand, an assessment was made in which phase of an AI development process a design pattern should be applied. Following the ML-Ops process [17], the development phases selected here are divided into three steps: (1) Design, (2) Development and (3) Roll-out and Operation. Secondly, the design patterns were assessed with regard to their main target group. A distinction was made between two groups, (1) AI Engineers and (2) End-Users and Management. In the third step, the design patterns allocated to the target group of End-Users and

Management were evaluated based on two variables. On the one hand, the relevance of a design pattern regarding the development and operation of HCAI applications in manufacturing should be assessed. On the other hand, the complexity of applying a design pattern in a manufacturing environment. Each pattern was then located on a graph along the two variables. In addition, at the end of each step, it was reflected whether further design patterns need to be added.

After the conclusion of the workshop, the results were consolidated. Design patterns that were evaluated and classified differently by the groups were discussed in an internal session consisting of two scientists from different disciplines (Group 3). In total, the evaluation of the design patterns was carried out by six experts for Industrial AI-based applications from different disciplines (Table 1).

Table 1. Description of the sample.

Expert	Professional background	Group
1	ML-Engineer for Industrial AI applications	Group 1
2	Digital Business Developer	Group 1
3	Expert for trustworthy AI in industrial applications	Group 2
4	Expert for Industrial AI-based systems	Group 2
5	Expert for acceptance of AI-based services	Group 3
6	Expert for development and quality assessment of smart services	Group 3

4 Results

At the beginning of the workshops, the 23 design patterns were explained to create an equal understanding of them among all participants. Following from this, the design principles were discussed in terms of their general applicability in an industrial environment. All experts agreed that applying the design patterns is theoretically achievable and that all patterns are basically relevant. The experts also mentioned the need for HCAI patterns for the industrial context and confirmed that their application would be beneficial. Only one of the 23 design patterns proved to be irrelevant for the industrial context. The experts consented that a use of the pattern "*Adding context from human sources.*" in an industrial setting would not be appropriate. According to the experts, parameters from the technical installations could be integrated as third-party sources. However, this would require a more detailed examination of the extent to which this could be usefully applied as a design pattern. For this reason, this design pattern is no longer considered in the subsequent presentation of results. Furthermore, the experts also noted that there might be differences depending on the IAI use case. Moreover, no major gaps could be identified and in consequence no further design principles are formulated. The design patterns can be considered as a starting point and provide a comprehensive orientation regarding which aspects should be considered when designing HCAI. However, it also

became clear that a one-to-one transfer is not possible, instead specific descriptions are required for the industrial application context.

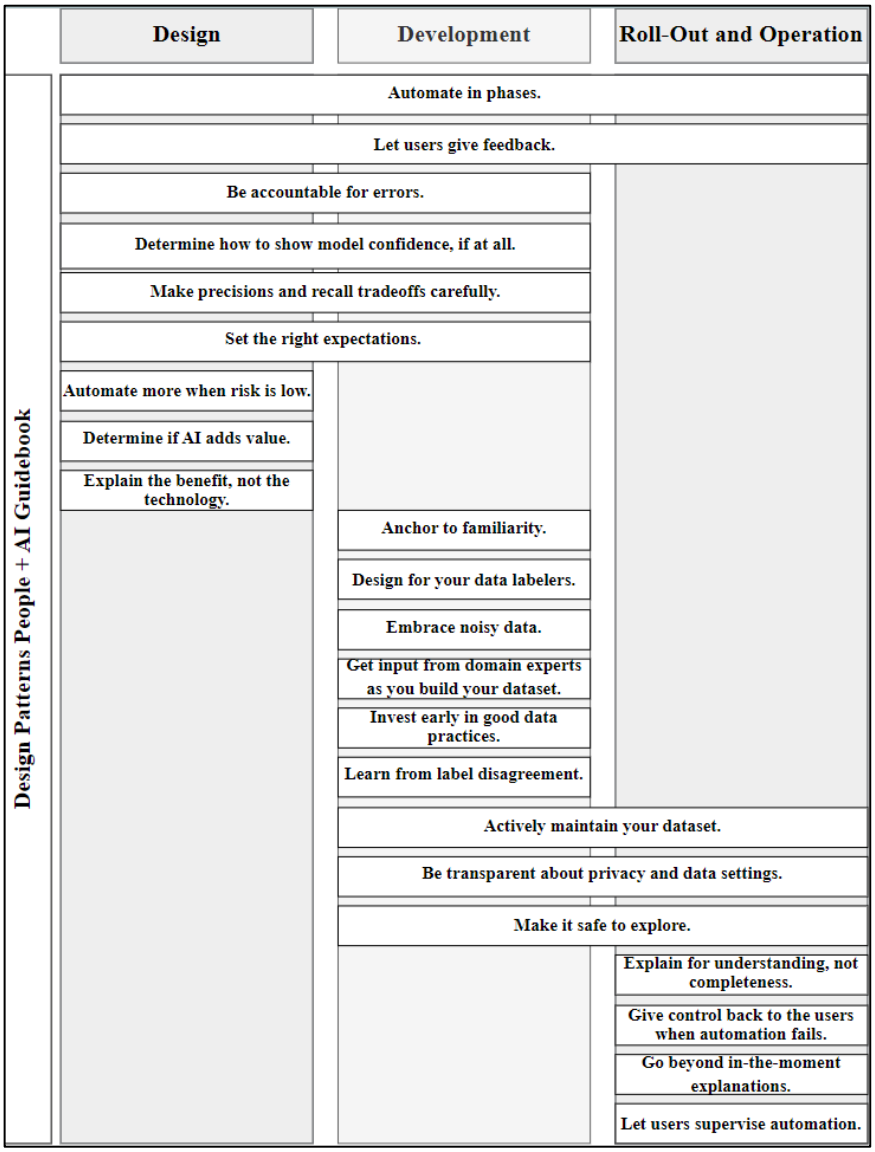


Fig. 1. Application of the design patterns of the People + AI Guidebook [10] in the development process of IAI-based services.

Answering the second research question: “When should the design principles be considered during the development process of IAI-based services?” the design patterns were located along the IAI development process, with each workshop group. A consolidated version of the allocation can be found in Fig. 1.

4.1 End-User Focused Design Patterns

In the second workshop phase, the application of the design patterns in the context of IAI-based services was analyzed in more detail. To answer the following research question "Which design principles are directly perceptible by the management and end-users?" the design patterns were evaluated in relation to their primary target group. The aim was to identify those patterns that end-users perceive directly in their everyday work with IAI-based services. Across the workshops, there was a strong consensus on which design patterns directly address end-users and influence their perception of and interaction with AI-based services. For five design patterns, differentiated opinions on the target group could be identified. For example, depending on the design of the development process, it is quite possible that the end-users will be directly affected by design patterns relating to data in their daily work (e.g., "*Actively maintain your dataset*.", "*Get input from domain experts as you build your dataset."*). Nevertheless, the experts agreed that this pattern is primarily tailored to the AI engineers and only has an indirect effect on the end-users through the increased quality of the AI-based service. Furthermore, it becomes clear that differentiating the pattern of "*Determine if AI adds value."* into a technical and a business level would be beneficial for the application in manufacturing environments.

The findings of the workshops are summarized in the table below, along with a few sample quotes from the workshops. Furthermore, the figures in the appendix illustrate which design principles were classified as end-user-centered in each workshop (Table 2).

The workshop results also indicate that the AI engineers are typically in charge of putting the design patterns into practice. According to the experts, this is not the case for four of the patterns; in these cases, the experts pointed out that the management is responsible. *Automate in phases:* Before the development and introduction of AI-based services, the management should develop an automation strategy for the company and based on this, decide in which automation steps an AI system will be rolled out. *Automate more when risk is low*: According to the experts, management is jointly accountable for analyzing risk and deciding whether to adopt AI projects based on the risk assessment. *Determine ifAI adds value*: The experts emphasize that management must decide whether to adopt IAI-based services on the predicted added value. *Make precision and recall tradeoffs carefully*: Is interpreted by the experts as a framework for the actual development, which must also be set by the management.

Based on the results of the second workshop phase, the patterns that were assessed as end-user-centered were further evaluated in the third phase. The experts assessed the patterns according to their application relevance and application complexity in the manufacturing environment. For 7 design principles, the relevance as well as the complexity of the application in the context of IAI-based services was assessed similarly (see Fig. 2).

Table 2. Experts' assessments of the main target group

AI Engineers	Management and End-Users
Actively maintain your dataset.	
The experts agree that primarily AI-engineers are targeted. Sample quote: "Patterns primarily influence the work of AI engineers, as good data management is important for them to develop and deliver high-quality AI systems." (Group 2)	End-users may also be targeted: "AI developers need support from production to collect data. If the end-users support this process and provide data themselves, they are directly affected by the pattern." (Group 1)
Anchor to familiarity	
/	The experts agree that primarily end-users are targeted. Sample quote: "… is relevant for users because they have to work with the system in the end." (Group 2)
Automate in phases.	
/	The experts agree that primarily end-users are targeted. Sample quote: "Especially relevant for end-users because they can slowly get used to the increasing automation and familiarize themselves with the new technology." (Group 1)
Automate more when risk is low.	
/	The experts agree that primarily management and end-users are targeted. Sample quote: "… particularly relevant for management, as they are finally in charge." (Group 2)
Be accountable for errors.	
The experts agree that primarily AI-engineers are targeted.	/
Be transparent about privacy and data settings.	
/	The experts agree that primarily management and end-users are targeted.
Design for your data labelers.	

(*continued)*

Table 2. (*continued*)

AI Engineers	Management and End-Users
The experts agree that primarily AI-engineers are targeted. Sample quote: "… is especially important for AI engineers to ensure that the data quality in the development process is correct." (Group 2)	/
Determine if AI adds value.	
"...when it comes to technical feasibility and an assessment of that, the principle primarily addresses AI engineers." (Group 1)	"...relevant for employees, whether their work is facilitated by the use of AI systems." (Group 2) "The management has to decide whether the implementation of an AI adds a value for the organization or not." (Group 2)
Embrace "noisy" data.	
The experts agree that primarily AI-engineers are targeted.	/
Explain for understanding, not completeness.	
/	The experts agree that primarily end-users are targeted.
Explain the benefit, not the technology.	
/	The experts agree that primarily end-users are targeted.
Get input from domain experts as you build your dataset.	
The experts agree that primarily AI-engineers are targeted. Sample quote: "…das soll man machen, aber Anwender und Management bekommen dies nicht unbedingt direkt mit." (Group 2)	End-users may also be targeted: Werden die Endanwender als Domain Experts einbezogen können diese hier direkt Einfluss nehmen auf die Qualität der Daten. (Group 1 und 3)
Give control back to the user when automation fails.	
/	The experts agree that primarily end-users are targeted.
Go beyond in-the-moment explanations.	
/	The experts agree that primarily end-users are targeted.
Invest early in good data practices.	

(*continued)*

Table 2. (*continued*)

AI Engineers	Management and End-Users
Invest early in good data practices.	
The experts agree that primarily AI-engineers are targeted.	End-users may also be targeted: "If the end-users support the data collection, they are directly affected and can actively influence the development process by providing data according to quality standards." (Group 1)
Learn from label disagreements.	
The experts agree that primarily AI-engineers are targeted.	/
Let users give feedback.	
	The experts agree that primarily end-users are targeted.
Let users supervise automation.	
/	The experts agree that primarily end-users are targeted.
Make it safe to explore.	
/	The experts agree that primarily end-users are targeted.
Make precision and recall tradeoffs carefully.	
/	The experts agree that primarily management and end-users are targeted. Sample quote: "Ratio of error rates is especially important for end-users and management." (Group 2)
Set the right expectations.	
"In development, it is important to have realistic expectations of an AI system, even within the team of AI engineers responsible for development." (Group 1)	The experts agree that primarily the management and end-users are targeted.

Note: The design patterns are the same as in the People + AI guidebook [10].

For example, both groups consider the principle "Let users give feedback" as very relevant, but also as highly complex to implement. Group 2 emphasizes that the complexity increases with the number of people involved in development, and that feedback made by end-users results in a less linear development process. Group 1 mentioned that depending on the AI-based service, it may be difficult to integrate feedback options into the production process accurately. At the same time, it must be guaranteed that the AI engineers take feedback into account and integrate it into the AI-based services. Both groups rated "*Give control back to the users when automation fails*" as one of the most important patterns. According to the experts, this is particularly relevant in the production

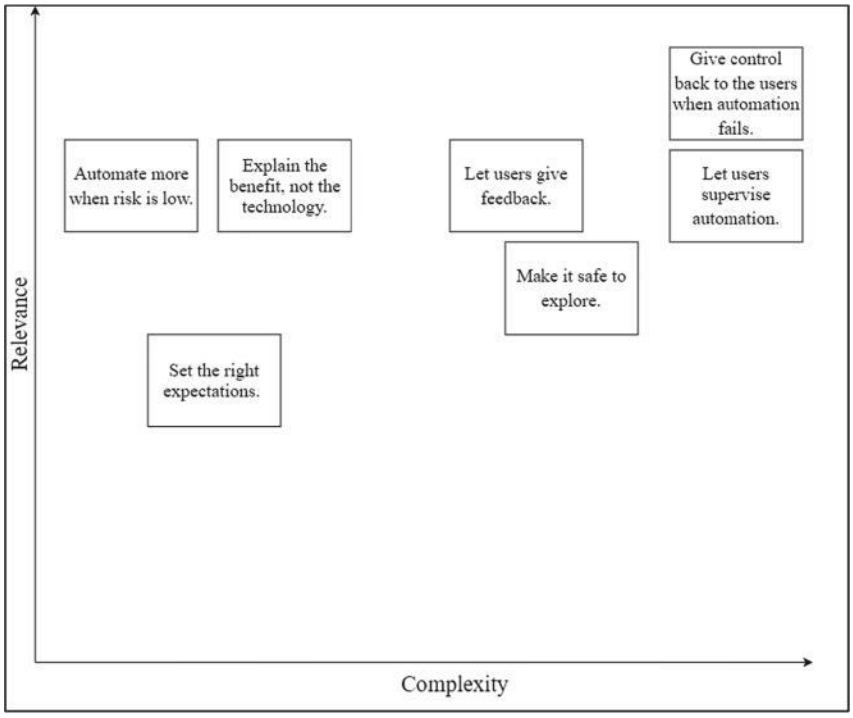


Fig. 2. Consolidated assessment of the relevance and complexity of the application of end-user centered design patterns of the People + AI Guidebook [10].

context because it is usually a prerequisite in manufacturing to have a back-up system. Designing this handover process is a challenge. According to the experts, complexity can change based on the use case, its importance, and its integration into the production process. According to the experts, the design pattern "*Let users supervise automation*" is equally complex to integrate into production processes, as these are currently standardized and offer little freedom for individual adjustments. The design principle "*Explain the benefit, not the technology*" is ranked relatively high on the relevance dimension by both workshop groups. Both groups justify this with the argument that in this way, end-users and management can be convinced to use an AI system. The complexity is rated as rather low by both groups. "*Make it safe to explore*" is classified as a pattern that is very important to build end-user trust in an AI-based service.

In contrast, the assessment of eight design patterns along the dimensions differs between the workshop groups. For example, the pattern "*Be transparent about privacy and data settings*" was assessed differently. Group 1 rated the principle as not very complex, but very relevant, whereas group 2 rated the relevance rather low, but the complexity high. "*Automate in phases*" is rated as rather relevant by both groups, but the complexity is rated differently. Group 1 explains that automation in phases can lead to additional development efforts. Group 2, on the other hand, justifies the low complexity by saying that a step-by-step approach to maximum automation can reduce complexity during development. Group 1 said the pattern "Go beyond in-the-moment explanations"

is less relevant, but also less complex in application. Group 2 assesses the application as very complex, since a comprehensible and more in-depth preparation of the contents from the systems would require additional effort. However, this varies depending on the application. Since detailed explanations can build trust, the experts rated the pattern as relevant. "*Explain for understanding, not completeness*" is considered very important by both groups, but the assessment of complexity varies. Group 2 clarified in the workshop that it is in any case more important to provide a simple explanation than the more in-depth explanations. "*Determine how to show model confidence, if at all*" was rated by group 2 as less complex to apply and was placed in the middle of the scale for determining relevance. The first group, on the other hand, rated it as very relevant and placed the complexity in the middle. Both groups pointed out that this can help to increase trust in an AI-based service. The appendix contains a presentation of the assessments per workshop.

5 Discussion and Conclusion

5.1 Discussion, Limitations and Further Need for Research

The use of HCAI is a promising approach for the manufacturing industry to create AI systems that are accepted by employees as they are perceived as safe and trustworthy. Looking at the requirements defined by Hoffmann et al. (2021) for IAI solutions, some requirements can be met by using Designing Patterns. For example, the recommended "stepwise introduction" of IAI solutions is addressed with the design pattern "*Automate in phases*" [12]. However, it can also be seen that the patterns examined go beyond the described requirements and address additional aspects. The design patterns also offer a good opportunity to overcome existing challenges. Among other things, the use of design patterns promotes a user-centered development process, strengthens trust in AI applications, leads to realistic expectation management and can provide orientation in the development process [5].

