Trustworthy Artificial Intelligence Systems Engineering

Konzeption und Implementierung vertrauenswürdiger KI-Systeme

Inauguraldissertation

zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften
des Fachbereichs Wirtschaftswissenschaften
der Universität Osnabrück

vorgelegt von

Jonas Rebstadt
M. Sc. Cognitive Science

Osnabrück, Juni 2023

Dekan: Prof. Frank Teuteberg

Referenten: Prof. Dr. Oliver Thomas
Prof. Dr. Frank Teuteberg

Tag der Disputation: 05. Juni 2023

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Teil A – Dachbeitrag

1 Ausgangslage

In den vergangenen Jahren wurden zahlreiche Ansätze aus den Bereichen der Künstliche Intelligenz[1] (KI) und des Maschinelles Lernen[2] (ML) in die Praxis übertragen und für spezifische Problemstellungen adaptiert (Perrault et al. 2019). Die resultierenden KI-Systeme wurden kontinuierlich stärker in existierende Prozesse integriert und schaffen bereits konkrete Mehrwerte für Unternehmen (Teodorescu et al. 2021), in der Forschung (Marx 2022) und für die Gesellschaft (Hamamoto et al. 2020).

Bei der Etablierung dieser Systeme können sich jedoch sowohl neue Herausforderungen ergeben als auch existierende Probleme, wie die Diskriminierung bestimmter Gruppen, systematisch verstärkt werden (Mayer et al. 2020; Teodorescu et al. 2021). Beispiele hierfür finden sich unter anderem bei KI-Systemen im Bereich Recruiting, die Frauen, bestimmte ethnische Gruppen oder Menschen mit Behinderung benachteiligen (Barocas et al. 2017) aber auch bei Chatbots mit rassistischen, sexistischen und antisemitischen Tendenzen (Wolf et al. 2017). Vor weiteren organisatorischen und kulturellen Herausforderungen stehen zahlreiche Unternehmen durch die notwendige Akzeptanz von Nutzenden bei der Einbindung von KI-Anwendungen in die operativen Prozesse und somit der Realisierung betriebswirtschaftlicher Potenziale (Fernández-Loría et al. 2020). Besonders deutlich wird dies bei Unternehmen aus stark regulierten Domänen wie der Wirtschaftsprüfung (Issa et al. 2016; Gierbl et al. 2020; Munoko et al. 2020). Aufgrund der Notwendigkeit, im Zweifelsfall alle Entscheidungen juristisch verteidigen zu können, wird die Integration von KI-Anwendungen trotz beträchtlicher Potenziale (Issa et al. 2016; Kokina, Davenport 2017; Downar, Fischer 2019) bisher weitgehend verhindert (Issa et al. 2016; Gierbl et al. 2020; Munoko et al. 2020).

Mögliche Ansatzpunkte zum Heben dieser Potenziale bieten Prinzipien und Konzepte, die unter den Begriffen der vertrauenswürdigen oder auch der ethischen KI subsumiert werden (Floridi et al. 2018; Hochrangige Expertengruppe für künstliche Intelligenz 2018; The House of Lords 2018; Larsson 2020). Bisher erfolgt nach Jobin et al. (2019) jedoch weder eine einheitliche Nennung noch eine entsprechende Ausgestaltung von Prinzipien über existierende Richtlinien zu vertrauenswürdiger KI hinweg.

In der von Jobin et al. (2019) durchgeführten Studie haben sich mit Transparenz, Rechtssicherheit und Fairness, Unbedenklichkeit, Rechenschaftspflicht sowie Datenschutz fünf Prinzipien herausgestellt, die weitestgehend einheitlich genannt werden (Jobin et al. 2019). Aber nicht nur die Identifikation und Benennung der relevanten Prinzipien, sondern auch ihre inhaltliche Ausgestaltung wird bisher nicht einheitlich vorgenommen. Exemplarisch hierfür sind die zahlreichen teils synonym genutzten Begriffe (u. a. Erklärbarkeit, Interpretierbarkeit, Transparenz, Verständlichkeit oder Verstehbarkeit) zur Transparenz von KI-Systemen (Moore, Swartout 1988; Dhurandhar et al. 2017; Lipton 2018; Tomsett et al.

[1] Die Definition von Künstlicher Intelligenz stellt aufgrund ihrer häufig vagen und sich über die Zeit wandelnden Ausprägung eine Herausforderung dar (McCorduck, Cfe 2004). Aufgrund dessen wird im Zuge dieser Dissertation auf die relativ allgemeine Definition von Russell und Norvig (2010) zurückgegriffen, die Künstliche Intelligenz als Maschinen beschreibt, welche kognitive Funktionen, wie Lernen oder Problemlösen, approximieren, die primär dem menschlichen Denken zugeschrieben werden.

[2] Für diese Dissertation wird auf die Definition von Mitchell (1997) zurückgegriffen, um Machine Learning als eines der zentralen Teilgebiete der Künstlichen Intelligenz auszugestalten. Hiernach lernen Computerprogramme in Bezug auf eine Aufgabe und eine Performanz-Metrik durch Erfahrung, wenn sich die Performanz in der Aufgabe bei wachsender Erfahrung verbessert (Mitchell 1997).

2018; Clinciu, Hastie 2019; Rudin 2019) zu nennen. Als Grundlage für eine einheitliche Nutzung von KI wird in dieser Dissertation auf den Ethikleitlinien für vertrauenswürdige KI der Europäischen Union (Hochrangige Expertengruppe für künstliche Intelligenz 2018) aufgesetzt und basierend auf dem aktuellen Forschungsstand um weitere Facetten erweitert. Für die Transparenz von KI-Systemen wird hierzu die von Rudin (2014) eingeführte Differenzierung von inhärent interpretierbaren und (post hoc) erklärbaren Modellen genutzt, bei denen mithilfe eines weiteren Modells Informationen über das zur Vorhersage genutzte Modell gesammelt werden. Diese Erklärungsansätze können hierbei einzelne Vorhersagen beschreiben (lokale Erklärbarkeit) oder auch Informationen über die Entscheidungsfindung des Gesamtmodells liefern (globale Erklärbarkeit) (Guidotti et al. 2018).

Schlussfolgernd sind bei der Entwicklung von KI-Systemen neben zunehmenden Genauigkeiten der Vorhersage weitere Faktoren oder Prinzipien hinsichtlich der Vertrauenswürdigkeit zu berücksichtigen, um (1) den regulatorischen Ansprüchen in verschiedenen Branchen gerecht zu werden und (2) die Akzeptanz von potenziellen Nutzenden zu steigern. Trotz existierender Ansätze zu Prinzipien wie Transparenz, Datenschutz und Nichtdiskriminierung ist hierbei weiterhin eine hohe Heterogenität festzustellen, die eine konkrete Anwendbarkeit erschwert.

2 Motivation und Zielsetzung

Die Sicherstellung eines vertrauenswürdigen Einsatzes von KI ist in vielen Branchen zentral, um KI-Systeme erfolgversprechend nutzen und die Gesellschaft vor potenziellen Gefahren durch unethische KI-Systeme schützen zu können (Munoko et al. 2020). Dies kann jedoch ein komplexes Spannungsfeld zwischen den entwickelten Prinzipien und der angestrebten Förderung von KI-basierten Innovationen erzeugen (Morley et al. 2020). Ein domänenübergreifende Herausforderung stellt hierbei laut Miller und Coldicott (2019) die hohe Abstraktionsebene der entwickelten Prinzipien dar. Laut ihrer Studie wünschen sich 79 % aller befragten technischen Angestellten einen Transfer der theoretischen Prinzipien auf praktisch anwendbare Vorgehensweisen und Best Practices (Miller, Coldicott 2019).

In den vergangenen Jahren wurden bereits erste Vorhaben initiiert, um die Prinzipien weiter zu konkretisieren, wobei hauptsächlich die *Assessment List for Trustworthy Artificial Intelligence* (ALTAI) (Europäische Kommission 2020) sowie der *VDE SPEC 90012* (VDE 2022) Ansätze bieten. In beiden Fällen wurden Gestaltungsvorgaben abgeleitet und in Form von Fragebögen strukturiert, sodass eine Bewertung sowie langfristig eine Zertifizierung von KI-Systemen ermöglicht werden kann (Europäische Kommission 2020; VDE 2022). Um eine intuitive Hilfestellung zu bieten, ist es jedoch notwendig, diese weiter zu operationalisieren und in etablierte Vorgehensweisen und Entwicklungsprozesse von Unternehmen einzubetten (Miller, Coldicott 2019). Damit diese darüber hinaus auch in hochregulierten Domänen wie Wirtschaftsprüfung oder Smart Living einsetzbar sind, muss auf die individuellen Anforderungen der Domäne explizit eingegangen werden (Munoko et al. 2020). Für die Wirtschaftsprüfung stellen sich (Munoko et al. 2020) hierzu zwei Aspekte heraus, die eine Entwicklung von KI aktuell hemmen: (1) die fehlende Transparenz aktueller KI-Systeme und (2) unterschiedliche Anforderungen und Implikationen der beteiligten Akteure. Das Ökosystem Smart Living demgegenüber adressiert mit der intelligenten Unterstützung des Alltags einen der privatesten Bereiche des menschlichen Lebens und erfordert die Erhebung sensibler Daten sowie deren Verarbeitung unabhängig von Geschlecht oder Ethnie (Kortum et al. 2020; Senden, Xenidis 2020). Entsprechend stellen der Schutz dieser sensiblen Daten sowie das Verhindern von Diskriminierung zentrale Herausforderungen für den Einsatz von KI im Bereich Smart Living dar (Senden, Xenidis 2020).

Im Zuge dieser Dissertation werden drei dieser Herausforderungen aus den exemplarischen Bereichen Wirtschaftsprüfung und Smart Living in den Vordergrund gestellt. Für Transparenz, Datenschutz und Nichtdiskriminierung werden die bisher meist abstrakt beschriebenen Konzepte instanziiert und praktisch anwendbare Vorgehensweisen und Best Practices abgeleitet, um das Spannungsfeld zwischen den branchenspezifischen Anforderungen und den Potenzialen beim Einsatz von KI zu reduzieren und die Grundlage für ein etabliertes KI-Management zu legen. Hierzu werden IT-Artefakte konzeptioniert, prototypisch implementiert und in den Anwendungsdomänen Smart Living und Wirtschaftsprüfung evaluiert.

3 Einordnung

Die Wirtschaftsinformatik versteht sich als interdisziplinäre Realwissenschaft an der Schnittstelle zwischen der Betriebswirtschaftslehre und der Informatik (Thomas 2006, S. 10). Zentraler Gegenstandsbereich der Wirtschaftsinformatik sind Informationssysteme. Diese können dabei als soziotechnische Systeme verstanden werden, die sich aus Menschen (personelle Aufgabenträger), Informations- und Kommunikationstechnik sowie Organisation (Geschäftsprozesse, Funktionen, Strukturen und Management) und ihrer Interaktion zusammensetzen (Thomas 2006, S. 44; Österle et al. 2010).

Ausgehend vom Informationssystem als Erkenntnisgegenstand hat sich eine methodenpluralistische Erkenntnisstrategie entwickelt, die im angloamerikanischen Information Systems Research eher verhaltenswissenschaftlich ausgeprägt ist, während in der deutschsprachigen Wirtschaftsinformatik konstruktionsorientierte Ansätze vorherrschend sind (Wilde, Hess 2007). Die resultierenden Erkenntnisziele unterscheiden sich entsprechend. Während in der gestaltungsorientierten Wirtschaftsinformatik Handlungsanleitungen (normative, praktisch verwendbare Aussagen) sowie Innovationen zu instanziierten Informationssystemen im Vordergrund stehen, werden in der verhaltensorientierten Wirtschaftsinformatik vorwiegend informationssystembezogene Phänomene untersucht, um Ursache-Wirkungs-Zusammenhänge zu entdecken (Österle et al. 2010, S. 666-667). Die beschriebenen Erkenntnisse können dabei sowohl induktiv als auch deduktiv gewonnen werden (Thomas 2006, S.16).

In der gestaltungsorientierten Wirtschaftsinformatik wird jedoch primär auf deduktive Verfahren zurückgegriffen, um Wissen zu generieren (Wilde, Hess 2007; Österle et al. 2010; Gregor, Hevner 2013). Nach Gregor, Hevner (2013) lässt sich dieses Wissen für gestaltungsorientierte Ansätze wie folgt in deskriptives und präskriptives Wissen differenzieren:

- *Deskriptives Wissen* beinhaltet primär Phänomene, basierend auf Beobachtung, Klassifizierung, Messung und Katalogisierung sowie daraus resultierende Naturgesetze, Regelmäßigkeiten, Gesetzmäßigkeiten, Schemata oder Theorien.
- *Präskriptives Wissen* umfasst Konstrukte (Konzepte, Symbole), Modelle (Repräsentationen, Semantik/Syntax), Methoden (Algorithmen, Techniken), Instanziierungen von Systemen, Produkten oder Prozessen und Designtheorien.

Im Fokus dieser Dissertation stehen Informationssysteme, die um KI-basierte Ansätze ergänzt werden – im Folgenden werde diese als KI-Systeme bezeichnet. Durch das Anreichern der Systeme mit KI-Ansätzen ergeben sich, wie in Kapitel 2 beschrieben, jedoch vor allem in hochregulierten Domänen wie der Wirtschaftsprüfung Herausforderungen, die über die rein technische Ausgestaltung hinausgehen (Munoko et al. 2020). Vor allem im Themenkomplex zu vertrauenswürdiger und ethischer KI existiert bereits aus verschiedenen Disziplinen eine breite theoretische Basis zur Auflösung des resultierenden

Spannungsfeldes. Jedoch sind die entwickelten Prinzipien, wie in Kapitel 2 beschrieben, in den meisten Fällen auf einem Abstraktionsniveau, das eine praktische Anwendbarkeit bisher verhindert und somit wirtschaftliche und gesellschaftliche Potenziale ungenutzt lässt. Damit diese Potenziale gehoben werden können, sind konkrete Vorgehensweisen und Best Practices notwendig. Um hier anzusetzen und konkrete praktische sowie theoretische Erkenntnisse zu generieren, wird in dieser Dissertation primär auf konstruktionsorientierte Methoden und deduktive Ansätze zurückgegriffen und damit primär präskriptives Wissen generiert (Gregor, Hevner 2013).

Auf Basis der hier erfolgten Einordnung existieren nach Eberhard (1999) verschiedene Erkenntnisinteressen, die im Folgenden detaillierter adressiert werden.

4 Methodik

4.1 Forschungsfragen und Erkenntnisinteresse

Grundsätzlich existieren verschiedene Erkenntnisinteressen, die mithilfe entsprechend ausgestalteter Forschungsfragen adressiert werden können. Nach Eberhard (1999, S. 16 - 19) führen individuelle, kollektive und gesellschaftliche Probleme zu drei zentralen Interessen, dem *phänomenalen*, dem *kausalen* und dem *aktionalen Erkenntnisinteresse*, die mithilfe von Erkenntniswegen, und -angeboten zurück in die Praxis überführt werden können. Das *phänomenale Erkenntnisinteresse* betrachtet faktische Gegebenheiten basierend auf ihren Merkmalen und Eigenschaften, umgangssprachlich formulierbar mit den Fragen „Was ist los?“ oder „Was geschieht?“. Dahingegen fokussiert das *kausale Erkenntnisinteresse* die Ursachen der Phänomene, adressierbar mit den Fragen „Warum ist das so?“ oder „Warum geschieht es?“. Das abschließende *aktionale Forschungsinteresse* adressiert explizite Handlungsmöglichkeiten und Ansätze zur strategischen Beeinflussung der Phänomene, was mithilfe der Frage „Was ist zu tun?“ ausgedrückt werden kann.

Aufsetzend auf dem hauptsächlich verfolgten gestaltungsorientierten Forschungsparadigma und der beschriebenen Ausgangslage wird in dieser Dissertation vorrangig ein phänomenales und ein aktionales Forschungsinteresse verfolgt. Der Hauptforschungsgegenstand der Arbeit ist hierbei die Implementierung und die unternehmensbezogene Einbettung von vertrauenswürdigen KI-Systemen. Daraus resultiert die folgende Leitfrage:

FF: Wie können KI-Systeme vertrauenswürdig ausgestaltet werden, um sowohl regulatorischen Anforderungen zu genügen als auch die Nutzerakzeptanz für einen erfolgreichen Einsatz sicherzustellen?

