Künstliche Intelligenz erklären, verstehen, nutzen
–
Anforderungen an Transparenz und ihr Einfluss auf die Nutzung von KI-Entscheidungsunterstützungssystemen

Von der Philosophischen Fakultät der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades einer Doktorin der Philosophie genehmigte Dissertation

vorgelegt von

Johanna Miriam Werz

Berichterinnen: Prof. Dr. Ingrid Isenhardt
Prof. Dr. Martina Ziefle

Tag der mündlichen Prüfung: 01.04.2025

Danksagung

Diese Arbeit ist im Rahmen meiner Tätigkeit am Lehrstuhl für Informationsmanagement im Maschinenbau (IMA) und am Werkzeugmaschinenlabor WZL der RWTH Aachen University entstanden. Dabei wird eine solche Arbeit nie durch eine Person allein geschrieben, sondern ich konnte auf viele tolle Menschen zählen.

Zunächst möchte ich Prof. Dr. Ingrid Isenhardt danken für die Betreuung der Dissertation. Ihre umfangreiche Unterstützung, der wertschätzende Austausch, das zum richtigen Zeitpunkt Infragestellen sowie Anspornen haben die Arbeit ermöglicht. Für die unkomplizierte Zweitbetreuung danke ich ganz herzlich Prof. Dr. Martina Ziefle sowie für die freundliche Übernahme des Prüfungsvorsitzes PD Dr. Malte Persike.

Ich danke all den großartigen Leuten aus dem Institut, ihr seid klasse. Ein besonderer Dank geht an Prof. Dr. Valerie Varney für den Anschub und das verlässliche Anfeuern. Danke, Dr. Esther Borowski, für die Motivation, dein Verständnis und die Unterstützung in den letzten Jahren, ohne die ich Diss, Arbeit und Familie nicht unter einen Hut bekommen hätte. Meiner besten Co-Teamleiterin Dr. Lea Daling gebührt tausend Dank für das Rückenfreihalten, viele inspirierende Gespräche und die beste Kameradinnenschaft! Danke an mein tolles Team: Mit euch macht Arbeit Spaß. Danke an Dr. Nina Collienne für deine Zeit und das stets offene Ohr und Danke an Dr. Sarah Müller-Abdelrazeq: Deine Unterstützung war Gold wert! Vielen Dank, Dr. Manuela Maschke, der besten Mentorin, die ich mir wünschen konnte für die Motivation und den Schub zum genau richtigen Moment. Ich danke den vielen studentischen Hilfskräfte und Abschlussarbeiterinnen, auf deren verlässliche und großartige Hilfe ich zählen konnte. Hervorheben möchte ich Jennifer Klütsch, Marisa Tenbrock, Jacqueline Engels, Vasilena Koleva, Inga Nießen und Jashandeep Kaur.

Ein riesiges Danke geht an meine wunderbaren Freundinnen und Freunde. Ihr wart da zum Nerven stärken, mit Leben ablenken, Studien testen, inspirieren, diskutieren und korrigieren und habt mir verziehen, dass ich so wenig Zeit hatte. Besonders viel Dank für die Unterstützung bei dieser Arbeit gebührt Karlotta, Julsi und Simon.

Der Dank an meine Familie ist unermesslich. Danke an meine Eltern für ihre bedingungslose Unterstützung und für Abendessen-Diskussionen, bei denen schlechte Argumente nicht zählten – ihr habt schon früh eine Wissenschaftlerin aus mir gemacht. Meiner Schwester Sophia kann ich niemals genug danken: Ich kann mir kein Leben ohne dich vorstellen.

Danke, Konrad, dass du mich erinnerst, was wirklich im Leben wichtig ist. Jens, danke, dass du das alles mitgemacht hast, danke, dass du es weiterhin mitmachst und danke, dass du bei mir bist. Ohne dich wäre ich verloren.

I. Inhaltsverzeichnis

II. Abbildungsverzeichnis

III. Tabellenverzeichnis

IV. Abkürzungsverzeichnis

A-g	Transparenzart: Akkuratheit global (Bedingung Forschungsfrage c)
A-lo	Transparenzart: Akkuratheit lokal (Bedingung Forschungsfrage c)
AI	Artificial Intelligence
AI HLEG	AI High Level Expert Group
ANOVA	Analysis of Variance
ATAS	Attitudes Towards Algorithms-Skala
DSGVO	Datenschutzgrundverordnung
DSS	Decision Support System
ELM	Elaboration Likelihood Model
ERS	Exploratory Research Space
F-g	Transparenzart: Funktionsweise global (Bedingung Forschungsfrage c)
F-lo	Transparenzart: Funktionsweise lokal (Bedingung Forschungsfrage c)
FAIRWork	Forschungsprojekt „Flexibilization of complex Ecosystems using Democratic AI-based Decision and Recommendation Systems at Work“
G-Bedingung	Genauigkeitsbedingung (Forschungsfrage a)
GG	Greenhouse-Geisser
GI	Gesellschaft für Informatik
HA	Bedingung hohe Akkuratheit von 90,1% (Forschungsfrage a)
HCIC	Human-Computer Interaction Center
IEEE	Institute of Electrical and Electronics Engineers
IMA	Lehrstuhl für Informationsmanagement im Maschinenbau
JAS	Judge-Advisor-System
KG	Kontrollbedingung keine Genauigkeitsinformation (Forschungsfrage a)
KI	Künstliche Intelligenz
KUT	Kontrollüberzeugung im Umgang mit Technik
LIME	Local Interpretable Model-agnostic Explanations
M	Mittelwert
ML	Machine Learning
MRU	Motivation to Reduce Uncertainty-Theorie
n	Stichprobengröße
NA	Bedingung niedrige Akkuratheit von 78,9% (Forschungsfrage a)
NFC-K	Kurzskala des Need for Cognition
O-T	Ohne Transparenz (Bedingung Forschungsfrage c)

p	Irrtumswahrscheinlichkeit
R-1	Kurzskala zur Risikobereitschaft
SD	Standardabweichung
TAIGERS	Forschungsprojekt „Transparency in AI: Considering Explainability, User and System Factors"
TAM	Technology Acceptance Modell
URT	Uncertainty Reduction Theory
UTAUT	Unified Theory of Acceptance and Use of Technology
WOA	Weight of Advice
XAI	Explainable AI

Zusammenfassung

Trotz der zunehmenden Zahl an Unterstützungssystemen mit künstlicher Intelligenz (KI) für den Privatgebrauch wurde KI-Transparenz lange Zeit vor allem aus technischer Perspektive erforscht. Studienergebnisse mit Endnutzenden zeigen jedoch, dass Systemtransparenz nicht automatisch zu Systemakzeptanz führt. Es stellt sich also die Frage, wie sich Transparenz von KI-Entscheidungsunterstützungssystemen auf die Nutzung dieser Systeme durch Endnutzende auswirkt.

Im Rahmen der vorliegenden Dissertation wurde diese Forschungsfrage anhand von drei Studien mit einem Mixed Method-Ansatz untersucht. Die erste Studie, ein quantitatives Onlineexperiment mit n = 169 Teilnehmenden, analysierte, wie Akkuratheitsangaben die Nutzung eines Algorithmus nach einem Fehler beeinflussen. Die zweite Studie, qualitative Fokusgruppendiskussionen mit n = 26 Teilnehmenden, identifizierte Anforderungen an KI-Transparenz aus Sicht von Endnutzenden. Die dritte Studie, ein quantitatives Onlineexperiment mit n = 151 Teilnehmenden, verglich vier Transparenzarten hinsichtlich ihrer Wirkung auf Vertrauen und Nutzung der Algorithmen.

Die Ergebnisse zeigen, dass technische Erklärungen allein nicht ausreichen, um das Vertrauen in KI-Systeme zu stärken oder deren Nutzung zu fördern. Mehr als Erklärungen, wie eine KI funktioniert, sind Hintergrundinformationen über Entwickler*innen, die Motive von hinter der KI stehenden Institutionen oder externe Prüfungen entscheidend für die Vertrauensbildung. Akkuratheitsangaben haben einen begrenzten positiven Effekt auf die Nutzung, während Erklärungen, warum ein einzelnes Ergebnis zustande kam, besonders bei Fehlern gewünscht werden. Die Anforderungen sind abhängig von den Eigenschaften eines Systems, insbesondere davon, wie schwerwiegend Fehler wären, und von den Vorerfahrungen der Nutzenden. Wichtiger als eine detaillierte Transparenz ist es, das Verständnis der Nutzenden der Transparenzmaßnahmen sicherzustellen und Verlässlichkeit der KI zu vermitteln.

Die Arbeit betont die Bedeutung einer nutzendenzentrierten Entwicklung von KI-Transparenz, um der Individualität von Systemen und Nutzendengruppen gerecht zu werden. Neben weiteren Implikationen wurde eine Transparenzmatrix für Entwickler*innen ausgearbeitet, mit der sich die notwendigen Transparenzimplikationen auf Basis gegebener Systemeigenschaften identifizieren lassen. Auch für politische Entscheidungsträger*innen ergeben sich Implikationen zur Förderung der Transparenz in KI-Systemen. Darüber hinaus werden die Limitationen der Einzelstudien sowie der Gesamtarbeit diskutiert und weiterführende Fragen für zukünftige Forschung abgeleitet.

Keywords: Künstliche Intelligenz, Transparenz, Algorithm Aversion, Nutzendenanforderungen, Vertrauen, KI-Nutzung, rechtliche Vorgaben.

Abstract (Englisch)

Despite the increasing number of artificial intelligence (AI) systems for private usage, AI transparency has long been researched primarily from a technical perspective. However, study results with end users show that system transparency does not automatically lead to system acceptance. Therefore, the question arises of how transparency of AI decision support systems affects the use of these systems by end users.

In this dissertation, this research question was investigated using three studies with a mixed-method approach. The first study, a quantitative online experiment with *n* = 169 participants, analyzed how accuracy information about an algorithm influences the use of this algorithm after an error. The second study, qualitative focus group discussions with *n* = 26 participants, identified requirements for AI transparency from the perspective of end users. The third study, a quantitative online experiment with *n* = 151 participants, compared four different types of transparency regarding their effect on trust and use of the respective algorithms.

The results show that technical explanations alone are not sufficient to strengthen trust in AI systems or increase their usage. More than explanations of how an AI works, background information about developers, the motives of the institutions behind the AI or external audits help to build trust. Accuracy information has a limited positive effect on usage, while explanations about why a single result emerged are desirable when errors occur. The requirements towards AI transparency depend on the characteristics of the system, in particular how severe errors would be, and users' previous experience. More important than detailed transparency is ensuring that users understand the transparency measures and conveying the reliability of the AI-system.

The work emphasizes the importance of a user-centered development of AI transparency due to the individuality of systems and user groups. In addition to further implications, a transparency matrix for developers was elaborated, which can be used to identify the necessary transparency implications based on given system characteristics. Implications also arise for political decision-makers to promote transparency in AI systems. In addition, limitations of the individual studies and the overall work are discussed and follow-up questions for further research are derived.

Keywords: Artificial Intelligence, Transparency, Algorithm Aversion, User Demands, Trust, AI Usage, Legislative Regulations.

1. Einleitung

Wir stehen am Beginn einer neuen Welt. Die Technologien des maschinellen Lernens, der Spracherkennung und des Verstehens natürlicher Sprache erreichen einen neuen Höhepunkt ihrer Fähigkeiten. Das Ergebnis ist, dass wir bald künstlich intelligente Assistenten haben werden, die uns in jedem Aspekt unseres Lebens helfen.