The results presented here are the result of a qualitative study based on opinions of 6 experts. In order to check the validity of the results, it would be useful to evaluate the results with other experts in further research projects. Also, the design principles should be tested in application, either in experimental settings or in the direct production environment. In this way, deeper insights into the effects of the patterns could be gained and best-practices for application in practice could be elaborated. The experts made it clear that an application is not directly transferable, but that it would make sense to adapt it to the production context. In the current study, the necessary adaptation could not be dealt with in more detail. Furthermore, the assessment of the application relevance and application complexity is only based on the rankings of four experts. Again, further research is needed to validate our findings. It is important to keep in mind when interpreting the findings that the experts may rate relevance and importance differently depending on the particular AI use cases. The ratings given reflect their general view. In further research, the different patterns should be considered separately and detailed elaborations on the application in the manufacturing environment should be conducted. Decision support for AI engineers on which of the principles should be applied in the design of specific AI use cases would also be valuable.

Another potential limitation is that we only analyzed the design patterns of the People + AI Guidebook [10]. Other Design Guidelines, even if they are providing detailed descriptions like the Microsoft Guidelines for Human-AI Interaction [11] were not considered. There might be further design guidelines, principles, and patterns that are also applicable in the manufacturing environment and offer added value, as well as going beyond the patterns contained in the Guidebook. However, it would have been beyond the scope of this research to examine further guidelines. Furthermore, it could not be considered within the scope of this study which specific types of human intelligence are supported by AI and whether different principles can be derived from this for the design of industrial AI based service applications [18].

Nevertheless, the design patterns considered are very focused on the IAI system itself, but the IAI-based service consists of further components. In particular, the processes before and after the AI application should also be considered in the design in an industrial context, since contextualizing factors form the framework for quality percep-tion of AI-based services [19]. Last but not least, the use of design patterns should be aligned with common validation procedures and transferred into a holistic and harmo-nized approach for design and testing activities during development [20]. In the near future, the regulatory framework should also be taken into account when designing design patterns. Discussions about the safeguarding and certification of AI-based ser-vices have gained relevance. The extent to which the design patterns already address these requirements should be examined in the future. The focus of this study was on industrial AI-based services, but the design of HCAI is also relevant in other disciplines. The extent to which design principles can be applied in other disciplines and whether these, e.g., lawyers or medical professionals, have different requirements for the design of AI-based services than production employees should be investigated in future studies.

5.2 Conclusion and Practical Implications

The aim of the paper is to initially address an HCAI design for the manufacturing envi-ronment. The experts agreed that the need for human-centered IAI-based services is high and generally consider the application of the design patterns is beneficial. By allocating the design patterns along the AI development process, AI engineers are provided with a practical guide for the future design of IAI-based services. During the development and implementation of IAI-based services, it is particularly important to involve the end-users in the process [21]. In addition, management support must also be ensured to successfully shape the digital transformation [5]. The use of design patterns can help developers to address these requirements. Across the workshops, the experts identified 15 of the relevant 22 design patterns as having a direct impact on end-users' or management's perception of IAI-based services:

1. Anchor to familiarity
2. Automate in phases
3. Automate more when risk is low
4. Be transparent about privacy and data settings
5. Determine how to show model confidence, if at all
6. Determine if AI adds value

7. Explain for understanding, not completeness
8. Explain the benefit, not the technology
9. Give control back to the user when automation fails
10. Go beyond in-the-moment explanations
11. Let users give feedback
12. Let users supervise automation
13. Make it safe to explore
14. Make precision and recall tradeoffs carefully
15. Set the right expectations

The application of the design patterns is considered very relevant in most cases, whereas the complexity of the application varies between the patterns. The remaining 7 design patterns address the AI engineers themselves in particular. To design reliable, safe, and trustworthy AI systems, these principles should also be taken into account, but they are not directly perceptible to the end-user after implementation.

Appendix

The experts assessed the patterns according to their application relevance and application complexity in the manufacturing environment (Figs. 3 and 4).

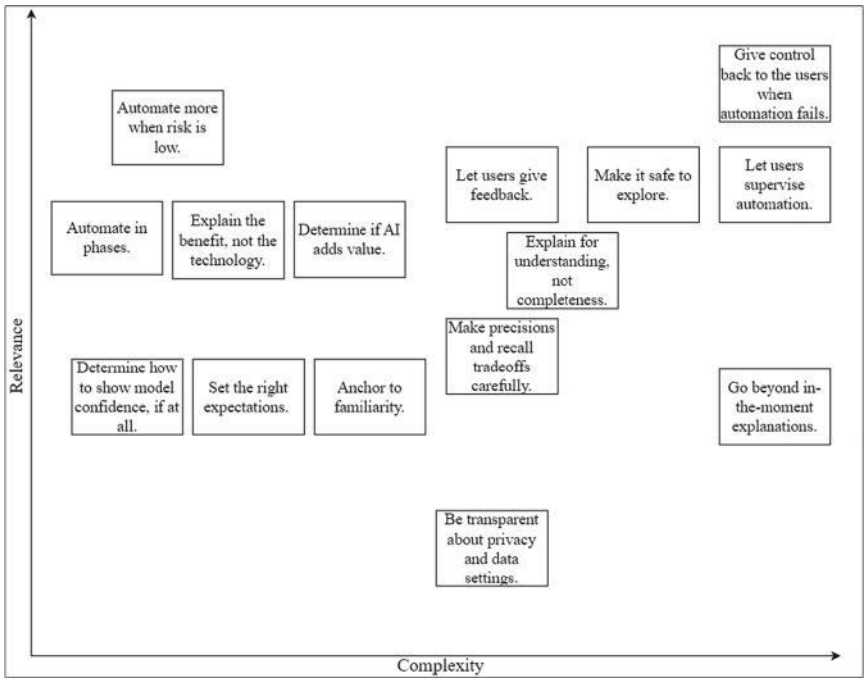


Fig. 3. Workshop 1: Assessment of the relevance and complexity of the application of end-user-centered design patterns of the People + AI Guidebook [10].

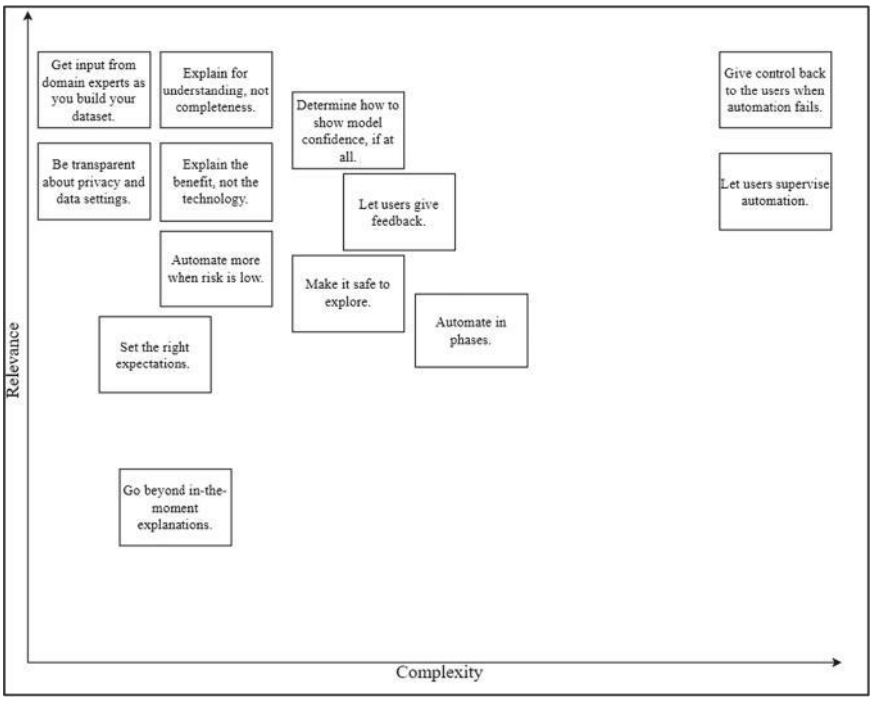


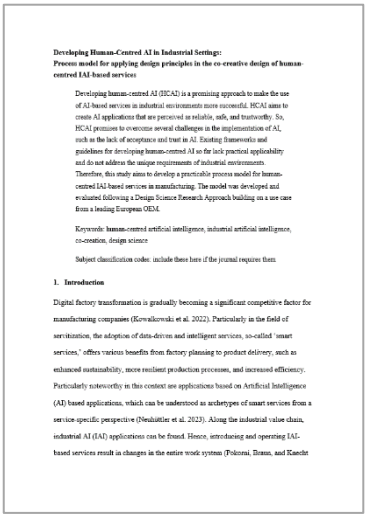
Fig. 4. Workshop 2: Assessment of the relevance and complexity of the application of end-user-centered design patterns of the People + AI Guidebook [10].

References

1. Amershi, S., et al.: Guidelines for human-AI interaction. In: Brewster, S., Fitzpatrick, G., Cox, A., Kostakos, V. (eds.) Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI 2019: CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, 04 May 2019–09 May 2019, pp. 1–13. ACM, New York (2019). https://doi.org/10.1145/3290605.3300233
2. Xu, W., Dainoff, M.J., Ge, L., Gao, Z.: Transitioning to human interaction with AI systems: new challenges and opportunities for HCI professionals to enable human-centered AI. Int. J. Hum. Comput. Interact. (2022). https://doi.org/10.1080/10447318.2022.2041900
3. Shneiderman, B.: Human-Centered AI. Oxford University Press, Oxford (2022)
4. Pokorni, B., Braun, M., Knecht, C.: Human-centred AI applications in production. Practical experience and guidelines for operational implementation strategies. Fraunhofer IAO (2021). (in German). http://publica.fraunhofer.de/dokumente/N-6249564.html.
5. Kutz, J., Neuhüttler, J., Spilski, J., Lachmann, T.: Implementation of AI technologies in manufacturing - success factors and challenges. In: The Human Side of Service Engineering. 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022), 24–28 July 2022. AHFE International (2022). https://doi.org/10.54941/ahfe1002565
6. Shneiderman, B.: Human-centered artificial intelligence: reliable, safe & trustworthy. Int. J. Hum. Comput. Interact. (2020). https://doi.org/10.1080/10447318.2020.1741118

7. Lütge, C., et al.: Automotive. AI4People-ethical guidelines for the automotive sector: fundamental requirements & practical recommendations for industry and policymakers. In: Floridini, L. (ed.) AI4People's 7 AI Global Frameworks
8. High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI, Brussels (2019). file:///C:/Users/kutz/Downloads/ai_hleg_ethics_guidelines_for_trustworthy_ai-en_87F84A41-A6E8-F38C-BFF661481B40077B_60419.pdf
9. Smit, K., Zoet, M., van Meerten, J.: A review of AI principles in practice. In: Pacific Asia Conference on Information Systems (2020)
10. Google PAIR: People + AI Guidebook. Designing human-centered AI products (2019). https://pair.withgoogle.com/guidebook/. Accessed 04 Nov 22
11. Microsoft: Guidelines for Human-AI Interaction. Microsoft HAX Toolkit (2022). https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/. Accessed 8 Feb 2023
12. Hoffmann, M.W., Drath, R., Ganz, C.: Proposal for requirements on industrial AI solutions. In: Beyerer, J., Niggemann, O., Maier, A. (eds.) Machine Learning for Cyber Physical Systems, pp. 63–72. Springer, Berlin Heidelberg (2021)
13. Kutz, J., Neuhüttler, J., Schaefer, K., Spilski, J., Lachmann, T.: Generic role model for the sys-tematic development of internal AI-based services in manufacturing. In: Bui, T.X. (ed.) Pro-ceedings of the 56th Annual Hawaii International Conference on System Sciences, Honolulu, HI, 3–6 January 2023, pp. 909–917 (2023)
14. Lee, J., Singh, J., Azamfar, M.: Industrial artificial intelligence (2019). http://arxiv.org/pdf/ 1908.02150v3
15. Wang, B., Xue, Y., Yan, J., Yang, X., Zhou, Y.: Human-centered intelligent manufacturing: overview and perspectives. Chin. J. Eng. Sci. (2020). https://doi.org/10.15302/J-SSCAE-2020.04.020
16. Abel, J., Hirsch-Kreinsen, H., Wienzek, T.: Acceptance of industry 4.0. Final report on an explorative empirical study of German industry. acatech – National Academy of Science and Engineering, Munic (2019). (in German)
17. Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., Plöd, M.: Machine learning operations (2022). https://ml-ops.org/content/mlops-principles. Accessed 10 Feb 2023
18. Huang, M.-H., Rust, R.T.: Artificial intelligence in service. J. Serv. Res. (2018). https://doi.org/10.1177/1094670517752459
19. Neuhüttler, J., Fischer, R., Ganz, W., Urmetzer, F.: Perceived quality of artificial intelligence in smart service systems: a structured approach. In: Shepperd, M., Brito e Abreu, F., Rodrigues da Silva, A., Pérez-Castillo, R. (eds.) QUATIC 2020. CCIS, vol. 1266, pp. 3–16. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58793-2_1
20. Neuhüttler, J., Hermann, S., Ganz, W., Spath, D., Mark, R.: Quality based testing of AI-based smart services: the example of Stuttgart airport. In: 2022 Portland International Conference on Management of Engineering and Technology (PICMET). 2022 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA, 07 August 2022–11 August 2022, pp. 1–10. IEEE (2022). https://doi.org/10.23919/PICMET 53225.2022.9882594
21. Lundborg, M., Gull, I.: Artificial intelligence in SMEs. In this way, AI becomes a game changer for small and medium-sized enterprises. A survey by Mittelstand-Digital Begleitforschung on behalf oft the Federal Ministry of Economic Affairs and Climate Action. wik consult, Bad Honnef (2021). (in German). https://www.mittelstand-digital.de/MD/Redaktion/DE/Publik ationen/ki-Studie-2021.pdf?__blob=publicationFile&v=5. Accessed 24 Jan 2022

3.5 Publikation 5: Developing Human-Centred AI in Industrial Settings: Process model for applying design principles in the co-creative design of human-centred IAI-based services



Developing Human-Centred AI in Industrial Settings: Process model for applying design principles in the co-creative design of human-centred IAI-based services

Developing human-centred AI (HCAI) is a promising approach to make the use of AI-based services in industrial environments more successful. HCAI aims to create AI applications that are perceived as reliable, safe, and trustworthy. So, HCAI promises to overcome several challenges in the implementation of AI, such as the lack of acceptance and trust in AI. Existing frameworks and guidelines for developing human-centred AI so far lack practical applicability and do not address the unique requirements of industrial environments. Therefore, this study aims to develop a practicable process model for human-centred IAI-based services in manufacturing. The model was developed and evaluated following a Design Science Research Approach building on a use case from a leading European OEM.

Keywords: human-centred artificial intelligence, industrial artificial intelligence, co-creation, design science

Subject classification codes: include these here if the journal requires them

1. Introduction

Digital factory transformation is gradually becoming a significant competitive factor for manufacturing companies (Kowalkowski et al. 2022). Particularly in the field of servitization, the adoption of data-driven and intelligent services, so-called 'smart services,' offers various benefits from factory planning to product delivery, such as enhanced sustainability, more resilient production processes, and increased efficiency. Particularly noteworthy in this context are applications based on Artificial Intelligence (AI) based applications, which can be understood as archetypes of smart services from a service-specific perspective (Neuhüttler et al. 2023). Along the industrial value chain, industrial AI (IAI) applications can be found. Hence, introducing and operating IAI-based services result in changes in the entire work system (Poksinski, Braun, and Knecht

Titel: tbd

Autoren: Janika Kutz, Jens Neuhüttler, Nicole Gladilov, Jan Spilski, Katharina Hölzle, Thomas Lachmann

Status: Eingereichtes Manuskript

Jahr der Veröffentlichung: unbekannt

Inhalt:

Eine Herausforderung der Umsetzung von HCAI in der industriellen Praxis sind fehlende Methoden und Vorgehensmodelle zur praktischen Anwendung. Außerdem sind bestehende Ansätze nicht angepasst an industrielle Rahmenbedingungen. Beide Herausforderungen sollen durch das „Vorgehensmodell zur Nutzung von Design-Prinzipien in der ko-kreativen Gestaltung menschenzentrierter KI-basierter Services“, das im Rahmen eines Design-Science basierten Forschungsansatzes entwickelt wurde, adressiert werden. Das Vorgehensmodell integriert durch seinen ko-kreativen Charakter verschiedene Stakeholder in den KI-Entwicklungsprozess und bildet eine Basis zur Auswahl und Integration von HCAI-Design-Prinzipien für industrielle KI-Use-Cases.

Einordnung in den Gesamtzusammenhang der Arbeit:

- Kontextspezifische Anpassung der allgemeinen Erkenntnisse im Bereich HCAI
- Vorgehensmodell zur Anwendung von HCAI-Gestaltungsmethoden in der industriellen Praxis wurde entwickelt

Developing Human-Centred AI in Industrial Settings:
Process model for applying design principles in the co-creative design of human-centred IAI-based services

Janika Kutz*[a,b], Jens Neuhüttler[a], Nicole Gladilov[a], Jan Spilski[b], Katharina Hölzle[a,c], Thomas Lachmann[b,d]

[a] *Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany;* [b] *Center for Cognitive Science, University of Kaiserslautern-Landau, Kaiserslautern, Germany;* [c] *Institute of Human Factors and Technology Management IAT, University of Stuttgart, Stuttgart, Germany;* [d] *Centro de Investigación Nebrija en Cognición, Universidad Nebrija, Nebrija, Spain*

Janika Kutz, Fraunhofer Institute for Industrial Engineering IAO (Research and Innovation Center for Cognitive Service Systems (KODIS)), Bildungscampus 9, 74076 Heilbronn | janika.kutz@iao.fraunhofer.de

Developing Human-Centred AI in Industrial Settings:
Process model for applying design principles in the co-creative design of human-centred IAI-based services

Developing human-centred AI (HCAI) is a promising approach to make the use of AI-based services in industrial environments more successful. HCAI aims to create AI applications that are perceived as reliable, safe, and trustworthy. So, HCAI promises to overcome several challenges in the implementation of AI, such as the lack of acceptance and trust in AI. Existing frameworks and guidelines for developing human-centred AI so far lack practical applicability and do not address the unique requirements of industrial environments. Therefore, this study aims to develop a practicable process model for human-centred IAI-based services in manufacturing. The model was developed and evaluated following a Design Science Research Approach building on a use case from a leading European OEM.