Aufgrund der Komplexität und der für zahlreiche Branchen individuellen Anforderungen an die Gestaltung von KI-Systemen wird die Forschungsfrage auf die Bereiche Smart Living und Wirtschaftsprüfung, die sich trotz vielfältiger Unterschiede beide durch hohe regulatorische und akzeptanzbezogene Ansprüche auszeichnen. Die dafür relevantesten Prinzipien werden in drei Teilfragen (FF1, FF2, FF3) adressiert und in Teilfrage 4 (FF4) in den Unternehmenskontext eingebettet, um ein explizites KI-Management unter Berücksichtigung branchenspezifischer Anforderungen zu etablieren. In der ersten Forschungsfrage wird die Transparenz als zentrales Prinzip vertrauenswürdiger KI-Systeme im Sinne eines phänomenalen sowie aktionalen Forschungsinteresses untersucht:

FF1: Welche Anforderungen haben die beteiligten Akteure in hochregulierten Domänen wie der Wirtschaftsprüfung an die Transparenz von KI-Algorithmen und wie können diese in KI-Systemen konkret instanziiert werden?

Aus FF1 haben sich divergierende Anforderungen an die Transparenz von KI-Algorithmen in der Wirtschaftsprüfung ergeben. Als zentral hat sich demzufolge eine rollenindividuelle Betrachtung erwiesen, um Anforderungen im Hinblick auf die notwendige Transparenz sorgfältig auflösen zu können. Die resultierenden Erkenntnisse wurden am Beispiel eines KI-Systems evaluiert, um konkrete Lösungsansätze zur transparenten Ausgestaltung von KI zu bieten. Diese und vergleichbare KI-Systeme basieren dabei sowohl für das Training als auch für die eigentliche Nutzung auf personenbezogenen oder auch unternehmenskritischen Daten. Um den dabei entstehenden Ansprüchen im Hinblick auf Datenschutz und Datensouveränität in komplexen Systemen gerecht zu werden, werden aufsetzend auf der folgenden Forschungsfrage mögliche Lösungsansätze betrachtet:

FF2: Wie können Datenschutz und -souveränität in KI-Systemen im Hinblick auf Regulatorik und Akzeptanz sichergestellt werden?

Hierbei haben sich zwei unterschiedliche Lösungsansätze als vielversprechend herausgestellt. Erstens kann die Offenlegung jeglicher auf den eigenen freigegebenen Daten basierenden Datenflüssen und Weiterverarbeitungsschritten technisch unterstützt werden. Hierdurch kann die grundsätzliche Datensouveränität aller an Datenökosystemen[3] partizipierenden Akteure sichergestellt werden. Dabei kann jedoch die für KI-Algorithmen essenzielle Datenbasis kritisch eingeschränkt werden. Aufgrund dessen wurde zweitens mit dem Similarity Preserving Hashing (SPH) ein Ansatz zur Anonymisierung von Daten für menschliche Nutzende bei gleichbleibenden von KI nutzbaren Ähnlichkeiten evaluiert. Allerdings werden im Bereich Smart Living die hierdurch entwickelbaren KI-Systeme von Menschen verschiedener Ethnien genutzt, wodurch die Diskriminierungsfreiheit einen zentralen Erfolgsfaktor sowohl aus gesellschaftlicher als auch aus wirtschaftlicher Sicht darstellt. Dies wurde in der folgenden Forschungsfrage weiter betrachtet:

FF3: Wie können KI-Systeme entwickelt werden, um deren diskriminierungsfreie Nutzung sicherzustellen?

Forschungsfrage 3 verfolgt ein aktionales Forschungsinteresse, das mithilfe einer Instanziierung sowie abgeleiteter Erkenntnisse beantwortet wurde. Hierzu wurden Implikationen auf das Design des KI-Systems betrachtet, wobei die Ableitung eines Vorgehensmodells zur diskriminierungsfreien Entwicklung von KI-Systemen im Fokus stand.

[3] Datenökosysteme stellen einen Zusammenschluss von Datenerstellern, -konsumenten und -intermediären als wesentliche Akteure dar, in denen Daten als zentrale Ressourcen generiert, konsumiert und verarbeitet werden (Oliveira, Lóscio 2018; Kortum et al. 2020).

Um die vertrauenswürdige Entwicklung von KI-Systemen in Unternehmen zu verankern, muss jedoch nicht nur eine Adaption des genutzten Vorgehensmodells erfolgen. Vielmehr sind für die langfristige Ausrichtung eines Unternehmens auf die Entwicklung und Nutzung von KI-Systemen tiefergehende Anpassungen im Unternehmen notwendig. Diese wurden in FF4 genauer betrachtet:

FF4: Wie können Unternehmensstrukturen auf die Entwicklung und Nutzung von KI-Systemen ausgerichtet werden, um branchenspezifische Anforderungen zur Vertrauenswürdigkeit sicherstellen zu können?

Um eine Quantifizierbarkeit der relevanten Unternehmensfacetten sowie ihrer Ausprägungen zu ermöglichen, wurde ein Reifegradmodell für die Wirtschaftsprüfung entwickelt. Im Folgenden werden die zur Beantwortung der Forschungsfragen relevanten Methoden dargestellt.

4.2 Methodenspektrum

Wie in Kapitel 3 eingeordnet, adressiert die deutschsprachige Wirtschaftsinformatik zwar eine methodenpluralistische Erkenntnisstrategie, es ist jedoch ein klarer Fokus auf konstruktionsorientierte Methoden feststellbar, die sich auch in dieser Dissertation widerspiegeln (Wilde, Hess 2007; Österle et al. 2010). Vor allem durch die prototypische Implementierung und Evaluation von IT-Systemen lassen sich, orientiert an Design-Science-Research-Vorgehensweisen (DSR) von Hevner et al. (2004) und Peffers et al. (2007), sowohl theoretische als auch praktische Implikationen für die Gestaltung von Modellen, Methoden oder Systemen ableiten (Wilde, Hess 2007). Hierzu wurde eine gezielte Auswahl an wissenschaftlichen Methoden genutzt, die in den individuellen Beiträgen genauer erläutert werden. Im Folgenden werden die relevantesten Methoden kurz beschrieben:

- *Literaturanalysen* adressieren die systematische Erhebung und Analyse des existierenden in (wissenschaftlichen) Publikationen expliziten Wissens mithilfe standardisierter Schritte (Rowley, Slack 2004). Hierzu werden der Umfang des Reviews (Cooper 1988) sowie die Schlüsselbegriffe für den Suchstring festgelegt (Rowley, Slack 2004), um anschließend die resultierenden Publikationen systematisch auszuwählen und zu analysieren (vom Brocke et al. 2009). Literaturanalysen wurden in dieser Dissertation sowohl zur Erhebung des aktuellen Wissensstandes und der Identifikation von Forschungslücken als auch zur Herleitung von Gestaltungswissen genutzt.

- Die *prototypische Implementierung* von IT-Systemen stellt neben der technischen Funktionalität (Thomas 2006) die Basis für die Entwicklung von Prinzipien und Theorien im Dreieck von Menschen, Informationstechnik und Organisation nach Österle et al. (2010) dar (Hevner et al. 2004). Die Implementierung kann systematisch in DSR-Vorgehensweisen (Peffers et al. 2007) eingebettet werden und generiert im Zusammenspiel mit Evaluationen vor allem präskriptives Wissen (Gregor, Hevner 2013). Als einer der Kernbestandteile dieser Dissertation finden sich prototypische Implementationen in vier der eingebrachten sechs Beiträge wieder.

- Mithilfe der *Demonstration* wird der Nutzen von Artefakten zur Lösung von Probleminstanzen gezeigt. In der Demonstration können Experimente, Simulationen, Fallstudien und weitere Aktivitäten eingesetzt werden (Peffers et al. 2007). In dieser Dissertation wurden Demonstrationen primär im Zuge des DSR-Vorgehens nach Peffers et al. (2007) verwendet.

- *Experteninterviews* stellen eine der bedeutsamsten und am häufigsten genutzten qualitativen Methoden zur Sammlung von Daten dar (Myers, Newman 2007). Hierzu kann eine Einbindung in verschiedenste Forschungsansätze und Methoden wie Fallstudien oder auch Aktionsforschung erfolgen (Myers, Newman 2007). Im Zuge der Dissertation wurden Experteninterviews sowohl bei der Erhebung von Anforderungen als auch bei der Evaluation von IT-Artefakten eingesetzt.
- Die *qualitative Inhaltsanalyse* ist eine quantifizierende Methode zur Textanalyse, die durch Gläser und Laudel (2009) in fünf Kernphasen strukturiert wurde. Bei dieser Dissertation wurden qualitative Inhaltsanalysen primär als Auswertungswerkzeug für Experteninterviews eingesetzt.
- Im Gegensatz zu Experteninterviews adressieren interaktive *Fokusgruppen* auf der Grundlage von Morgan (1996) sowie Sutton und Arnold (2013) explizit die Schöpfung von kreativen Potenzialen in kooperativen Situationen. Im Zuge der Dissertation wurden Fokusgruppen sowohl zur Erhebung von Anforderungen als auch zur Konzeptionierung von Artefakten und der Evaluation resultierender Konzepte genutzt.
- *Deduktive Analysen* stellen eine Methode der gestaltungsorientierten Wirtschaftsinformatik dar und können in formal-, konzeptionell- und argumentativ-deduktive Analysen differenziert werden (Wilde, Hess 2007). Für die Dissertation wurden primär argumentativ-deduktive Analysen verwendet.
- Die *Entwicklung von Reifegradmodellen* nach Becker et al. (2009) standardisiert die vormals häufig willkürlichen Entwicklungsprozesse für Reifegradmodelle als ein Baustein für ein erfolgreiches IT-Management. Im Zuge der Dissertation wurde das Vorgehen zur Entwicklung eines Reifegradmodells für den Einsatz von KI in der Wirtschaftsprüfung eingesetzt.

4.3 Forschungsplan

Auf Basis der in Kapitel 4.1 beschriebenen Forschungsfragen sowie den in Kapitel 4.2 präsentierten Methoden ist in Abb. 1 der Forschungsplan der vorliegenden Dissertation dargestellt. Der Forschungsplan strukturiert die Leitfrage in vier zentrale Forschungsfragen, die bei FF1 und FF2 in zwei weitere Teilprobleme aufgegliedert wurden. Für die jeweiligen Fragen wurden die Kernergebnisse zusammengefasst und die jeweilige Anwendungsbranche verdeutlicht. Die Forschungsbeiträge fokussieren dabei auf die Branchen Wirtschaftsprüfung und Smart Living. Einzig Unterfrage 2b wurde branchenunabhängig beantwortet, kann jedoch bei Bedarf auf die jeweiligen Branchen übertragen werden.

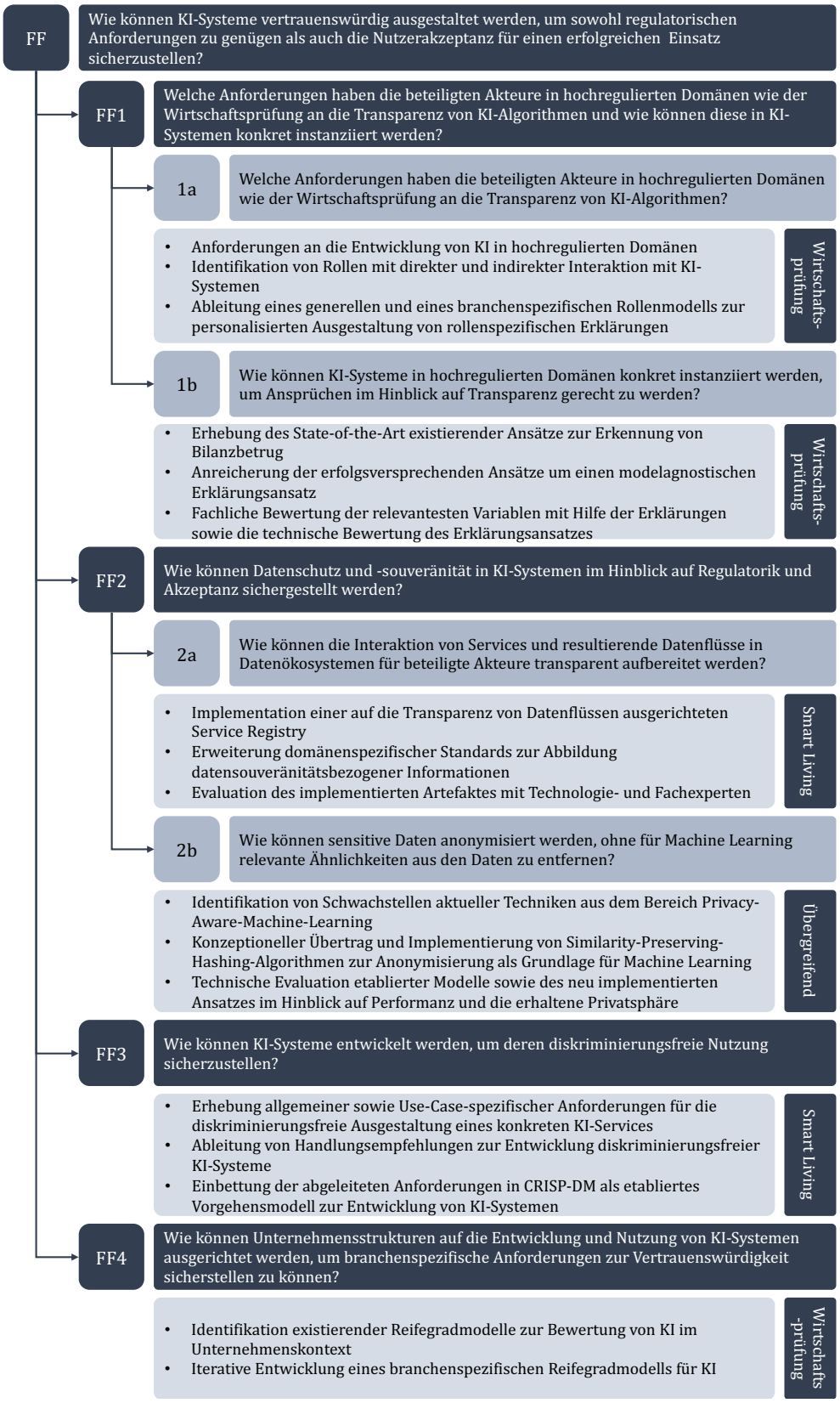

Abb. 1. Forschungsplan der Dissertation

5 Ergebnisse

5.1 Überblick

Die Dissertation umfasst insgesamt 22 Beiträge, wovon B1 bis B21 bereits veröffentlicht wurden, während sich B22 aktuell noch im Veröffentlichungsprozess befindet. Von den 22 Beiträgen werden die Beiträge B1 bis B6 in das Kernthema der Dissertation eingebracht. In Tab. 1 sind die genannten Beiträge aufgelistet und in Bezug zu den in Kapitel 4.1 vorgestellten Forschungsfragen gesetzt.