Amy Stapleton, AI Industry Analyst (Zitat von ca. 2019)

Nach Abschluss eines Videos auf YouTube werden uns automatisch weitere Videos empfohlen. Laut YouTube entstehen 70 % der angesehenen Inhalte auf der Plattform durch diese Empfehlungen (Solsman, 2018). Das vorgeschlagene Video passt inhaltlich zu dem gerade angesehenen, das scheint für uns Nutzende offensichtlich. Dass bei der Auswahl des Videovorschlags künstliche Intelligenz (KI) eingesetzt wird, die zahlreiche weitere Faktoren berücksichtigt und in kürzester Zeit verrechnet, um so autonom die Entscheidung für den nächsten Videovorschlag zu fällen, ist wahrscheinlich einem Teil der Nutzenden klar. Individuelle Videohistorie, Likes und gefolgte Kanäle scheinen berücksichtigt zu werden.

Nachträgliche Reue und Unwohlsein über die betrachteten Inhalte auf YouTube, die in einer Befragung als Folge der KI-Vorschläge berichtet werden, deuten darauf hin, dass das Wohlergehen der Nutzenden nicht im Fokus der Auswahl steht (Mozilla, 2021). Manche Expert*innen argumentieren, YouTube habe ein Interesse daran, unsere Verweildauer auf der Plattform zu maximieren, um uns möglichst viel Werbung zu zeigen. Eine KI, die mit dieser Zielvorgabe arbeitet, macht möglicherweise eher spektakuläre als harmlose Vorschläge (Yoo, 2018). Ebenso lässt sich YouTube dafür bezahlen, bestimmte Inhalte ganz konkreten Zielgruppen zu zeigen (Google, 2024). Die KI der Plattform also einen Anreiz, bei der Wahl der Nachfolgevideos wirtschaftliche Interessen des Unternehmens über die persönliche Präferenz der Nutzenden zu stellen.

Warum uns von YouTube ein Video vorgeschlagen wird, ist für uns Nutzende nicht nachvollziehbar. Es herrscht keine Transparenz darüber, wie die Auswahl-KI der Videoplattform ihre Entscheidungen trifft. Vielmehr ist die KI ein wohlgehütetes Geheimnis von YouTube (Albert, 2023). In ähnlicher Weise bleibt Facebook vage, wenn es um die Verwertung der Nutzendendaten oder die Identifizierung von Fakenews geht und X, vormals Twitter, hat ein ursprünglich öffentliches Projekt zur Moderation der Inhalte wieder eingestellt (Albert, 2023; Dang, 2023; Singh, 2020).

Obwohl seit Jahren die Zahl der KI-Systeme in unserem Arbeits- und Privatleben steigt und KI in immer mehr Bereichen Einfluss nimmt (Littman et al., 2021), wird für Endnutzende in den seltensten Fällen transparent, wo KI eingesetzt wird, wie sie ihre Ergebnisse produziert oder Daten verarbeitet. Neben

bewussten unternehmerischen Entscheidungen behindern auch Faktoren, die in der KI selbst liegen, diese Transparenz. Mit wachsender Komplexität der KI wird es selbst für Entwickler*innen zunehmend schwierig nachvollziehen, wie das System zu seinen Schlussfolgerungen gelangt (Larsson & Heintz, 2020). Zwar kann Auskunft darüber gegeben werden, wie zugrundeliegende Algorithmen und Modelle trainiert, also erstellt wurden. Aber da auf Grundlage dieser Trainingsprozesse selbstlernende Systeme mit vernetzten und vielschichtigen Entscheidungsknoten – sogenannte neuronale Netze – entstehen, können die genauen Prozesse, die KI-Entscheidungen zugrunde liegen, immer seltener nachvollzogen werden und es bleibt unklar, wieso auf eine Eingabe genau diese Antwort folgt (Arrieta et al., 2020). Das, was heutzutage als künstlich intelligent bezeichnet wird, sei es zur Bilderstellung (Dall-E), zum Schreiben von Gedichten (Haiku) oder zur Texterstellung im Chat (ChatGPT), ist in seiner Funktionsweise überwiegend eine Blackbox (Chiba, 2022; OpenAI, 2022, 2023).

Zum einen haben Ereignisse, in denen solche künstlich intelligenten Systeme problematische – z. B. rassistische, sexistische oder unsoziale – oder schlicht falsche Ergebnisse produzieren, Forderungen nach mehr Transparenz dieser Systeme zur Folge (Dastin, 2018; Hundt et al., 2022; O'Neil, 2016; Starr, 2014). Dahinter steckt die Annahme, durch Einsichten in und über die Systemprozesse Fehler erkennen und beseitigen zu können: Transparenz als Voraussetzung für Systemverbesserungen für Entwickler*innen (Ananny & Crawford, 2018).

Zum anderen gilt Transparenz als Zielzustand, wenn KI-Systeme Entscheidungen (mit-)treffen, die Menschen betreffen. Dabei muss die menschliche Entscheidungsautonomie bei der Benutzung von KI als Teil der Menschenwürde gewahrt bleiben, argumentiert z. B. die AI High Level Expert Group der Europäischen Kommission in ihrer „Ethical guidelines for trustworthy AI" (AI HLEG, 2019). Um dieser ethischen Richtlinie zu folgen, bedarf es also Transparenz. So erhöhen neue Gesetze und Gesetzesinitiativen den Druck auf Unternehmen, in ihren KI-Systemen Transparenz herzustellen. Beim gerade verabschiedeten AI Act der EU geht es nicht mehr (nur) um Transparenz für Entwickler*innen, sondern um die Anforderung, Endnutzenden Einblicke in die Systeme zu ermöglichen, die sie nutzen (Verordnung über künstliche Intelligenz, 2024). In ähnlicher Weise verfolgt auch die DSGVO Transparenz: Es geht um Transparenz algorithmischer Anwendungen als Recht der Nutzenden (Felzmann et al., 2019). Und auch der 2020 ratifizierte Medienstaatsvertrag, der nun auch Medienplattformen wie Google News oder YouTube umfasst, fordert in Paragraph 85 von den Anbietern Transparenz ein, beispielsweise bezüglich der Kriterien, nach denen bestimmte Inhalte angezeigt oder sortiert werden (Medienstaatsvertrag, 2020).

Die Zahl der Studien zu transparenter KI hat seit 2015 massiv zugenommen. Technische Disziplinen, insbesondere Informatik und Computerwissenschaften, konzentrieren sich dabei vor allem auf Transparenz zur Systemverbesserung. Unter den Begriffen Explainability bzw. XAI (Explainable AI) und

Interpretability untersuchen sie Methoden, um KI-Blackboxen für Entwickler*innen transparent zu gestalten (Arrieta et al., 2020; Felzmann et al., 2019; Miller, 2019; Molnar, 2019; Samek et al., 2019). In den letzten Jahren rückten immer mehr auch die Endnutzenden von KI-Systemen in den Fokus, unter der Annahme, „that by building more transparent, interpretable, or explainable systems, users will be better equipped to understand and therefore trust the intelligent agents" (Miller, 2018, S. 3).

Mit dem zunehmenden Interesse an den Endnutzenden und der erkannten Lücke zwischen Explainability für Expert*innen und dem Anspruch, Transparenz für Endnutzende zu gestalten, stieg zuletzt auch die Zahl der sozialwissenschaftlichen Studien über transparente KI und Algorithmen (z. B. Bertino et al., 2019; Felzmann et al., 2019; Herm et al., 2023; Larsson & Heintz, 2020; Wanner et al., 2022). Die Effekte von transparenter KI auf Variablen wie Zufriedenheit mit, Verständnis von einem System oder auf Vertrauen zeigten sich jedoch sehr uneinheitlich. So hebt Springer (2019) hervor: „In some settings there are expected benefits: transparency improves algorithmic perceptions because users may better understand system behavior" (S. 101), während in anderen Kontexten „transparency can have other quite paradoxical effects. Transparency may cause users to have worse perceptions of a system, trusting it less because the transparency led them to question the system even when it was correct" (Springer, 2019, S. 101).

Einerseits ist diese Uneinheitlichkeit auf das sehr breite und vielfältige Verständnis des Begriffs Transparenz zurückzuführen (siehe Kapitel 2.2). Darüber hinaus besteht die Annahme, dass Kontextfaktoren und Systemeigenschaften einen großen Einfluss auf die Effekte von Transparenz haben. Zum Dritten sind abhängige Variablen wie Vertrauen oder Zufriedenheit sicherlich zentral, um die Einstellung gegenüber KI-Systemen zu erheben. Eigentlich von Interesse ist jedoch die tatsächliche Nutzung dieser Systeme – und diese Variable wurde bisher vergleichsweise selten untersucht.

Es gilt also die Frage zu klären: **Wie wirkt sich Transparenz von KI-Entscheidungsunterstützungssystemen auf die Nutzung dieser Systeme durch Endnutzende aus?**

1.1. Forschungsfragen der Arbeit

Zur Beantwortung dieser zentralen Fragestellung werden drei Perspektiven und Methoden gewählt, um verschiedene Aspekte von KI-Transparenz zu beleuchten. Ziel ist es, mithilfe dieser drei Perspektiven ein breites Verständnis des Untersuchungsgegenstands zu erlangen, der an ausgewählten Stellen in die Tiefe geht und gleichzeitig der Breite des Themas gerecht wird. Die drei gewählten Perspektiven zeichnen sich durch verschiedene Fragestellungen und Forschungsdesigns aus und tragen je einen Teil zur Antwort auf die übergeordnete Forschungsfrage (FF) bei.

In der ersten Forschungsfrage a, im Folgenden mit „**Fehlerfall**“ betitelt, wird der Fokus auf den speziellen Effekt der Algorithm Aversion gelegt, also auf die Ablehnung des Algorithmus durch die Nutzenden nach einem erlebten Fehlerfall (Dietvorst et al., 2014). Der Effekt tritt selbst dann auf, wenn die Ablehnung der algorithmischen Hilfe in Folge eines erlebten Fehlers zu schlechteren Ergebnissen führt als ohne Nutzung (Dietvorst et al., 2014). Da aus diesem Grund Algorithm Aversion eine Gefahr für gutes Entscheiden darstellt, sind Maßnahmen zur Überwindung notwendig. Als ein Grund für Algorithm Aversion gilt, dass Nutzende von Algorithmen der falschen Annahme folgen, Algorithmen arbeiteten fehlerfrei (Prahl & Swol, 2017). Information über die in Wahrheit eingeschränkte Akkuratheit der von Algorithmen produzierten Ergebnisse könnte demnach der Überwindung von Algorithm Aversion dienen. Das Maß der algorithmischen Akkuratheit ist ein grundlegender, technisch einfach zu ermittelnder Fakt über ein KI-System, der zur Transparenz eines KI-Systems beiträgt. Beim Ansatz (a) wird die Frage adressiert:

> **FF (a) „Fehlerfall“: Inwieweit führen Angaben von Akkuratheit eines Algorithmus dazu, dass dieser auch nach einem Fehlerfall genutzt wird?**

Mehr zu den Hintergründen und der Herleitung der Forschungsfrage findet sich in Kapitel 2.1.4. Die Forschungsfrage verfolgt also einen sehr engen Fokus auf ein spezifisches KI-Nutzungsphänomen, das auch nur bezüglich einer Art der Transparenz untersucht wird. Dies geschieht in einer experimentellen Onlinestudie, also mit **quantitativem** Design.