Keywords: human-centred artificial intelligence, industrial artificial intelligence, co-creation, design science

Subject classification codes: include these here if the journal requires them

1. Introduction

Digital factory transformation is gradually becoming a significant competitive factor for manufacturing companies (Kowalkowski et al. 2022). Particularly in the field of servitization, the adoption of data-driven and intelligent services, so-called ‘smart services,’ offers various benefits from factory planning to product delivery, such as enhanced sustainability, more resilient production processes, and increased efficiency. Particularly noteworthy in this context are applications based on Artificial Intelligence (AI) based applications, which can be understood as archetypes of smart services from a service-specific perspective (Neuhüttler et al. 2023). Along the industrial value chain, industrial AI (IAI) applications can be found. Hence, introducing and operating IAI-based services result in changes in the entire work system (Pokorni, Braun, and Knecht 2021). This transformation often leads to uncertainty among employees, which results in diverse

challenges like a lack of acceptance or even refusal of AI-based services (Kutz et al. 2022). Therefore, companies need to actively address these challenges to ensure that IAI-based services are accepted and used as intended. It is a necessary prerequisite for exploiting the potential of data and IAI to manage digital factory transformation successfully.

Currently, the development process of IAI-based services is mainly technically driven (Cremers et al. 2019; Lee, Singh, and Azamfar 2019; Xu et al. 2021), where IAI is understood as something that 'needs to connect people and things through systems' in industrial settings (Lee 2020, 19). In order to add further perspectives to this strongly technical view, we understand IAI as an application in a service offering that is embedded in a socio-technical system (Michael Sony and Subhash Naik 2020; Pokorni, Braun, and Knecht 2021). This understanding of IAI indicates that, organisational and human requirements must be considered in addition to technological issues. Human-centred AI (HCAI) is a growing research field addressing the transformation of existing environments due to the emergence of AI systems and the challenges and difficulties associated with them (Amershi et al. 2019; Xu et al. 2021). Frameworks for HCAI, such as the two-dimensional framework of Shneiderman (2020) or the three-dimensional framework of Xu et al. (2021), as well as design principles for HCAI, such as the Ethics guidelines for trustworthy AI (High-Level Expert Group on Artificial Intelligence 2019), the design principles described in the People + AI Guidebook (Google PAIR 2019), or the HAX Toolkit (Microsoft 2021), provide a good basis for the development of human-centred IAI-based services. However, even though this is a possibility to overcome the aforementioned challenges, these approaches have not been applied much in the development of IAI-based services so far (Kutz et al. 2023a). Furthermore, they have not been developed specifically for use in a manufacturing environment or adapted to do so.

To integrate HCAI into the complex, highly standardised and efficiency-focused manufacturing process, existing approaches need to be adapted to current industrial conditions (Kutz et al. 2023a). Moreover, each type of IAI-based service has unique requirements and challenges, that must also be considered.

Incorporating AI design principles into the development process is an appropriate way to address the challenges described in the context of IAI. Therefore, methods for integrating HCAI design principles into existing development processes are needed for the development of IAI-based services. Furthermore, it is crucial to promote human-centred IAI development, placing end-users at the centre of IAI-based services and proactively involving them in the development process. Methods for fostering interdisciplinary collaboration in AI development are needed to achieve this (Xu and Gao 2023). In this context, the paper aims to outline a process model for the co-creative design of human-centred IAI based on the usage of design principles.

2. Related Work

2.1 Industrial AI-based Services in manufacturing

Manufacturing companies have faced a great need for change recently. On the one hand, this is caused by shifts in international supply chains, increased market requirements, increased international competition, and a shortage of skilled workers (Kagermann and Wahlster 2022). On the other hand, the dynamics of available digital technologies, particularly AI applications, have increased significantly (Kinkel, Baumgartner, and Cherubini 2022). The development of intelligent and adaptable production structures that make it possible to mass-produce high-quality products under competitive conditions even in the face of increased uncertainty and volatility is becoming increasingly important.

Against this backdrop, the development of internal, AI-based services to increase the quality and efficiency of industrial production processes based on production data is becoming increasingly important (Kohtamäki et al. 2020). In this context, we speak of IAI-

based services, i.e., an interactive process in which people use IAI applications to complete a specific task or create value. The fields of application of AI-based services for the optimisation of production processes and factories are diverse and range, for example, from plant maintenance and servicing, condition monitoring or quality assurance, to capacity control and decision support (Hoffmann, Drath, and Ganz 2021; Lee 2020).

However, although the application potential is diverse, operational use is not yet widespread or equally advanced (Hoffmann, Drath, and Ganz 2021; Feike, Bienzeisler, and Neuhüttler 2024). There are many reasons for this. These resource barriers include a lack of knowledge and experience in dealing with AI, a lack of resources for development and implementation, and a lack of access to the required data (Jan et al. 2023). In addition, technical barriers are associated with using of AI systems, such as assumed security risks of AI-based services (Brajovic et al. 2023). Moreover, organisational challenges to overcome include the consideration and involvement of all stakeholders, a lack of competencies among the workforce, uncertainties about development and approval processes, and unclear roles and responsibilities (Kutz et al. 2022). Finally, the lack of user and employee acceptance is also a key risk in the development of AI-based services. For instance, due to concerns and fears regarding the handling of sensitive data, a loss of privacy, a lack of perceived added value, or a perceived risk of surveillance (Neuhüttler 2022).

The various challenges described, combined with the considerable impact of IAI on work processes and task layouts, underline the importance of comprehensively considering users in developing AI-based services. So far, however, the development of AI-based services has been primarily technology-driven (Wang et al. 2020), and employees as a key factor are often neglected when optimising production systems (Reiman et al. 2023). All

things considered, there is a strong need for industrial transformation to be human-centred (Grosse et al. 2023; Breque, Nul, and Petridis 2021; Lu et al. 2022).

2.2 Human Centred AI

Human-Centred AI (HCAI) is a viable approach to developing AI-based services that are accepted by employees and perceived as reliable, safe, and trustworthy (Shneiderman 2020b; Xu 2019). HCAI is increasingly discussed in the scientific community and understood as a way to respond to major challenges associated with the use of AI-based services. HCAI design enables AI technologies to realise their full potential (Xu and Gao 2023) while ensuring that AI considers and supports human capabilities (Auernhammer 2020). The goal of HCAI is to create AI-based services 'that increase automation, while amplifying, augmenting, enhancing, and empowering people to innovatively apply systems and creatively refine them' (Shneiderman 2020b, 495).

In a comprehensive literature review, Capel and Brereton (2023) analysed 257 articles from the fields of HCAI as well as human-centred Machine Learning and identified four major research areas: (1) *Explainable and Interpretable AI*; (2) *Human-Centred Approaches to Design and Evaluate;* (3) *Humans Teaming with AI;* and (4) *Ethical AI*. Considering all identified areas of research, they define HCAI as the utilising of data to enable and empower the user while revealing its underlying values, biases, limitations, and the ethics of its data collection and algorithms to encourage ethical, interactive, and contestable use. In this article, we will follow this definition of HCAI.

The first frameworks outlining the dimensions of HCAI have been published in recent years. Three frameworks are outlined in more detail in the following: The to-dimensional framework by Shneiderman (2020b), the three-dimensional framework by Xu et al. (2019), and the 4P model of HCAI Design by Olsson and Väänänen (2021).

Shneiderman published a two-dimensional framework for the design of HCAI, which encompasses two essential elements: human control and computer automation. The aim is to design AI systems that enhance and support human capabilities while also ensuring a significant degree of human oversight and a high level of automation (Shneiderman 2020b).

A three-dimensional understanding of HCAI is published by Xu et al. (2019), to 'promote a comprehensive approach, ultimately providing people with safe, efficient, healthy, and satisfying HAI solutions' (Xu 2019, 44). The three dimensions of this HCAI framework are human, technology, and ethics. According to Xu et al. (2021), all three dimensions must be addressed in research and development of AI systems to achieve the general HCAI goal of reliable, safe, and trustworthy AI applications. Across all dimensions, the approach puts people at the centre of consideration. Moreover, a complementary understanding between human intelligence and machine intelligence is required, and the three dimensions of human, technology, and ethics are considered in their interdependence.

Olsson and Väänänen (2021) published the 4P model of HCAI Design, which provides a structure for addressing AI-specific design issues. The 4P are an acronym for product, people, principles, and process as perspectives in the design process building on the 4P model from marketing and service quality literature. The *product* perspective focuses on a change from designing reactive information tools to designing proactive agents and collaborative partners. Designing for and together with various user groups, considering how the system will affect stakeholders, and taking into account their requirements and values are all essential components of addressing the *people's* viewpoint (Ozmen Garibay et al. 2023; Olsson and Väänänen 2021) The *principles* perspective in AI design incorporates the values and fundamental propositions that designers use to solve

design problems. Principles of ethical AI, such as responsibility, fairness, and explainability, are important to consider throughout the HCAI design (Olsson and Väänänen 2021; Ozmen Garibay et al. 2023). The *process* perspective in AI design involves following human-centred design principles, including identifying user and stakeholder requirements, exploring solution alternatives, and conducting user-based testing. Ethical deliberation and consideration of quality criteria and values should be encouraged throughout the design process (Olsson and Väänänen 2021). This can be achieved through the integration of human-centred design phases into the traditional lifecycle of AI product development (Ozmen Garibay et al. 2023, 408).

All of these frameworks offer a good overview of HCAI and provide first development suggestions. However, concrete application of HCAI is still underexplored in HCAI research (Xu and Gao 2023; Hartikainen et al. 2022), and HCAI approaches are rarely used in the field of IAI (Kutz et al. 2023a). Therefore, this paper focuses on a specific implementation of HCAI in industrial environments.

2.3 AI Design Principles and Guidelines

A few guidelines and design concepts have already been established to support the development of HCAI. Essentially, a distinction can be made between high-level guidelines and more action-oriented design principles or design patterns (Kutz et al. 2023b). The ‘Ethics Guidelines for Trustworthy AI’ is a comprehensive high-level guideline from the High-Level Expert Group on Artificial Intelligence (High-Level Expert Group on Artificial Intelligence 2019). This guideline defines seven requirements for trustworthy AI: Human agency and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity; Non-discrimination and fairness; Environmental and societal well-being; Accountability. This guideline aims to foster the advancement of Trustworthy AI through a human-centred approach. Although these

general guidelines provide a good orientation, they are only, to a limited extent, appropriate in context-specific scenarios (Auernhammer 2020).

More action-oriented guidelines have been, published by technology companies such as Google and Microsoft. The People and AI Guidebook by Google's PAIR Group lists 23 design patterns along with guiding questions that can be considered in the development of AI applications. These patterns are described using pattern language and serve as concise expressions for important aspects that provide solutions to typical design problems (Google PAIR 2019).

Microsoft's guideline is based on a 2019 literature review that identified 18 Human-AI design principles (Amershi et al. 2019). Microsoft's design principles can be found in an online application and have been extended by various help functions. The so-called HAX Toolkit 'is for teams building user-facing AI products. It helps you conceptualise what the AI system will do and how it will behave' (Microsoft 2021).

An orientation towards guidelines and the use of design principles can be of considerable benefit for the design of HCAI. Especially the action-oriented design principles are considered applicable by IAI experts (Kutz et al. 2023a). However, the existing literature does not discuss the integration of these design principles into practice. High-level guidelines such as the EC's guideline for trustworthy AI (High-Level Expert Group on Artificial Intelligence 2019) are considered equally important but offer little potential for direct application. Rather the principles must be adapted to production-specific circumstances, as some requirements and ethical issues are very context-specific (Auernhammer 2020) and the known guidelines and principles are mostly general and poorly tailored to specific application scenarios (Kutz et al. 2023a; Hartikainen et al. 2022; Olsson and Väänänen 2021)

According to Shneiderman (2020a), it becomes apparent that depending on the content orientation of these principles, different organisational levels are responsible for adopting them. So, it is the responsibility of the development team to incorporate the use-case-specific application of design principles into the development process. The team level is embedded in the organisational level. The organisational level influences development practice at the team level, for example through the prevailing safety culture in the use of AI, management strategies and leadership behaviour. In addition, standards at the industry level and political regulations set the general framework and have an impact on the organisations (Shneiderman 2020a).

Despite the large number of frameworks, guidelines, and principles that have already been published by academia and industry, these are hardly used by AI engineers in companies (Hartikainen et al. 2022).

2.4 Co-Creative Service Design

HCAI aims to put people at the centre of AI applications. A central starting point for this is the involvement of stakeholders in the development process, which is discussed in science under the term co-creation. ‘The participation of stakeholders in the design and implementation of the AI system can provide essential perspectives in the re-design of the social organisation such as the work system’ (Auernhammer 2020, 1324–25). In addition, end-users' participation is defined as a success factor in implementing an IAI-based service (Kutz et al. 2022).

Co-creation is considered a key success factor in service development, helping companies to increase the perceived benefits and market success of services and shorten development times (Carbonell, Rodriguez-Escudero, and Pujari 2012; Jonas 2018; Mahr, Lievens, and Blazevic 2014). Although co-creation is conceptually discussed under different terms in various disciplines such as marketing and management theory,

psychology, and innovation management, all approaches take two key perspectives into account (Greve 2019): On the one hand, co-creation in development encompasses the involvement of various stakeholders across company boundaries (in the sense of open innovation). On the other hand, the involvement of co-creators who are themselves affected by the developed service. Since IAI-based services are developed within the company with the aim of supporting employees in production and beyond in their work, this article will focus in particular on the involvement of end-users. We define end-users as: Employees who use the IAI-based service to fulfil their work tasks.

Involving end-users in the development supports the identification of context-specific requirements as well as manifest and latent needs and fears (Witell, Gustafsson, and D. Johnson 2014; Ruiz-Alba et al. 2019). Particularly against the background of perceived risks and uncertainties, early identification and consideration in the development of AI-based services is an important measure for creating accepted and beneficial solutions (Auernhammer 2020). At the same time, however, this also represents one of the greatest challenges in the development of IAI-based services (Hartikainen et al. 2022).

In principle, end-users can be meaningfully involved in all development phases, from brainstorming and requirements analysis to testing prototypes and finished solutions. Based on an empirical study, Alam (2002) distinguishes four levels of co-creation based on the intensity of collaboration with end-users, which, similarly, can also be found in recent contributions to participatory AI development (Berditchevskaia, Peach, and Malliaraki 2021). According to Alam (2002), the first stage of co-creation is passive involvement (1). In this case, the company only accepts proactive suggestions from end-users, end-user involvement is correspondingly low, and no co-creation takes place. The second stage, information and feedback (2), describes the efforts of companies to actively participate in specific issues within the development process and thus represents a first step

towards co-creation. The third stage, comprehensive consultation (3) differs from this in that the development team makes extensive use of end-user resources throughout the entire development process as part of a defined process, e.g., as part of in-depth interviews, focus group surveys, or group discussions. The final stage, representation (4), describes the involvement of end-users in the service development team, which significantly increases the intensity of co-creation across all phases.

2.5 Purpose of Research

Current studies indicate that there is a great need for human-centred IAI-based services. In previous research, however, little attention is paid to the possibilities of adapting the mainly technical development process of IAI to a more human-centred one (Hoffmann, Drath, and Ganz 2021; Kutz et al. 2023a; Lee, Singh, and Azamfar 2019). In summary, this means that with the increasing importance of AI for manufacturing companies, there is a growing need for systematic approaches and methods to successfully develop human-centred IAI-based services and embed them in a specific application context. Therefore, the objective of this study is to develop a practicable process model that supports the design of human-centred IAI-based services in manufacturing integrating aspects co-creative service design, and AI design principles in one approach (fig. 1). The research domains under consideration should not be regarded as independent entities but rather as having areas of intersection. Therefore, it is appropriate to integrate these three viewpoints into an integrated method.

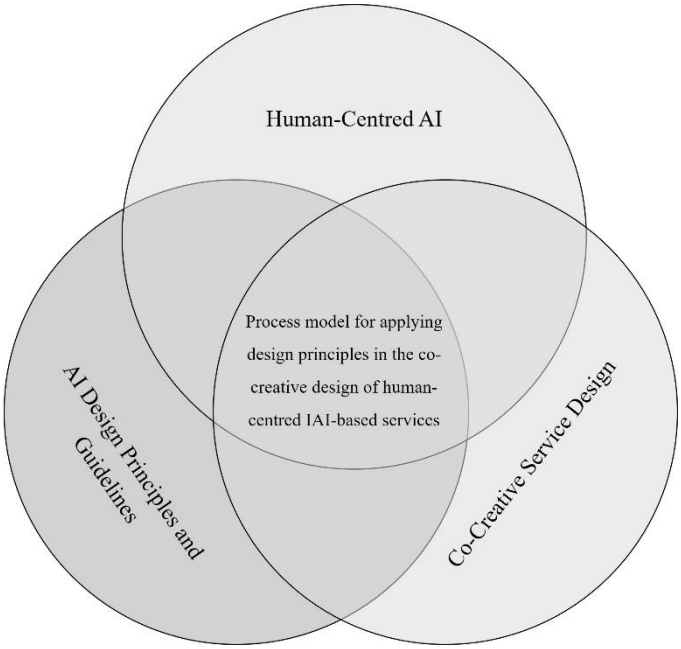


Figure 1. Overview of integrated research fields in the development of the process model.