Tab. 1. Überblick über die publizierten Forschungsbeiträge

			Ranking[4]			
#	*Publikationsorgan*	*Medium*	WK WI	VHB JQ3	*Bibliographische Informationen*	*FF*
B1	International Conference on Wirtschaftsinformatik (WI 2022)	Tagung	A	C	**Rebstadt, J.**; Remark, F.; Fukas, P.; Meier, P.; Thomas, O. (2022): Towards personalized explanations for AI systems: designing a role model for explainable AI in auditing. In: Wirtschaftsinformatik 2022 Proceedings. 2.	1
B2	International Conference on Advanced Information Systems Engineering (CAiSE 2022)	Tagung	B	C	Fukas, P.; **Rebstadt, J.;** Menzel, L.; Thomas, O. (2022): Towards Explainable Artificial Intelligence in Financial Fraud Detection: Using Shapley Additive Explanations to Explore Feature Importance. In: Franch, X; Poels, G; Gailly, F; Snoeck, M (Hrsg.), International Conference on Advanced Information Systems Engineering. Springer, Cham, S. 109-126.	1
B3	INFORMATIK 2021	Tagung	B	C	**Rebstadt, J.;** Kortum, H.; Hagen, S.; Thomas, O., (2021): Towards a transparency-oriented and integrating Service Registry for the Smart Living Ecosystem. In: Gesellschaft für Informatik e.V. (GI) (Hrsg.), INFORMATIK 2021. Gesellschaft für Informatik, Bonn, S. 1425-1438.	2
B4	INFORMATIK 2022	Tagung	B	C	Eleks, M.; **Rebstadt, J.;** Fukas, P.; Thomas, O., (2022): Learning without Looking: Similarity Preserving Hashing and Its Potential for Machine Learning in Privacy Critical Domains. In: Demmler, D.; Krupka, D.; Federrath., H. (Hrsg.), INFORMATIK 2022. Gesellschaft für Informatik, Bonn, S. 161-177.	2
B5	HMD Praxis der Wirtschaftsinformatik	Journal	B	D	**Rebstadt, J.;** Kortum, H.; Gravemeier, L. S.; Eberhardt, B.; Thomas, O. (2022): Non-Discrimination-by-Design: Handlungsempfehlungen für die Entwicklung von vertrauenswürdigen KI-Services. In: HMD Praxis der Wirtschaftsinformatik, Nr. 59(2), S. 495–511.	3
B6	European Conference on Information Systems (ECIS 2022)	Tagung	A	B	Fukas, P.; **Rebstadt, J.;** Remark, F.; Thomas, O. (2021): Developing an Artificial Intelligence Maturity Model for Auditing. In: European Conference on Information System (ECIS 2021), A Virtual AIS Conference, Research Paper. 133.	4
B7	INFORMATIK 2022	Tagung	B	C	Kortum, H.; **Rebstadt, J**.; Böschen, T.; Meier, P.; Thomas, O., (2022): Towards the Operationalization of Trustworthy AI: Integrating the EU Assessment List into a Procedure Model for the Development and Operation of AI-Systems. In: Demmler, D.; Krupka, D.; Federrath., H. (Hrsg.), INFORMATIK 2022. Gesellschaft für Informatik, Bonn, S. 283–299.	1, 3, 4
B8	Hawaii International Conference on System Sciences (HICCS 2022)	Tagung	B	C	Kortum, H.; **Rebstadt, J**.; Hagen, S.; Thomas, O., (2022): Integrating Data and Service Lifecycle for Smart Service Systems Engineering: Compilation of a Lifecycle Model for the Data Eco-system of Smart Living. In: Proceedings of the 55th Hawaii International Conference on System Sciences. 2022.	-

[4] Die Rankings der jeweiligen Beiträge wurden auf Basis der WI-Orientierungsliste der WKWI (WI-Journalliste 2008, Stand 2008-03-03, v39; WI-Liste der Konferenzen, Proceedings und Lecture Notes 2008, Stand 2008-03-03, v27) und des VHB-Jourqual 3 – Teilrating WI ermittelt.

#	Publikationsorgan	Medium	Ranking[4] WK WI	Ranking[4] VHB JQ3	Bibliographische Informationen	FF
B9	Hawaii International Conference on System Sciences (HICCS 2022)	Tagung	B	C	Kortum, H.; **Rebstadt, J.**; Thomas, O., (2022): Dissection of AI Job Advertisements: A Text Mining-based Analysis of Employee Skills in the Disciplines Computer Vision and Natural Language Processing. In: Proceedings of the 55th Hawaii International Conference on System Sciences. 2022.	-
B10	International Conference on Wirtschaftsinformatik (WI 2022)	Tagung	A	C	Kortum, H.; Fukas, P.; **Rebstadt, J.;** Eleks, M.; Nobakht Galehpardsari, M.; Thomas, O. (2022): Proposing a Roadmap for Designing Non-Discriminatory ML Services: Preliminary Results from a Design Science Research Project. In: Wirtschaftsinformatik 2022 Proceedings. 3.	3
B11	Dienstleistungsinnovationen durch Digitalisierung – Band 2: Prozesse – Transformation – Wertschöpfungsnetzwerke	Buchband	-	-	Brinker, J.; Kammler, F.; Hagen, S.; Remark, F.; **Rebstadt, J.;** Jasper, M.; Dollmann, T; Nüttgens, M.; Thomas, O. (2021): smartTCS – Eine Plattform zur flexiblen Einbindung von Kunden in technische Dienstleistungen für den Maschinen- und Anlagenbau. In: Beverungen, D., Schumann, J.H., Stich, V., Strina, G. (Hrsg.) Dienstleistungsinnovationen durch Digitalisierung. Springer Gabler, Berlin, Heidelberg, S. 439-482.	-
B12	Smart Glasses: Augmented Reality zur Unterstützung von Logistikdienstleistungen	Buchband	-	-	Straede, H.; **Rebstadt, J.**; Hucke, S.; Thomas, O. (2020): Logistische Prozesse in der erweiterten Realität: Konzeption und Implementierung eines Smart-Glasses-basierten Systems. In: Thomas, O.; Ickerott I. (Hrsg.): Smart Glasses: Augmented Reality zur Unterstützung von Logistikdienstleistungen, Springer Gabler, S. 106-118.	-
B13	HMD Praxis der Wirtschaftsinformatik	Journal	B	D	Kortum, H.; **Rebstadt, J.**; Gravemeier, L. S.; Thomas, O. (2021): Data-based Customer-Retention-as-a-Service: Induktive Entwicklung eines datenbasierten Geschäftsmodells auf Basis einer Fallstudie der Automobilbranche. In: HMD Praxis der Wirtschaftsinformatik, 58(3), 537–551.	-
B14	AI Perspectives	Journal	-	-	Barenkamp, M.; **Rebstadt, J.**; Thomas, O. (2020): Applications of AI in classical software engineering. In: AI Perspectives, 2(1), S. 1–15.	-
B15	Langfassung der Studie Audit Clouds	Buchband	-	-	Thomas, O.; Langhein, J.; Sack, M.; Feld, T.; Remark, F.; **Rebstadt, J.** (2019): Langfassung der Studie Audit Clouds – Analyse und Vergleich cloudbasierter Geschäftsmodelle in der Wirtschaftsprüfung – Explorative Datenanalyse, Technologieakzeptanzuntersuchung und qualitative Inhaltsanalyse. Institut der Wirtschaftsprüfer (IDW) Verlag GmbH.	-
B16	Die Wirtschaftsprüfung (WPg)	Journal	-	C	Thomas, O.; Langhein, J.; Sack, M.; Feld, T.; Remark, F.; **Rebstadt, J.** (2019): Audit Clouds – Analyse und Vergleich cloudbasierter Geschäftsmodelle in der Wirtschaftsprüfung. Die Wirtschaftsprüfung (WPg), 18(72), S. 964–975.	-
B17	Die Wirtschaftsprüfung (WPg)	Journal	-	C	Thomas, O.; Sack, M.; Langhein, J.; Pöhlmann, A.; Feld, T.; Remark, F.; **Rebstadt, J**. (2020): Audit Clouds – Akzeptanz cloudbasierter Geschäftsmodelle in der Wirtschaftsprüfung. Die Wirtschaftsprüfung (WPg), 1(73), S. 2–10	-
B18	International Journal of Semantic Computing	Journal	-	-	Hartmann, S.; Hallay, F.; Brinkschulte, L.; **Rebstadt, J.;** Gesterkamp, L.; Enders, A.; Kewitz, N.; Mertens, R. (2017): Aspect-oriented visual ontology editor (avoned): Visual language, aspect-oriented editing concept and implementation. In: International Journal of Semantic Computing, 11(02), S. 229–274.	-
B19	International Conference on Semantic Computing (ICSC 2017)	Tagung	-	-	Gesterkamp, L.; **Rebstadt, J.**; Mertens, R. (2017): Tackling Complex Ontologies with AVOnEd—Aspect-Oriented Visual Ontology Editor. In: 2017 IEEE 11th International Conference on Semantic Computing (ICSC). IEEE, S. 268-269.	-
B20	SEMANTiCS 2016	Tagung	-	-	**Rebstadt, J.**; Brinkschulte, L.; Enders, A.; Mertens, R. (2016): A Visual Language for OWL Lite Editing. In: SEMANTiCS (Posters, Demos, SuCCESS).	-
B21	International Conference on Semantic Computing (ICSC)	Tagung	-	-	Brinkschulte, L.; Enders, A.; **Rebstadt, J.**; Mertens, R. (2016): Aspect-oriented mind mapping and its potential for ontology editing. In: 2016 IEEE Tenth International Conference on Semantic Computing (ICSC). IEEE, S. 194-201.	-
B22	Die Wirtschaftsprüfung (WPg)	Journal	-	C	**Rebstadt, J.**; Fukas, P.; Remark, F.; Thomas, O.; Sack, M.; 1,2 Pöhlmann, A. (2023): Vertrauenswürdigkeit und Transparenz: kritische Erfolgsfaktoren für den Einsatz von Künstlicher Intelligenz in der Abschlussprüfung In: Die Wirtschaftsprüfung (WPg), 12(76), S. 665–673.	1, 2

Nach Österle et al. (2010) liegt der Wissensbestand der Wirtschaftsinformatik jedoch nicht nur in der wissenschaftlichen Literatur. Vielmehr liegt ein Großteil des Wissens in Form von Informationssystemen, Software, organisatorischen Lösungen oder auch Methoden in der Wirtschaft. Entsprechend dieser Differenzierung haben Becker et al. (2004, S. 335 - 347) zentrale Forschungsziele und Aufträge der Wirtschaftsinformatik definiert. Hierzu wird zwischen Erkenntniszielen und Gestaltungszielen unterschieden. Während sich das Erkenntnisziel auf das Verständnis und mögliche Prognosen zur Veränderung gegebener Sachverhalte bezieht, setzen Gestaltungsziele auf die erkenntnisgeleitete Forschung auf, um Sachverhalte zu verändern oder sogar neu zu erschaffen (Becker et al. 2004, S. 346). Darüber hinaus werden zwei inhaltliche Schwerpunkte differenziert und als Aufträge formuliert. Im methodischen Auftrag werden Techniken zur Beschreibung, Entwicklung, Einführung und Nutzung von Informationssystemen sowie die Entwicklung von Methoden subsumiert (Becker et al. 2004, S. 347). Der inhaltlich-funktionale Auftrag adressiert hingegen die explizite Ausgestaltung von Informationssystemen für ausgewählte Branchen (Becker et al. 2004, S. 347). Entsprechend der dabei aufgespannten Matrix von Zielsetzungen und Aufträgen werden in Abb. 2 die eingebrachten Beiträge eingeordnet.

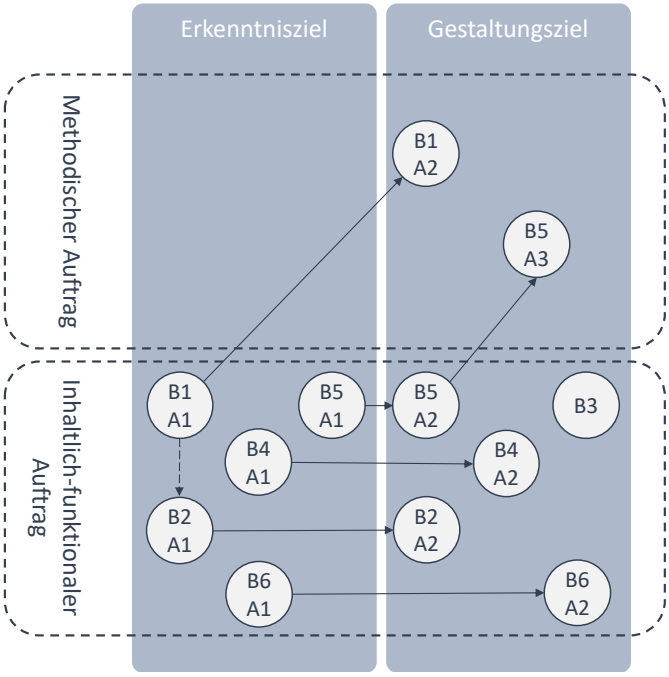

Abb. 2. Einordnung der eingebrachten Beiträge (Bx) und ihrer Kernartefakte (Ax) in Abhängigkeit von Ziel und Aufgabe in Anlehnung an Becker et al. (2004)

Im Folgenden werden die eingebrachten Beiträge auf Basis ihrer Ziele und Aufträge dargestellt:

(1) Im ersten Beitrag wurden mithilfe von Experteninterviews Anforderungen an transparente KI-Systeme in der Wirtschaftsprüfung erhoben (B1 A1). Um diese Anforderungen zu adressieren, wurde anhand einer systematischen Literaturrecherche und Fokusgruppeninterviews iterativ ein generalistisches Rollenmodell entwickelt, evaluiert und für die Wirtschaftsprüfung instanziiert (B1 A2).

(2) Aufbauend auf den Erkenntnissen des ersten Beitrages wurde ein exemplarisches erklärbares KI-System zur Erkennung von Bilanzbetrug prototypisch in der Wirtschaftsprüfung implementiert. Dazu wurde auf die Fundierung einer systematischen Literaturrecherche zurückgegriffen (B2 A1) und eine technische Evaluation mithilfe von etablierten Metriken durchgeführt (B2 A2).

(3) Beitrag 3 untersucht die transparente Orchestrierung von Services in komplexen Datenökosystemen. Mithilfe einer Erweiterung des WOT-Standards[5] sowie der prototypischen Implementierung einer Service Registry wird ein Ansatz zur technischen Unterstützung bei der Orchestrierung von Services und die Nachvollziehbarkeit von Datenflüssen durch Akteure des Ökosystems evaluiert (B3).

(4) In Beitrag 4 werden verschiedene Ansätze zur Sicherstellung von Privacy Aware Machine Learning (PAML) literaturbasiert analysiert (B4 A1). Um die identifizierten Schwachstellen zu adressieren, wurde SPH aus der Datenforensik übertragen. Eine Auswahl der identifizierten Ansätze wurde prototypisch implementiert und bezüglich der resultierenden Modellgüte und erhaltener Informationen mittels Mutual Information evaluiert (B4 A2).

(5) Im fünften Beitrag wurden literaturbasiert Anforderungen und Vorgehensweisen zur Entwicklung von diskriminierungsfreien KI-Systemen erhoben (B5 A1) und am Beispiel des intelligenten Gebäudepförtners prototypisch instanziiert (B5 A2). Die dabei erhobenen praktischen Anforderungen wurden mit den Erkenntnissen aus der Literatur kombiniert, in Handlungsempfehlungen überführt und in das etablierte Vorgehensmodell CRISP-DM (Shearer et al. 2000; Wirth 2000) eingebettet (B5 A3).

(6) In Beitrag 6 wurde iterativ ein Reifegradmodell für KI in der Wirtschaftsprüfung entwickelt (B6 A2). Dafür wurden, aufsetzend auf Becker et al. (2009), in einem multimethodischen Ansatz Erkenntnisse aus Experteninterviews und einer systematischen Literaturrecherche zusammengeführt (B6 A1).

Hinsichtlich der betrachteten Zielsetzungen und der Aufträge wurde das Vorgehen in dieser Dissertation auf den idealtypischen Erkenntnisprozess von Österle et al. (2010) ausgerichtet. Der Prozess beinhaltet dabei mit Analyse, Entwurf, Evaluation und Diffusion vier Phasen. Die Initiierung in der Analyse erfolgt auf Basis einer Problemstellung, die in der Praxis oder in der Wissenschaft aufkommen kann (Österle et al. 2010). Basierend auf einem hieraus abgeleiteten Forschungsplan werden die zur Problemlösung benötigten Artefakte anhand anerkannter Methoden hergeleitet und im Anschluss evaluiert (Österle et al. 2010). Den Abschluss des Erkenntnisprozesses stellt die Diffusion dar. Hierbei wird versucht, die Ergebnisse für alle potenziell interessierten Gruppen aufzubereiten und ihnen Zugang zu ermöglichen. Dies kann in Form von Publikationen, aber auch mithilfe von Förderanträgen, Implementierungen in privaten Betrieben oder auch in Spin-offs erfolgen (Österle et al. 2010). In Abb. 3 wurden die eingebrachten Beiträge dieser Dissertation in die beschriebenen Phasen eingeordnet und im Hinblick auf ihre Anwendungsdomänen differenziert.

[5] Das Web of Things (WOT) bietet, wie definiert in *https://www.w3.org/WoT/documentation/*, standardisierte Technologie-Bausteine, um die Entwicklung von IOT-Anwendungen zu vereinfachen. Hierdurch wird die Flexibilität, Interoperabilität sowie die Wiederverwendung von etablierten Standards und Tools ermöglicht (Raggett 2015).