Über den speziellen Effekt der Algorithm Aversion hinaus besteht die Herausforderung, dass das Verständnis von Transparenz insbesondere für Laien noch nicht ausreichend geklärt ist. Trotz der gestiegenen Anzahl von Studien zur Transparenz von KI für Endnutzende stellt der Großteil der Untersuchungen quantitative Ansätze dar (Springer, 2019). Eine qualitative Betrachtung dessen, was Laien über KI-Systeme wissen möchten, hinsichtlich welcher Aspekte sie sich Transparenz wünschen und welches Verständnis von Transparenz dahintersteckt, ist noch nicht geklärt. Mögliche Einflüsse könnten Kontext und technologischen Eigenschaften zukommen. Entsprechend widmet sich der zweite Ansatz, der im Folgenden als **„Nutzendenanforderungen“** bezeichnet wird, der Forschungsfrage:

> **FF (b) „Nutzendenanforderungen“: Welche Anforderungen an Transparenz in KI für Endnutzende bestehen und inwiefern sich diese nach Eigenschaften der KI unterscheiden.**

Zu diesen Eigenschaften zählen beispielsweise der Anwendungskontext oder die Relevanz, die mögliche Fehler der Anwendung auf das eigene Leben haben. Die Herleitung dieser Forschungsfrage erfolgt in Kapitel 2.2.2. Es handelt sich beim Ansatz (b) um eine **qualitative** Fokusgruppenstudie, die KI-Transparenz in ihrer Breite untersucht.

Daneben unterscheidet die wissenschaftliche Literatur – und, wie Forschungsfrage (b) zu Nutzendenanforderungen annimmt, auch Endnutzende – verschiedene (technisch mögliche) Arten der Transparenz in künstliche Intelligenz. Gleichzeitig besteht eine Forschungslücke hinsichtlich der tatsächlichen Nutzung von transparenter KI, wie schon bei Forschungsfrage (a) „Fehlerfall" aufgeführt. Häufiger wurden bisher Effekte von transparenten Systemen auf Vertrauen und Akzeptanz, oft über Skalen, erhoben. Während Forschungsfrage (a) „Fehlerfall" einen sehr engen Fokus setzt und den Effekt von Akkuratheit im Fehlerfall untersucht, widmet sich Forschungsfrage (c), die im Folgenden mit „**Transparenzarten**" zusammengefasst wird, mit einem vergleichenden Design vier verschiedenen Arten der Transparenz und untersucht quantitativ:

> **FF (c) „Transparenzarten": Wie wirken sich verschiedene Arten der KI-Transparenz auf Vertrauen und Nutzung eines Systems aus?**

Dabei werden vier Arten unterschieden, die sich nach Umfang der Erklärung – ganzes System oder einzelnes Ergebnis – und Inhalt – Funktionalität des Systems oder seine Akkuratheit – einteilen lassen. Mehr Details zu verschiedenen Transparenzarten sowie die Einordnung der Forschungsfrage sind in Kapitel 2.2.4 zu finden. Ansatz (c) wird mit einem experimentellen, hypothesentestenden Design umgesetzt, also als **quantitative** Studie.

Diesen Forschungsfragen widmet sich die vorliegende Arbeit und trägt durch die verschiedenen Ansätze dazu bei, die Anforderungen, die aus Endnutzendensicht an Transparenz in KI gestellt werden und werden sollten, aufzudecken. Dies kann einerseits Entwickler*innen bei der Gestaltung von KI-Systemen zu einer verbesserten Nutzendenzentrierung verhelfen. Andererseits liefert die Arbeit Empfehlungen für politisch Gestaltende, die auf nationaler Ebene wie EU-Ebene KI-Regularien entwerfen.

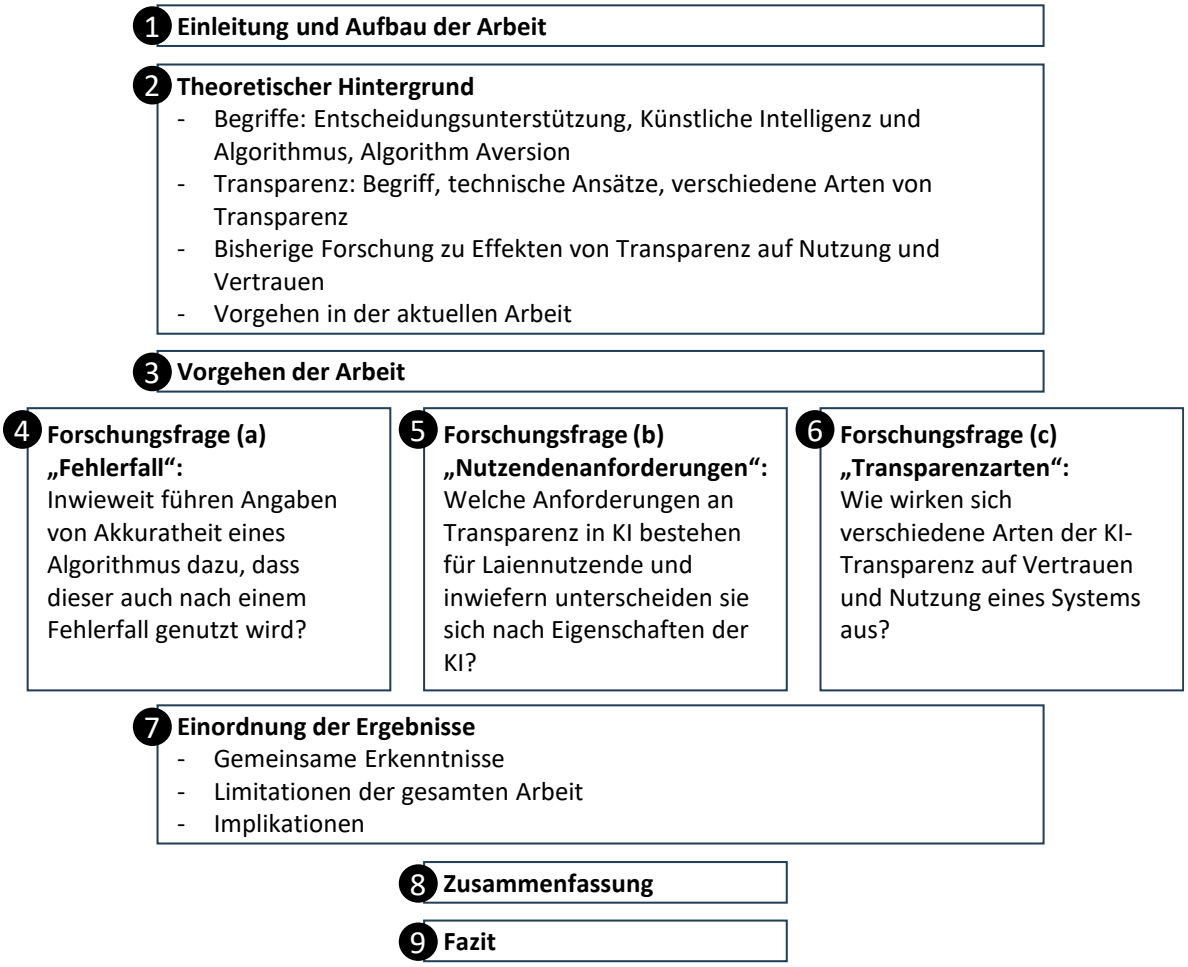
1.2. Aufbau der Arbeit

Im folgenden Kapitel 2 werden zunächst der wissenschaftliche Hintergrund und relevante wissenschaftliche Vorarbeiten der Arbeit dargelegt. Dabei werden außerdem wichtige Begriffe hergeleitet und ihr Gebrauch für den weiteren Verlauf geklärt, darunter Entscheidungsunterstützungssystem, künstliche Intelligenz und Transparenz. Außerdem werden die Forschungsfragen der Arbeit eingeordnet, ihr Hintergrund und ihre Relevanz dargelegt. Ebenso findet im folgenden Kapitel eine ausführliche Darstellung bereits stattgefundener Nutzungsstudien zu transparenter KI statt, um den aktuellen Forschungsstand des Feldes zu beschreiben. Anschließend folgt im Kapitel 3 eine Beschreibung des Vorgehens, das zur Untersuchung der Forschungsfragen gewählt wurde.

Aufbauend auf dem wissenschaftlichen Hintergrund, den Fragestellungen und dem dargelegten Vorgehen im Rahmen der Dissertationsarbeit folgen in den Kapiteln 4 bis 6 die Auseinandersetzungen mit den Forschungsfragen (a), (b) und (c). Zunächst setzt sich Kapitel 4 mit einem quantitativen Experiment mit der Frage auseinander, welchen Einfluss Akkuratheitsangaben über KI-Systeme auf die Nutzung dieser Systeme nach Fehlern haben (Forschungsfrage (a) „Fehlerfall“). Kapitel 5 setzt einen sehr viel breiteren Ansatz an und erforscht mithilfe von Fokusgruppen qualitativ die Anforderungen von Laien an KI-Systeme in Abhängigkeit von verschiedenen Systemeigenschaften (Forschungsfrage (b) „Nutzendenanforderungen“). In Kapitel 6 wird wiederum ein quantitatives Experimentaldesign dargelegt, mit dessen Hilfe der Effekt von vier verschiedenen Transparenzarten auf die Nutzung von und das Vertrauen in ein KI-System untersucht wurde (Forschungsfrage (c) „Transparenzarten“). Der Aufbau dieser drei Kapitel besteht jeweils aus einer kurzen theoretischen Einführung, der Beschreibung der Methodik, dem Bericht der Ergebnisse, einer forschungsfragenspezifischen Diskussion, Implikationen und Fazit.

Anschließend an die Auseinandersetzung mit den einzelnen Forschungsfragen folgen in Kapitel 7 die Einordnung der vorangegangenen Ergebnisse hinsichtlich der übergeordneten Forschungsfrage sowie eine gemeinsame Diskussion, in deren Rahmen Implikationen für Praxis und Theorie sowie Limitationen der Arbeit dargelegt werden. Nach einer Zusammenfassung in Kapitel 8 folgt in Kapitel 9 das Fazit mit Ausblick.

***Abbildung 1**: Übersicht über den Aufbau der vorliegenden Arbeit; die Zahlen stehen für die Kapitelnummern.*



2. Theoretischer Hintergrund

Um die Fragestellungen einzuordnen und die im folgenden Kapitel gewählte Methodik herzuleiten, werden in diesem Kapitel Begriffsdefinitionen vorgenommen, bestehende Forschung zu diesen dargelegt und die Herleitung der einzelnen Fragestellung erläutert. Dazu erfolgt im nächsten Abschnitt 2.1 die Herleitung des Begriffes „künstliche Intelligenz“ und damit zusammenhängender Konzepte. Dazu wird zunächst der Begriff KI zusammen mit verwandten Konzepten eingeordnet (2.1.1), anschließend ein Ziel von KI, die Entscheidungsunterstützung von Nutzenden, dargelegt (2.1.2) sowie die an KI beteiligten Zielgruppen aufgezeigt (2.1.3). Zuletzt erfolgt die Auseinandersetzung mit dem Konzept der Algorithm Aversion in Kapitel 2.1.4. Ausgehend von diesem Effekt ergibt sich in diesem Kapitel auch die erste **Forschungsfrage (a) „Fehlerfall“**.