3. Method

The aim of this study is to develop a process model that supports manufacturing companies in the development of HCAI. The process model was developed using the principles of method engineering, characterised by the systematic development of reusable and adaptable methods, techniques, and tools for the development of information systems (Brinkkemper 1996). So-called reference process models play a key role , as they provide a generic representation of knowledge about tasks and activities that can be adapted to different development projects and application areas (Goldkuhl and Karlsson 2020). They serve to collect and structure knowledge on a specific topic and can thus be used for knowledge exchange between different users and application areas (Feike et al. 2023). Recurring use generates empirical knowledge and enables low-cost implementation. The process model development procedure follows a Design Science Research (DSR) approach (fig. 2). Accordingly, a multi-stage research process was initiated that lasted about 8 months, from January 2023 to August 2023.

The need for research was defined based on previous research findings (see sec. 1 & 2).

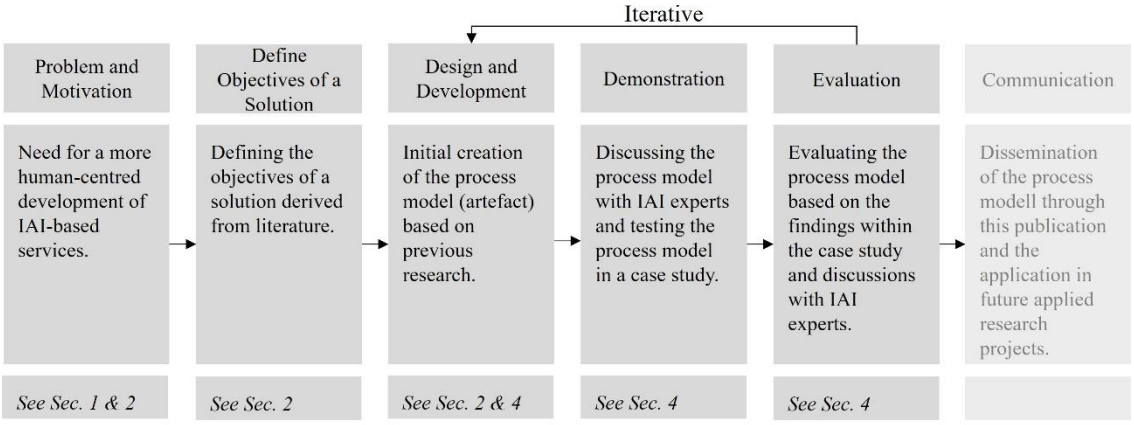


Figure 2. Visualisation of the research process following Peffers et al. 2007

Afterwards, potential solutions were identified through desk research (see sec. 2).

The process model was then developed in an iterative process through the design and development phase, the demonstration phase, and the evaluation phase. Based on relevant findings from the literature and considering the defined objectives, an initial concept of the process model for integrating HCAI into the IAI development process was created.

Expert opinions on the model were then sought and incorporated into the model in a first iteration loop. In the second iteration, then, an IAI use case at a leading European original equipment manufacturer (OEM) in the automotive sector served as the foundation for the testing and evaluation of the process model. For this case study, a suitable IAI use case was chosen in collaboration with IAI experts from the OEM. The chosen IAI use case is a service for automatic quality control of weld seams in the field of body construction. The objective of the use case is to determine the quality of weld seams by analysing process data. It was still in its early phases of development, despite the fact that its core technological feasibility had already been demonstrated. However, more phases of development were still required, especially for usage in the operating manufacturing process. This involves, for example, technical validation under production settings and the

adaptation of existing manufacturing and working processes. After preliminary discussions with two members of the use case's core development team (an expert in welding technology and an ML engineer), the concept was adapted to the selected IAI use case. Afterwards, selected process steps were tested with a group of seven OEM representatives who will be affected by the introduction of the IAI use case in the future. The participants of the OEM were rather diverse, including:

- domain expert from the field of body construction and maintenance (N = 4)
- potential end-user of the IAI use case (N = 1)
- expert for quality assurance and management (N = 1)
- ML-engineer of the use case (N = 1)

Following the testing, an evaluation took place with the participants. To validate the process model, the results were reflected in a subsequent meeting with the technical experts, and the experiences were classified regarding future IAI development projects. Afterwards, the initial concept was modified and transformed into a comprehensive process model.

4. Results and building the process model

4.1 Design and Development of the process model

The first version of the process model was conceptualised based on the perspectives of the 4P model of AI Design (Olsson and Väänänen 2021), which supports the open exchange of diverse stakeholders who will be affected by the introduction of an AI use case in the future. By integrating co-creative elements into the process, the aim is to ensure that the interests of various stakeholder groups are considered in IAI development. Co-creation is also intended to ensure that the design principles for HCAI are considered in the process. Therefore, the core of the process is the co-creative prioritisation, selection, concretisation,

implementation, and evaluation of HCAI design principles tailored to the requirements of the use case. These tasks are integrated into the process model in the second phase as well as the third phase (task 7 – task 12). Task 6, "Identification of challenges and desires of stakeholder groups" was also previously implemented to ensure the concerns and wishes and the resulting requirements for the implementation of the use case and the selection of context-specific HCAI design principles. To inform the participants in advance, a general introduction to the upcoming process steps is required, which is included in the process model as task 5. Derived from the literature, a process model consisting of eight tasks was designed in phase two of the DSR approach.

4.2 Demonstration of the process model

In the first iteration, the process model initially conceptualised during the design and development phase was presented to two IAI experts. The suggestions of the company's AI specialists were also integrated into the process model. For example, it was determined that a brief introduction to the use case as well as an introduction to the field of HCAI are necessary to give all involved stakeholders a common basis of information. These suggestions were integrated and made part of task 5 and task 7.

In addition, the experts also made it clear that the implementation of the approach is only possible if the time required to involve end-users remains within a manageable amount of time. Moreover, it was recommended to integrate a continuous dialogue with diverse stakeholders, with a particular focus on potential end-users, into the ongoing development process. Nevertheless, in the preliminary discussions with the developers responsible for the selected IAI use case, positive feedback was expressed on the general structure of the concept and its co-creative nature. All in all, the preliminary coordination with two experts from the core development team of the selected IAI use case proved to be valuable for the successful execution of the subsequent process steps. Therefore, it was

integrated into the process model in the form of an initialisation and mobilisation phase (task 1 – task 2).

The subsequent testing phase tested selected steps of the process model, particularly those focused on co-creatively selecting and specifying HCAI design principles. In our case study, it was determined at the organisational level that the design principles of the People + AI guidebook (Google PAIR 2019) are used as the basis for a co-creative prioritisation and selection of principles. The following four HCAI design principles from the guidebook were prioritised and selected for further implementation in our case study: (1) "Give control back to the user when automation fails", (2) "Anchor on familiarity", (3) "Make it safe to explore", and (4) "Determine if AI adds value". During the test phase, it became clear that some of the defined design principles have to be considered by company-specific agreements. Concerning our test case, this pertains to design principles regarding privacy and data security. By including the issue of company agreements as an influencing factor at the organisational level, this insight has been considered.

The test-phase also demonstrated the feasibility of early stakeholder involvement in a concise format in an industrial setting. It is important to schedule the workshop well in advance to ensure that people working in production-related areas can attend.

4.3 Evaluation of the process model

In general, the process model received positive ratings from the participants and the IAI experts. The following points were particularly emphasised by them:

- The process should be conducted at an early stage of the development of an IAI-based service.

- The early involvement of various stakeholders and the opportunity for joint discussions regarding the specific IAI use case were particularly effective and well-received.
- Presentation of the topic HCAI as well as the design principles is useful to sensitise tech-savvy people to the need for a more human-centred design of IAI-based services.
- Exchanging and engaging with different stakeholders about the IAI use case was considered important for a comprehensive understanding and future successful implementation.
- The IAI use case's risks, obstacles, and criticisms were openly addressed through the co-creative process.
- Prioritisation is essential and deliberate since, in industrial practice, not all HCAI design principles can be considered simultaneously.
- It is necessary to adapt selected design principles to a specific use case.
- It was acknowledged that to overcome biases and operational blindness, an external workshop moderation by an HCAI expert is helpful.

5. Process model for applying design principles in the co-creative design of human-centred IAI-based services

The developed approach consists of three levels (team, organisation, and industry), which are based on Shneiderman's (2020) model of governance structures for HCAI.

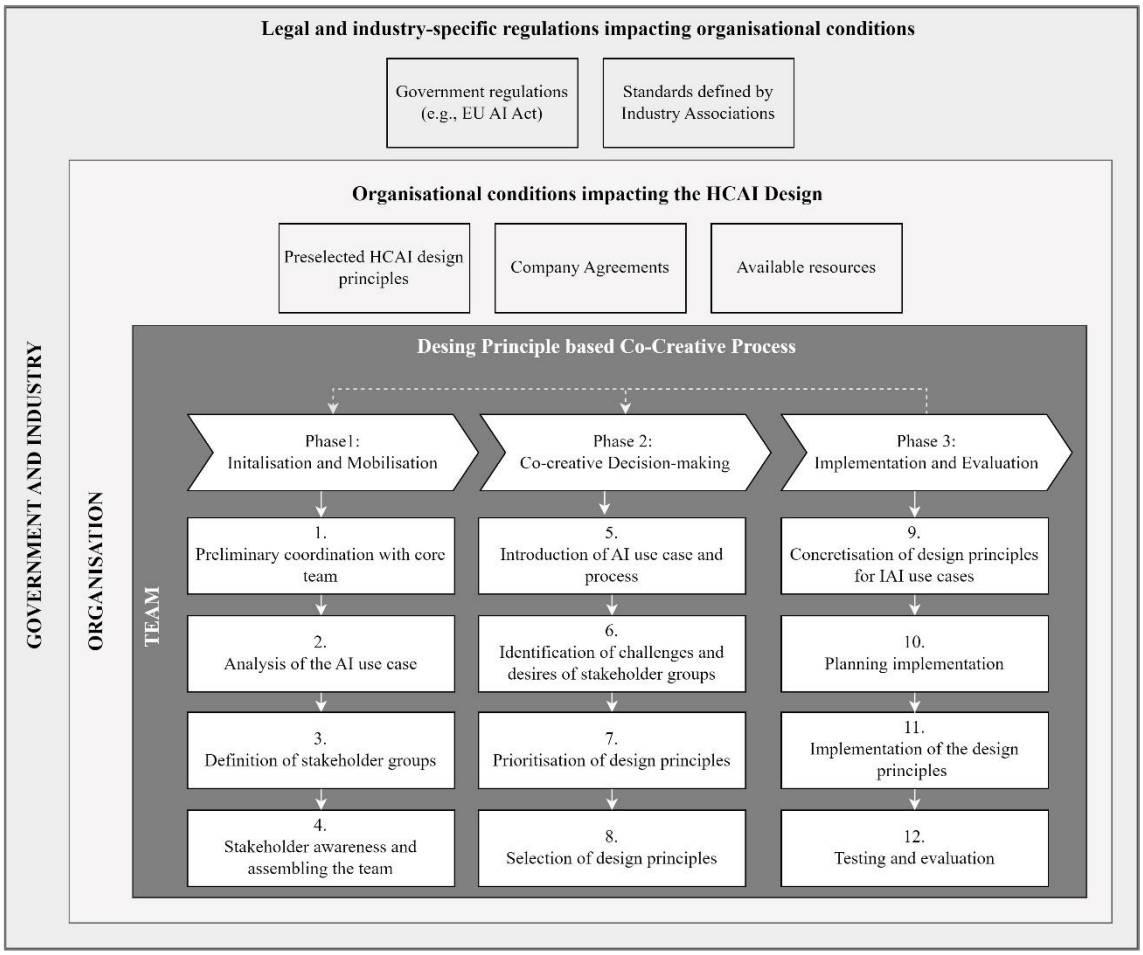


Figure 3. Process model for applying design principles in the co-creative design of human-centred IAI-based services.

The co-creative process for designing human-centred IAI is at the core of the approach, which is applied at the team level. This process is divided into three phases: (1) Initialisation and mobilisation; (2) Co-creative decision-making; and (3) Implementation and Evaluation. Each phase in turn consists of four successive tasks derived from the literature or based on the results of the case study. The process steps are described in more detail in the following chapters (5.1 – 5.3).

The implementation of this process is influenced by the organisational and industry level, either directly or indirectly. The organisational level establishes the necessary conditions for the execution of a co-creative development process, and company agreements made there can directly influence the selection and application of design

principles. Furthermore, at the organisational level, we recommend making a preselection from the abundance of already published design principles, which are assessed as particularly relevant for the company context (Kutz et al. 2023a). On the other hand, the organisation must comply with and ensure the implementation of applicable government requirements and industry standards. This includes, for example, regulations related to the governance of AI.

5.1 Phase 1. Initialisation and mobilisation

The aim of the first phase is to create awareness of the necessity of HCAI. The process is initialised, and the core team and members of designated stakeholder groups are mobilised to engage in the co-creative process.

Table 2. Description Phase 1: Initialisation and mobilisation

Task	Description
1. Preliminary coordination with core team	The aim is to initialise the process. This entails informing and explaining the proposed approach to the core development team and clarifying the added value and necessity of HCAI development methods.
2. Analysis of the IAI use case	Together with the main development team, the current development status of the AI use case is analysed in order to gain a better understanding of the initial situation.
3. Definition of stakeholder groups	The aim is to identify a broad range of stakeholder groups who are directly or indirectly affected by the IAI use case through their different job roles.
4. Stakeholder awareness and assembling the team	The stakeholder groups must be informed about the process, and their awareness of it must be raised. Based on the defined stakeholder groups, a diverse team is assembled for collaboration in further process steps.

5.2 Phase 2. Co-creative decision-making

The second phase provides a platform for selected employees, in the following named as participants, to provide feedback, express concerns, and ask questions related to the IAI use case.

This ensures that different perspectives are considered and addressed during the further development process. Table 3 describes the tasks included in the second process step.

Table 3. Description Phase 2: Co-creative decision-making

Task	Description
5. Introduction of AI use case and process	The aim is to improve understanding of the IAI use case and the process. This can be achieved through a presentation of the use case and a joint discussion with the participants. The next steps of the co-creative process are also explained.
6. Identification of challenges and desires of stakeholder groups	In a collaborative brainstorming session, challenges and desires related to the AI use case are identified with the participants.
7. Prioritisation of Design Principles	Firstly, an introduction to the field of HCAI is given, and the design principles are presented. Afterwards, the HCAI design principles are prioritised collaboratively by evaluating and ranking them based on relevance, feasibility, and impact on achieving a human-centred IAI design.
8. Selection of design principles	Based on the prioritisation and in comparison, with the previously identified challenges and desires, design principles are selected collaboratively for implementation. It is essential to ensure that a realistically implementable number of principles is selected.

Our test and evaluation phase has shown that the tasks described in phase 2 can be realised in a half-day workshop. This enables the participation of as many different stakeholder groups as possible, as the time required for the individual participants is feasible, which also gives direct employees the opportunity to participate.

5.3 Phase 3. Implementation and Evaluation

The aim of the third phase is to implement the design principles. For this, it is necessary to make use-case-specific adjustments and to specify the principles with regard to concrete implementation.

Table 4. Description Phase 3: Implementation and Evaluation

Task	Description
9. Concretisation of design principles for IAI use case	The selected design principles are concretised to the IAI use case by identifying use case-specific requirements for the technical and organisational implementation. The various stakeholder perspectives should also be considered in this step.
10. Planning implementation	The chosen design principles are translated into actionable steps. This includes identifying the necessary tasks, resources, and strategies to ensure the successful integration of the design principles into the IAI use case.
11. Implementation of the design principles	The selected design principles are implemented and integrated into the IAI or its development process. If required, this can also be realised in several stages, from prototype to production readiness.
12. Testing and Evaluation	The implementation of the design principles should be tested and evaluated in terms of their intended effect. An ongoing evaluation with various stakeholders to ensure that the IAI-based system remains human-centred throughout its lifecycle is recommended.

The process model presented extends over three phases and comprises a total of twelve sequential tasks. After completing a phase or after completing a task, it is possible to re-enter an earlier process step and adjust if necessary. At the same time, the experiences from the process steps should be reflected back to the organisational level in order to obtain feedback on the appropriateness of the pre-selected design principles, existing company agreements, and resource requirements.

6. Discussion and Conclusion

The integration of a service-oriented perspective on value creation has become increasingly important in the manufacturing sector recently. Moreover, the introduction of AI technologies in particular has brought a fundamental change in the servitization of manufacturing processes (Neuhüttler et al. 2023). There are many advantages associated with the utilisation of IAI-based services. For instance, they promote productivity, strengthen resilience, and enhance work procedures. However, AI technologies can only

realise their full potential if they are successfully used in the production process. Nevertheless, there are currently numerous obstacles that make this challenging (Kutz et al. 2022; Jan et al. 2023). One way to address these challenges is to transform the traditionally technology-driven development process of IAI-based services into a more human-centred one. So, the aim of this study was to define a process model to integrate aspects of the HCAI design into a traditionally technology-driven development process of IAI-based services. Based on a design science research approach, the ‘Process model for applying design principles in the co-creative design of human-centred IAI-based services’ was developed. The approach was developed by incorporating literature from the disciplines of HCAI, Co-Creative Service Design, and AI Design Principles and Guidelines, as well as the outcomes of the case study.