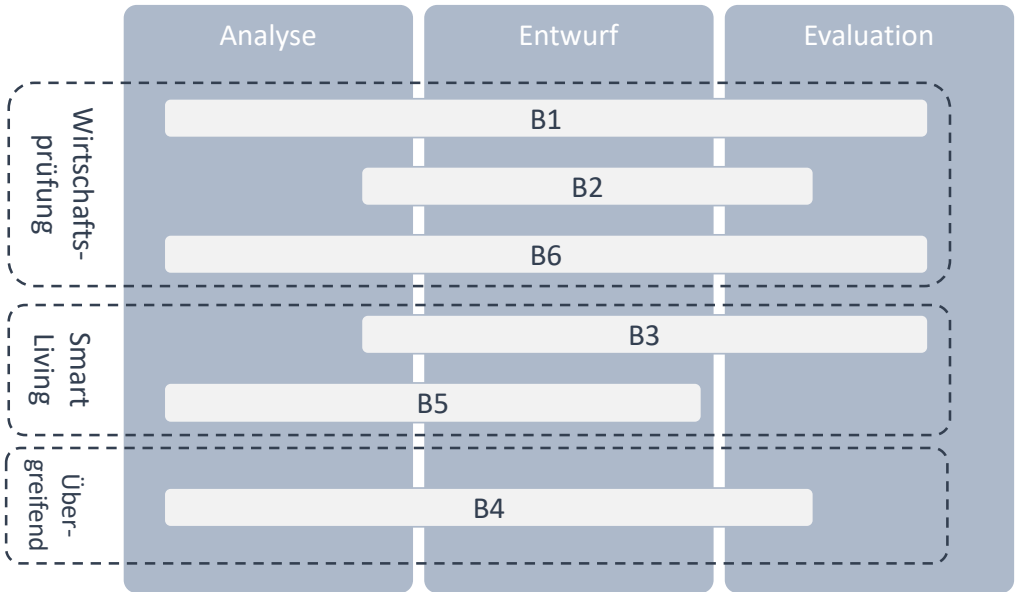

Abb. 3. Einordnung der eingebrachten Beiträge in den Erkenntnisprozess in Anlehnung an Österle et al. (2010)

Die Ergebnisse, die im Zuge dieser Dissertation entstanden sind, wurden neben den wissenschaftlichen Aufsätzen, die im folgenden Kapitel detaillierter erläutert werden, erweitert und zu potenziell interessierten Anspruchsgruppen getragen. Erstens wurden die Ansätze und Ideen zu vertrauenswürdiger KI und PAML weiterentwickelt, um in bereits betrachteten Domänen wie der Wirtschaftsprüfung und der Wohnungswirtschaft die betrachteten Ansätze zu nutzen und Problemstellungen bei der Bereitstellung von Daten zu adressieren. Zweitens wurden diese mit der Medizin auf eine weitere hochregulierte Domäne übertragen. Hieraus haben sich zwei drittmittelgeförderte Forschungsprojekte ergeben. Im Kontext des BMWK-Projektes *SECAI* werden Ansätze des PAML genutzt, um Heizungen und Heizverhalten intelligenter steuern zu können. Bei *KardioInterakt* (BMBF) hingegen werden diese Ansätze in den Medizinbereich übertragen, um die Nachsorge von Kardiokrankheiten sensorgestützt zu verbessern und eine kontaktreduzierte Weiterbetreuung zu ermöglichen. Drittens wurden auf Basis des Technologietransferprozessmodells von Scheer (1993) die erzielten Ergebnisse über ein Technologie-Spin-off in Beratungs- und Implementierungsprojekten mit operativ agierenden Unternehmen sowie übergeordneten Berufsständen direkt in die Praxis der Wirtschaftsprüfung und der Wohnungswirtschaft zurückgegeben.

5.2 Zentrale Ergebnisse der Beiträge

In den folgenden Abschnitten werden die zentralen Ergebnisse sowie die Implikationen der eingebrachten Beiträge dargestellt. Aufbauend auf dem in Kapitel 4.3 aufgestellten Forschungsplan werden das genutzte Vorgehen sowie die resultierenden Artefakte vorgestellt. Bei existierenden Verbindungen zwischen aufeinander folgenden Beiträgen werden zudem die jeweiligen Verknüpfungen aufgezeigt.

5.2.1 Erhebung und rollenspezifische Differenzierung von Anforderungen zum Einsatz transparenter KI

Die Entwicklung sowohl zunehmend leistungsfähigerer Hardware als auch neuer vielversprechender ML-Methoden hat in zahlreichen Branchen für eine stärkere Durchdringung

von KI-Methoden gesorgt (Perrault et al. 2019). Trotz eines großen Potenzials zur technischen Unterstützung bis hin zur (Teil-)Automatisierung von ganzen Prozessschritten bleibt die Wirtschaftsprüfung beim praktischen Einsatz jedoch hinter anderen Branchen zurück (Issa et al. 2016; Munoko et al. 2020). Ein zentraler Faktor hierfür sind branchenspezifische Gesetze und Standards, wodurch die eingesetzten Softwaresysteme strengen Anforderungen unterliegen, wie die notwendige Transparenz der getroffenen Entscheidungen. In Beitrag 1 dieser Dissertation wurden mithilfe von national sowie international tätigen Wirtschaftsprüfern in Experteninterviews Anforderungen zum Einsatz von KI in der Jahresabschlussprüfung erhoben und im Hinblick auf die zentrale Transparenz der Entscheidungsfindung kategorisiert. Diese Anforderungen sind in Abb. 4 visualisiert.

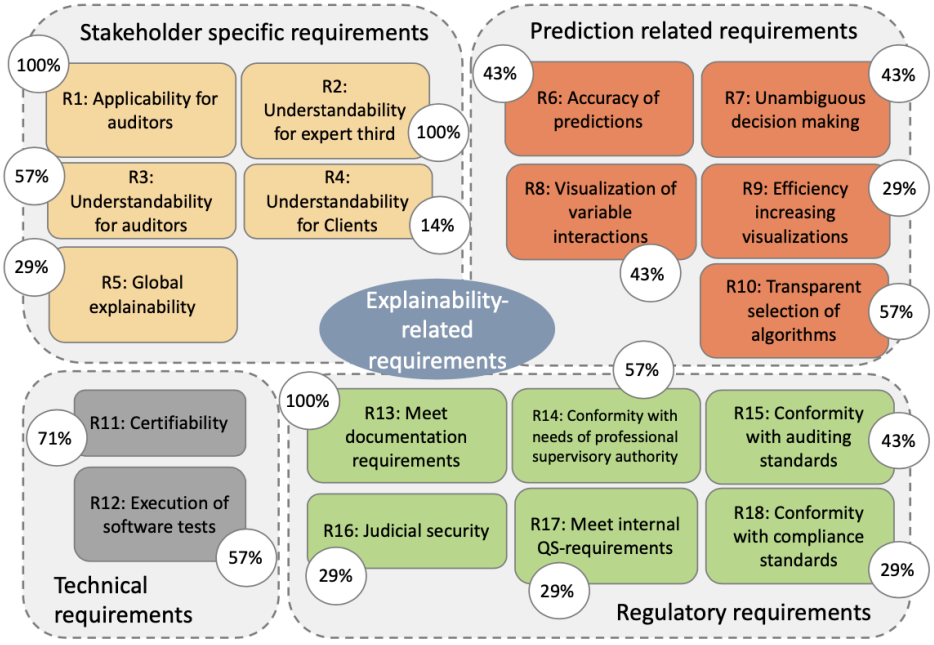

Abb. 4. Transparenzbezogene Anforderungen zum Einsatz von KI in der Wirtschaftsprüfung (Rebstadt et al. 2022b)

Während die meisten Anforderungen in der Entwicklung von KI-Systemen mit entsprechenden Vorgehensweisen direkt adressierbar sind, wurden bei der Ausgestaltung der Erklärungsansätze rollenspezifische Diskrepanzen deutlich. Diese ergeben sich sowohl aus den unterschiedlichen Zielsetzungen bei der direkten oder auch indirekten Nutzung von KI-Systemen als auch durch das unterschiedliche technische sowie fachliche Know-how der Nutzenden. Um hierbei die Anforderungen aller potenziell relevanten Rollen adäquat abbilden zu können, wurden basierend auf der existierenden Literatur sowohl direkt als auch indirekt mit KI-Systemen interagierende Rollen erhoben und bezüglich der Bereitstellung von Erklärungen in einem Rollenmodell strukturiert. In drei Fokusgruppen wurden diese Interaktionen evaluiert und für die Wirtschaftsprüfung instanziiert. Das resultierende wirtschaftsprüfungsbezogene Rollenmodell in Abb. 5 beinhaltet hierbei aufgrund der existierenden Regulatorik und der gelebten Praxis einzelne Vereinfachungen, aber auch angepasste Kommunikationsstrukturen, die sich potenziell auch auf die Ausgestaltung der Erklärungen auswirken.

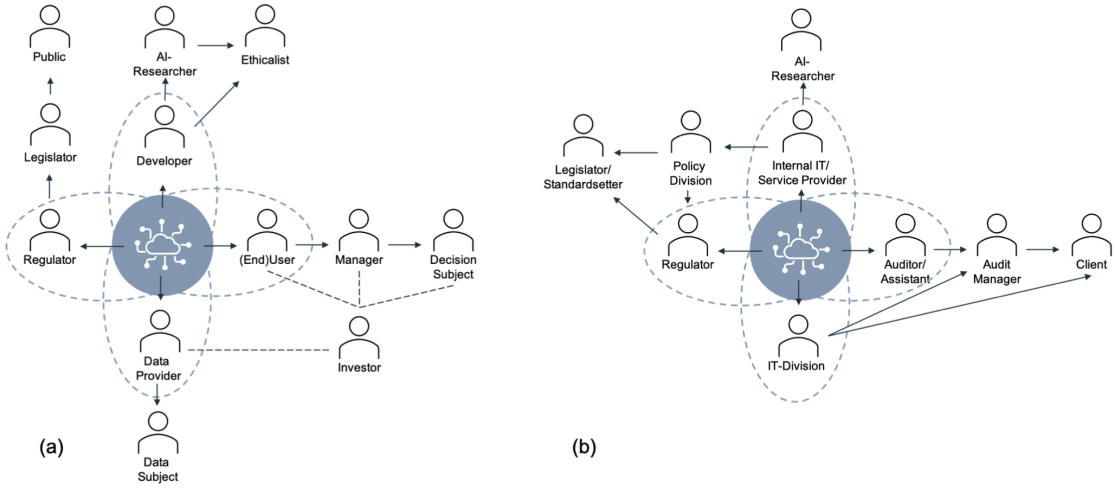

Abb. 5. Direkt oder indirekt mit KI-Systemen interagierende Rollen im Allgemeinen (a) und im Bereich der Wirtschaftsprüfung (b) (Rebstadt et al. 2022b)

Aufsetzend auf dem entwickelten Rollenmodell können anwendungsfallspezifisch alle relevanten Akteure identifiziert und personalisierte Erklärungen entwickelt werden. Hierdurch kann sowohl der Bedarf nach interpretierbaren Modellen für potenzielle Regulatoren als auch nach unter Zeitdruck intuitiv verständlichen Erklärungen spezifischer Entscheidungen erfüllt werden, welche die Grundlage für die Arbeit der Prüfenden mit dem System darstellt.

5.2.2 Prototypische Implementierung eines transparenten Ansatzes zur Fraud Detection in der Wirtschaftsprüfung

Den Erkenntnissen aus Beitrag 1 zufolge wurde in Beitrag 2, strukturiert durch das CRISP-DM-Vorgehensmodell, ein KI-System exemplarisch für den Use Case der Erkennung von Bilanzbetrug in der Wirtschaftsprüfung instanziiert. Im Fokus stand dabei erstens die Genauigkeit der Vorhersagen und zweitens die Gestaltung eines Systems, das sowohl für die Entwickelnden als auch für die Prüfenden eine transparente Entscheidungsfindung sicherstellt. Hierzu wurden die aktuell in der Literatur vertretenen KI-Ansätze systematisch erhoben und relevante Inputwerte identifiziert. Basierend auf den Rohdaten der United States Securities and Exchange Commission sowie aggregierten Finanzkennzahlen wurden drei in der Literatur als vielversprechend identifizierte KI-Ansätze prototypisch implementiert. Um die drei Ansätze, logistische Regression, Support-Vektor-Maschinen und eXtreme Gradient Boosting, post hoc einheitlich erklären zu können und eine Vergleichbarkeit sicherzustellen, wurde auf *SHapley Additive exPlanations* (SHAP) als modellagnostisches Verfahren zurückgegriffen (Lundberg, Lee 2017). Bedingt durch die existierenden Kombinationsmöglichkeiten der identifizierten KI-Ansätze und die zur Verfügung stehenden Daten wurden sechs verschiedene Modelle trainiert und evaluiert und mit einem Baseline-Modell mit zufälliger Zuordnung der vorhergesagten Variable (Fraud: Ja/Nein) verglichen. Die Ergebnisse der Evaluation sind in Abb. 6 gegenübergestellt.

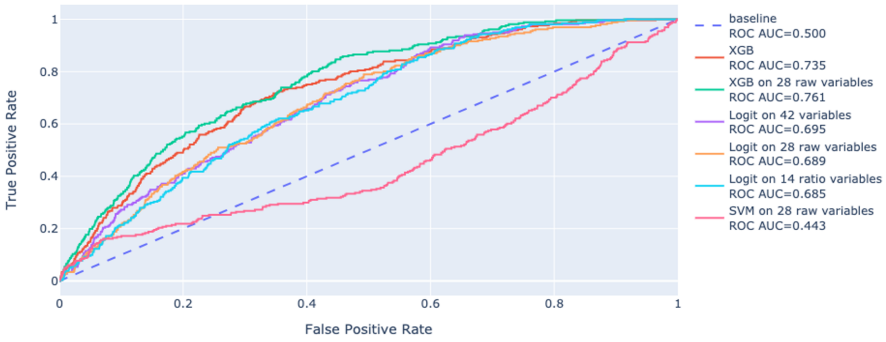

Abb. 6. Receiver-Operating-Characteristic-Kurven der betrachteten Machine-Learning-Modelle (Fukas et al. 2022)

Neben der Sicherstellung einer positiven Performanz wurden mithilfe der erhobenen SHAP-Werte Ausprägungen und Relevanz der genutzten Input-Variablen verglichen. Hierdurch kann bei der Erkennung von Bilanzbetrug zukünftig die Erhebung der für die Modelle notwendigen Datenbasis bei den jeweiligen Mandanten auf die relevantesten Attribute beschränkt werden, um im Zielbild ein optimales Kosten-Nutzen-Verhältnis zwischen den notwendigen Aufwänden und der erzielten Genauigkeit sicherzustellen. Die globalen Ergebnisse der auf SHAP-basierten Auswertung sind in Abb. 7 dargestellt.