Das darauffolgende Kapitel 2.2 geht auf den Begriff der Transparenz ein, der zunächst in seinem geschichtlichen Verständnis eingeführt (2.2.1) und dann vor dem Hintergrund algorithmischer Transparenz näher ausdifferenziert wird. Dazu gehört die Erläuterung verschiedener Arten der Transparenz (2.2.2), woraus sich **Forschungsfrage (b) „Nutzendenanforderungen“** ableitet. Anschließend erfolgt die Auseinandersetzung mit Akkuratheit als Gegenläufer wie auch Bestandteil von Transparenz (2.2.3) sowie mit technischen Ansätzen von Transparenz in KI (2.2.4), aus denen **Forschungsfrage (c) „Transparenzarten“** folgt. Der Abschnitt endet mit einer Darlegung des Transparenzverständnisses für die vorliegende Arbeit (2.2.5).

Angesichts der Forschungsfragen, die sich mit der Nutzung von Transparenz befassen, folgt in Kapitel 2.3 eine Übersicht über bestehende Nutzungsstudien zu transparenter KI. Dabei werden zunächst die häufig und für die vorliegende Arbeit relevanten abhängigen Variablen Nutzung und Vertrauen eingeordnet und voneinander abgegrenzt (Kapitel 2.3.1). Anschließend wird explizit auf Studien zu den Effekten von Akkuratheitsangaben eingegangen (Kapitel 2.3.2), um in Kapitel 2.3.3 ausführlich die Effekte von weiteren, klassischen Transparenzarten auf die Nutzung aufzuzeigen.

Das vorliegende Kapitel endet mit einer Zusammenfassung der Darstellung der Fragestellungen im Abschnitt 2.4, bevor im nächsten Kapitel 3 das Vorgehen zur Untersuchung der drei Forschungsfragen dargelegt wird.

2.1. Künstliche Intelligenz

Die Zahl der Anwendungen, die KI nutzen, nimmt auch im privaten Bereich stetig zu. Gleichzeitig ist der Blick der Gesellschaft auf KI sehr uneinheitlich und reicht von Ablehnung und Skepsis bis hin zu großer Hoffnung (Brauner et al., 2023, 2024). Dabei wird der Begriff KI, wie es bei Trendthemen häufig der Fall ist, sehr uneinheitlich verwendet, je nach Anwendungsdomäne, Disziplin und Zeitpunkt

(Larsson & Heintz, 2020). Deshalb erfolgt im nächsten Kapitel 2.1.1 zunächst die Herleitung des Begriffs KI sowie weiterer verwandter Begrifflichkeiten wie Machine Learning und Algorithmus, um abschließend ihre Verwendung im Rahmen der Arbeit darzulegen.

Die Ziele von KI sind vielfältig und reichen von der Automatisierung ganzer Prozesse bis hin zur Unterstützung bei Text- oder Bildgenerierung. Als Werkzeug für den Menschen ist das Ziel von KI seit langem die Unterstützung und Verbesserung von Entscheidungen. Eine Auseinandersetzung mit diesem Ziel und den daraus für die Arbeit resultierenden Einschränkungen geschieht in Kapitel 2.1.2. Anschließend folgt eine Darlegung, welche Zielgruppen für KI unterschieden und welche im Rahmen der Arbeit betrachtet werden (Kapitel 2.1.3). Zum Abschluss wird in Kapitel 2.1.4 auf einen speziellen, im Anwendungsbereich der KI-Entscheidungsunterstützung relevanten Effekt eingegangen: Algorithm Aversion. Dabei wird auch die erste Forschungsfrage der Arbeit abgeleitet, bei der dieser Effekt im Fokus stand.

2.1.1. Der Begriff und seine Einordnung

John McCarthy, der die Forschung zu KI 1956 begründete, bezeichnete KI als „the science and engineering of making intelligent machines, especially intelligent computer programs“ (McCarthy, 2017; McCarthy et al., 1956). Wenn auch bis heute gültig, beinhaltet diese Definition das Problem, dass auch der Begriff „Intelligenz“ auf verschiedenen Weisen definiert wird (Legg & Hutter, 2007). Betrachtet man die Systeme, die als KI bezeichnet werden, so ist der Begriff ähnlich unstet: Was als Gegenstand der KI gilt, hat sich seit den 50ern ständig verändert (Littman et al., 2021). Eine Erklärung hierfür liefert der sogenannte „KI-Effekt“ (AI effect), wonach KI nur bis zu dem Zeitpunkt als solche bezeichnet wird, bis sie das Problem, für das KI als notwendig empfunden wird, löst (erste Beschreibung des Effekts durch Simon, 1955). Demzufolge ist eine KI, sobald sie entwickelt ist und funktioniert, schlicht eine weitere Softwareanwendung (siehe auch P. Stone et al., 2016).

Heutige Definitionen bezeichnen KI als eine Reihe von Technologien wie Spracherkennung, Bildklassifizierung oder selbstlernende Systeme. Damit wäre KI der Überbegriff über verschiedene Methoden, die zum Lösen besonders komplexer Probleme verwendet werden. Die häufigsten Methoden werden Supervised, Unsupervised oder Reinforcement Learning genannt und unter dem Begriff „Machine Learning“ (ML) zusammengefasst (Goodfellow et al., 2016). KI bezeichnet ML-Modelle, die komplex und vielschichtig sind und eine Blackbox bilden, deren Prozesse für den Menschen nicht mehr (einfach) nachzuvollziehen sind (Herm et al., 2023; Larsson & Heintz, 2020). Dass sich die Definition von KI im Wandel befindet, ständig dem neuesten Stand anpasst und deshalb immer wieder uneinheitlich ist, sei nicht nur als Nachteil zu betrachten, sondern ermöglichte laut der „One hundred year study on Artificial Intelligence“ „the field to grow, blossom, and advance at an ever-accelerating pace“ (P. Stone et al., 2016, S. 12).

Algorithmus, im Gegenzug, ist in den Computerwissenschaften klarer definiert und bezeichnet eine schrittweise Anleitung zum Lösung eines konkreten Problems (Knuth, 1997). In den Sozialwissenschaften wird der Begriff Algorithmus sehr viel weniger eng benutzt und umfasst recht allgemein Software, künstliche Intelligenz oder Computer-basierte („embedded“) technische Systeme, mit denen Menschen über Bildschirme in Interaktion treten. Algorithmen lassen sich also von Robotern oder virtuellen Realitäten abgrenzen (Glikson & Woolley, 2020). Während in sozialwissenschaftlichen Untersuchungen die Interaktion mit Algorithmen im Fokus steht, sind die technischen, dahinterliegenden Funktionsweisen von Algorithmen weniger von Interesse oder definiert (siehe z. B. den Begriff Algorithm Aversion; Dietvorst et al., 2014).

In den technischen Disziplinen, die für diese Arbeit relevant sind, wird der Begriff Algorithmus überwiegend synonym verwendet mit ML-Modellen (z. B. Goodfellow et al., 2016; Molnar, 2019). Im Zusammenschluss oder in größerem Umfang wird auch von ML-Systemen gesprochen (z. B. Carvalho et al., 2019). Dabei werden im Bereich Machine Learning flache und tiefe ML-Modelle unterschieden (Goodfellow et al., 2016). Während flache ML-Modelle einfacher gebaut sind und lineare Regressionen oder Entscheidungsbäume bezeichnen, sind „Deep Machine Learning Models“ vielschichtiger und komplexer. Sie bilden dabei sogenannte neuronale Netze – englisch Deep Neural Networks –, die nicht per se verständlich, sondern Blackboxen sind (Goodfellow et al., 2016; Herm et al., 2023; Rudin, 2019). Der Begriff der neuronalen Netze deutet die Anlehnung der Technologie an die Idee des menschlichen Gehirns an. Hier schließt sich der Kreis zum Begriff künstliche Intelligenz (McCarthy et al., 1956).

In der vorliegenden Arbeit werden die beiden Begriffe KI und Algorithmus weitgehend synonym verwendet, um computerbasierte komplexe technische Systeme und Software zu beschreiben, mit denen Nutzende über einen Bildschirm interagieren und die sie bei Entscheidungen unterstützen. Von Algorithmen ist dabei insbesondere im Zusammenhang mit den konkreten sozialwissenschaftlichen Untersuchungen die Sprache, bei denen die genaue Funktionsweise weniger im Fokus steht. Hingegen bezeichnet KI die Systeme, die als Blackboxen anzusehen und nicht von sich aus verständlich sind.

Eine weitere wichtige Einschränkung des Begriffs kommt durch die Anwendung, für die KI eingesetzt wird, zustande. Das für diese Arbeit relevante Verständnis von KI zur menschlichen Entscheidungsunterstützung ergibt sich durch die Idee, künstliche Intelligenz zu nutzen, um die menschliche Intelligenz zu erweitern. Dieses Ziel von KI wird im folgenden Kapitel näher erläutert.

2.1.2. Ein Ziel von KI: Entscheidungsunterstützung

Menschen treffen ständig Entscheidungen[1]. Besonders bei komplexen, wichtigen Fragestellungen machen wir uns diese bewusst, denken nach, schreiben Pro-Contra-Listen oder fragen andere um Hilfe. Ein Großteil der Entscheidungen wird allerdings unbewusst und automatisch getroffen. Dies steht im starken Kontrast zur Annahme des Menschen als homo economicus, also der Vorstellung des Menschen als rationales nutzenmaximierendes Wesen. Diese Annahme wurde spätestens durch die Arbeit von Amos Tversky und Daniel Kahnemann widerlegt, die mit dem Wirtschaftsnobelpreis ausgezeichnet wurde (Kahneman et al., 1982; Tversky & Kahneman, 1973). Die aus ihrer Arbeit entstandene Verhaltensökonomie zeigt auf, dass Menschen nicht nutzenmaximierend handeln, sondern Heuristiken und Abkürzungen beim Entscheiden nutzen. Während diese unter den Bedingungen von Zeit- und Ressourcenknappheit, denen das menschliche Gehirn unterliegt, sicherlich sinnvoll sind, führen sie nicht in jedem Fall zu den rational besten oder richtigen Ergebnissen. Der Reiz, die begrenzte Rationalität des Menschen zu erweitern und so zu besseren Ergebnissen zu gelangen, ist schon immer groß, weshalb Menschen Werkzeuge nutzen, um ihre Fähigkeiten zu erweitern, dazuzulernen, bessere Entscheidungen zu treffen. Ein besonders vielversprechendes Werkzeug stellen technische Systeme dar, die große Mengen an Daten durchsuchen, ordnen, emotionslos beurteilen oder in Zusammenhang bringen können: Algorithmen bzw. in ihrer komplexen Form KI. Sie zur Unterstützung von Menschen bei Entscheidungen zu nutzen und so Beschränkungen des menschlichen Gehirns zu erweitern, ist zur Erlangung bestmöglicher Entscheidungen also attraktiv.

Der für diese Systeme eingeführte Begriff „Entscheidungsunterstützungssystem“, „Decision Support System“ (DSS), bezeichnete dabei lange Zeit Systeme, die im Managementbereich eingesetzt wurden, um Berichte zu erstellen, Produktplanung und Marketingmaßnahmen zu skizzieren oder Investitionen zu bewerten. Während DSS in der Arbeitswelt seit den 1970ern an Popularität gewannen, entstand mit zunehmend intelligenten Systemen in den 1990ern eine Brache, die – auch heute noch – Business Intelligence genannt wird (Power, 2002; Turban et al., 2011). Die Effektivität dieser Systeme bei der Unterstützung oder gar zum Ersatz von Menschen zeigte sich früh: Schon in den 1950ern übertrafen algorithmische Vorhersagen zum Verhalten von Patient*innen, die damals auf einfachen Regressionen basierten, die Vorhersagen von Psycholog*innen (Meehl, 1954). Eine spätere Metaanalyse bestätigte diese Überlegenheit in einer Vielzahl von Feldern: Medizin, Bildung bis hin zu Managemententscheidungen (Grove et al., 2000).