The process model allows for human-centred IAI design by integrating this approach to the usually technical development process. This creates an incorporates development practice that includes both points of view (Auernhammer 2020). The novelty of the approach is that it can be understood as a method for the realisation of the practical application of HCAI in the manufacturing environment. This addresses two weaknesses of previous HCAI frameworks: lack of practicality (Capel and Brereton 2023; Xu and Gao 2023) and lack of contextualisation (Kutz et al. 2023a; Olsson and Väänänen 2021).

The process model is based on our understanding of IAI as part of a socio-technical system in which IAI is understood as a tool that improves human capabilities and their working conditions. This is accompanied by our understanding of HCAI design, which can be fostered through co-creation (Auernhammer 2020; Witell, Gustafsson, and D. Johnson 2014; Ruiz-Alba et al. 2019). Accordingly, the process model is designed to engage diverse stakeholders by integrating them into the process of HCAI design. The systematic identification of stakeholder groups affected by the AI use case and the involvement of

representatives from these groups in the process ensures that diverse perspectives are considered. So, their feedback, concerns, and demands related to the IAI use case can be considered. The implementation of the process model can be seen as a practical guide to realising one of the identified success factors, ‘Co-determination and participation of end-users in development and implementation’ of a successful AI implementation in an industrial setting (Kutz et al. 2022).

Phase 2 and phase 3 of the process model focus on identifying and implementing appropriate design principles for HCAI for specific IAI use cases. According to the IAI experts, the prioritisation and selection of certain principles can ensure their implementation, as the given conditions in practice do not allow for the implementation of all principles, and not every use case requires the consideration of the same principles. This underlines the necessity of decomposing current comprehensive frameworks into practical methodologies (Xu and Gao 2023).

In general, it can be determined that the process model described is highly applicable in an industrial environment. Firstly, it is based on a real-life use case and considers the experience of IAI experts. Secondly, because it systematically reduces the overwhelming number of principles to a more manageable level, with a focus on the most relevant principles in line with the specific use case. Nevertheless, the process model has only been tested in one use case so far. Accordingly, the process model should be evaluated on the basis of its application in other use cases. In addition, a complementary study would be appropriate to validate the approach based on expert opinion. In addition, future research should investigate which factors at the organisational and team level are facilitators or barriers to the implementation of the process model. This will enable targeted measures to be taken to promote its actual application at manufacturing companies.

Moreover, the approach in its current version addresses the design process in particular. The actual application of selected design principles is not part of the approach. Further research is needed to determine the feasibility of applying specific design principles in an industrial setting (Kutz et al. 2023a).

Furthermore, it is worth noting that the process model focuses on the team level, and the organisational level is not extensively developed. To enhance the effectiveness and inclusivity of the approach, it would be beneficial to further develop and incorporate co-creative procedures at the organisational level as well. A key aspect that should be addressed is the formalisation of a process for the collaborative pre-selection of HCAI design principles within an organisation. By engaging stakeholders from different departments, including representatives from user groups, domain experts, IAI experts, management, and ethics experts, a more diverse range of perspectives and expertise can be included. This would contribute to a more comprehensive and inclusive approach to human-centred IAI design.

As well, there will be a need in the future to analyse governmental regulations regarding their effect on the design of HCAI. In the published proposal of the EU AI ACT (European Commission 2021), defined requirements like transparency and provision of information to users (Article 13) and human oversight (Article 14) can influence the development process by ensuring that human-centred considerations are integrated into AI systems. The extent to which future certification processes will include evidence of the HCAI design methods cannot be determined at this time. However, in the future, it might be necessary to integrate the requirements for external audits into the framework.

The study's focus was on IAI-based services, although HCAI design is equally important in other disciplines. Future research should study the extent to which the process model can be implemented in other disciplines, e.g., GovTech or education.

To summarise, the process model presented is a way of extending the primarily technical development process of IAI applications to include human-centred aspects, with the aim for creating IAI-based services that are usable and align with human values, fostering trust and acceptance. The process model can be understood as a guide for integrating HCAI design principles into IAI-based services, while considering the needs of different stakeholders. In this way, it contributes to making the results of research in the field of HCAI more accessible to practical application.

References

Alam, I. 2002. “An Exploratory Investigation of User Involvement in New Service Development.” *Journal of the Academy of Marketing Science* 30 (3): 250–61. doi:10.1177/0092070302303006.

Amershi, Saleema, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh et al. 2019. “Guidelines for Human-AI Interaction.” In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, edited by Stephen Brewster, Geraldine Fitzpatrick, Anna Cox, and Vassilis Kostakos, 1–13. New York, NY, USA: ACM.

Auernhammer, Jan. 2020. “Human-centered AI: The role of Human-centered Design Research in the development of AI.” In *Proceedings of DRS 2020, Vol. 1: Synergy, Situations*, edited by Stella Boess, Ming Cheung, and Rebecca Cain. London: Design Research Society.

Berditchevskaia, Aleks, Kathy Peach, and Eirini Malliaraki. 2021. “Participatory AI for humanitarian innovation: a briefing paper.” Accessed September 10, 2023. https://media.nesta.org.uk/documents/Nesta_Participatory_AI_for_humanitarian_innovation_Final.pdf.

Brajovic, Danilo, Niclas Renner, Vincent P. Goebels, Philipp Wagner, Benjamin Fresz, Martin Biller, Mara Klaeb, Janika Kutz, Jens Neuhuettler, and Marco F. Huber. 2023. “Model Reporting for Certifiable AI: A Proposal from Merging EU Regulation into AI Development.” http://arxiv.org/pdf/2307.11525v1.

Breque, Maija, Lars de Nul, and Athanasios Petridis. 2021. *Industry 5.0: Towards a sustainable, human-centric and resilient European industry.* R&I Paper Series, policy brief. Luxembourg: Publications Office of the European Union.

Brinkkemper, Sjaak. 1996. “Method engineering: engineering of information systems development methods and tools.” *Information and Software Technology* 38 (4): 275–80. doi:10.1016/0950-5849(95)01059-9.

Capel, Tara, and Margot Brereton. 2023. “What Is Human-Centered About Human-Centered AI? a Map of the Research Landscape.” In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, edited by Albrecht Schmidt, 1–23. ACM Digital Library. New York,NY,United States: Association for Computing Machinery.

Carbonell, Pilar, Ana I. Rodriguez-Escudero, and Devashish Pujari. 2012. “Performance effects of involving lead users and close customers in new service development.” *Journal of Services Marketing* 26 (7): 497–509. doi:10.1108/08876041211266440.

Cremers, Armin, Alex Englander, Markus Gabriel, Michael Mock, Maximilian Poretschkin, Julia Rosenzweig, Frauke Rostalski et al. 2019. “Vertrauenswürdiger Einsatz von Künstlicher Intelligenz: Handlungsfelder aus philosophischer, ethischer, rechtlicher und technologischer Sicht als Grundlage für eine Zertifizierung von Künstlicher Intelligenz.” Accessed September 14, 2023. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf.

European Commission. 2021. “Proposal for a Regulation of the European Parliament and the Council: Laying down harmonised rules on artificial intelligence.” Accessed February 09, 2022. file:///C:/Users/kutz/Downloads/regulation_ai_875509BF-C386-0D30-2CB7E56A798BA4EA_75788.pdf.

Feike, Maximilian, Bernd Bienzeisler, and Jens Neuhüttler. 2024. “Künstliche Intelligenz aus Sicht von Unternehmen.”.

Goldkuhl, Göran, and Fredrik Karlsson. 2020. “Method Engineering as Design Science.” *JAIS* 21 (5): 1237–78. doi:10.17705/1jais.00636.

Google PAIR. 2019. “People + AI Guidebook: Designing human-centered AI products.” Accessed 04.11.22. https://pair.withgoogle.com/guidebook/.

Greve, Katharina. 2019. “Facilitating Co-Creation in Living Labs.” Apollo - University of Cambridge Repository.

Grosse, Eric H., Fabio Sgarbossa, Cecilia Berlin, and W. P. Neumann. 2023. “Human-centric production and logistics system design and management: transitioning from Industry 4.0 to Industry 5.0.” *International Journal of Production Research* 61 (22): 7749–59. doi:10.1080/00207543.2023.2246783.

Hartikainen, Maria, Kaisa Väänänen, Anu Lehtiö, Saara Ala-Luopa, and Thomas Olsson. 2022. “Human-Centered AI Design in Reality: a Study of Developer Companies’ Practices.” In *Nordic Human-Computer Interaction Conference (NordiCHI’22)*, 1–11. ACM Digital Library. New York,NY,USA: ACM.

High-Level Expert Group on Artificial Intelligence. 2019. “Ethics Guidelines for Trustworthy AI.” Accessed February 09, 2024. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Hoffmann, Martin W., Rainer Drath, and Christopher Ganz. 2021. “Proposal for requirements on industrial AI solutions.” In *Machine Learning for Cyber Physical Systems*, edited by Jürgen Beyerer, Oliver Niggemann, and Alexander Maier, 63–72: Springer Berlin Heidelberg.

Jan, Zohaib, Farhad Ahamed, Wolfgang Mayer, Niki Patel, Georg Grossmann, Markus Stumptner, and Ana Kuusk. 2023. “Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities.” *Expert Systems with Applications* 216: 119456. doi:10.1016/j.eswa.2022.119456.

Jonas, Julia M. 2018. *Stakeholder Integration in Service Innovation.* Wiesbaden: Springer Fachmedien Wiesbaden.

Kagermann, Henning, and Wolfgang Wahlster. 2022. “Ten Years of Industrie 4.0.” *Sci* 4 (3): 26. doi:10.3390/sci4030026.

Kinkel, Steffen, Marco Baumgartner, and Enrica Cherubini. 2022. “Prerequisites for the adoption of AI technologies in manufacturing – Evidence from a worldwide sample of manufacturing companies.” *Technovation* 110: 102375. doi:10.1016/j.technovation.2021.102375.

Kohtamäki, Marko, Vinit Parida, Pankaj C. Patel, and Heiko Gebauer. 2020. “The relationship between digitalization and servitization: The role of servitization in capturing the financial potential of digitalization.” *Technological Forecasting and Social Change* 151: 119804. doi:10.1016/j.techfore.2019.119804.

Kowalkowski, Christian, Bård Tronvoll, David Sörhammar, and Alexey Sklyar. 2022. “Digital Servitization : How Data-Driven Services Drive Transformation.” https://www.diva-portal.org/smash/record.jsf?pid=diva2:1623889. Accessed December 31, 2021.

Kutz, Janika, Jens Neuhüttler, Bernd Bienzeisler, Jan Spilski, and Thomas Lachmann. 2023a. “Human-Centered AI for Manufacturing – Design Principles for Industrial AI-Based Services.” In *Artificial Intelligence in HCI*. Vol. 14050, edited by Helmut Degen and Stavroula Ntoa, 115–30. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland.

Kutz, Janika, Jens Neuhüttler, Jan Spilski, and Thomas Lachmann. 2022. “Implementation of AI Technologies in manufacturing - success factors and challenges.” In *The Human Side of Service Engineering*. AHFE International: AHFE International.

Kutz, Janika, Jens Neuhüttler, Jan Spilski, and Thomas Lachmann. 2023b. “AI-based Services - Design Principles to Meet the Requirements of a Trustworthy AI.” In *The Human Side of Service Engineering*. AHFE International: AHFE International.

Lee, Jay. 2020. *Industrial AI.* Singapore: Springer Singapore.

Lee, Jay, Jaskaran Singh, and Moslem Azamfar. 2019. “Industrial Artificial Intelligence.” http://arxiv.org/pdf/1908.02150v3.

Lu, Yuqian, Hao Zheng, Saahil Chand, Wanqing Xia, Zengkun Liu, Xun Xu, Lihui Wang, Zhaojun Qin, and Jinsong Bao. 2022. “Outlook on human-centric manufacturing towards Industry 5.0.” *Journal of Manufacturing Systems* 62: 612–27. doi:10.1016/j.jmsy.2022.02.001.

Mahr, Dominik, Annouk Lievens, and Vera Blazevic. 2014. “The Value of Customer Cocreated Knowledge during the Innovation Process.” *Journal of Product Innovation Management* 31 (3): 599–615. doi:10.1111/jpim.12116.

Michael Sony, and Subhash Naik. 2020. “Industry 4.0 integration with socio-technical systems theory: A systematic review and proposed theoretical model.” *Technology in Society* 61. doi:10.1016/j.techsoc.2020.101248.

Microsoft. 2021. “Guidelines for Human-AI Interaction: Microsoft HAX Toolkit.” Accessed February 08, 2024. https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/.

Neuhüttler, Jens. 2022. “Ein Verfahren zum Testen der wahrgenommenen Qualität in der Entwicklung von Smart Services.” Fraunhofer-Gesellschaft.

Neuhüttler, Jens, Maximilian Feike, Janika Kutz, Christian Blümel, and Bernd Bienzeisler. 2023. “Digital Factory Transformation from a Servitization Perspective: Fields of Action for Developing Internal Smart Services.” *Sci* 5 (2): 22. doi:10.3390/sci5020022.

Olsson, Thomas, and Kaisa Väänänen. 2021. “How does AI challenge design practice?” *Interactions)* 28 (4): 62–64. doi:10.1145/3467479.

Ozmen Garibay, Ozlem, Brent Winslow, Salvatore Andolina, Margherita Antona, Anja Bodenschatz, Constantinos Coursaris, Gregory Falco et al. 2023. “Six Human-Centered

Artificial Intelligence Grand Challenges." *International Journal of Human–Computer Interaction* 39 (3): 391–437. doi:10.1080/10447318.2022.2153320.

Peffers, Ken, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. 2007. "A Design Science Research Methodology for Information Systems Research." *Journal of Management Information Systems* 24 (3): 45–77. doi:10.2753/MIS0742-1222240302.

Pokorni, Bastian, Martin Braun, and Christian Knecht. 2021. "Menschenzentrierte KI-Anwendungen in der Produktion: Praxiserfahrungen und Leitfaden zu betrieblichen Einführungsstrategien." http://publica.fraunhofer.de/dokumente/N-6249564.html.

Reiman, Arto, Jari Kaivo-oja, Elina Parviainen, Esa-Pekka Takala, and Theresa Lauraeus. 2023. "Human work in the shift to Industry 4.0: a road map to the management of technological changes in manufacturing." *International Journal of Production Research*, 1–18. doi:10.1080/00207543.2023.2291814.

Ruiz-Alba, José L., Anabela Soares, Miguel A. Rodríguez-Molina, and Dolores M. Frías-Jamilena. 2019. "Servitization strategies from customers' perspective: the moderating role of co-creation." *JBIM* 34 (3): 628–42. doi:10.1108/JBIM-02-2017-0028.

Shneiderman, Ben. 2020a. "Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems." *ACM Trans. Interact. Intell. Syst.* 10 (4): 1–31. doi:10.1145/3419764.

Shneiderman, Ben. 2020b. "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy." *International Journal of Human–Computer Interaction* 36 (6): 495–504. doi:10.1080/10447318.2020.1741118.

Wang, Baicun, Yuan Xue, Jianlin Yan, Xiaoying Yang, and Yuan Zhou. 2020. "Human-Centered Intelligent Manufacturing: Overview and Perspectives." *Chinese Journal of Engineering Science* 22 (4): 139. doi:10.15302/J-SSCAE-2020.04.020.

Witell, Lars, Anders Gustafsson, and Michael D. Johnson. 2014. “The effect of customer information during new product development on profits from goods and services.” *European Journal of Marketing* 48 (9/10): 1709–30. doi:10.1108/EJM-03-2011-0119.

Xu, Wei. 2019. “Toward human-centered AI: A Perspective from Human-Computer-Interaction.” *interactions* 26 (4): 42–46. doi:10.1145/3328485.

Xu, Wei, Marvin Dainoff, Liezhong Ge, and Zaifeng Gao. 2021. “From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centered AI.” *ArXiv* abs/2105.05424. https://api.semanticscholar.org/CorpusID:234469954.

Xu, Wei, and Zaifeng Gao. 2023. “Enabling Human-Centered AI: A Methodological Perspective.” http://arxiv.org/pdf/2311.06703.pdf.