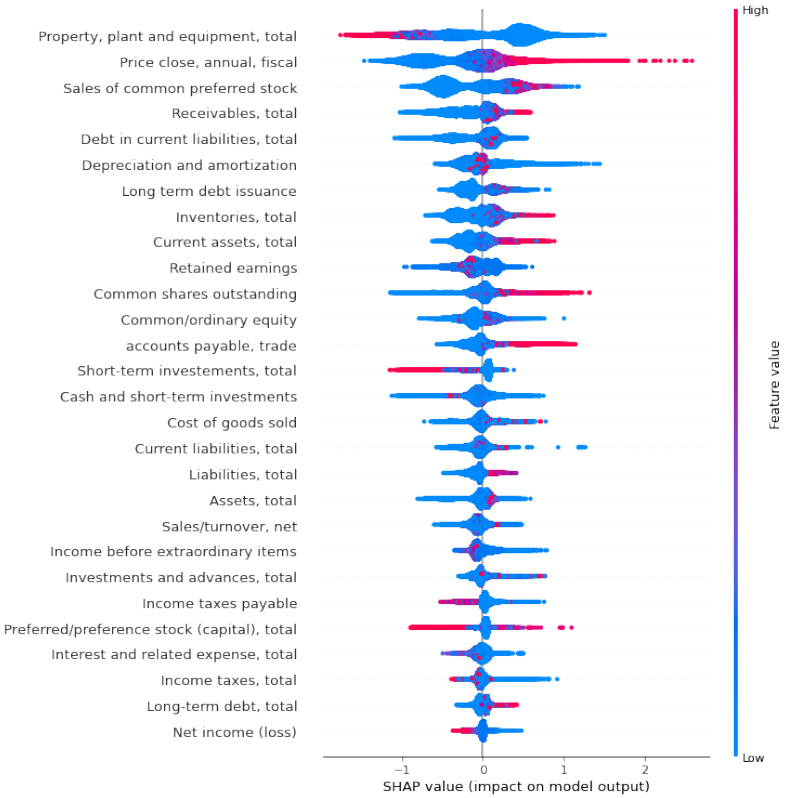

Abb. 7. Zusammenfassung der SHAP-basierten Relevanzen der 28 Basisvariablen im XGBoost-28-Modell (Fukas et al. 2022)

Somit konnten in Beitrag 2 sowohl Implikationen für die Weiterentwicklung bei der Erkennung von Bilanzbetrug als auch die Bereitstellung einer transparenten Entscheidungsfindung für potenzielle Nutzende abgeleitet werden. Systeme, wie der hier exemplarisch

instanziierte KI-Service, setzen dabei sowohl zum Training als auch in der eigentlichen Nutzung auf personenbezogenen oder auch unternehmenskritischen Daten auf. Um den dabei entstehenden Ansprüchen im Hinblick auf Datenschutz und Datensouveränität in komplexen Systemen gerecht zu werden, wurden in den Beiträgen B3 und B4 konkrete technische Lösungsansätze implementiert und evaluiert.

5.2.3 Implementation und Evaluation einer Service Registry zur transparenten Aufbereitung von Serviceinteraktionen und Datenflüssen in Datenökosystemen

Um die Entwicklung von neuen Services entsprechend branchenspezifischen Restriktionen und akteurspezifischen Anforderungen zu adressieren, wurde in Beitrag 3 die Orchestrierung von Services und Daten in komplexen Datenökosystemen betrachtet. Am Beispiel des Smart-Living-Ökosystems wurden dafür zwei zentrale Zielsetzungen verfolgt. Erstens wurde die Darstellung existierender Services im Ökosystem sowie die Schaffung von Interoperabilität zwischen Services mit vergleichbarem Wertversprechen behandelt. Darüber hinaus sollte die Offenlegung und Verfolgbarkeit von Datenflüssen zwischen Services sichergestellt werden, um eine Basis für die individuelle Datensouveränität sowie die Einhaltung von Datenschutzanforderungen zu legen. Hierfür wurde aufsetzend auf dem WOT-Standard eine Selbstbeschreibung konzipiert und um Anforderungen aus der GAIA-X-Selbstbeschreibung[6] und der Wohnungswirtschaft erweitert. Bei den dabei ergänzten Attributen stehen vor allem zentrale Aspekte zur Sicherstellung datenschutzbezogener Anforderungen im Vordergrund. Zur Pflege und zur intuitiven Bereitstellung der in der Selbstbeschreibung enthaltenen Informationen wurde in Beitrag 3 die prototypische Implementierung einer Service Registry ausgearbeitet und mithilfe von technischen und fachlichen Experten evaluiert. Der Webservice stellt hierbei sowohl die in Abb. 8 exemplarisch dargestellte Weboberfläche zur Bedienbarkeit durch menschliche Akteure als auch eine API zur direkten technischen Ansprechbarkeit bereit.

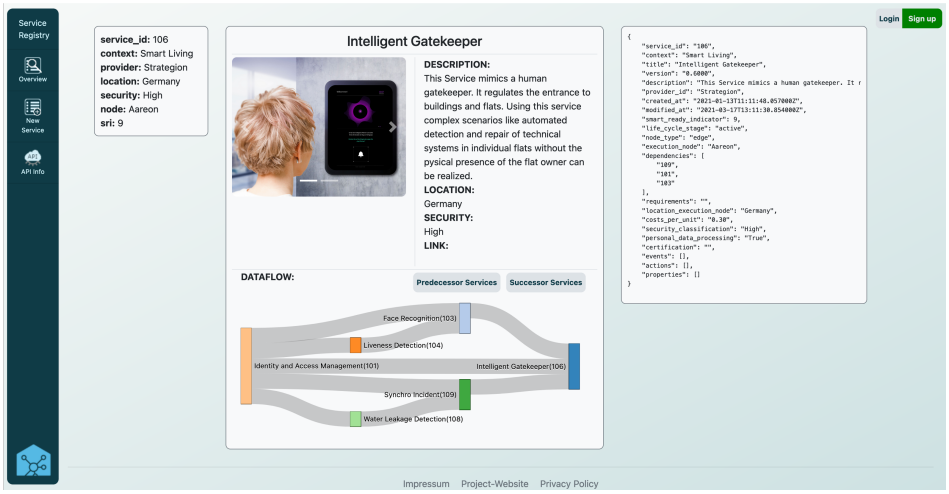

Abb. 8. Oberfläche der Service Registry: Service-Details einschließlich einer Datenflussdarstellung komplexer Service-Abhängigkeiten (Rebstadt et al. 2021)

[6] GAIA-X fokussiert die Entwicklung einer Infrastruktur, die auf die digitale Souveränität und einen souveränen Datenaustausch ausgerichtet ist. GAIA-X-Selbstbeschreibungen adressieren dabei eine maschinenlesbare und maschinenauswertbare Beschreibung verbunden mit einer aussagekräftigen Semantik (Otto et al. 2021).

Auf diese Weise konnte sowohl für Datenanbietende als auch für -konsumierende ein praktischer Mehrwert erzielt werden, indem Informationen über die Weiterleitung der eigenen Daten sowie über die Herkunft und die mögliche Qualität technisch aufgearbeitet und per Weboberfläche und API bereitgestellt werden. So können auf einer organisatorischen Ebene die Themen Datensouveränität und Datenschutz adressiert werden.

5.2.4 Transfer, Implementierung und Evaluation von Similarity Preserving Hashing als vielversprechender Ansatz für Privacy Aware Machine Learning

Der im vorherigen Beitrag dargestellte Lösungsansatz bietet zwar eine organisatorische Sicherstellung von Datensouveränität und Datenschutz, dieser kann jedoch die für KI-Algorithmen essenzielle Datenbasis, bedingt durch die individuellen Entscheidungen der Datenanbieter, kritisch einschränken. Um dem entgegenzuwirken, wurden in Beitrag 4 technische Lösungsansätze betrachtet, die das Potenzial bieten, dem Spannungsfeld zwischen der Notwendigkeit von Daten zum Training von Maschine-Learning-Ansätzen bei gleichzeitig hohen Datenschutzanforderungen bestimmter Branchen entgegenzuwirken. Hierzu wurden mithilfe einer systematischen Literaturrecherche wissenschaftlich etablierte Ansätze identifiziert, einhergehende Problemstellungen aufgearbeitet und Anforderungen zur Auflösung des Spannungsfeldes abgeleitet. Als zentrale Lösungsansätze wurden dabei, wie in Abb. 9 dargestellt, Differential Privacy (Dwork, Roth 2013), homomorphe Verschlüsselung (Rivest et al. 1978), k-Anonymity (Sweeney 2002) und die Generierung synthetischer Daten (Park et al. 2018) tiefergehend betrachtet. Aufgrund der verbleibenden Problemstellungen bei allen etablierten Verfahren wurde mit SPH ein weiterer vielversprechender Lösungsansatz identifiziert. SPH stellt hierbei eine Unterkategorie des Hashings dar, die klassischen Hashing-Ansätzen neue Merkmale hinzufügt, um in der Datenforensik zum Vergleich von Dokumenten eingesetzt zu werden (Gayoso Martínez et al. 2014). SPH reduziert dabei die Lesbarkeit durch menschliche Akteure, erhält jedoch für das Machine Learning relevante Ähnlichkeiten.

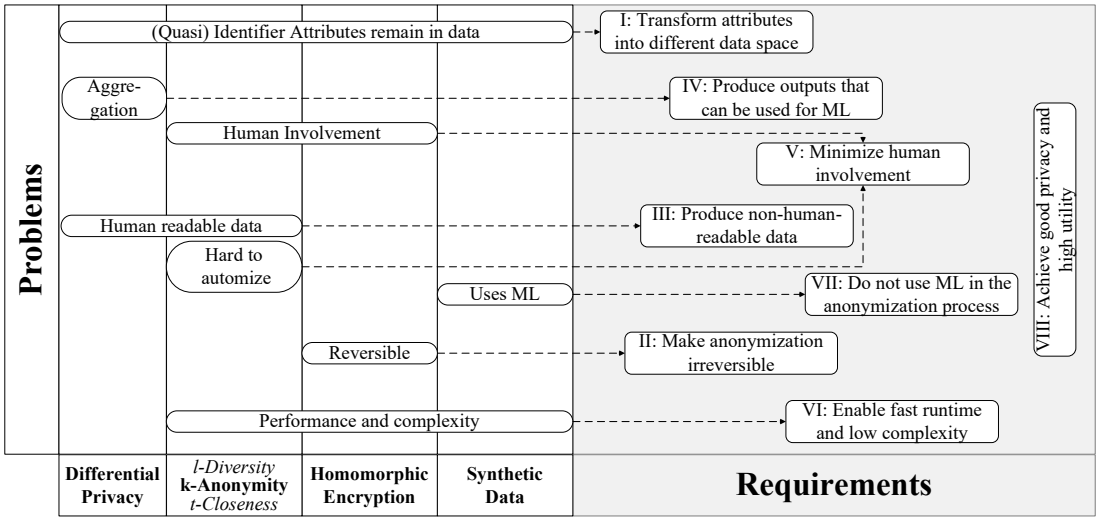

Abb. 9. Anforderungen an SPH-basierte Anonymisierung (Eleks et al. 2022)

Aufsetzend auf den festgestellten Anforderungen wurden mit Sdhash, Ssdeep und bbHash drei der identifizierten Ansätze prototypisch implementiert. Die Ansätze wurden anschließend im Hinblick auf den erzielten Nutzen und die Privatsphäre technisch evaluiert

und für zwei ML-Ansätze mit k-Anonymity, der Generierung synthetischer Daten sowie ML auf Basis der Rohdaten verglichen.

Zur Messung der Privatsphäre, approximiert durch den verbleibenden Informationsgehalt, wurde dabei auf Mutual Information und zur Evaluation des erzielten Nutzens auf den F1-Score als etablierte Kennzahlen zurückgegriffen. Die Zielsetzung ist dabei die Erzielung eines möglichst hohen F1-Scores bei gleichbleibend kleiner Mutual Information. Wie in Abb. 10 aufgezeigt, stellten sich dabei die drei neu etablierten SPH-Ansätze als erfolgversprechend heraus. Sdhash erreicht für neuronale Netze sogar vergleichbare F1-Werte, wie das auf den Rohdaten trainierte Baseline-Modell.

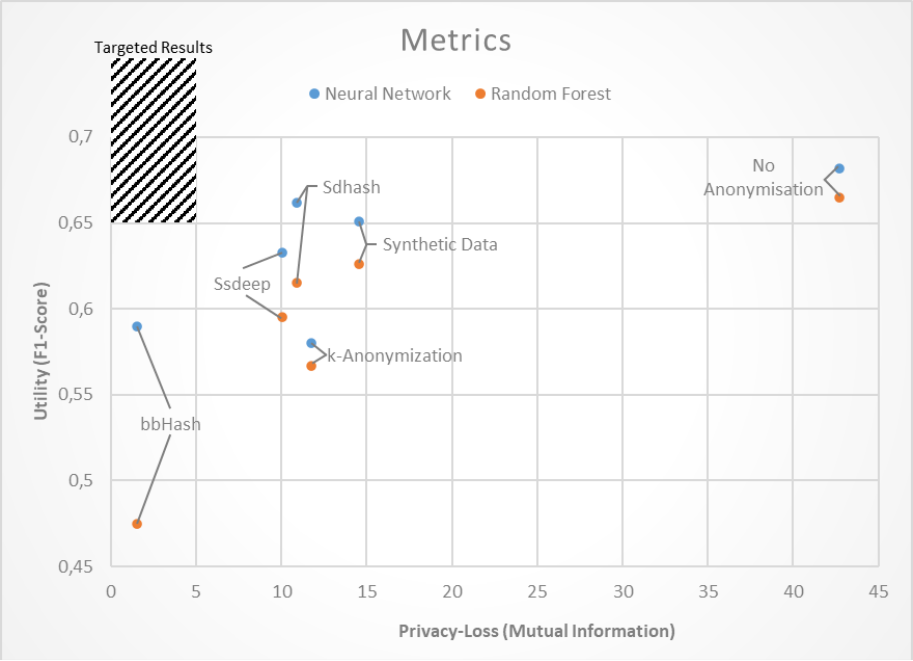

Abb. 10. Datenschutz und Nutzwertmetriken der evaluierten Algorithmen (Eleks et al. 2022)

In Beitrag 4 wurden somit, auf der Grundlage der Erhebung von Anforderungen an PAML-Algorithmen, Erkenntnisse aus der Datenforensik auf die Anonymisierung angewandt, um erkannte Lücken in der aktuellen Literatur zu schließen. Zudem werden konkrete Algorithmen sowohl aus Sicht des Nutzens als auch der Privatsphäre bewertet und verglichen, um eine schnelle praktische Anwendung zu ermöglichen.

5.2.5 Reduktion von Diskriminierung in prototypischen KI-Systemen und Ableitung von Handlungsempfehlungen zur Entwicklung von diskriminierungsfreien KI-Systemen

Im Kontext von Smart Living werden KI-Systeme von Menschen verschiedener Ethnien genutzt, wodurch die Diskriminierungsfreiheit einen zentralen Erfolgsfaktor sowohl aus gesellschaftlicher als auch aus wirtschaftlicher Sicht darstellt. Die Sicherstellung der Diskriminierungsfreiheit in entwickelten KI-Anwendungen und die übergreifende Berücksichtigung im Entwicklungsprozess ist der Gegenstand von Beitrag 5. Die Ausgangsbasis hierfür stellen zwei Kernkomponenten dar: Zum einen die literaturbasierte Erhebung existierender Anforderungen und Vorgehensweisen und zum anderen die Adaption einer prototypischen Implementierung inklusive der Ableitung fachlicher Anforderungen aus dem Bereich Smart Living. Im Zuge der prototypischen Implementierung wurde der intelligente

Gebäudepförtner als KI-basiertes, auf Gesichtserkennung aufsetzendes Zutrittskontrollsystem (Kortum et al. 2020) adaptiert, um (1) algorithmisch bei den eingesetzten ML-Ansätzen robuster gegenüber Diskriminierung zu werden als auch (2) möglichen Einschränkungen von Nutzenden durch das Systemdesign mit abzubilden. Die Ergebnisse aus der literaturbasierten Erhebung sowie die Erkenntnisse aus der prototypischen Implementierung wurden zusammengeführt und in die in Tab. 2 dargestellten Handlungsempfehlungen überführt.