Ein Effekt, der sich bei Entscheidungsunterstützung allerdings immer wieder zeigt und bereits viel Forschung provozierte, ist der des Advice Discounting. Damit wird die nicht- oder nur teilweise

[1] In populärwissenschaftlichen Artikeln kursieren Zahlen von 20.000 oder gar 35.000 Entscheidungen pro Tag, auch wenn diese Zahlen nicht wissenschaftlich belegt werden (Hoomans, 2015; Utikal, 2020).

Befolgung eines Ratschlags beschrieben, obwohl dieser das Ergebnis verbessert hätte (Yaniv & Kleinberger, 2000). Eine prominente Erklärung für Advice Discounting basiert auf der Annahme der Informationsasymmetrie. Danach haben Nutzende zwar Informationen über die eigenen Entscheidungsrechtfertigungen, nicht aber über die des Ratgebers. Informationen über Entscheidungsfindung oder -sicherheit können hilfreich sein, diese Asymmetrie auszugleichen (Pálfi et al., 2022; Yaniv, 2004; Yaniv & Kleinberger, 2000). Welche Informationen hierzu den Unterschied machen, wurde in zahlreichen Studien untersucht. Die Studien zu automatischen bzw. künstlichen Beratungssystemen verbindet die grundlegende Annahme, mehr Transparenz führe zu einem besseren und überlegteren Umgang mit den Systemen.

Während der Begriff DSS als Containerbegriff verschiedenste Bezeichnungen umfasst, die sich je nach Anwendungsfall unterscheiden, ist der Fokus der Entscheidungsunterstützung in der vorliegenden Arbeit nicht auf den Arbeitseinsatz, sondern auf alltägliche Entscheidungen gelegt. Für die hier diskutierten Systeme sollen drei Eigenschaften gelten, die schon in den 1980ern als für Entscheidungsunterstützungssysteme charakterisierend festgelegt wurden:

1. Sie sollen den Entscheidungsprozess **vereinfachen**.
2. Sie automatisieren keine Entscheidungen, sondern **unterstützen** den Prozess und
3. **verändern** sich passend an die Anforderungen der Nutzenden (Alter, 1980).

Die vorliegende Arbeit konzentriert sich auf Interaktionen und Systeme, bei denen der **Mensch als finale*r Entscheider*in** tätig ist. Es handelt sich bei der untersuchten KI also um Entscheidungsunterstützungssysteme. Jedoch werden solche Systeme betrachtet, die zur Unterstützung in **alltäglichen Entscheidungen** dienen. Dies folgt hauptsächlich dem Zweck, keine arbeitsrechtlichen, -organisationalen oder -systemischen Einflüsse berücksichtigen zu müssen, sondern Endnutzende zu untersuchen, die weder KI- noch Domänen-Expert*innen sind. Ein Bezug der Ergebnisse auf Arbeitskontexte könnte unter Berücksichtigung weiterer Einflussfaktoren im Anschluss an die vorliegende Arbeit stattfinden.

2.1.3. Zielgruppen von KI

Mit der Zunahme und Popularität von KI steigt auch die Zahl der an ihr beteiligten Gruppen. Waren früher nur KI-Expert*innen mit KI beschäftig und gleichzeitig auch die einzigen Nutzenden, hat sich das mit der großen Verbreitung von KI geändert. Die Zielgruppe spielt eine entscheidende Rolle bei der Untersuchung der Rahmenbedingungen und Anforderungen an KI-Systeme (Bertrand et al., 2022; Bhatt et al., 2020; Gerlings et al., 2022; van Nuenen et al., 2020). In der Forschung werden mehrere Zielgruppen unterschieden, wobei sich darunter Folgende identifizieren lassen (Bertrand et al., 2022;

Bhatt et al., 2020; Brasse et al., 2023; Herm et al., 2023; Hind et al., 2019; Larsson et al., 2019; Mohseni et al., 2021; van Nuenen et al., 2020):

- Entwickler*innen bzw. KI-Expert*innen
- Politiker*innen und Jurist*innen
- Datenexpert*innen
- Domänen-Expert*innen
- KI-Noviz*innen

KI-Expert*innen schreiben die Algorithmen, entwickeln ML-Modelle und wollen diese möglichst akkurat gestalten. Sie verstehen mathematische Prozesse und Abhängigkeiten. Ihr Hauptinteresse an transparenter KI liegt darin, ihre Systeme durch die Nachvollziehbarkeit verbessern zu können (Bertrand et al., 2022; Brasse et al., 2023; van Nuenen et al., 2020). **Politiker*innen und Jurist*innen** haben dagegen wenig technische Expertise, aber die Aufgabe, Prozesse zu verstehen, um sie in Vertretung von dritten Betroffenen regulieren und kontrollieren zu können. Sie sind also einerseits auf Zuarbeit und Gutachten von KI-Expert*innen angewiesen, aber müssen andererseits als Übersetzer*innen zwischen ethischen Regulierungen und mathematisch-, KI-basierter Fairness fungieren (Hind et al., 2019; Larsson, 2019; van Nuenen et al., 2020). Zu dieser Kategorie lassen sich auch „Examiner" zuordenen, also Personen, die KI nach vorgegebenen Kriterien untersuchen und beispielsweise auditieren. Hier ist der Übergang zu KI-Expert*innen fließend (Cambria et al., 2023).

Datenexpert*innen hingegen wenden KI-Systeme in der Forschung oder Industrie an. Entsprechend kennen sie sich nicht mit den technischen Hintergründen der Systeme aus, haben aber ein Interesse daran, die Modell-Daten-Passung zu verbessern, um so exaktere Aussagen zu generieren (Mohseni et al., 2021). Ihre Abgrenzung zu **Domänen-Expert*innen** ist fließend, da sich unter Domänen-Expert*innen auch solche aus dem Datenbereich befinden können. Gleichermaßen sind unter Domänen-Expert*innen aber auch Expert*innen anderer Bereiche zu finden, wie z. B. Prozess-Optimierer*innen oder HR-Verantwortliche, die sich gut mit ihrer Domäne auskennen und beispielsweise (sehr) falsche Ergebnisse erkennen, die aber üblicherweise nicht mit KI-Entwicklung, -Verbesserung oder -Implementierung betraut sind. Domänen-Expert*innen sind häufig Endnutzende von KI im Arbeitskontext (Bertrand et al., 2022; Brasse et al., 2023; Herm et al., 2023). Entsprechend ist auch hier der Übergang zu **KI-Noviz*innen** fließend, wobei mit dieser Nutzendengruppe eher Endnutzende aus privater Nutzung bezeichnet werden. Sie haben besonders wenig Wissen über Technik und KI, wissen teilweise nicht einmal, dass sie ein KI-System nutzen. Für sie sind Nutzen und Bedienbarkeit von höchster Bedeutung (Mohseni et al., 2021). In der Gegenüberstellung mit Expert*innen wird diese Nutzendengruppe auch häufig als Laien bezeichnet (Ribes et al., 2021; Shulner-Tal et al., 2023).

Die vorliegende Arbeit befasst sich mit den Auswirkungen transparenter KI auf die Nutzung der Systeme durch Endnutzende. Dabei steht kein expliziter Arbeitskontext im Fokus, also keine spezielle Domänen-Expertise, vielmehr wurde bei Auswahl und Design der Studien darauf geachtet, Expertise als Faktor weitestgehend auszuschließen. Die Studien befassen sich also mit Endnutzenden, die in erster Linie **KI-Noviz*innen bzw. Laien** sind.

2.1.4. Algorithm Aversion

Betrachtet man die Nutzung von KI-Systemen durch Laien, ergeben sich zwei Tendenzen. Einerseits zeigt sich zu großes Vertrauen, Overtrust, in die Systeme und eine mangelnde kritische Auseinandersetzung mit ihnen: ein Effekt, der insbesondere bei Hypes einzelner Systeme wie beispielsweise ChatGPT deutlich wird (García et al., 2021; Grigutyté, 2023; Wang et al., 2023). Andererseits findet sich seit längerem der stabile Effekt der Algorithm Aversion, bei dem meist die Nutzung eines menschlichen mit einem algorithmischen Ratgeber verglichen wird. Während er in seinem ursprünglichen Verständnis noch die grundsätzliche Ablehnung von Algorithmen bezeichnete, wird er heute spezifischer verwendet (siehe Filiz et al., 2023, für eine Übersicht verschiedener Definitionen). Denn Studien zeigen inzwischen uneinheitliche Ergebnisse bezüglich der Ablehnung von KI: So kommt es bei der Nutzung von KI immer häufiger zu „Algorithm Appreciacion“ oder zumindest einer gleichhäufigen Nutzung von menschlichem und algorithmischen Rat. Insbesondere bei objektiven oder utilitaristischen Fragestellungen werden Algorithmen menschlichen Ratgebern vorgezogen (B. Berger et al., 2021; Logg et al., 2019).

Was über die allgemeine Ablehnung von KI hinaus heute häufig mit dem Begriff Algorithm Aversion bezeichnet wird, ist der Rückgang der Nutzung, der sich zeigt, nachdem Nutzende ein fehlerhaftes Ergebnis des Systems erfahren haben(Burton et al., 2020; Dietvorst et al., 2014; Mahmud et al., 2022). Besonders problematisch ist: Algorithm Aversion zeigt sich auch, wenn das Ablehnen der KI-Unterstützung zu insgesamt schlechteren Urteilen führt, wie wenn Nutzende allein oder mithilfe eines menschlichen Ratgebers entscheiden. Außerdem scheint der Effekt bei wichtigen Entscheidungen zuzunehmen (Filiz et al., 2023). Der Effekt, den Algorithmus wegen eines vorherigen Fehlers abzulehnen, führt also zu schlechteren Ergebnissen. Aus diesem Grund stellt Algorithm Aversion eine Gefahr für gutes Entscheiden dar und Maßnahmen zur Überwindung sind vonnöten.

Zahlreiche Studien zeigen Maßnahmen, die zur Überwindung von Algorithm Aversion ergriffen werden können. Haben Versuchspersonen die Möglichkeit, präsentierte Algorithmen-Vorschläge zu verändern, nutzen sie einen Algorithmus nach einem Fehler mehr und sind mit dem Ergebnis zufriedener, im Vergleich dazu, wenn sie ihn nicht anpassen dürfen und seinen Vorschlag einfach übernehmen müssen. Diese Kontrolle, die die Versuchspersonen gewinnen, führte insgesamt zu besseren Entscheidungen als ohne Algorithmennutzung (Dietvorst et al., 2018). In einer Studie von

Berger et al. zeigte sich eine geringere Algorithm Aversion nach einer Fehlererfahrung, wenn die Versuchspersonen im Anschluss an den Fehler eine Verbesserung des Algorithmus beobachten konnten; wenn der Algorithmus durch den Fehler also dazulernte (B. Berger et al., 2021). Damit wurden die Potentiale heutiger KI-Systeme in der Studie berücksichtigt (Lohoff & Rühr, 2021). Tatsächlich wird als ein Grund für das Auftreten von Algorithm Aversion die Annahme der Nutzenden aufgeführt, Algorithmen würden im Gegensatz zu Menschen nicht aus ihren Fehlern lernen (Reich et al., 2022). Nutzende nähmen bei einem Fehler an, „that a single mistake signals that the algorithm is irrevocably flawed or simply broken" (Reich et al., 2022, S. 2). Diese Annahme ins rechte Licht zu rücken, bietet also einen Ansatz zur Überwindung von Algorithm Aversion. Eine Möglichkeit hierfür, so postuliert die vorliegende Arbeit, stellt Information über die Algorithmusakkuratheit dar. Transparenz darüber, dass sich KI irren kann, weil sie keine 100 %, sondern möglicherweise nur 80 % Sicherheit bietet, könnte die falsche Annahme, ein einziger Fehler offenbare ein defektes System, regulieren.