4 Diskussion

4.1 Theoretische und praktische Implikationen

Industrielle KI-basierte Services versprechen ein hohes Potenzial für produzierende Unternehmen, u. a. im Hinblick auf Effizienzsteigerung, Ressourceneinsparung und Qualitätssteigerung (Mockenhaupt, 2021). Doch trotz der vielen Potenziale werden diese Anwendungen nicht flächendeckend in der Industrie eingesetzt (Pokorni et al., 2021). Dies geht einher mit früheren Erfahrungen in Forschungsprojekten mit industriellen Unternehmen. Basierend auf dieser Erkenntnis wurde in einer initialen, explorativen Analyse die Ausgangssituation im Kontext KI-basierter Services analysiert (Kutz et al., 2022). Die Studienergebnisse haben die aus der Literatur bekannten Hindernisse der KI-Entwicklung und -Implementierung wie die fehlende Einbindung der Mitarbeitenden in den Entwicklungsprozess und einen Mangel an KI-Kompetenzen und Akzeptanz der KI-Systeme bestätigt. Bis dato wenig untersucht hingegen sind Faktoren, die die Entwicklung, die Implementierung und den Betrieb industrieller KI-basierter Services positiv beeinflussen. Die identifizierten Erfolgsfaktoren (siehe Publikation 1) zeigen Möglichkeiten eines erfolgreichen industriellen KI-Designs über die Ebenen Mensch, Technik und Organisation innerhalb eines sozio-technischen Systems auf (Ulich, 2013). Dazu zählen beispielsweise die Partizipation der Endanwenderinnen und Endanwender im Entwicklungsprozess, die Verdeutlichung des Mehrwerts von KI-Technologien sowie die Verwendung von KI-Demonstratoren. Herausforderungen und Erfolgsfaktoren konnten auf allen drei Ebenen identifiziert werden, was wiederum die Relevanz einer ganzheitlichen Betrachtung KI-basierter Services bestätigt (Gabriel et al., 2022; Hoffmann et al., 2021; Pokorni et al., 2021; Winkelhaus et al., 2021).

Um diese ganzheitliche Betrachtung sowie den operativen Einsatz KI-basierter Services nachhaltig zu fördern, bedarf es zur Unterstützung einer erfolgreichen Gestaltung Methoden, Prozesse und Werkzeuge, die die Herausforderungen und Erfolgsfaktoren adressieren (Böhmann et al., 2018; Hunke & Schüritz, 2019). Die vorliegende Arbeit leistet insbesondere einen Beitrag dazu, den bisher primär technisch getriebenen Entwicklungsprozess (Cremers et al., 2019; Lee, 2020) um Ansätze zu erweitern, die Herausforderungen und Erfolgsfaktoren auf den Ebenen Mensch und Organisation adressieren. Damit wird in der vorliegenden Arbeit auch der im Ansatz der ‚Industrie 5.0' formulierte Bedarf einer menschengerechteren Ausgestaltung der industriellen Transformation behandelt (Breque et al., 2021).

Basierend auf Ergebnissen anwendungsnah ausgerichteter Forschungsprojekte wurden ausgewählte Herausforderungen und Erfolgsfaktoren in handlungsorientierte Modelle zur Anwendung in Unternehmen überführt. Die Erfahrung zeigt, dass es für die industrielle Praxis entscheidend ist, effiziente Lösungen zu entwickeln, die sich leicht in bestehende Prozesse integrieren lassen. Mit dem gewählten anwendungsorientierten Entwicklungsansatz der vorgestellten Modelle soll genau dies gewährleistet werden.

Das entwickelte „Generische Rollenmodell zur systematischen Entwicklung interner KI-basierter Services in der Produktion" (siehe Publikation 2) unterstützt durch seine interdisziplinäre Ausrichtung die notwendige ganzheitliche Betrachtung der Entwicklung und Implementierung KI-basierter Services (Hoffmann et al., 2021; Pokorni et al., 2021) sowie die Einbindung verschiedener Stakeholder über diverse Unternehmensbereiche (z. B. Instandhaltung, Endanwenderinnen und Endanwender, Change Management) hinweg. Damit leistet das Rollenmodell primär einen Beitrag, die identifizierte Herausforderung der „undefinierten Rollen und Verantwortlichkeiten" zu bewältigen. Darüber hinaus unterstützt es Entwicklerinnen und Entwicklern, alle relevanten Stakeholder im Entwicklungsprozess zu berücksichtigen. Außerdem wird durch die Anwendung des Rollenmodells die Mitbestimmung

und Beteiligung der Endnutzenden und anderer Betroffener gestärkt, was wiederum das menschenzentrierte KI-Design unterstützt. Das Rollenmodell gibt zusätzlich Auskunft darüber, welche Rollen in einem ko-kreativen Entwicklungsansatz für industrielle KI integriert werden sollten, und trägt somit dazu bei, den definierten Forschungsbedarf von Hieber et al. (2023) zu erfüllen.

HCAI ist ein Ansatz, der in der Wissenschaft seit einigen Jahren vermehrt Aufmerksamkeit genießt, jedoch in der Umsetzung wenig Anwendbarkeit findet (Hartikainen et al., 2022). Zurückzuführen ist dies u. a. auf zwei zentrale Herausforderungen von HCAI: Es fehlen konkrete Methoden, Prozesse und Modelle zur Umsetzung in der Praxis (Bingley et al., 2023; Hartikainen et al., 2022; Xu et al., 2023). Die bisherigen Arbeiten sind meist branchenunabhängig (Auernhammer, 2020). Beide Herausforderungen werden in der vorliegenden Arbeit adressiert. Zum einen wurde ein Vorgehensmodell entwickelt, welches die tatsächliche Anwendung von HCAI in realen KI-Entwicklungsprojekten unterstützt, und zum anderen wird versucht, die Themen industrielle KI und HCAI gemeinsam zu betrachten, indem bisher domänenunabhängige Ansätze von HCAI auf den industriellen Kontext angewandt wurden. Insgesamt haben die vorliegenden Studien gezeigt, dass HCAI ein vielversprechender Ansatz ist, den Menschen als zentralen Bestandteil industrieller KI-basierter Services zu verstehen und bei der Entwicklung zu berücksichtigen.

Im Kontext HCAI fokussiert die vorliegende Arbeit Leitlinien und Prinzipien, um das menschenzentrierte Design praxisnah zu unterstützen. Dafür wurde untersucht, inwiefern bereits vorhandene handlungsorientierte Design-Prinzipien (Google PAIR, 2019; Microsoft, o. J.) dafür geeignet sind, allgemeine Anforderungen an eine vertrauenswürdige KI, wie beispielsweise von der EU formuliert (Hochrangige Expertengruppe für künstliche Intelligenz, 2019), zu adressieren. Die Untersuchung hat gezeigt, dass dies eingeschränkt möglich ist. Es bedarf jedoch insbesondere in Bezug auf die Anforderungen „Vielfalt, Nichtdiskriminierung und Fairness“, „Gesellschaftliches und ökologisches Wohlergehen“ sowie „Rechenschaftspflicht“ ergänzender Design-Prinzipien, um die Anforderungen vollständig in Handlungsanweisungen zu überführen und den Anforderungen vollumfänglich nachzukommen (Kutz, Neuhüttler, Spilski, & Lachmann, 2023). Dennoch ist nach Einschätzungen von Expertinnen und Experten für industrielle KI die Anwendung handlungsorientierter Design-Prinzipien ein vielversprechender Ansatz zur Umsetzung von HCAI in der Industrie (Kutz, Neuhüttler, Bienzeisler, et al., 2023). Auch die von Hoffmann et al. (2021) definierten Anforderungen an industrielle KI sind in Teilen durch die Anwendung der Design-Prinzipien im *People + AI Guidebook* des Google PAIR adressierbar. Dies gilt beispielsweise für die notwendige schrittweise Einführung KI-basierter Services sowie die stufenweise Erhöhung des Grads der Automatisierung und die Anforderungen an einfache Erklärungen der KI-Systeme bzw. deren Outputs. Eine systematische Untersuchung, inwiefern bestehende Design Patterns die definierten Anforderungen erfüllen können, ist nicht Gegenstand der vorliegenden Arbeit. Ergänzend zu den hier durchgeführten Studien würde eine solche Untersuchung in zweierlei Hinsicht einen Mehrwert bieten: Zum einen könnten Lücken in den Design-Prinzipien zur Erfüllung von Anforderungen im industriellen Kontext aufgezeigt werden, und zum anderen könnten die Anforderungen an industrielle KI-Systeme durch Aspekte erweitert werden, die in den Design-Prinzipien Berücksichtigung finden, jedoch nicht in den Anforderungen inkludiert sind.

Die Einbindung der Mitarbeitenden in die Gestaltung KI-basierter Services wird in verschiedenen Publikationen zu HCAI immer wieder als Erfolgsfaktor genannt (Auernhammer, 2020; Berditchevskaia et al., 2021; Plattform Lernende Systeme, 2019; Xu et al., 2023). Über alle im Rahmen dieser Arbeit durchgeführten Studien hinweg hat sich diese Erkenntnis

bestätigt. Ko-kreative Entwicklungsprozesse sind eine Methode, um die Einbindung von Mitarbeitenden zu unterstützen (Hieber et al., 2023; Russo-Spena & Mele, 2012). Sie zeichnen sich durch ein partizipatives Vorgehen aus. Das entwickelte „Vorgehensmodell zur Nutzung von Design-Prinzipien in der ko-kreativen Gestaltung menschenzentrierter KI-basierter Services" kombiniert deshalb Ko-Kreation und die Anwendung handlungsorientierter Design-Prinzipien, die als wertvoll durch Expertinnen und Experten für industrielle KI-Anwendungen beschrieben wurden, in einem Ansatz (siehe Publikation 5). Diese Integration zweier zentraler Konzepte (Design-Prinzipien und Partizipation/Ko-Kreation) von HCAI in einem Ansatz geht über bisherige Forschungsarbeiten in dem Bereich hinaus.

Durch die praxisnahe Konzeption leistet das vorgestellte Modell insbesondere einen Beitrag dazu, den Erfolgsfaktor „Nutzendenzentriertes KI-Design" zu stärken, da es als Werkzeug zur Umsetzung dessen fungiert. Gleichzeitig führt die Anwendung auch dazu, die Mitbestimmung und Beteiligung der Endnutzenden an der Entwicklung zu stärken. Ergänzend kann die Einbindung der Mitarbeitenden in den KI-Entwicklungsprozess die Kompetenzen im Bereich KI bei den beteiligten Personen stärken und somit zur Qualifizierung beitragen. Durch die Integration der HCAI-Design-Prinzipien in das Vorgehensmodell soll die Anwendung dieser im industriellen Kontext gestärkt werden. In Abhängigkeit der zur Umsetzung ausgewählten Design-Prinzipien können weitere Herausforderungen und Erfolgsfaktoren wie das Erwartungsmanagement, das Vertrauen sowie der Abbau von Ängsten adressiert werden.

Ferner verdeutlichen sowohl die Arbeit von Shneiderman (2020a) als auch der von Xu et al. (2024) vorgestellte hierarchische Ansatz für intelligente sozio-technische Systeme, dass nicht allein Entwicklungsteams verantwortlich für die Umsetzung von HCAI sind. Organisationale Strukturen, Industriestandards sowie Regulierungen und ethische Grundsätze bilden die Strukturen und den Rahmen zur Umsetzung. Das „Vorgehensmodell zur Nutzung von Design-Prinzipien in der ko-kreativen Gestaltung menschenzentrierter KI-basierter Services" konzentriert sich auf die Umsetzung von HCAI auf Teamebene. Berücksichtigt wird jedoch im Modell, dass die Organisation durch ihre vorgegebenen Rahmenbedingungen einen Einfluss auf die tatsächliche Anwendung von HCAI hat und politische sowie branchenspezifische Regulierungen wiederum die Rahmenbedingungen auf Organisationsebene beeinflussen. Durch die definierten drei Ebenen (Team, Organisation und Industrie) im Modell bietet das Konzept außerdem einen ersten Ansatz, notwendige Führungs- und Verwaltungsstrukturen für HCAI in Unternehmen (Ozmen Garibay et al., 2023) umzusetzen. Der ko-kreative Ansatz auf Teamebene sollte ergänzend auf die Organisations- und Industrieebene übertragen werden, um auch auf diesen Ebenen durch Partizipation einen Beitrag zur Entwicklung reliabler, sicherer und vertrauenswürdiger KI-basierter Services zu leisten.

Beide im Rahmen der vorliegenden Arbeit entwickelten Modelle hängen eng miteinander zusammen. Zum einen kann das Rollenmodell als zentrales Element einer menschenzentrierten KI-Entwicklung interpretiert werden, da es den Rahmen für die Einbindung diverser Stakeholder im KI-Entwicklungsprozess vorgibt. Zum anderen kann das Rollenmodell auch als Hilfestellung verwendet werden, um die passenden Stakeholder für die Beteiligung im entwickelten Vorgehensmodell zu identifizieren. Zusammenfassend leisten beide Modelle einen Beitrag dazu, die Entwicklung menschenzentrierter KI-basierter Services in der Industrie zu unterstützen und somit die Akzeptanz und Vertrauenswürdigkeit dieser zu fördern.

4.2 Limitationen und Ausblick

Im Folgenden wird auf die Limitationen der vorliegenden Arbeit eingegangen, die über die bereits in den Forschungsarbeiten diskutierten hinausgehen oder sich aus dem Kontext der Gesamtarbeit ergeben. Darüber hinaus wird auf bestehende Forschungslücken eingegangen und ein Ausblick auf weitere ergänzende bzw. notwendige Forschungsarbeiten gegeben.

Ziel der Arbeit ist es, für ausgewählte Erfolgsfaktoren und Herausforderungen praxisnahe Ansätze zu schaffen, die zu einem optimierten Einsatz KI-basierter Services in der Industrie führen. Deswegen basieren die konzipierten Modelle auf Ergebnissen anwendungsnaher und qualitativ ausgerichteter Forschungsprojekte in Zusammenarbeit mit Expertinnen und Experten für industrielle KI-basierte Services. Jedoch konnte für beide entwickelten Modelle in den laufenden Forschungsprojekten die Wirksamkeit in der Anwendung nicht systematisch evaluiert werden. In weiteren Schritten ist es daher notwendig, das Rollenmodell sowie das Vorgehensmodell über verschiedene KI-Use-Cases hinweg anzuwenden und auf deren Wirksamkeit hin zu überprüfen. Darüber hinaus basieren die Analysen der vorliegenden Arbeit auf internen KI-basierten Services, d. h. im Fokus stehen KI-basierte Services, die innerhalb eines Unternehmens entwickelt, implementiert und bereitgestellt werden. Neben Eigenentwicklungen gibt es jedoch auch viele Services, die extern bezogen werden. Insbesondere für kleine und mittelständische Unternehmen ist die Bedeutung externer KI-Lösungen relevant (Lundborg & Gull, 2021). In diesem Zusammenhang ist anzumerken, dass die hier entwickelten Konzepte nicht direkt anwendbar sind bei Bezug von KI-Lösungen externer Anbieter. Eine Überprüfung der Konzepte sowie eine entsprechende Adaption auf externe Lösungen sollten daher Gegenstand künftiger Forschungs- und Entwicklungsprojekte sein. Dies sollte in zweierlei Hinsicht passieren: Zum einen aus Unternehmensperspektive als Kunde und letztlich Nutzender, zum anderen aus Perspektive der Anbieter, die KI-basierte Services in Form von Dienstleistungen an ihre Kunden verkaufen.

Des Weiteren sollten auch die identifizierten Herausforderungen und Erfolgsfaktoren Gegenstand weiterer Forschungsarbeiten sein. Durch den gewählten qualitativen, auf Experteninterviews beruhenden Forschungsansatz wurden einige Faktoren (z. B. im Hinblick auf Kompetenzen, Erwartungen und Stakeholdermanagement) sowohl als Herausforderung als auch als Erfolgsfaktoren identifiziert. Eine detaillierte Analyse der einzelnen Faktoren wäre sinnvoll, um hier eine eindeutige Zuordnung gewährleisten zu können. Darüber hinaus sollten die Faktoren in Anlehnung an die Zwei-Faktoren-Theorie von Herzberg (Nerdinger, 2014) auf ihre Wirkung hin überprüft werden. Eine Analyse dessen, welche Faktoren als Mindestanforderungen zu verstehen sind und welche tatsächlich eine Zufriedenheit im Umgang mit KI-basierten Services bei Mitarbeitenden hervorrufen, wäre für künftige KI-Entwicklungsprojekte hilfreich.

Überdies können durch die untersuchten Ansätze nicht alle in der initialen Analyse identifizierten Herausforderungen und Erfolgsfaktoren adressiert werden. Insbesondere technische Herausforderungen sind nicht Gegenstand der Arbeit, sollten jedoch im Sinne eines ganzheitlichen Ansatzes unter Berücksichtigung menschenzentrierter Aspekte künftig tiefergehend analysiert werden. Entsprechend gilt dies auch für Herausforderungen und Erfolgsfaktoren, die den Ebenen Mensch und Organisation zugeordnet sind und nicht im Rahmen dieser Arbeit adressiert wurden. Dazu zählen z. B.: Data Governance, Unternehmenskultur und Kommunikation im Zusammenhang mit KI. Eine Herausforderung soll an dieser Stelle besonders hervorgehoben werden: Kooperationsschwierigkeiten zwischen IT- und Produktionsabteilungen. Diese wurde in den verschiedenen Studien immer wieder genannt, war aber kein fokussierter Bestandteil der vorgestellten Studien. Eine positive Kooperationskultur zwischen diesen Abteilungen ist jedoch zentral für die Praxiswirksamkeit

der entwickelten Modelle und sollte daher unbedingt Gegenstand zukünftiger Forschungsprojekte sein. Ebenso wurden Kompetenzdefizite der Mitarbeitenden im Kontext KI als zentrale Herausforderung identifiziert. Die Integration verschiedener Mitarbeitenden in den KI-Entwicklungsprozess kann zu einem Kompetenzerwerb im Bereich KI führen, jedoch nicht flächendeckend für alle Mitarbeitenden. Doch ebendies wurde als Erfolgsfaktor („Qualifikationen und Kompetenzen im Bereich KI über alle Stakeholder hinweg") hervorgehoben. Ein erster Ansatz, um dies zu gewährleisten, ist das Konzept der „digital versierten Belegschaft", das sowohl das Vertrautsein mit neuen digitalen Technologien als auch die Kompetenz, diese zu nutzen, umfasst (Krcmar, 2022). In einem 2023 konzipierten Modell werden entlang der Ebenen des sozio-technischen Systems Faktoren beschrieben, die zu einer individuellen und situationsspezifischen Förderung der digitalen Versiertheit in Unternehmen beitragen (Kutz, Hieber, et al., 2023). Dieses Verständnis der digitalen Versiertheit kann auf den Umgang mit KI-Technologien adaptiert werden und zur Stärkung einer KI-Versiertheit bei Mitarbeitenden beitragen. Eine weitere Forschungsarbeit, die sich ebenfalls mit der Kompetenzentwicklung im Kontext von KI befasst und den Erfolgsfaktor „Nutzung von Demonstratoren" aufgreift, beschreibt, wie interaktive Lernformate und insbesondere die Nutzung von KI-Demonstratoren zur Förderung des Verständnisses von KI im Arbeitskontext beitragen (Gladilov et al., im Druck). Beide aufgegriffenen Publikationen können ergänzend zu den hier vorgestellten Studien herangezogen werden, um den Einsatz KI-basierter Services in der Produktion ganzheitlich zu stärken.