Tab. 2. Aus praktischer und theoretischer Perspektive abgeleitete Handlungsempfehlungen (Rebstadt et al. 2022a)

Handlungs-empfehlung	Beschreibung	Zugrundeliegende literaturbasierte (L) und praktische (P) Maßnahmen
H1	Akquisition möglichst balancierter Datensätze in Bezug auf Subgruppen	L1
H2	Analyse des Problems auf mögliche Diskriminierungsrisiken	L2
H3	Identifikation potenziell diskriminierter Subgruppen	L3
H4	Identifikation potenziell diskriminierender Variablen und Proxy-Variablen	L4
H5	Definition einer quantifizierbaren Metrik für die Nicht-Diskriminierung	L5
H6	Diskriminierungsfreie Objektivierung der Zielvariable	L6
H7	Untersuchung der Datengrundlage auf Über- oder Unterrepräsentation von Subgruppen	L7
H8	Entfernen von potenziell diskriminierenden Variablen und Proxy-Variablen	L8
H9	Definition von Kriterien für nichtdiskriminierende Algorithmenauswahl	L9, P4
H10	Auswahl von Algorithmen entsprechend der in H9 definierten Kriterien	L10, P4
H11	Integration von nichtdiskriminierenden Kriterien in Optimierungsmetrik und Modellparameter	L11, P4, P5
H12	Ergänzung entwickelter KI-Modelle um direkte Anpassung des Outputs	L12, P4, P5
H13	Quantitative Evaluation auf Basis der entwickelten Nichtdiskriminierungs-Metrik (H5)	L13, P6
H14	Kontinuierliche Bewertung des Modells in Hinblick auf die Nichtdiskriminierungs-Metrik (H5)	L14
H15	Etablierung einer Feedbackschleife für potenzielle Diskriminierung bei der Anwendung von KI-Systemen	L15
H16	Etablierung eines Audit-Verfahrens für den gesamten Entwicklungs- und Anwendungsprozess von KI-Systemen	L16
H17	Bewusste Zusammenstellung diverser, inklusiver Teams	L17
H18	Sensibilisierung und Schulung der Teams bezüglich (Nicht)Diskriminierungsthematik	L18
H19	Schaffung eines grundsätzlichen Verständnisses des Gesamtsystems bei Nutzenden	P1
H20	Sicherstellung von Transparenz bei den durch das KI-System getroffenen Entscheidungen	P2
H21	Schaffung von Transparenz in Bezug auf die durch das KI-System verwendeten Datenquellen	P3
H22	Balancierte Ausgestaltung des Test-Datensatzes in Bezug auf die identifizierten Subgruppen	P7
H23	Integration der KI-Komponente in ein diskriminierungsfreies Interface	P8
H24	Integration der KI-Komponente in ein gegenüber fehlerhaftem Nutzungsverhalten robusten Interface	P9
H25	Ermöglichen von alternativen Lösungsansätzen im Falle von Fehlfunktion oder Nutzungsfehlverhalten	P10

Um die abgeleiteten Handlungsempfehlungen intuitiv in die Entwicklung von KI-Systemen integrieren zu können und eine direkte Verknüpfung der jeweiligen Phase mit den für sie relevanten Handlungsempfehlungen zu ermöglichen, wurden diese in das etablierte CRISP-DM-Vorgehensmodell eingebettet. Das resultierende Modell ist in Abb. 11 visualisiert.

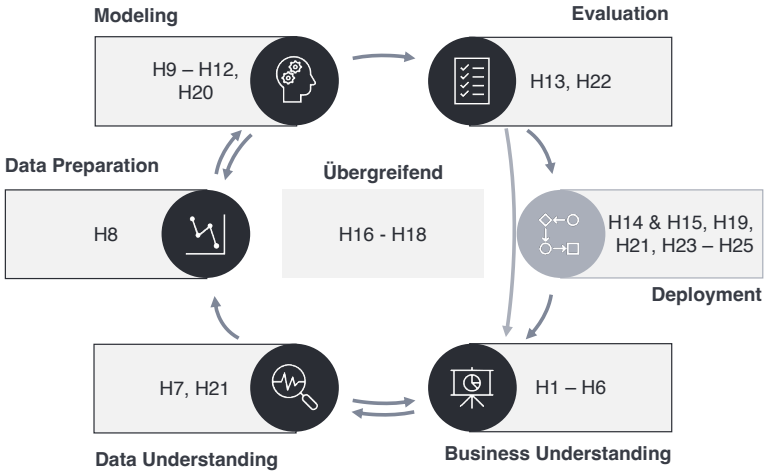

Abb. 11. Einordnung der Handlungsempfehlungen in den CRISP-DM-Zyklus (Rebstadt et al. 2022a)

Mithilfe von Beitrag 5 wurde somit eine prozessuale Grundlage für ein zentrales Prinzip vertrauenswürdiger KI gelegt. Um die vertrauenswürdige Entwicklung von KI-Systemen übergreifend in Unternehmen zu verankern, muss jedoch nicht nur eine Adaption des genutzten Vorgehensmodells erfolgen. Vielmehr sind für die langfristige Ausrichtung eines Unternehmens auf die Entwicklung und Nutzung von KI-Systemen tiefergehende Anpassungen im Unternehmen notwendig.

5.2.6 Konzeptionierung eines Reifegradmodells für KI in der Wirtschaftsprüfung unter besonderer Berücksichtigung von Ethik und Regulatorik

Beitrag 6 adressiert diese langfristige Ausrichtung des Unternehmens mithilfe von Reifegradmodellen. Durch die Entwicklung eines Reifegradmodells für KI in der Wirtschaftsprüfung als stark regulierte Domäne wird die Verankerung und die Messbarkeit der für die Etablierung von KI zentralen Dimensionen adressiert. In einem iterativen Verfahren wurden hierzu basierend auf dem Vorgehen von Becker et al. (2009) existierende Modelle in der Literatur verglichen und ein initiales Reifegradmodell abgeleitet. Mithilfe einer qualitativen Studie wurde dieses Modell auf die Wirtschaftsprüfung übertragen und evaluiert. Das hieraus resultierende Modell mit acht verschiedenen Dimensionen und fünf verschiedenen Reifegraden ist in Abb. 12 dargestellt.

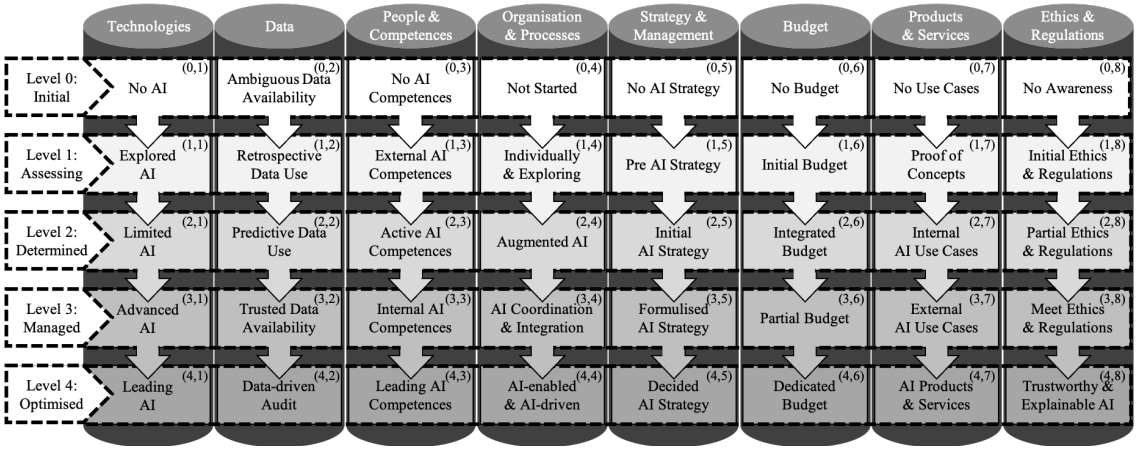

Abb. 12. Resultierendes Reifegradmodell für KI in der Wirtschaftsprüfung (Fukas et al. 2021)

Das Reifegradmodell liefert eine Basis für Wirtschaftsprüfungsunternehmen auf ihrem Weg zu einer KI-gestützten Organisation, um den Einsatz von KI in der Wirtschaftsprüfung strategisch voranzutreiben und dabei insbesondere die ethischen und regulatorischen Herausforderungen des Berufsstandes zu berücksichtigen.

5.3 Theoretische Implikationen

In hochregulierten Branchen wie der Wirtschaftsprüfung ist in den vergangenen Jahren ein Spannungsfeld zwischen den akzeptanzbezogenen und vor allem regulatorischen Anforderungen sowie den Potenzialen technischer Innovationen durch KI entstanden (Morley et al. 2020), welches durch Prinzipien vertrauenswürdiger KI reduziert werden kann. Um diese zu konkretisieren und zu operationalisieren, wurden im Zuge dieser Dissertation Ansätze zu drei zentralen Prinzipien (*Transparenz* – FF1, *Datenschutz und -souveränität* – FF2, *Nichtdiskriminierung* – FF3) und deren Einbettung in das KI-Management (FF4) untersucht und resultierend theoretische und praktische Implikationen abgeleitet, die in den folgenden beiden Kapiteln dargestellt werden.

Im Zuge von FF1 wurden sowohl die Notwendigkeit als auch die Ausprägung von Transparenz untersucht und somit die Verbindung von domänen- und KI-bezogenen Anforderungen adressiert. Hierbei haben sich in Bezug auf die Transparenz akteurspezifische Differenzen ergeben, die durch einen einheitlichen transparenzfördernden Ansatz nicht ohne Weiteres erfüllbar sind. Um die Anforderungen aller beteiligten Akteure wie Wirtschaftsprüfer, Entwickelnde, Mandanten oder auch Regulierungsbehörden zusammenzubringen, kann das in Beitrag 1 entwickelte Rollenmodell und das Konzept personalisierter Erklärungen eine theoretische Grundlage bieten. In Beitrag 2 wurde ein entsprechender Erklärungsansatz prototypisch am Beispiel eines Modells zur Identifikation von Bilanzbetrug instanziiert. Die Instanziierung trägt hierbei als erste Anwendung von SHAP als modellagnostisches Verfahren für die Erkennung von Bilanzbetrug zum einen zum Verständnis von Transparenz in der Wirtschaftsprüfung bei, zum anderen liefern die Erklärungen auch bedeutsame Erkenntnisse zur betriebswirtschaftlichen Relevanz und die direktionalen Auswirkungen der betrachteten Variablen auf die Vorhersage von Bilanzbetrug. Somit erweitert die durchgeführte Analyse existierende Studien wie die von Bao et al. (2020) um relevante Erkenntnisse.

Zur Sicherstellung von Datenschutz und -souveränität bieten Ansätze zu PAML eine Basis für das Training sowie für die operative Nutzung von KI- und ML-Anwendungen. Jedoch

weisen die aktuell in der Literatur dargestellten Algorithmen zentrale Schwächen auf. Im Zuge von FF2 wurde hierzu in Beitrag 4 mit SPH ein bisher nur aus dem Bereich der Datenforensik bekannter Ansatz auf den betrachteten Anwendungsfall übertragen, implementiert und evaluiert, um Erkenntnisse über das Potenzial von SPH für PAML abzuleiten und Lücken in der aktuellen Literatur zu schließen.

Neben der Adressierung der Prinzipien selbst wurde im Zuge von FF4 die Einbettung von vertrauenswürdiger KI in das KI-Management durch die Entwicklung eines domänenspezifischen KI-Reifegradmodells betrachtet. Das in Beitrag 6 entwickelte KI-Reifegradmodell berücksichtigt explizit die für die Wirtschaftsprüfung zentralen Punkte Ethik und Regulatorik und stellt somit das erste KI-Reifegradmodell für die Wirtschaftsprüfung dar. Die Relevanz der expliziten Berücksichtigung von Ethik und Regulatorik im KI-Management geht jedoch über die Wirtschaftsprüfung als Anwendungsdomäne hinaus. Wie durch die Datenethikkommission der Bundesregierung festgestellt wurde, müssen mit der zunehmenden Omnipräsenz von algorithmischen und vor allem KI-basierten Systemen Regeln für die Entwicklung und den Einsatz dieser etabliert werden (Datenethikkommission der Bundesregierung 2019). Überdies zeigt die Studie von Alsheibani et al. (2020), dass staatliche Regulierungen einen positiven Einfluss auf die Adoption von KI haben können. In Anbetracht der strengen regulatorischen Anforderungen in der Wirtschaftsprüfung kann das entwickelte KI-Reifegradmodell folglich eine Vorreiterrolle für die Einbeziehung von Ethik und Regulatorik in die Bewertung des KI-Reifegrads einer Organisation einnehmen.

5.4 Praktische Implikationen

Entsprechend des im deutschsprachigen Raum vorherrschenden gestaltungsorientierten Forschungsparadigmas (Wilde, Hess 2006) fokussiert diese Dissertation die Entwicklung von IT-Artefakten und die Generierung von präskriptivem Wissen, aus dem sich praktische Implikationen ableiten lassen (Gregor, Hevner 2013).

Im Zuge von FF1 ergeben sich praktische Implikationen sowohl durch das entwickelte Rollenmodell in Beitrag 1 als auch aus der prototypischen Implementierung in Beitrag 2. Beitrag 1 bietet mit dem Ansatz zu personalisierten Erklärungen und dem zugrundeliegenden Rollenmodell einen Startpunkt für die Entwicklung von KI-Systemen mit akteurspezifischen und sich gegebenenfalls sogar widersprechenden Anforderungen an die Transparenz der genutzten KI-Algorithmen. Dabei wird auf die Erkenntnisse von Oh et al. (2020) und Tintarev et al. (2016) aufgebaut und mithilfe des Rollenmodells eine Möglichkeit geschaffen, die formulierten Anforderungen an die Transparenz akteurspezifisch zu differenzieren und Widersprüche aufzulösen. Darüber hinaus können personalisierte Erklärungen eine Grundlage für die erhöhte Fairness und die Reduktion von Diskriminierung in den eingesetzten KI-Systemen bieten und gezielte Rückmeldung von begründeten Einwänden durch Nutzende ermöglichen.

Darauf aufsetzend wurde in Beitrag 2 ein transparentes KI-System zur Erkennung von Bilanzbetrug im Zuge der Jahresabschlussprüfung instanziiert. Hierdurch haben sich sowohl fachliche Mehrwerte bei der Auswahl der relevanten Variablen zur Erkennung von Finanzbetrug als auch technische Implikationen zur Ausgestaltung von transparenten KI-Systemen ergeben. Neben diesen technischen und fachlichen Erkenntnissen hat sich die Transparenz als wesentlicher Baustein zur Überschreitung der Line of Governance[7]

[7] Die Line of Governance (LoG) stellt die für eine operative Nutzung zu überschreitende Schwelle im Entwicklungsprozess von KI-Systemen dar (Thomas et al. 2021).

herausgestellt, was eine explizite Ausrichtung der in den Unternehmen eingesetzten Vorgehensweisen zur Entwicklung von KI-Systemen ermöglicht.

Sowohl im Training als auch beim Einsatz solcher Systeme wird jedoch in vielen Fällen auf personenbezogene oder auch unternehmenskritische Daten aufgebaut. Um den dabei entstehenden Ansprüchen im Hinblick auf Datenschutz und -souveränität gerecht zu werden, können sowohl organisatorische als auch technische Maßnahmen eingesetzt werden. Durch die Entwicklung einer Service Registry wurde in Beitrag 3 die technische Unterstützung der organisationalen Sicherstellung von Datensouveränität und Datenschutz adressiert, indem (1) Informationen über die Weiterleitung der eigenen Daten, aber auch (2) über die Herkunft der Daten und die mögliche Datenqualität technisch aufgearbeitet und per Weboberfläche und API bereitgestellt werden. Auf diese Weise konnte sowohl für Datenanbietende als auch für Datenkonsumierende ein praktischer Mehrwert erzielt werden und mit der Bereitschaft zum Teilen von Daten ein zentrales Hemmnis in Datenökosystemen mit einem konkreten technischen Ansatz adressiert werden. In Beitrag 4 wurden darüber hinaus mit SPH vielversprechende Algorithmen für PAML aus dem Bereich der Datenforensik identifiziert, prototypisch implementiert und evaluiert. Hierdurch werden neue Möglichkeiten für die praktische Anwendung von ML auf sensible Datenquellen aufgezeigt und die Grundlage für eine schnelle praktische Anwendbarkeit geschaffen.

Um im Zuge der Entwicklung von KI-Systemen mit der Nichtdiskriminierung auch das dritte in dieser Dissertation betrachtete Prinzip zu berücksichtigen, wurden in Beitrag 5 sowohl Wissen zur Ausgestaltung von praktisch eingesetzten KI-Systemen gesammelt als auch Handlungsempfehlungen für die Entwicklung diskriminierungsfreier KI-Systeme abgeleitet. Damit eine intuitive Nutzbarkeit für Praktiker sichergestellt werden kann, wurden diese darüber hinaus mit CRISP-DM in ein etabliertes Vorgehensmodell zur Entwicklung von KI-Systemen eingebettet. Neben den explizit betrachteten Prinzipien haben sich auch bei der Einbettung in das IT-Management praktische Implikationen ergeben. Das in Beitrag 6 entwickelte KI-Reifegradmodell kann hierbei als Ausgangspunkt für die Verbesserung des strategischen Managements von Wirtschaftsprüfungsgesellschaften im Hinblick auf KI dienen und Wirtschaftsprüfer dabei unterstützen, KI-Ansätze langfristig und zielgerichtet in ihre Tätigkeiten zu integrieren.