In einer weiteren Studie, die die Nutzung von Algorithmenhinweisen bei Schachzügen untersuchte, zeigten Versuchspersonen hohes initiales Vertrauen in den Algorithmus, das bei einem Fehler rapide abbaute und anschließend nur langsam wieder zunahm. Jedoch beschränkte sich das verlorene Vertrauen nicht nur auf den Algorithmus, sondern die Versuchspersonen wurden auch selbst unsicherer und verloren Vertrauen in ihre eigene Leistung (Chong et al., 2022). Vorangehende Information über die Akkuratheit des Algorithmus könnte also nicht nur das (möglicherweise) überhöhte initiale Vertrauen in den Algorithmus, sondern auch die Verunsicherung der Versuchspersonen regulieren.

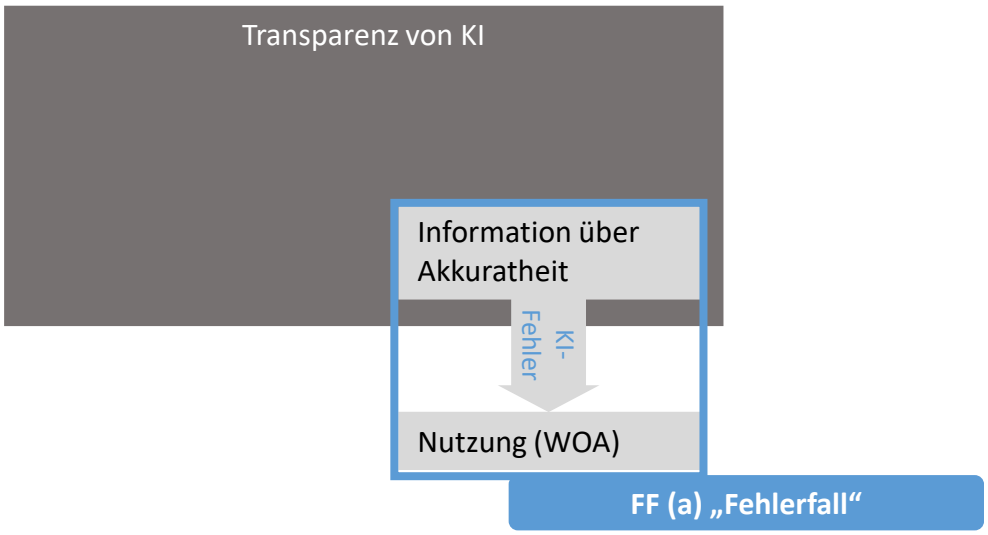
Die emotionalen Prozesse hinter Algorithm Aversion untersuchte eine Studie, in der verschiedene hypothetische Situationen einer Interaktion mit Algorithmen zu bewerten waren. Die algorithmischen Fehlerfälle lösten intensivere negative Gefühle und Verhaltensintentionen aus als menschliche Fehler, führten jedoch zu weniger Schuldzuschreibungen als bei menschlichen Fehlern (Renier et al., 2021). Die Autoren ziehen daraus die Schlussfolgerung, Fehler würden als menschlich, nicht algorithmisch angesehen. Diese Annahme stützt, dass „participants with the human advisor framing initially expected a significantly lower forecasting accuracy (mean = 10.91% absolute error, n = 61, SD = 11.62) compared to the algorithmic advisor framing (mean = 8.46% absolute error, n = 65, SD = 13.07)" (Daschner & Obermaier, 2022, S. 91).

Diese Übererwartung an KI, also die Annahme, KI und Algorithmen arbeiteten fehlerfrei, dient paradoxerweise als eine weitere Erklärung für Algorithm Aversion (Dzindolet et al., 2003; Prahl & Swol, 2017). Bei einer Verletzung dieser Makellosigkeitserwartung sind Nutzende abgeschreckt und glauben, das System sei gänzlich disfunktional. Dies ist deshalb problematisch, da KI die Realität nur in Modellen nachbilden kann und daher irgendwann zwangläufig Fehler produziert. Folgt man der Argumentation

der Makellosigkeitserwartung, könnte also Information über die tatsächlich eingeschränkte Akkuratheit von Algorithmen der Überwindung von Algorithm Aversion dienen. In einem ausführlichen Review zum Effekt argumentieren auch Burton und Kolleg*innen, ein Ansatz zur Überwindung von Algorithm Aversion sei „Algorithm Literacy“, also Wissen über Funktionsweise, Einsatz und Grenzen von Algorithmen (Burton et al., 2020). Es stellt sich die Frage, inwiefern eine Akkuratheitsangabe die Übererwartungen mäßigen und dadurch Algorithm Aversion abschwächen kann. Wie Abbildung 2 schematisch dargestellt, gilt es also zu klären:

> **Forschungsfrage (a) „Fehlerfall“: Inwieweit führen Angaben von Akkuratheit eines Algorithmus dazu, dass dieser auch nach einem Fehlerfall genutzt wird?**

***Abbildung 2**: Schematische Darstellung von Forschungsfrage (a) im Rahmen der Gesamtarbeit; der Weight of Advice (WOA) ist dabei die Operationalisierung von Nutzung.*



In einer Studie untersuchten Daschner und Obermaier (2022) den Effekt der Akkuratheit relativ zur Teilnehmenden-Akkuratheit. Sie gaben in mehreren Runden Rückmeldung zur prozentualen Akkuratheit der Versuchspersonen und manipulierten die Akkuratheit eines Beratungssystems (vs. eines menschlichen Ratgebers) als besser, gleich oder schlechter als die der Versuchspersonen. Tatsächlich ergab sich keine Reduzierung der Algorithmennutzung, wenn deren Akkuratheit besser oder gleich war als die der Nutzenden, sondern nur, wenn der Algorithmus eine geringere Akkuratheit aufwies als die Nutzenden (Daschner & Obermaier, 2022). Ebenso zeigte eine Studie „stated accuracy does have a significant effect” sowohl in der Nutzung der Model-Ratschläge als auch im berichteten Vertrauen (Yin et al., 2019, S. 2). Gleichzeitig untersuchten die beiden Studien nur den Effekt einer vorgegebenen Akkuratheitsangabe auf die generelle Nutzung eines Algorithmus, nicht auf Algorithm Aversion, wie sie heutzutage verstanden wird, also nach einem Fehlerfall. Ob die in den Studien ermittelten positiven Effekte auch im Fehlerfall wirken, wird in der entsprechenden Studie in Kapitel 4 untersucht.

Zunächst folgen jedoch die Herleitung des Begriffs Transparenz, seine geschichtliche Einordnung sowie seine Bedeutung im Zusammenhang mit KI im nachfolgenden Kapitel. Die vorliegende Arbeit versteht Akkuratheit als offenzulegende Information über KI und damit als Bestandteil von Transparenz. Im Rahmen des Kapitels zu Akkuratheit 2.2.3 wird jedoch auch die gegenteilige Sicht von Akkuratheit als Gegenspieler von Transparenz erläutert.

2.2. Transparenz

Dieses Kapitel nimmt eine Einordnung des Begriffs Transparenz vor, der sich von einem politischen Verständnis in ein vielfach genutztes, idealisiertes Konzept gewandelt hat, das schwer zu fassen ist. Diese Entwicklung des Begriffs und aktuelle politische Auseinandersetzungen mit KI-Transparenz finden sich in Kapitel 2.2.1. Anschließend gibt Kapitel 2.2.2 eine Übersicht über verschiedene Konzepte, die dem Oberbegriff Transparenz häufig zugeordnet werden, und ordnet sie ein.

Darauf folgt in Kapitel 2.2.3 die Auseinandersetzung mit dem besonderen Ansatz der Akkuratheit, die, wie gerade beschrieben, sowohl als Gegenspieler als auch als Bestandteil von Transparenz angesehen werden kann. Entwickler*innen, die im Bereich Explainability arbeiten, postulieren, Transparenz erhöhe Vertrauen, Akzeptanz und Nutzung von KI und helfe Endnutzenden, KI-Skepsis zu überwinden. Welche technischen Ansätze zu KI-Transparenz bestehen und mit welchen Ansätzen diese aktuell hergestellt wird, beschreibt Kapitel 2.2.4. Die Erläuterung von Transparenz schließt mit der Darlegung des Transparenzverständnisses in der vorliegenden Arbeit (Kapitel 2.2.5).

2.2.1. Geschichte des Begriffs

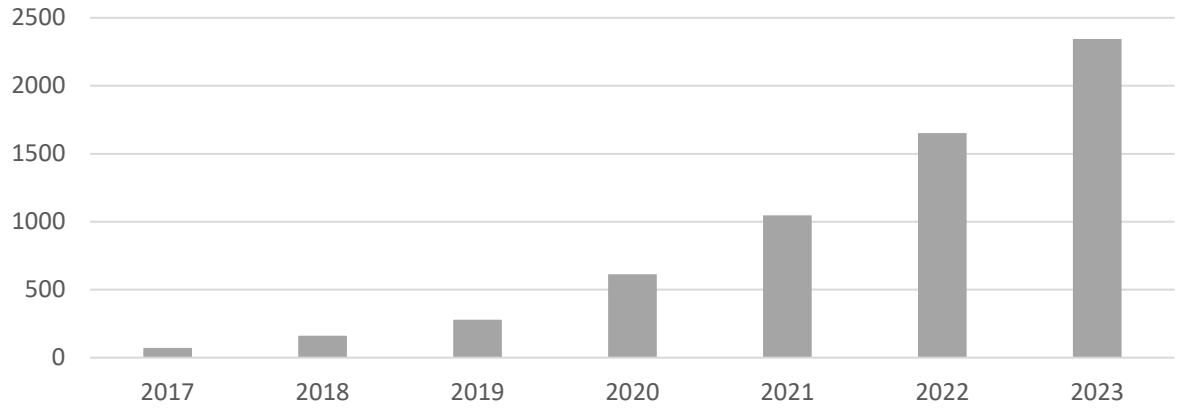
Als der Begriff Transparenz in den 1990er Jahren an Popularität gewann, bezeichnete er in erster Linie ein wirtschaftspolitisches Verständnis (Forssbaeck & Oxelheim, 2014; Larsson & Heintz, 2020) und stellte „a public value embraced by society to counter corruption“ dar (Ball, 2009, S. 293). Im weiteren Verlauf wurde Transparenz häufig mit Begriffen wie Offenheit in Verbindung gebracht und einem politisch-wirtschaftlichen Verständnis von Transparenz, beispielsweise Offenheit in Prozessen oder Politik (Felzmann et al., 2019; Flyverbom, 2016; Fung et al., 2007). In Bezug auf Technologie steht Offenheit für positive Attribute wie „offene Daten“ oder „Open Source“ (Larsson & Heintz, 2020). Gleichzeitig birgt der Begriff der Offenheit im Sinne von Transparenz auch ein Paradox, bei dem „a genuinely well-intended discourse of openness may lead to unintended consequences“ (Larsson & Heintz 2020, S. 6). Als Gefahr sehen sie den Missbrauch von offenen Daten oder die – absichtliche oder unabsichtliche – Verwirrung von Nutzenden durch die Veröffentlichung von zu viel und unstrukturierter Information. Viele Forschende äußern Kritik an der Vorstellung, sämtliche Komponenten eines Algorithmus öffentlich preiszugeben, da dadurch sensible Daten freigelegt oder Unternehmensinformationen von Wettbewerbern ausgenutzt werden könnten (de Laat, 2018).