In dieser Arbeit wird Ganzheitlichkeit primär im Zusammenhang des sozio-technischen Systems verstanden. Es ist jedoch wichtig, dieses Verständnis in zukünftigen Forschungsarbeiten zu erweitern. Dabei sollte die Perspektive des KI-Lebenszyklus von der Ideengenerierung bis zur Außerbetriebnahme eines KI-Systems einbezogen werden. Der KI-Lebenszyklus ist geprägt durch die Heterogenität KI-basierter Services (Bienzeisler et al., 2023) und erfordert daher die Fähigkeit, auf Veränderungen der KI in allen Ebenen des sozio-technischen Systems reagieren zu können. Die vorgestellten Modelle in dieser Arbeit konzentrieren sich hauptsächlich auf die Entwicklungsphase eines KI-basierten Services. Daher ist es notwendig, die Modelle zu erweitern und weitere Methoden in Unternehmen einzuführen, um auch die späteren Phasen im KI-Lebenszyklus aus einer menschenzentrierten Perspektive zu berücksichtigen und die Akzeptanz der Systeme langfristig sicherzustellen.

Die in dieser Arbeit näher betrachteten Designprinzipien des *People + AI Guidebook* (Google PAIR, 2019) wurden dem KI-Lebenszyklus von der Designphase bis zum Betrieb zugeordnet. Dabei hat sich gezeigt, dass durch die Prinzipien die verschiedenen Phasen bereits gut abgedeckt werden können und eine generelle Anwendbarkeit bestehender HCAI-Design-Prinzipien auch im industriellen Kontext möglich ist. Es ist dennoch erforderlich, die Prinzipien phasenspezifisch zu beschreiben, um sicherzustellen, dass die Umsetzung dem Entwicklungsstand der KI entspricht. Darüber hinaus müssen die Prinzipien an die industrielle Praxis angepasst werden, und es sollten Best-Practices bereitgestellt werden, die als Vorlage zur Umsetzung zukünftiger Anwendungen dienen können. Überdies gibt es nach derzeitigem Kenntnisstand keine Studien, die die Wirksamkeit der Gestaltungsprinzipien belegen. Künftige Untersuchungen sollten sowohl die Wirksamkeit einzelner Prinzipien als auch die Wirksamkeit der Prinzipien in ihrer Wechselwirkung analysieren. Anzumerken ist außerdem, dass die bestehenden Prinzipien insbesondere auf KI als technische Anwendung fokussieren. Es bedarf daher weiterer Prinzipien, die auch die vor- und nachgelagerten Prozesse umfassen und die KI aus einer ganzheitlichen Dienstleistungsperspektive betrachten (Neuhüttler et al., 2020). Darüber hinaus wurde aufgezeigt, dass mit den bestehenden handlungsorientierten Prinzipien nicht alle Anforderungen an eine vertrauenswürdige KI erfüllt werden können. Entsprechend braucht es auch hier eine Erweiterung bestehender Prinzipien, um vollumfänglich den seitens

der Europäischen Union definierten Anforderungen (Hochrangige Expertengruppe für künstliche Intelligenz, 2019) gerecht zu werden. Dieser Bedarf betrifft alle KI-basierten Services, auch über Anwendungen im industriellen Kontext hinaus. Über die vorgestellten Studien hinweg hat sich zudem gezeigt, dass es in der industriellen Praxis kaum möglich ist, alle definierten Prinzipien innerhalb eines KI-basierten Services anzuwenden. Aus diesem Grund basiert das entwickelte Vorgehensmodell auf einem ko-kreativen Ansatz, um konsensbasiert zu einer anwendungsfallspezifischen Priorisierung der umzusetzenden Prinzipien zu gelangen. Eine auf Best-Practices basierende Entscheidungsunterstützung bei der Auswahl geeigneter Gestaltungsprinzipien wäre ergänzend hilfreich. Hierfür wäre ein KI-basiertes Tool zur Unterstützung des Entwicklungsprozesses auf Basis unternehmensübergreifender Daten denkbar. Ein solches Tool könnte auch den Umgang mit widersprüchlichen Anforderungen an menschenzentrierte und vertrauenswürdige KI-Systeme unterstützen (Xu et al., 2023).

Mit der Verabschiedung des EU AI Act (European Commission, 2021) entstehen neue rechtliche Anforderungen an KI-basierte Services. Hier bedarf es künftiger Analysen, inwiefern bestehende Design-Guidelines und -Prinzipien herangezogen werden können, um die im EU AI Act definierten Anforderungen zu erfüllen.

Die verschiedenen Studien zu Designprinzipien haben gezeigt, dass im Allgemeinen die Anwendung von handlungsorientierten Prinzipien, die über allgemeine Grundsätze hinausgehen, einen großen Mehrwert für die menschenzentrierte Entwicklung von KI-Anwendungen darstellt. Anzumerken ist in diesem Zusammenhang, dass bisher definierte Prinzipien dieser Art insbesondere von großen Technologiekonzernen präsentiert werden. Hier ist die Forschung gefordert, Gestaltungsprinzipien für eine ganzheitliche KI-Entwicklung bereitzustellen, die aus einer forschungsgeleiteten, von wirtschaftlichen Eigeninteressen unabhängigen Perspektive entwickelt werden.

Auf einige zentrale Herausforderungen der anwendungsnahen KI-Forschung, die sich in den verschiedenen Arbeiten herauskristallisiert haben, soll abschließend noch eingegangen werden. Zum einen ist der Einsatz von KI-basierten Services in der unternehmensinternen Nutzung stark an monetäre Effekte gebunden. Demzufolge stehen bei der industriellen Entwicklung technische Aspekte im Fokus, die Kern der KI-Entwicklung sind. Effekte menschenzentrierter Entwicklungsaspekte können hingegen kaum in ökonomischen Effekten abgebildet werden und entfalten ihre Wirkung primär auf anderen Ebenen wie der Mitarbeitendenzufriedenheit oder der Akzeptanz. Hier bedarf es einer Sensibilisierung auf Managementebene für die Relevanz der KI-Forschung auf allen Ebenen des sozio-technischen Systems. Zum anderen schreitet die Entwicklung der KI-Technologien rasch voran. Dies birgt das Risiko, dass die technische Entwicklung und die Diskussion ethischer, menschenzentrierter Aspekte nicht parallel verlaufen. Diese Herausforderung betrifft sowohl Unternehmen als auch Politik und Forschung. Daher ist es wichtig, anwendungsnahe KI-Forschung zu betreiben, die sich mit den konkreten Herausforderungen und Auswirkungen in der Praxis auseinandersetzt. Durch enge Zusammenarbeit und regelmäßigen Austausch zwischen Forschung und Industrie können relevante Themen identifiziert und gemeinsame Lösungsansätze entwickelt werden. Dies ermöglicht es, KI-Technologie verantwortungsvoll einzusetzen und mögliche Risiken frühzeitig zu identifizieren und zu adressieren. Damit einher geht aber auch die Notwendigkeit, die gängigen empirischen Forschungsmethoden an die in der Industrie vorherrschenden Rahmenbedingungen anzupassen und die Forschungslandschaft für die neu entstandenen Bedürfnisse zu sensibilisieren.

5 Fazit

Künstliche Intelligenz wird zunehmend zum integralen Bestandteil von Produktionsumgebungen und ist somit wichtiger Treiber für die digitale Fabriktransformation (Bienzeisler et al., 2023). Um diesen Wandel erfolgreich gestalten zu können, bedarf es eines ganzheitlichen Entwicklungsansatzes für industrielle KI-basierte Services als Teil sozio-technischer Arbeitssysteme (Winkelhaus et al., 2021). Die vorliegende Arbeit integriert fünf Studien, die jeweils einen Beitrag zur Unterstützung einer erfolgreichen Entwicklung von KI-basierten Services entlang der Ebenen Mensch, Technik und Organisation leisten. Beantwortet wird zum einen die Frage, welche Herausforderungen und Erfolgsfaktoren im Zusammenhang mit der Entwicklung, der Implementierung und dem Betrieb industrieller KI-basierter Services entlang dieser drei Ebenen Mensch, Technik und Organisation bestehen. Zum anderen die Frage, welche spezifischen Methoden und Modelle zur Lösung ausgewählter Herausforderungen bzw. zur Stärkung ausgewählter Erfolgsfaktoren beitragen können und wie solche Methoden und Modelle im industriellen Kontext gestaltet sein können. Die vorliegende Arbeit fokussiert dabei auf organisations- und menschzentrierte Aspekte. Aus diesem Grund wurden Ansätze zur Gestaltung menschenzentrierter KI analysiert und anschließend auf den industriellen Kontext übertragen. Durch die Überprüfung der Anwendbarkeit bestehender HCAI-Gestaltungsansätze und deren Kontextualisierung im industriellen Umfeld sowie die Entwicklung anwendungsorientierter Modelle zur menschzentrierten Gestaltung von industriellen KI-basierten Services leistet die Arbeit einen Beitrag zur Lösung zentraler Herausforderungen des Forschungsfeldes HCAI.

Entwickelt wurden zum einen das „Generische Rollenmodell zur systematischen Entwicklung interner KI-basierter Services in der Produktion“ und zum anderen das „Vorgehensmodell zur Nutzung von Design-Prinzipien in der ko-kreativen Gestaltung menschenzentrierter KI-basierter Services“. Die entwickelten Vorgehensmodelle können zukünftig den bisher technisch geprägten Entwicklungsprozess industrieller KI-basierter Services unterstützen (Pokorni et al., 2021) und somit einen Beitrag zu einer ganzheitlichen Entwicklungsperspektive leisten. Zusammenfassend kann gesagt werden, dass die vorliegende Arbeit dazu beiträgt, Entwicklerinnen und Entwickler bei der Gestaltung menschenzentrierter KI-basierter Services für die Industrie zu unterstützen.

6 Literaturverzeichnis

Abel, J., Hirsch-Kreinsen, H., & Wienzek, T. (2019). *Akzeptanz von Industrie 4.0: Abschlussbericht zu einer explorativen empirischen Studie über die deutsche Industrie.* acatech – Deutsche Akademie der Technikwissenschaften.

Adler, R., Bunte, A., Burton, S., Großmann, J., Jaschke, A., Kleen, P., Lorenz, J. M., Ma, J., Markert, K., Meeß, H., Meyer, O., Neuhüttler, J., Philipp, P., Poretschkin, M., Rennoch, A., Roscher, K., Sperl, P., Usländer, T., Weicken, E., . . . Tcholtchev, N. V. (2022). *Deutsche Normungsroadmap Künstliche Intelligenz: Ausgabe 2.* DIN e. V., DKE deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE. https://publica.fraunhofer.de/entities/publication/8648d5c2-5c85-4f19-9ba3-26db005e2640/details https://doi.org/10.24406/publica-1632

Ahmad, K., Abdelrazek, M., Arora, C., Bano, M., & Grundy, J. (2023, January 25). *Requirements Practices and Gaps When Engineering Human-Centered Artificial Intelligence Systems*. http://arxiv.org/pdf/2301.10404.pdf

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300233

Auernhammer, J. (2020). Human-centered AI: The role of Human-centered Design Research in the development of AI. In S. Boess, M. Cheung, & R. Cain (Eds.), *Proceedings of DRS 2020, Vol. 1: Synergy, Situations.* Design Research Society. https://doi.org/10.21606/drs.2020.282

Berditchevskaia, A., Peach, K., & Malliaraki, E. (September 2021). *Participatory AI for humanitarian innovation: a briefing paper*. https://media.nesta.org.uk/documents/Nesta_Participatory_AI_for_humanitarian_innovation_Final.pdf

Bienzeisler, B., Neuhüttler, J., & Kutz, J. (2023). Von der digitalen Fabriktransformation zur digitalen Servicetransformation – das Beispiel KI-Services. In M. Bruhn & K. Hadwich (Eds.), *Forum Dienstleistungsmanagement: Band 1. Innovationsperspektive - Digitalisierungsperspektive - Nachhaltigkeitsperspektive* (pp. 427–439). Springer Gabler. https://doi.org/10.1007/978-3-658-41813-7_15

Bingley, W. J., Curtis, C., Lockey, S., Bialkowski, A., Gillespie, N., Haslam, S. A., Ko, R. K., Steffens, N., Wiles, J., & Worthy, P. (2023). Where is the human in human-centered AI? Insights from developer priorities and user experiences. *Computers in Human Behavior*, *141*, 107617. https://doi.org/10.1016/j.chb.2022.107617

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. In *ACM Digital Library, Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3551624.3555290

BMW Group. (2020, October 12). *Sieben Grundsätze für KI: BMW Group legt Ethik-Kodex für den Einsatz von Künstlicher Intelligenz fest [*Press release].

Böhmann, T., Leimeister, J. M., & Möslein, K. (2018). The New Fontiers of Service Systems Engineering. *Business & Information Systems Engineering*, *60*(5), 373–375. https://doi.org/10.1007/s12599-018-0553-1

Breque, M., Nul, L. de, & Petridis, A. (2021). *Industry 5.0: Towards a sustainable, human-centric and resilient European industry*. https://op.europa.eu/en/publication-detail/-/publication/468a892a-5097-11eb-b59f-01aa75ed71a1/

Bundesministerium für Bildung und Forschung, Referat Künstliche Intelligenz (Ed.). (n. d.). *ki-strategie-deutschland*. https://www.ki-strategie-deutschland.de/home.html

Bundesministerium für Bildung und Forschung, Referat Künstliche Intelligenz (Ed.). (2018). *Strategie Künstliche Intelligenz der Bundesregierung*. https://www.ki-strategie-deutschland.de/home.html

Bundesministerium für Bildung und Forschung, Referat Künstliche Intelligenz (Ed.). (2023). *BMBF-Aktionsplan Künstliche Intelligenz: Neue Herausforderungen chancenorientiert angehen.*

Bundesministerium für Wirtschaft und Energie (Ed.). (2015). *Industrie 4.0 und Digitale Wirtschaft: Impulse für Wachstum, Beschäftigung und Innovation*. https://www.bmwk.de/Redaktion/DE/Publikationen/Industrie/industrie-4-0-und-digitale-wirtschaft.pdf?__blob=publicationFile&v=3

Capel, T., & Brereton, M. (2023). What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In A. Schmidt (Ed.), *ACM Digital Library, Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–23). Association for Computing Machinery. https://doi.org/10.1145/3544548.3580959

Cremers, A., Englander, A., Gabriel, M., Mock, M., Poretschkin, M., Rosenzweig, J., Rostalski, F., Sicking, J., Volmer, J., Voosholz, J., Voss, A., & Wrobel, S. (2019). *Vertrauenswürdiger Einsatz von Künstlicher Intelligenz: Handlungsfelder aus philosophischer, ethischer, rechtlicher und technologischer Sicht als Grundlage für eine Zertifizierung von Künstlicher Intelligenz.* Fraunhofer IAIS; Universität Bonn; Universität Köln. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf

Diemer, J., Elmer, S., Gaertler, M., Gamer, T., Görg, C., Grotepass, J., Kalhoff, J., Kramer, S., Legat, C., Meyer-Kahlen, J.-P., Nettsträter, A., Niehöster, O., Schmidt, B., Schweichhart, K., Ulrich, M., Weitschat, R., Winter, J., & Industrie 4.0, P. (2020). *KI in der Industrie 4.0: Orientierung, Anwendungsbeispiele, Handlungsempfehlungen*. https://elib.dlr.de/138923/

European Commission. (2021). *Proposal for a Regulation of the European Parliament and the Council: Laying down harmonised rules on artificial intelligence*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Fischer, S., Koch, T., Habenicht, T., & Becker, A. (2022). *Künstliche Intelligenz (KI) in der Industrie - ein kurzer Überblick*. https://www.bmwk.de/Redaktion/DE/Publikationen/Industrie/ki-in-der-industrie.pdf?__blob=publicationFile&v=4

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. (2018). An Introduction to Deep Reinforcement Learning. *Foundations and Trends in Machine Learning*, *11*(3-4), 219–354. https://doi.org/10.1561/2200000071

Gabriel, S., Bentler, D., Grote, E.-M., Junker, C., Wendischhoff, D. M. zu, Bansmann, M., Latos, B., Hobscheidt, D., Kühn, A., & Dumitrescu, R. (2022). Requirements analysis for an intelligent workforce planning system: a socio-technical approach to design AI-based systems. *Procedia CIRP*, *109*, 431–436. https://doi.org/10.1016/j.procir.2022.05.274

Gladilov, N., Kutz, J., Anduschus, P.-O., Hieber, N., Göbels, V. P., & Neuhuettler, J. (in press). Mit spielerischen Demonstratoren das Verständnis für KI-Anwendungen im Arbeitskontext fördern. In Gesellschaft für Arbeitswissenschaft e.V. (Chair), *70. Frühjahrskongress 2024,* Stuttgart.