Insgesamt haben sich auch über die eingebrachten Beiträge hinaus Implikationen ergeben, die sich in den in Kapitel 5.1 dargestellten Beiträgen widerspiegeln. Neben der Dissemination über wissenschaftliche Publikationen stellt die explizite Weiternutzung der erzielten Ergebnisse in einer technologieorientierten Spin-off-Unternehmung entsprechend dem Modell zum Technologietransferprozess von Scheer (1993) eine weitere Besonderheit dieser Dissertation dar. Im Zuge dessen erfolgte sowohl direkt mit Unternehmen als auch mit Berufsverbänden eine praxisorientierte Dissemination von Erkenntnissen, die auf den Ergebnissen dieser Dissertation aufsetzen.

5.5 Limitationen

Alle in dieser Dissertation eingebrachten Forschungsergebnisse (vgl. Kapitel 5) wurden mithilfe von anerkannten Forschungsmethoden und -ansätzen (vgl. Kapitel 4.2) hergeleitet. Die dabei entstandenen Beiträge haben eine doppelblinde Begutachtung durchlaufen und wurden in ausgewiesenen Publikationsorganen der Wirtschaftsinformatik nach dem Ranking des Verbandes der Hochschullehrer (VHB, JOURQUAL-Ranking) und der im VHB organisierten Wissenschaftlichen Kommission für Wirtschaftsinformatik (WKWI) publiziert. Nach WKWI werden dabei alle eingebrachten Beiträge mindestens in der Kategorie ‚B‘

eingeordnet. Hervorzuheben sind darüber hinaus die Beiträge B1 und B6, die mit der WI und der ECIS auf Konferenzen der Kategorie ‚A' nach WKWI-Ranking publiziert wurden.

Trotz dieser Ausgangsvoraussetzungen unterliegt diese Dissertation methodischen, aber auch technischen Limitationen. Aufgrund der im Zuge der einzelnen Forschungsbeiträge verschiedenartigen Zielsetzungen zur Evaluation haben sich unterschiedlich ausgeprägte Evaluationsansätze ergeben. In den Beiträgen B2 und B4 wurde hierzu eine technische, rein metrik-basierte Evaluation durchgeführt. Somit lässt sich zwar bei Beitrag B4 durch den ausbleibenden Bezug auf eine spezifische Anwendungsdomäne eine gewisse Generalisierbarkeit erreichen, jedoch können hierdurch wesentliche Erkenntnisse bei der Interaktion von potenziellen Nutzenden mit den Systemen nicht miterfasst werden. Sowohl bei Experteninterviews als auch in den Fokusgruppengesprächen stellt zudem sowohl die Anzahl als auch die Auswahl der eingesetzten Experten eine Limitation der Beiträge B1, B3, B5 und B6 dar. Es wurde versucht, diesen Nachteil durch eine ausgewogene und gezielte Auswahl von Befragten auszugleichen, die die jeweilige Domäne aus verschiedenen Blickwinkeln betrachten können und über umfassende Berufserfahrung verfügen. Darüber hinaus beschränken sich sowohl die Instanziierungen als auch die zumeist formativ ausgerichteten Evaluationen (Venable et al. 2016) in allen Beiträgen außer B4 ausschließlich auf eine ausgewählte Anwendungsdomäne, wodurch die abgeleiteten Erkenntnisse nur eingeschränkt generalisierbar sind. Neben diesen Limitationen des vorwiegend qualitativen Forschungsdesigns können auch Teile der Explikation fehlen, da IT-Artefakte selbst nicht vollständig publiziert wurden. Auch wenn bei implementierungsintensiven Forschungsarbeiten in der Regel eine Publikation des Source Codes über Plattformen wie GitHub erfolgen sollte (Wattanakriengkrai et al. 2022), konnte dies aufgrund von Unternehmensgeheimnissen bei Praxispartnern oder Kunden nicht flächendeckend erfolgen.

6 Zusammenfassung

Die Ergebnisse dieser Dissertation adressieren das Spannungsfeld zwischen regulatorischen Anforderungen und der Förderung von Innovationen (Morley et al. 2020), das die Grundlage für das Heben betriebswirtschaftlicher Potenziale durch den Einsatz von KI in hochregulierten Branchen darstellt. Hierfür wurde im Zuge der eingebrachten Forschungsbeiträge die folgende zentrale Forschungsfrage adressiert: „Wie können KI-Systeme vertrauenswürdig ausgestaltet werden, um sowohl regulatorischen Anforderungen zu genügen als auch die Nutzerakzeptanz für einen erfolgreichen Einsatz sicherzustellen?"

Die Forschungsfrage wurde am Beispiel von drei zentralen Prinzipien vertrauenswürdiger KI und ihrer Einbettung in das Management von KI-Systemen beantwortet. Hierzu wurden bisher meist abstrakte Prinzipien (Miller, Coldicott 2019) konkretisiert, mithilfe von prototypischen Implementierungen instanziiert und Handlungsempfehlungen abgeleitet. Neben der Erarbeitung zentraler Erkenntnisse für den Einsatz von KI in der Wirtschaftsprüfung wurde auch die Grundlage für eine branchenübergreifende Betrachtung gelegt. Aufgrund seines hochgradig regulierten Charakters können Erkenntnisse aus Domänen wie der Wirtschaftsprüfung als Beispiel für zahlreiche andere Bereiche dienen, da, wie bereits angestoßen durch die Europäische Union (Europäische Kommission 2020) oder den VDE (VDE 2022), Regulatorik beim Einsatz von KI im Allgemeinen eine zunehmend bedeutende Rolle spielen wird. Um dabei Innovationen nicht zu hemmen, sondern einen positiven Einfluss auf die Adoption von KI durch staatliche Regulierungen erzielen zu können, bietet diese Dissertation erfolgversprechende technische sowie methodische Ansatzpunkte.

7 Literatur

Alsheibani, S.; Messom, C.; Cheung, Y. (2020): *Re-thinking the Competitive Landscape of Artificial Intelligence*. Proceedings of the 53rd Hawaii International Conference on System Sciences (3):5861–5870.

Bao, Y.; Ke, B.; Li, B.; Yu, Y.J.; Zhang, J. (2020): *Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach*. Journal of Accounting Research 1(58):199–235.

Barocas, S.; Hardt, M.; Narayanan, A. (2017): *Fairness in machine learning*. Nips tutorial (1):2.

Becker, J.; Holten, R.; Knackstedt, R.; Niehaves, B. (2004): *Epistemologische Positionierungen in der Wirtschaftsinformatik am Beispiel einer konsensorientierten Informationsmodellierung*. In: Frank, U. (Hrsg.): Wissenschaftstheorie in Ökonomie und Wirtschaftsinformatik: Theoriebildung und -bewertung, Ontologien, Wissensmanagement. Wiesbaden, Deutscher Universitätsverlag, 335–366.

Becker, J.; Knackstedt, R.; Pöppelbuß, J. (2009): *Developing Maturity Models for IT Management*. Business & Information Systems Engineering 3(1):213–222.

vom Brocke, J.; Niehaves, B.; Simons, A.; Riemer, K. (2009): *Reconstructing the Giant : On the Importance of Rigour in Documenting the Literature Search Process*. Proceedings of the European Conference on Information Systems 2009. Verona, 161.

Clinciu, M.A.; Hastie, H.F. (2019): *A Survey of Explainable AI Terminology*. Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019):8–13.

Cooper, H.M. (1988): *Organizing knowledge syntheses: A taxonomy of literature reviews*. Knowledge in Society 1(1):104–126.

Datenethikkommission der Bundesregierung (2019): *Gutachten der Datenethikkommission*. https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.html.

Dhurandhar, A.; Iyengar, V.; Luss, R.; Shanmugam, K. (2017): *TIP: Typifying the Interpretability of Procedures*. https://arxiv.org/abs/1706.02952, Abruf am 25.10.2022.

Downar, B.; Fischer, D. (2019): *Wirtschaftsprüfung im Zeitalter der Digitalisierung*. Handbuch Industrie 4.0 und Digitale Transformation. Wiesbaden, Springer Fachmedien, 753–779.

Dwork, C.; Roth, A. (2013): *The algorithmic foundations of differential privacy*. Foundations and Trends in Theoretical Computer Science 3–4(9):211–487.

Eberhard, K. (1999): *Einführung in die Erkenntnis- und Wissenschaftstheorie: Geschichte und Praxis der konkurrierenden Erkenntniswege*. 2., durchg. Auflage. Stuttgart, Kohlhammer.

Eleks, M.; Rebstadt, J.; Fukas, P.; Thomas, O. (2022): *Learning without Looking : Similarity Preserving Hashing and Its Potential for Machine Learning in Privacy Critical Domains Privacy Aware Machine Learning with Similarity Preserving*. In: D. Demmler, D. Krupka, H.F. (Hrsg.): INFORMATIK 2022. Bonn, Gesellschaft für Informatik, 161–177.

Europäische Kommission (2020): *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342, Abruf am 25.10.2022.

Fernández-Loría, C.; Provost, F.; Han, X. (2020): *Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach*. arXiv preprint arXiv:2001.07417 (2020):1–33.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; Schafer, B.; Valcke, P.; Vayena, E. (2018): *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*. Minds and Machines 4(28):689–707.

Fukas, P.; Rebstadt, J.; Menzel, L.; Thomas, O. (2022): *Towards Explainable Artificial Intelligence in Financial Fraud Detection: Using Shapley Additive Explanations to Explore Feature Importance*. In: Franch, X; Poels, G; Gailly, F; Snoeck, M (Hrsg.), International Conference on Advanced Information Systems Engineering :109–126.

Fukas, P.; Rebstadt, J.; Remark, F.; Thomas, O. (2021): *Developing an Artificial Intelligence Maturity Model for Auditing*. ECIS 2021, A Virtual AIS Conference.

Gayoso Martínez, V.; Hernández Álvarez, F.; Hernández Encinas, L. (2014): *State of the Art in Similarity Preserving Hashing Functions*. The 2014 International Conference on Security and Management (SAM'14):139–145.

Gierbl, A.S.; Schreyer, M.; Leibfried, P.; Borth, D. (2020): *Künstliche Intelligenz in der Prüfungspraxis - Eine Bestandsaufnahme aktueller Einsatzmöglichkeiten und Herausforderungen*. Expert Focus 09(2020):612–617.

Gläser, J.; Laudel, G. (2009): *Experteninterviews und qualitative Inhaltsanalyse: als Instrumente rekonstruierter Untersuchungen*. 3. Auflage. VS Verlag.

Gregor, S.; Hevner, A.R. (2013): *Positioning and Presenting Design Science Research for Maximum Impact*. MIS Quarterly 37(2):337–355.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. (2018): *A survey of methods for explaining black box models*. ACM Computing Surveys 5(51):1–45.

Hamamoto, R.; Suvarna, K.; Yamada, M.; Kobayashi, K.; Shinkai, N.; Miyake, M.; Takahashi, M.; Jinnai, S.; Shimoyama, R.; Sakai, A.; Takasawa, K.; Bolatkan, A.; Shozu, K.; Dozen, A.; Machino, H.; Takahashi, S.; Asada, K.; Komatsu, M.; Sese, J.; Kaneko, S. (2020): *Application of artificial intelligence technology in oncology: Towards the establishment of precision medicine*. Cancers 12(12):1–32.

Hevner, A.R.; March, S.T.; Park, J.; Ram, S. (2004): *Design science in information systems research*. MIS Quarterly: 1(28):76–105.

Hochrangige Expertengruppe für künstliche Intelligenz, E.K. (2018): *ETHIK-LEITLINIEN FÜR EINE VERTRAUENSWÜRDIGE KI*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, Abruf am 25.10.2022.

Issa, H.; Sun, T.; Vasarhelyi, M.A. (2016): *Research Ideas for Artificial Intelligence in Auditing: The Formalization of Audit and Workforce Supplementation*. Journal of Emerging Technologies in Accounting 2(13):1–20.

Jobin, A.; Ienca, M.; Vayena, E. (2019): *The global landscape of AI ethics guidelines*. Nature Machine Intelligence 9(1):389–399.

Kokina, J.; Davenport, T.H. (2017): *The Emergence of Artificial Intelligence: How Automation is Changing Auditing*. Journal of Emerging Technologies in Accounting 1(14):115–122.

Kortum, H.; Gravemeier, L.S.; Zarvic, N.; Feld, T.; Thomas, O. (2020): *Engineering of Data-Driven Service Systems for Smart Living: Application and Challenges*. IFIP Advances in Information and Communication Technology. Springer, 592: , 291–298.

Larsson, S. (2020): *On the Governance of Artificial Intelligence through Ethics Guidelines*. Asian Journal of Law and Society 3(7):437–451.

Lipton, Z.C. (2018): *The mythos of model interpretability*. Communications of the ACM 10(61):35–43.

Lundberg, S.M.; Lee, S.I. (2017): *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems Section 2(2017-Decem):4766–4775.

Marx, V. (2022): *Method of the Year: protein structure prediction*. Nature Methods 1(19):5–10.

Mayer, A.S.; Strich, F.; Fiedler, M. (2020): *Unintended Consequences of Introducing AI Systems for Decision Making*. MIS Quarterly Executive 4(19):239–257.

McCorduck, P.; Cfe, C. (2004): *Machines Who Think*. Natick, A K Peters/CRC Press.

Miller, C.; Coldicott, R. (2019): *People, power and technology: The tech workers' view,* Abruf am 25.10.2022.

Mitchell, T. (1997): *Machine Learning*. Portland, McGraw-Hill.

Moore, J.D.; Swartout, W.R. (1988): *Explanation in Expert Systems: A Survey*. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Morgan, D.L. (1996): *Focus groups as qualitative research.* 16.

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 4(26):2141–2168.

Munoko, I.; Brown-Liburd, H.L.; Vasarhelyi, M. (2020): *The Ethical Implications of Using Artificial Intelligence in Auditing*. Journal of Business Ethics 167:209-234 .

Myers, M.D.; Newman, M. (2007): *The qualitative interview in IS research: Examining the craft*. Information and Organization 1(17):2–26.

Oh, C.; Kim, S.; Choi, J.; Eun, J.; Kim, S.; Kim, J.; Lee, J.; Suh, B. (2020): *Understanding how people reason about aesthetic evaluations of artificial intelligence*. Proceedings of the 2020 ACM Designing Interactive Systems Conference:1169–1181.

Oliveira, M.I.S.; Lóscio, B.F. (2018): *What is a data ecosystem?* ACM International Conference Proceeding Series. New York, New York, USA, Association for Computing Machinery, 1–9.

Österle, H.; Becker, J.; Frank, U.; Hess, T.; Karagiannis, D.; Krcmar, H.; Loos, P.; Mertens, P.; Oberweis, A.; Sinz, E.J. (2010): *Memorandum zur gestaltungsorientierten Wirtschaftsinformatik*. Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung 6(62):664–672.

Otto, B.; Eitel, A.; Schleimer, A.M.; Lange, C. (2021): *GAIA-X and IDS*. https://s.fhg.de/naw, Abruf am 25.10.2022.

Park, N.; Mohammadi, M.; Gorde, K.; Jajodia, S.; Park, H.; Kim, Y. (2018): *Data synthesis based on generative adversarial networks*. Proceedings of the VLDB Endowment 10(11):1071–1083.

Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. (2007): *A design science research methodology for information systems research*. Journal of Management Information Systems 3(24):45–77.

Perrault, R.; Shoham, Y.; Brynjolfsson, E.; Clark, J.; Etchemendy, J.; Grosz Harvard, B.; Lyons, T.; Manyika, J.; Carlos Niebles, J.; Mishra, S. (2019): *'The AI Index 2019 Annual Report'*. Stanford, Stanford University.

Raggett, D. (2015): *The web of things: Challenges and opportunities*. Computer 5(48):26–32.

Rebstadt, J.; Kortum, H.; Gravemeier, L.S.; Eberhardt, B.; Thomas, O. (2022a): *Non-Discrimination-by-Design: Handlungsempfehlungen für die Entwicklung von vertrauenswürdigen KI-Services*. HMD Praxis der Wirtschaftsinformatik 2(59):495–511.

Rebstadt, J.; Kortum, H.; Hagen, S.; Thomas, O. (2021): *Towards a transparency-oriented and integrating Service Registry for the Smart Living Ecosystem*. Lecture Notes in In: Gesellschaft für Informatik e.V. (GI) (Hrsg.), INFORMATIK 2021. Gesellschaft für Informatik, P-314:1425–1438.