Ein weiterer Bestandteil von Transparenz, insbesondere im sozialwissenschaftlichen Diskurs, ist die Rechenschaftspflicht (accountability). Dieser Begriff ist schwer zu fassen, da er sowohl einen Aspekt als auch ein Ergebnis von Transparenz bezeichnet (Ananny & Crawford, 2018; Felzmann et al., 2020; Wieringa, 2020). Rechenschaftspflicht ist insofern mit Transparenz verknüpft, „that any observation into a system's logic provides insight, and this insight creates knowledge, which in turn is a precondition for holding systems accountable" (Felzmann et al., 2020, S. 3338). In dieser Hinsicht ist Transparenz aus sozialrechtlicher Perspektive ein Mittel, „for keeping society ‚in-the-loop'", wie Larsson argumentiert (2019, S. 574). Gleichzeitig kann die Rechenschaftspflicht auch das Ergebnis eines transparenten Systems sein, so dass sich Nutzende in der Kontrolle und verantwortlich für das System fühlen (Rader et al., 2018). In Bezug auf KI ist das Thema Rechenschaftspflicht häufig eines, das unter juristischen, ethischen und philosophischen Gesichtspunkten betrachtet wird: Wenn KI autonom handelt, wer ist dann für ihr Handeln verantwortlich? Und kann KI oder Technologie überhaupt für ihr Handeln verantwortlich sein?

Aufgrund der Herausforderungen in Bezug auf die Rechenschaftspflicht betont Flyverbom (2016), dass Transparenz als eine „form of visibility management" (S. 110) betrachtet werden sollte, das, „like other visibility practices […] involves decisions about what to disclose and to whom and what to keep out of sight" (S. 112). Nach Stohl et al. (2016) besteht diese Sichtbarkeit aus drei Attributen: „availability of information, approval to disseminate information, and accessibility of information to third parties" (S. 123). Obwohl grundsätzlich dann Transparenz angenommen wird, wenn alle drei Attribute erfüllt sind, betonen Stohl und ihre Kollegen ein daraus resultierendes „transparency paradox": „when there is an abundance of information available, it is often difficult to obtain useful, relevant information." (S. 134). Daher sollten weder Sichtbarkeit noch Transparenz das ultimative Ziel sein, sondern Sichtbarkeit gesteuert werden, um die effektive Nutzung von Informationen zu verbessern.

Mit der Digitalisierung und zuletzt insbesondere der Zunahme technischer Systeme, erweiterte sich das Verständnis des Begriffs Transparenz zu einer Voraussetzung für einen ethischen und verantwortungsvollen Umgang mit Daten (Larsson & Heintz, 2020). Ab den 2010er Jahren nahm die Zahl der Studien zu ethischer KI und maschinellem Lernen massiv zu und Transparenz in der KI sowie die Forschung zu XAI gewannen an Popularität (Larsson et al., 2019; siehe Abbildung 3).

***Abbildung 3**: Die Zahl der Veröffentlichungen auf EBSCOhost zu „explainability" oder „interpretability" zusammen mit „ai" oder „artificial intelligence", abgerufen über scopus für die Jahre 2017 bis 2023*



Anmerkung. Stand Juni 2024.

Definitionen von transparenter KI umfassen KI-Mechanismen und ihre zugrundeliegende Logik bis hin zur Möglichkeit, Einblicke in die Blackbox zu erhalten, Systeme zu verbessern, Rechenschaftspflicht herzustellen und Diskriminierung zu verhindern (Ananny & Crawford, 2018). So wurde Transparenz zu einem „modern, surprisingly complex […] ideal" (Koivisto, 2016, S. 2): Transparenz künstlicher Intelligenz gilt als Vorrausetzung für menschliche Entscheidungsautonomie und damit als herzustellenden Zielzustand, um ethische Anforderungen – und neu aufgestellte rechtliche Rahmenbedingungen – zu erfüllen (AI HLEG, 2019; Jobin et al., 2019; Digital Services Act, 2022). Ebenso vertreten immer mehr Forschende und Praktiker*innen die Annahme, auch Unternehmen gingen bei der Nutzung intransparenter Algorithmen ein Risiko ein. Sie könnten dann nicht nachvollziehen, welche Entscheidungen ein System tatsächlich trifft, inwiefern es Verzerrungen ausschließt und welchen zugrundeliegenden Interessen es dient (Diakopoulos, 2016; Rudin, 2019). Gleichzeitig und parallel zur Annahme, politische Transparenz führe zu mehr Vertrauen in Parlamente und Demokratie, wird argumentiert, Transparenz in KI stärke das Vertrauen von Nutzenden in KI-Systeme und damit ihre Akzeptanz und Nutzung (Felzmann et al., 2019; Jobin et al., 2019; Shulner-Tal et al., 2023). Ergänzend dazu soll KI-Transparenz als Mittel dienen, Informationen über Algorithmen zu erhalten, die „monitoring, checking, criticism, or intervention by interested parties" ermöglichen (Diakopoulos & Koliska, 2017, S. 811). All diese Perspektiven erklären jedoch nicht, was Transparenz „means, to whom it related, and to what extent it is beneficial" (Felzmann et al., 2020, S. 3336). Der Begriff und das Verständnis von Transparenz bleiben also „quite malleable and therefore […] can mean all things to all people" (Fox, 2007, S. 664). Das zeigt sich auch daran, dass Transparenz mit weiteren Konzepten wie „explainability, interpretability, openness, accessibility, and visibility, among others" in Verbindung gebracht wird (Felzmann et al., 2020, S. 3335).

Dabei hat sich die Forderung nach Transparenz als Anforderung an KI seit den 1990er und 2000er Jahren maßgeblich verschärft. Denn mit der gestiegenen Komplexität von KI und den damit

verbundenen Modellen – insbesondere Deep Learning und dadurch entstehende Deep Neural Networks – sind Blackbox-Systeme entstanden, die von sich aus keine Erklärungen liefern oder ihre innenliegenden Prozesse nachvollziehbar machen (können) (Arrieta et al., 2020). In ihrer ersten Veröffentlichung im Jahr 2019 kam die Hochrangige Expertengruppe für künstliche Intelligenz der EU-Kommission (AI High Level Expert Group: AI HLEG) zu dem Schluss, „those [legal regimes] regarding transparency, traceability and human oversight are not specifically covered under current legislation“ (AI HLEG, 2019, S. 9). Im Zuge ihrer Arbeit entwickelte die AI HLEG eine Bewertungsliste für vertrauenswürdige KI, die sieben Anforderungen umfasst:

1. Human Agency and Oversight,
2. Technical Robustness and Safety,
3. Privacy and Data Governance,
4. Transparency,
5. Diversity, Non-discrimination and Fairness,
6. Societal and Environmental Well-being und
7. Accountability (AI HLEG, 2020).

Transparenz umfasst drei Elemente: “1) traceability, 2) explainability and 3) open communication about the limitations” (S. 14). **Traceability** bezeichnet die Anforderung, die Herkunft und Verarbeitung der Daten und KI-Modelle sowie Ergebnisse zu dokumentieren und nachvollziehbar zu machen. **Explainability** umfasst die benötigte Erklärbarkeit, die gegenüber Nutzenden des Systems hergestellt und regelmäßig geprüft werden soll. Eine **offene Kommunikation** bedeutet, Nutzenden mitzuteilen, *dass* sie mit einem System interagieren, seine Fähigkeiten, Risiken und auch die Akkuratheit des Systems zu kommunizieren sowie adäquates Trainingsmaterial zu seiner Nutzung bereitzustellen.

Aufbauend auf den Vorarbeiten der AI HLEG wurden seit 2020 verschiedene politische Initiativen der Europäischen Union und Kommission angestoßen. Ein Ergebnis der Arbeit ist der Digital Service Act, der mehr Sicherheit für Nutzende von Online-Diensten gewährleisten und die Durchsetzung von Grundrechten möglich machen soll. Im ersten Schritt wurden sehr große Online-Plattformen verpflichtet, verschiedene Nutzungszahlen offenzulegen und eine Risikoeinschätzung vorzunehmen (Digital Services Act, 2022). Parallel wurde seit 2021 an einem Gesetz für die Regulierung von KI gearbeitet, das 2024 vom Europäischen Rat angenommen wurde. Dieses unterscheidet zwischen KI-Systemen mit hochrisikoreichem, risikoreichem und allgemeinem Verwendungszweck. Abhängig davon ergeben sich unterschiedlich strenge Regularien und Meldepflichten. Generell hinaus müssen KI-Systeme, die mit Menschen interagieren oder zur Erkennung von Emotionen oder biometrischer Kategorisierung eingesetzt werden, sich als solche zu erkennen geben. Ebenso müssen künstlich erstellte Medieninhalte oder Texte als solche gekennzeichnet werden (Verordnung über künstliche

Intelligenz, 2024; Artikel 50). Während KI für besonders gefährliche Zwecke, z. B. zur Täuschung oder Manipulation, zum sozialen Profiling oder zur biometrischen Kategorisierung von Personen oder Gruppen, verboten werden, gilt für risikoreiche KI-Dienste die Pflicht zur Registrierung in einer europäischen Datenbank und der Angabe umfangreicher Informationen. Für alle Systeme gilt das Recht „auf Erläuterung der Entscheidungsfindung im Einzelfall“ (Artikel 86). Darüber hinaus verweist das Gesetz auf die oben genannten Veröffentlichungen der AI HLEG zu vertrauenswürdiger KI:

> „Transparenz bedeutet, dass KI-Systeme so entwickelt und verwendet werden, dass sie angemessen nachvollziehbar und erklärbar sind, wobei den Menschen bewusst gemacht werden muss, dass sie mit einem KI-System kommunizieren oder interagieren, und dass die Betreiber ordnungsgemäß über die Fähigkeiten und Grenzen des KI-Systems informieren und die betroffenen Personen über ihre Rechte in Kenntnis setzen müssen. [….] Die Anwendung dieser Grundsätze sollte, soweit möglich, in die Gestaltung und Verwendung von KI-Modellen einfließen.“ (Verordnung über künstliche Intelligenz, 2024; Absatz 27)

Obwohl die bestehenden Anforderungen zunehmen, spezifizieren sie jedoch selten, was unter Transparenz zu verstehen ist. Im Rahmen des neuen EU-Gesetzes werden diverse EU-Institutionen zur Prüfung von KI eingesetzt. Auch enthält es obige Definition von Transparenz, die für Endnutzende zu gelten hat. Was aber „ordnungsgemäß […] in Kenntnis setzen“ bedeutet, bleibt offen. Ähnliches trifft auch auf die Allgemeine Datenschutzgrundverordnung (DSGVO) zu, wie Felzmann et al. feststellen, die das „level of detail [of] a ‚meaningful‘ explanation“ vermissen lasse (2019, S. 3). Und auch der 2020 ratifizierte deutsche Medienstaatsvertrag, der nun auch Medienplattformen wie Google News oder YouTube betrifft, fordert in Paragrafen 85 von den Anbietern Transparenz. Dies beinhaltet die Frage, warum wer welche Inhalte angezeigt bekommt oder in welcher Reihenfolge (Medienstaatsvertrag, 2020). Doch wie geprüft wird, ob diese Inhalte verständlich sind oder wie sie präsentiert werden müssen, bleibt offen. In der Forschung zu KI-Transparenz und insbesondere bei der KI-Entwicklung gestaltet es sich deshalb schwierig, die erforderlichen Anforderungen zu berücksichtigen, selbst wenn dies angestrebt wird (Felzmann et al., 2019; Larsson et al., 2019; Larsson & Heintz, 2020).