Göbels, V. P., Fischer-Pressler, D., Kutz, J., & Bienzeisler, B. (in press). *Innovation im Blick: Wertschöpfung durch Data Pooling: Herausforderungen und Erfolgsfaktoren.* Fraunhofer IAO.

Google PAIR. (2019). *People + AI Guidebook: Designing human-centered AI products*. https://pair.withgoogle.com/guidebook/

Harlacher, M., Feggeler, N., Peifer, Y., & Ottersböck, N. (2023). Produzierendes Gewerbe auf internationalem Niveau: Ergebnisse der Online-Befragung zum Thema „Künstliche

Intelligenz in produzierenden Unternehmen". *Zeitschrift Für Wirtschaftlichen Fabrikbetrieb*, *118*(3), 173–177. https://doi.org/10.1515/zwf-2023-1012

Hartikainen, M., Väänänen, K., Lehtiö, A., Ala-Luopa, S., & Olsson, T. (2022). Human-Centered AI Design in Reality: A Study of Developer Companies' Practices. In *ACM Digital Library, Nordic Human-Computer Interaction Conference (NordiCHI'22)* (pp. 1–11). ACM. https://doi.org/10.1145/3546155.3546677

Hermann, M., Pentek, T., & Otto, B. (2016). Design Principles for Industrie 4.0 Scenarios. In *2016 49th Hawaii International Conference on System Sciences (HICSS)* (pp. 3928–3937). IEEE. https://doi.org/10.1109/HICSS.2016.488

Hieber, N., Feike, M., & Prochazka, V. (2023). Towards a co-creative approach to interactively develop digital services. In *AHFE International, The Human Side of Service Engineering*. AHFE International. https://doi.org/10.54941/ahfe1003131

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hochrangige Expertengruppe für künstliche Intelligenz. (2019). Ethik-Leitlinien für eine vertrauenswürdige KI. https://op.europa.eu/de/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-de/format-PDF https://doi.org/10.2759/22710

Hoffmann, M. W., Drath, R., & Ganz, C. (2021). Proposal for requirements on industrial AI solutions. In J. Beyerer, O. Niggemann, & A. Maier (Eds.), *Machine Learning for Cyber Physical Systems* (pp. 63–72). Springer Berlin Heidelberg.

Hunke, F., & Schüritz, R. (2019). Smartere Produkte durch analysebasierte Dienstleistungen – Ein methodisches Werkzeug zur strukturierten Entwicklung. *HMD Praxis Der Wirtschaftsinformatik*, *56*(3), 514–529. https://doi.org/10.1365/s40702-019-00531-8

IBM Deutschland GmbH (Ed.). (n. d.). *KI-Ethik*. https://www.ibm.com/de-de/impact/ai-ethics

ifaa - Institut für angewandte Arbeitswissenschaft e. V. (n. d.). *ifaa-Studie: Künstliche Intelligenz in produzierenden Unternehmen*. https://www.arbeitswissenschaft.net/angebote-produkte/studien/kwh-ue-alf-ki-studie-ergebnisse

Jan, Z., Ahamed, F., Mayer, W., Patel, N., Grossmann, G., Stumptner, Markus, & Kuusk, A. (2023). Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. *Expert Systems with Applications*, *216*, 119456. https://doi.org/10.1016/j.eswa.2022.119456

Kämpf, T., & Langes, B. (2023). Wissenschaftliche Perspektive: Was wir über die nachhaltige Gestaltung von KI wissen: Künstliche Intelligenz und der Wandel der Arbeitswelt: Warum wir einen neuen Leitstern brauchen. In T. Kämpf, B. Langes, L. C. Schatilow, & H.-J. Gergs (Eds.), *Human Friendly Automation: Arbeit und Künstliche Intelligenz neu denken* (1. Auflage, pp. 38–54). Frankfurter Allgemeine Buch.

KATI. Fraunhofer INT. KATI search (fraunhofer.de)

Krcmar, H. (2022). Digitale Transformation: Deutschland im Vergleich. *Ifo Schnelldienst*, *75*(2), 20–23. https://www.ifo.de/DocDL/sd-2022-02-czernich-falck-pfaffl-etal-digitale-tansformation.pdf

Kutz, J., Göbels, V. P., Brajovic, D., Fresz, B., Renner, N., Omri, S., Neuhüttler, J., Huber, M., & Bienzeisler, B. (2023). *KI-Zertifizierung und Absicherung im Kontext des EU AI Act: Herausforderungen und Bedürfnisse aus Sicht von Unternehmen.* https://doi.org/10.24406/PUBLICA-1875

Kutz, J., Hieber, N., Neuhüttler, J., Petzolt, S., & Hölzle, K. (2023). Individuelle und situationsspezifische Faktoren zur Unterstützung der digitalen Transformation in Unternehmen - Ein Modell zur Beschreibung der digitalen Versiertheit (B.5.5). In Gesellschaft für Arbeitswissenschaft e.V. (Ed.), *Nachhaltig Arbeiten und Lernen - Analyse und Gestaltung lernförderlicher und nachhaltiger Arbeitssysteme und Arbeits- und Lernprozesse.* GfA Press.

Kutz, J., Neuhüttler, J., Bienzeisler, B., Spilski, J., & Lachmann, T. (2023). Human-Centered AI for Manufacturing – Design Principles for Industrial AI-Based Services. In H. Degen & S. Ntoa (Eds.), *Lecture Notes in Computer Science, Artificial Intelligence in HCI* (pp. 115–130). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35891-3_8, reproduced with permission from Springer Nature

Kutz, J., Neuhüttler, J., Schaefer, K., Spilski, J., & Lachmann, T. (2023). Generic Role Model for the Systematic Development of Internal AI-based Services in Manufacturing. In T. X. Bui (Chair), *Proceedings of the 56th Annual Hawaii International Conference on System Sciences: January 3-6, 2023*. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/ea841716-73a6-4ea8-9a4e-858e8f498d6d/content

Kutz, J., Neuhüttler, J., Spilski, J., & Lachmann, T. (2022). Implementation of AI Technologies in manufacturing - success factors and challenges. In *AHFE International, The Human Side of Service Engineering.* AHFE International. https://doi.org/10.54941/ahfe1002565

Kutz, J., Neuhüttler, J., Spilski, J., & Lachmann, T. (2023). AI-based Services - Design Principles to Meet the Requirements of a Trustworthy AI. In C. Leitner, J. Neuhüttler, C. Bassano, & D. Satterfield (Eds.), *AHFE International, The Human Side of Service Engineering.* AHFE International. https://doi.org/10.54941/ahfe1003107

Lee, J. (2020). *Industrial AI: Applications with sustainable performance*. Springer. https://doi.org/10.1007/978-981-15-2144-7

Lee, J., Singh, J., & Azamfar, M. (2019). *Industrial Artificial Intelligence*. http://arxiv.org/pdf/1908.02150v3

Lundborg, M., & Gull, I. (2021). *Künstliche Intelligenz im Mittelstand. So wird KI für kleine und mittlere Unternehmen zum Game Changer: Eine Erhebung der Mittelstand-Digital Begleitforschung im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz.* wik consult. https://www.mittelstand-digital.de/MD/Redaktion/DE/Publikationen/ki-Studie-2021.pdf?__blob=publicationFile&v=5

Mazarakis, A., Bernhard-Skala, C., Braun, M., & Peters, I. (2023). What is critical for human-centered AI at work? - Toward an interdisciplinary theory. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1257057

Microsoft. (n. d.). *Guidelines for Human-AI Interaction: Best practices for designing AI user experiences*. https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/

Mockenhaupt, A. (2021). *Digitalisierung und Künstliche Intelligenz in der Produktion*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-32773-6

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *26*, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Nerdinger, F. W. (2014). Arbeitsmotivation und Arbeitszufriedenheit. In F. W. Nerdinger, G. Blickle, & N. Schaper (Eds.), *Arbeits- und Organisationspsychologie* (pp. 419–440). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41130-4_24

Neuhüttler, J., Fischer, R., Ganz, W., & Urmetzer, F. (2020). Perceived Quality of Artificial Intelligence in Smart Service Systems: A Structured Approach. In M. Shepperd, F. Brito e Abreu, A. Da Rodrigues Silva, & R. Pérez-Castillo (Eds.), *Communications in Computer and Information Science. Quality of Information and Communications Technology* (Vol. 1266, pp. 3–16). Springer International Publishing. https://doi.org/10.1007/978-3-030-58793-2_1

OECD (Ed.). (2023). *Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449.*

Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., Havens, J. C., Jirotka, M., Kacorri, H.,

Karwowski, W., Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., . . . Xu, W. (2023). Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of Human–Computer Interaction*, *39*(3), 391–437. https://doi.org/10.1080/10447318.2022.2153320

Peres, R. S., Jia, X., Lee, J., Sun, K., Colombo, A. W., & Barata, J. (2020). Industrial Artificial Intelligence in Industry 4.0 - Systematic Review, Challenges and Outlook. *IEEE Access*, *8*, 220121–220139. https://doi.org/10.1109/ACCESS.2020.3042874

Plattform Lernende Systeme (Ed.). (n. d.). *KI Konkret. Künstliche Intelligenz - einfach erklärt: Was ist KI?* https://www.ki-konkret.de/was-ist-ki.html

Plattform Lernende Systeme (Ed.). (2019). *Arbeit, Qualifizierung und Mensch-Maschine Interaktion – Whitepaper der Arbeitsgruppe Arbeit/Qualifikation, Mensch- Maschine-Interaktion*.

Pokorni, B., Braun, M., & Knecht, C. (2021). *Menschenzentrierte KI-Anwendungen in der Produktion: Praxiserfahrungen und Leitfaden zu betrieblichen Einführungsstrategien.* Fraunhofer IAO. http://publica.fraunhofer.de/dokumente/N-6249564.html

Robert Bosch GmbH. (n. d.). *KI-Kodex: Ethische Leitlinien für Künstliche Intelligenz*. https://www.bosch.com/de/stories/ethische-leitlinien-fuer-kuenstliche-intelligenz/

Russo-Spena, T., & Mele, C. (2012). "Five Co-s" in innovating: a practice-based view. *Journal of Service Management*, *23*(4), 527–553. https://doi.org/10.1108/09564231211260404

Schaller, D., Wohlrabe, K., Wolf, A., Demary, V., Mertens, A., Fregin, M.-C., Stops, M., Gillhuber, A., Heckmann, J., & Grunwald, W. A. (2023). Künstliche Intelligenz: Chance oder Gefahr? Wie verändert der Einsatz von KI unsere Gesellschaft? *Ifo Schnelldienst*, *76*(8), 3–28. https://www.ifo.de/publikationen/2023/aufsatz-zeitschrift/kuenstliche-intelligenz

Scheuer, D. (2020). *Akzeptanz von Künstlicher Intelligenz: Grundlagen intelligenter KI-Assistenten und deren vertrauensvolle Nutzung*. Springer Vieweg. https://doi.org/10.1007/978-3-658-29526-4

Shneiderman, B. (2020a). Bridging the Gap Between Ethics and Practice:: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, *10*(4), 1–31. https://doi.org/10.1145/3419764

Shneiderman, B. (2020b). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Subramonyam, H., Im Jane, Seifert, C., & Adar, E. (2022, July 5). *Human-AI Guidelines in Practice: Leaky Abstractions as an Enabler in Collaborative Software Teams*. https://arxiv.org/pdf/2207.01749

Tombeil, A.-S., Kremer, D., Neuhüttler, J., Dukino, C., & Ganz, W. (2020). Potenziale von Künstlicher Intelligenz in der Dienstleistungsarbeit. In M. Bruhn & K. Hadwich (Eds.), *Forum Dienstleistungsmanagement. Automatisierung und Personalisierung von Dienstleistungen* (pp. 135–154). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-30168-2_5

Ulich, E. (2013). Arbeitssysteme als soziotechnische Systeme-eine Erinnerung. *Journal Psychologie Des Alltagshandelns*, *6*(1), 4–12.

VDMA, Fraunhofer-Institut für Gießerei-,Composite- und Verarbeitungstechnik IGCV, & Institut für Werkzeugmaschinen und Betriebswissenschaften (iwb), Technische Universität München (Eds.). (2020). *Leitfaden Künstliche Intelligenz - Potenziale und Umsetzungen im Mittelstand*. https://publica.fraunhofer.de/handle/publica/300499

Winkelhaus, S., Sutter, A., Grosse, E., & Morana, S. (2021). Soziotechnische Systeme: Der Mensch in der Industrie 4.0. *Industrie 4.0 Management*, *2021*(3), 45–48. https://doi.org/10.30844/I40M_21-3_S45-48

Xu, W. (2019). Toward human-centered AI: A Perspective from Human-Computer-Interaction. *Interactions*, *26*(4), 42–46. https://doi.org/10.1145/3328485

Xu, W., Dainoff, M., Ge, L., & Gao, Z. (2021). From Human-Computer Interaction to Human-AI Interaction: New Challenges and Opportunities for Enabling Human-Centered AI. *ArXiv*, *abs/2105.05424*. https://api.semanticscholar.org/CorpusID:234469954

Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2022). Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. *International Journal of Human–Computer Interaction*, 1–25. https://doi.org/10.1080/10447318.2022.2041900

Xu, W., & Gao, Z. (2023). *Enabling Human-Centered AI: A Methodological Perspective*.

Xu, W., & Gao, Z. (2024). *An intelligent sociotechnical systems (iSTS) framework: Toward a sociotechnically-based hierarchical human-centered AI approach*.

Xu, W., Gao, Z., & Dainoff, M. (2023). *An HCAI Methodological Framework: Putting It Into Action to Enable Human-Centered AI.* arXiv. http://arxiv.org/pdf/2311.16027v3

Anhang

Anhang A: Lebenslauf

Persönliche Daten

Name	Janika Kutz

Bildungsweg

Seit 09/2021	**Promotion Psychologie,** *Rheinlandpfälzische Technische Universität Kaiserslautern-Landau* Forschungsschwerpunkte: • Human-Centered AI • Technologieakzeptanz • Vertrauen in KI
10/2016 – 03/2020	**Masterstudium Psychologie**, *Universität Koblenz-Landau* • Profil: AOW-Psychologie • Profil: Klinische Psychologie • Wahlfach: Internetbasierte Forschungsmethoden • Masterarbeit: Modellierung von Veränderungen in den Ergebnissen studentischer Lehrveranstaltungsevaluationen und den Ausprägungen auf Studierendenvariablen über die Zeit – Eine RI-CLPM Analyse
10/2013 – 02/2017	**Bachelorstudium Psychologie,** *Universität Koblenz-Landau* • Anwendungsfächer: Klinische Psychologie und Prävention, Pädagogische Psychologie, Wirtschaftspsychologie • Bachelorarbeit: Merkmale von Schulinspektionen als Einflussfaktoren auf das persönliche Anschlusshandeln von Lehrkräften und Schulleitungen nach Schulinspektionen

Berufserfahrung

Seit 05/2023	**Teamleitung,** *Public Service Innovation, Forschungs- und Innovationszentrum Kognitive Dienstleistungssysteme, Fraunhofer IAO* • Akquise und Betreuung von anwendungsorientierten Forschungsprojekten mit dem Fokus auf die Entwicklung, Testung und Evaluation smarter Services im öffentlichen Raum • Führung eines interdisziplinär zusammengesetzten Teams bestehend aus 6 wissenschaftlichen Mitarbeitenden
04/2020 – 04/2023	**Wissenschaftliche Mitarbeiterin,** *Forschungs- und Innovationszentrum Kognitive Dienstleistungssysteme, Fraunhofer IAO* • Forschungsschwerpunkte im Bereich Technologieakzeptanz, menschenzentriertes Design sowie Absicherung von KI-basierten Services • Mitarbeit in verschiedenen Forschungs- und Entwicklungsprojekten im öffentlichen Sektor: ScooP, LikeBike, H_2-Hafen und H_2-Innovationslabor • Leitung verschiedener Teilprojekte innerhalb der AI25: Digitalversierte Belegschaft, EdgeCloud for Production, Industrial Computer Vision, Smartes Shopfloor Management • Unterstützung bei der Vorbereitung, Durchführung und Auswertung verschiedener Workshops und Veranstaltungen
09/2018 – 03/2020	**Wissenschaftliche Hilfskraft,** *Zentrum für Methoden, Diagnostik und Evaluation der Universität Koblenz-Landau* • Mitarbeit in einem Forschungsprojekt zu universitärer Lehrevaluation • Literaturrecherche • Fragebogendesign • Datensammlung,-aufbereitung und -auswertung (Excel, Mplus, SPSS und R)

Anhang B: Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbst angefertigt und alle benutzten Hilfsmittel in der Arbeit angegeben habe, dass ich diese Dissertation nicht schon als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht, und dass weder die gleiche noch eine andere Abhandlung der Dissertation bei einer anderen Universität oder einem anderen Fachbereich der RPTU Kaiserslautern-Landau veröffentlicht wurde.

Janika Kutz
Heilbronn, 15.02.2024