Rebstadt, J.; Remark, F.; Fukas, P.; Meier, P.; Thomas, O. (2022b): *Towards Personalized Explanations for AI Systems: Designing a Role Model for Explainable Al in Auditing*. 17th International Conference on Wirtschaftsinformatik 2022 Proceedings February 2022:1–18.

Rivest, R.; Shamir, A.; Adleman, L. (1978): *On Data Banks and Privacy Homomorphisms*. Foundations of secure computation 4(11):169–180.

Rowley, J.; Slack, F. (2004): *Conducting a literature review*. Management Research News 6(27):31–39.

Rudin, C. (2014): *Algorithms for interpretable machine learning*. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining :1519–1519.

Rudin, C. (2019): *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature Machine Intelligence 5(1):206–215.

Russell, S.J.; Norvig, P. (2010): *Artificial Intelligence: A Modern Approach*. 3. Auflage. Upper Saddle River, Pearson Education.

Scheer, A.-W. (1993): *Wirtschaftsinformatik im Unternehmen 2000*. Wirtschaftsinformatik ’93. Heidelberg, Physica-Verlag HD:53–67.

Senden, L.; Xenidis, R. (2020): *EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination*. In: Ulf Bernitz et. al (Hrdg.) General Principles of EU law and the EU Digital Order (Kluwer Law International, 2020):151-182.

Shearer, C.; Watson, H.J.; Grecich, D.G.; Moss, L.; Adelman, S.; Hammer, K.; Herdlein, S. a (2000): *The CRISP-DM model: The New Blueprint for Data Mining*. Journal of Data Warehousing.

Sutton, S.G.; Arnold, V. (2013): *Focus group methods: Using interactive and nominal groups to explore emerging technology-driven phenomena in accounting and information systems*. International Journal of Accounting Information Systems 2(14):81–88.

Sweeney, L. (2002): *A model for protecting privacy*. Ieee S&P ’02 5(10):1–14.

Teodorescu, M.H.M.; Morse, L.; Awwad, Y.; Kane, G.C. (2021): *Failures of fairness in automation require a deeper understanding of human–ml augmentation*. MIS Quarterly: Management Information Systems 3(45):1483–1499.

The House of Lords (2018): *Select Committee on Artificial Intelligence AI in the UK : ready , willing and able?* https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf, Abruf am 25.10.2022.

Thomas, O. (2006): *Management von Referenzmodellen — Entwurf und Realisierung eines Informationssystems zur Entwicklung und Anwendung von Referenzmodellen*. Berlin, Logos-Verlag.

Thomas, O.; Bruckner, A.; Leimkühler, M.; Remark, F.; Thomas, K. (2021): *Konzeption, Implementierung und Einführung von KI-Systemen in der Wirtschaftsprüfung*. Die Wirtschaftsprüfung (WPg) 74(09):551–562.

Tintarev, N.; O’Donovan, J.; Felfernig, A. (2016): *Introduction to the special issue on human interaction with artificial advice givers*. ACM Transactions on Interactive Intelligent Systems 4(6):1–12.

Tomsett, R.; Braines, D.; Harborne, D.; Preece, A.; Chakraborty, S. (2018): *Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems*. arXiv preprint arXiv:1806.07552.

VDE, V. der E.E.I. e. V. (2022): *VCIO based description of systems for AI trustworthiness characterisation VDE SPEC 90012 V1.0 (en)*. Offenbach am Main.

Venable, J.; Pries-Heje, J.; Baskerville, R. (2016): *FEDS: A Framework for Evaluation in Design Science Research*. European Journal of Information Systems 1(25):77–89.

Wattanakriengkrai, S.; Chinthanet, B.; Hata, H.; Kula, R.G.; Treude, C.; Guo, J.; Matsumoto, K. (2022): *GitHub repositories with links to academic papers: Public access, traceability, and evolution*. Journal of Systems and Software 183:1–28.

Wilde, T.; Hess, T. (2006): *Methodenspektrum der Wirtschaftsinformatik: Überblick und Portfoliobildung*. Arbeitsbericht 2006/2: Institut für Wirtschaftsinformatik und Neue Medien der Ludwig-Maximilians-Universität München 2:1–14.

Wilde, T.; Hess, T. (2007): *Forschungsmethoden der Wirtschaftsinformatik Eine empirische Untersuchung*. Wirtschaftsinformatik 4(49):280–287.

Wirth, R. (2000): *CRISP-DM : Towards a Standard Process Model for Data Mining*. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining:29–39.

Wolf, M.J.; Miller, K.W.; Grodzinsky, F.S. (2017): *Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications*. The ORBIT Journal 2(1):1–12.

Teil B – Einzelbeiträge

Beitrag 1: Towards Personalized Explanations for AI Systems: Designing a Role Model for Explainable AI in Auditing

Titel	Towards Personalized Explanations for AI Systems: Designing a Role Model for Explainable AI in Auditing
Autoren	**Jonas Rebstadt**, Florian Remark, Philipp Fukas, Pascal Meier, Oliver Thomas
Publikationsorgan	Wirtschaftsinformatik 2022
Ranking	WKWI: A / VHB JQ3: C
Status	Veröffentlicht
Bibliographische Information	Rebstadt, J., Remark, F., Fukas, P., Meier, P., & Thomas, O. (2022). Towards personalized explanations for AI systems: designing a role model for explainable AI in auditing. In: Wirtschaftsinformatik 2022 Proceedings. 2.
Zusammenfassung	Due to a continuously growing repertoire of available methods and applications, Artificial Intelligence (AI) is becoming an innovation driver for most industries. In the auditing domain, initial approaches of AI have already been discussed in scientific discourse, but practical application is still lagging behind. Caused by a highly regulated environment, the explainability of AI is of particular relevance. Using semi-structured expert interviews, we identified stakeholder specific requirements regarding explainable AI (XAI) in auditing. To address the needs of all involved stakeholders a theoretical role model for AI systems has been designed based on a systematic literature review. The role model has been instantiated and evaluated in the domain of financial statement auditing using focus groups of domain experts. The resulting model offers a foun- dation for the development of AI systems with personalized explanations and an optimized usage of existing XAI methods.
Identifikation	AIS-eLibary: https://aisel.aisnet.org/wi2022/ai/ai/2
Link	https://aisel.aisnet.org/wi2022/ai/ai/2
Copyright	Copyright is retained by the authors.

Tab. 3. Factsheet Beitrag 1

Beitrag 2: Towards Explainable Artificial Intelligence in Financial Fraud Detection: Using Shapley Additive Explanations to Explore Feature Importance

Titel	Towards Explainable Artificial Intelligence in Financial Fraud Detection: Using Shapley Additive Explanations to Explore Feature Importance
Autoren	Philipp Fukas, **Jonas Rebstadt**, Lukas Menzel, Oliver Thomas
Publikationsorgan	CAiSE 2022: Advanced Information Systems Engineering
Ranking	WKWI: B / VHB JQ3: C
Status	Veröffentlicht
Bibliographische Information	Fukas, P.; Rebstadt, J.; Menzel, L.; Thomas, O. (2022): Towards Explainable Artificial Intelligence in Financial Fraud Detection: Using Shapley Additive Explanations to Explore Feature Importance. In: Franch, X; Poels, G; Gailly, F; Snoeck, M (Hrsg.), International Conference on Advanced Information Systems Engineering. Springer, Cham, S. 109-126.
Zusammenfassung	As the number of organizations and their complexity have increased, a tremendous amount of manual effort has to be invested to detect financial fraud. Therefore, powerful machine learning methods have become a critical factor to reduce the workload of financial auditors. However, as most machine learning models have become increasingly complex over the years, a significant need for transparency of artificial intelligence systems in the accounting domain has emerged. In this paper, we propose a novel approach using Shapley additive explanations to improve the transparency of models in the field of financial fraud detection. Our information systems engineering procedure follows the cross industry standard process for data mining including a systematic literature review of machine learning methods in fraud detection, a systematic development process and an explainable artificial intelligence analysis. By training a downstream Logistic Regression, Support Vector Machine and eXtreme Gradient Boosting classifier on a dataset of publicly traded companies convicted of financial statement fraud by the United States Securities and Exchange Commission, we show how the key items for financial statement fraud detection and their directionality can be identified using Shapley additive explanations. Finally, we contribute to the current state of research with this work by increasing model transparency and by generating insights on important financial statement fraud detection variables.
Identifikation	DOI: 10.1007/978-3-031-07472-1_7
Link	https://link.springer.com/chapter/10.1007/978-3-031-07472-1_7
Copyright	

Tab. 4. Factsheet Beitrag 2

Beitrag 3: Towards a transparency-oriented and integrating Service Registry for the Smart Living Ecosystem

Titel	Towards a transparency-oriented and integrating Service Registry for the Smart Living Ecosystem
Autoren	**Jonas Rebstadt**, Henrik Kortum, Simon Hagen, Oliver Thomas
Publikationsorgan	Informatik 2021
Ranking	WKWI: B / VHB JQ3: C
Status	Veröffentlicht
Bibliographische Information	Rebstadt, J.; Kortum, H.; Hagen, S.; Thomas, O., (2021). Towards a transparency-oriented and integrating Service Registry for the Smart Living Ecosystem. In: Gesellschaft für Informatik e.V. (GI) (Hrsg.), INFORMATIK 2021. Gesellschaft für Informatik, Bonn, S. 1425-1438.
Zusammenfassung	Many domains are increasingly dominated by interdependent services and data exchange between different actors, leading to the emergence of data ecosystems. As a result, service engineers are increasingly tasked with integrating existing service components and data sources into service systems and orchestrating them. In complex areas such as smart living, these tasks are even more difficult by the particular relevance of individual data protection requirements and the low fault tolerance of security-related systems. To address these issues, a central service registry for the domain smart living has been prototypically developed and evaluated, focusing especially on the transparency of data flows and the technical exchangeability of service components. In this way, added value is achieved for data providers and for data users by providing information on the forwarding of their own data as well as on the origin of the data and possible data quality.
Identifikation	DOI: 10.18420/informatik2021-118
Link	https://dl.gi.de/handle/20.500.12116/37623
Copyright	

Tab. 5. Factsheet Beitrag 3

Beitrag 4: Learning without Looking: Similarity Preserving Hashing and Its Potential for Machine Learning in Privacy Critical Domains

Titel	Learning without Looking: Similarity Preserving Hashing and Its Potential for Machine Learning in Privacy Critical Domains
Autoren	Marian Eleks, **Jonas Rebstadt**, Philipp Fukas, Oliver Thomas
Publikationsorgan	Informatik 2022
Ranking	WKWI: B / VHB JQ3: C
Status	Veröffentlicht
Bibliographische Information	Eleks, M.; Rebstadt, J.; Fukas, P.; Thomas, O., (2022). Learning without Looking: Similarity Preserving Hashing and Its Potential for Machine Learning in Privacy Critical Domains. In: Demmler, D.; Krupka, D.; Federrath., H. (Hrsg.), INFORMATIK 2022. Gesellschaft für Informatik, Bonn, S. 161-177.
Zusammenfassung	Machine Learning is frequently ranked as one of the most promising technologies in several application domains but falls short when the data necessary for training is privacy-sensitive and can thus not be used. We address this problem by extending the field of Privacy Aware Machine Learning with the application of Similarity Preserving Hashing algorithms to the task of data anonymization in a Design Science Research approach. In this endeavor, novel anonymization algorithms made to enable Machine Learning on anonymized data are designed, implemented, and evaluated. Throughout the Design Science Research process, we present a collection of issues and requirements for Privacy Aware Machine Learning algorithms along with three Similarity Preserving Hashing-based algorithms to fulfil them. A metric-based comparison of established and novel algorithms as well as new arising opportunities for Machine Learning on sensitive data are also added to the current knowledge base of Information Systems research.
Identifikation	DOI: 10.18420/inf2022_16
Link	https://dl.gi.de/handle/20.500.12116/39513
Copyright	

Tab. 6. Factsheet Beitrag 4

Beitrag 5: Non-Discrimination-by-Design: Handlungsempfehlungen für die Entwicklung von vertrauenswürdigen KI-Services

Titel	Non-Discrimination-by-Design: Handlungsempfehlungen für die Entwicklung von vertrauenswürdigen KI-Services
Autoren	**Jonas Rebstadt**, Henrik Kortum, Laura Sophie Gravemeier, Birgid Eberhardt, Oliver Thomas
Publikationsorgan	HMD Praxis der Wirtschaftsinformatik
Ranking	WKWI: B / VHB JQ3: D
Status	Veröffentlicht
Bibliographische Information	Rebstadt, J., Kortum, H., Gravemeier, L. S., Eberhardt, B., & Thomas, O. (2022). Non-Discrimination-by-Design: Handlungsempfehlungen für die Entwicklung von vertrauenswürdigen KI-Services. In: HMD Praxis der Wirtschaftsinformatik, 59(2), S. 495-511.
Zusammenfassung	Neben der menschen-induzierten Diskriminierung von Gruppen oder Einzelpersonen haben in der jüngeren Vergangenheit auch immer mehr KI-Systeme diskriminierendes Verhalten gezeigt. Beispiele hierfür sind KI-Systeme im Recruiting, die Kandidatinnen benachteiligen, Chatbots mit rassistischen Tendenzen, oder die in autonomen Fahrzeugen eingesetzte Objekterkennung, welche schwarze Menschen schlechter als weiße Menschen erkennt. Das Verhalten der KI-Systeme entsteht hierbei durch die absichtliche oder unabsichtliche Reproduktion von Vorurteilen in den genutzten Daten oder den Entwicklerteams. Da sich KI-Systeme zunehmend als integraler Bestandteil sowohl privater als auch wirtschaftlicher Lebensbereiche etablieren, müssen sich Wissenschaft und Praxis mit den ethischen Rahmenbedingungen für deren Einsatz auseinandersetzen. Daher soll im Kontext dieser Arbeit ein wirtschaftlich und wissenschaftlich relevanter Beitrag zu diesem Diskurs geleistet werden, wobei am Beispiel des Ökosystems Smart Living auf einen sehr privaten Bezug zu einer diversen Bevölkerung bezuggenommen wird. Im Rahmen der Arbeit wurden sowohl in der Literatur als auch durch Expertenbefragungen Anforderungen an KI-Systeme im Smart-Living-Ökosystem in Bezug auf Diskriminierungsfreiheit erhoben, um Handlungsempfehlungen für die Entwicklung von KI-Services abzuleiten. Die Handlungsempfehlungen sollen vor allem Praktiker dabei unterstützen, ihr Vorgehen zur Entwicklung von KI-Systemen um ethische Faktoren zu ergänzen und so die Entwicklung nicht-diskriminierender KI-Services voranzutreiben.
Identifikation	DOI: https://doi.org/10.1365/s40702-022-00847-y
Link	https://link.springer.com/article/10.1365/s40702-022-00847-y
Copyright	Copyright is retained by the authors.

Tab. 7. Factsheet Beitrag 5

Beitrag 6: Developing an Artificial Intelligence Maturity Model for Auditing

Titel	Developing an Artificial Intelligence Maturity Model for Auditing
Autoren	Philipp Fukas, **Jonas Rebstadt**, Florian Remark, Oliver Thomas
Publikationsorgan	European Conference on Information Systems (ECIS 2021)
Ranking	WKWI: A / VHB JQ3: B
Status	Veröffentlicht
Bibliographische Information	Fukas, P.; Rebstadt, J.; Remark, F.; Thomas, O. (2021). Developing an Artificial Intelligence Maturity Model for Auditing. In: European Conference on Information System (ECIS 2021), A Virtual AIS Conference, Research Paper. 133.
Zusammenfassung	Artificial Intelligence (AI) is increasingly being used in various domains including highly regulated areas such as auditing. Although the use of AI in auditing may seem promising at the first glance, there are a number of implications that have so far prevented its broad application. By proposing the first Auditing Artificial Intelligence Maturity Model (A-AIMM), we assess the adoption and diffusion of AI in auditing by considering audit specific requirements. The resulting model contains eight different dimensions and five different maturity levels that foster audit firms in becoming AI-enabled organisations by providing recommendations for the further use of AI with their current capabilities. The development procedure represents a Design Science Research approach including a systematic literature review, a qualitative survey with audit experts and an iterative development process.
Identifikation	ISBN: 978-1-7336325-6-0
Link	https://aisel.aisnet.org/ecis2021_rp/133
Copyright	Copyright is retained by the authors.

Tab. 8. Factsheet Beitrag 6