Die Frage, was eine sinnvolle Erklärung darstellt, gilt es also zu beantworten, um sie bei Design, Entwicklung und Einsatz von KI berücksichtigen zu können. Eine Auseinandersetzung mit den Begriffen und der bereits bestehenden Forschung aus dem Bereich Transparente KI findet in den folgenden Kapiteln statt.

2.2.2. Arten von Transparenz in KI

Transparenz in KI bezeichnet die Möglichkeit, zu verstehen und zu erklären, wie KI-Modelle Entscheidungen treffen. Unter dem Begriff KI-Transparenz existieren vielfältige Konzepte, die auf

unterschiedliche Weisen untersucht und umgesetzt und häufig austauschbar genutzt werden. In ihrer mehrere Disziplinen umfassenden Übersicht strukturieren Mohseni und ihre Kollegen (2021) die wichtigsten dieser Begriffe: Sie unterscheiden zwischen den Konzepten „Intelligible Systems“ (verständliche Systeme) und „Transparent AI“ (transparente KI), denen sie verschiedene Ansätze zuordnen (siehe Tabelle 1).

***Tabelle 1**: Übersicht und Klassifizierung verschiedener Konzepte von transparenter KI bzw. verständlichen Systemen*

Intelligible Systems (verständliche Systeme)	**Hauptkonzept**
Verständlichkeit (Understandability/intelligibility)	Erwünschte Eigenschaften
Vorhersagbarkeit (Predictability)	
Vertrauenswürdigkeit (Trustworthyness)	Erwünschtes Ergebnis
Verlässlichkeit (Reliability)	
Sicherheit (Safety)	
Transparente KI (Transparent AI)	**Hauptkonzept**
Interpretierbare KI (Interpretable AI)	Praktische Ansätze
Erklärbare KI (Explainable AI)	
Interpretierbarkeit (Interpretability)	Erwünschte Eigenschaften
Erklärbarkeit (Explainability)	
Rechenschaftspflichtige KI (Accountable AI)	Erwünschtes Ergebnis
Faire KI (Fair AI)	

Anmerkung. Tabelle nach Mohseni et al., 2021, S. 8, eigene Übersetzung.

Die Autor*innen unterscheiden zunächst die beiden Hauptkonzepte „transparente KI“ und „verständliche Systeme“. Dabei stellt transparente KI die KI-basierte Klasse von **verständlichen Systemen** dar und damit die technische Umsetzung von diesen. **Verständlichkeit** und **Vorhersagbarkeit** sind zwei Unterkategorien von verständlichen Systemen, die in psychologischer und nutzendenzentrierter Forschung besonders von Interesse sind. Es geht dabei um durch Nutzende wahrgenommene Eigenschaften, an denen ein verständliches System gemessen werden kann. **Transparente KI** hingegen lässt sich durch die technisch verstandenen Ansätze **„Interpretierbarkeit“** und **„Erklärbarkeit“** beschreiben. Der Weg zu diesen Ansätzen führt über **interpretierbare und erklärbare KI**. Interpretierbare KI und erklärbare KI sind somit erforderlich, um **rechenschaftspflichtige und faire KI** zu erreichen. Um eine KI zu entwickeln, die als verständliches System gilt, sind sie außerdem nötig, um die Eigenschaften des Systems – **Verständlichkeit** und **Vorhersagbarkeit** – herzustellen, durch die das erwünschte Ergebnis – **Vertrauenswürdigkeit**, **Verlässlichkeit** und **Sicherheit** – geschaffen werden kann. Vor diesem Hintergrund betonen die Autor*innen, was auch für diese Arbeit zentral ist: Durch die Hinzunahme von KI ergäben sich auch für die Kategorien der

verständlichen Systeme neue Fragen, „that were not necessarily problematic in intelligible rule-based systems but now require closer attention from research communities“ (Mohseni et al., 2021, S. 7).

Einerseits ermöglicht diese Unterteilung von Mohseni et al. eine Unterteilung in subjektive Systemeigenschaften, an denen ein übergeordnetes verständliches System gemessen werden kann, sowie die Unterordnung von transparenter KI und ihrer technisch zu verstehenden Konzepte. Andererseits ist diese, ebenso wenig wie andere Einteilungen von KI-Transparenz nicht verbreitet. Andere Autor*innen nutzen und verstehen die Begrifflichkeiten unterschiedlich oder sehr viel weniger systematisch. So fordert beispielsweise die Europäische Union im Digital Service Act von Online-Medienanbietern, diese haben ein „adequate level of transparency and accountability“ herzustellen (Digital Services Act, 2022, Abs. 49). Auch gelten die Kriterien FAT – fair, accountable, transparent – im politisch-rechtlichen Zusammenhang häufig als allgemeine Anforderungen an KI (Larsson, 2019).

In einer weiteren häufig vorgenommenen Unterscheidung bezeichnet Interpretierbarkeit die Art der Transparenz, die von sich aus vorhanden ist, weil die Prozesse und Entscheidungen eines Systems nachvollzogen werden können. Die Sprache ist dabei oftmals von opaken Systemen: Entscheidungsregeln sind nachvollziehbar, verständliche Entscheidungsbäume liegen vor (Arrieta et al., 2020; Mohseni et al., 2021). Entsprechend bezeichnet Erklärbarkeit/Explainability eine nachträglich hinzugefügte Transparenz, weil Systeme von sich aus nicht erklärbar sind und erst durch nachträglich ergänzte Modelle (Post-Hoc Explainability) erklärbar werden. Dies ist bei komplexen Systemen wie Deep Neural Networks nötig (Herm et al., 2023; Mohseni et al., 2021; Murdoch et al., 2019; Rudin, 2019). Auch wenn diese Unterscheidung sehr prominent ist, ist sie doch in erster Linie eine technische, die lediglich das System betrachtet (weitere technische Ansätze zu transparenter KI finden sich in Kapitel 2.2.4). Verschiedene Autor*innen nutzen die Begriffe auf andere Weise und bezeichnen beispielweise Interpretability als Zustand des Systems und Explainability, abhängig vom jeweiligen Empfänger (Brasse et al., 2023; Molnar, 2019). Aber auch hier herrscht Uneinigkeit über die genauen Begrifflichkeiten.

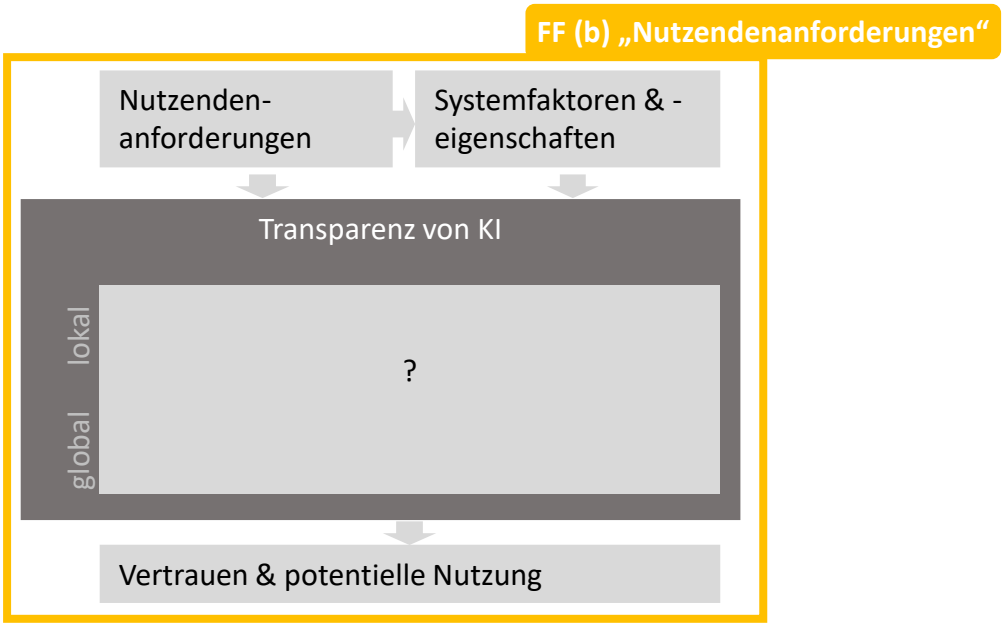
Eine weitere Argumentationslinie entfernt sich ganz von technischen Auseinandersetzungen und fordert, Transparenz ausgehend von den Nutzenden zu definieren: Verschiedene Zielgruppen haben unterschiedliche Ansprüche, müssen entsprechend unterschieden und abhängig davon Transparenz gewährleistet werden (Gerlings et al., 2022; van Nuenen et al., 2020). So verstehen Entwickler*innen, Politik, Jurist*innen und Endnutzende verschiedene Dinge unter transparenter KI und haben verschiedene Bedarfe nach und an Transparenz. Van Nuenen und Kollegen betonen: „The goal of transparency, we should not forget, is human understanding“ (2020, S. 7). Demzufolge bedarf es neben einer technischen und algorithmischen Transparenz für Endnutzende also anderer Informationen, um Verständnis zu fördern, Vertrauen herzustellen und am Ende eine informierte Nutzung zu ermöglichen.

Angesichts der Unklarheit, welche Informationen eine solche Transparenz umfassen sollte und der dadurch unklaren Anforderungen, die Endnutzenden an KI-Transparenz stellen, gilt für die vorliegende Arbeit zu klären (siehe Abbildung 4):

> **Forschungsfrage (b) „Nutzendenanforderungen“: Welche Anforderungen an Transparenz in KI für Endnutzende bestehen und inwiefern unterscheiden sich diese nach Eigenschaften der KI?**

Als Eigenschaften der KI werden dabei diejenigen Systemfaktoren bezeichnet, durch die sich ein System auszeichnet, also beispielsweise der Anwendungskontext oder die Relevanz, die mögliche Fehler des Systems auf das Leben der Nutzenden haben. Im Rahmen einer qualitativen Studie wird untersucht, welche Anforderungen für die Transparenz von KI erfüllt sein müssen, damit Endnutzende Vertrauen in die KI haben und bereit sind, sie zu nutzen. Es wird auch analysiert, inwieweit verschiedene Eigenschaften des Systems diese Anforderungen beeinflussen.

Abbildung 4: *Schematische Darstellung der Forschungsfrage (b) im Rahmen der Gesamtarbeit; die Unterscheidung von globaler und lokaler künstlicher Intelligenz (KI) folgt in Kapitel 2.2.4.*



Zunächst wird jedoch auf einen speziellen Fall von Transparenz eingegangen. Akkuratheit wird sowohl als Gegenspieler als auch als Teil von Transparenz verstanden und muss entsprechend näher erläutert werden.

2.2.3. Akkuratheit

Wie beschrieben, unterscheidet eine prominente Einteilung von KI-Transparenz Interpretierbarkeit und Erklärbarkeit. Dabei gilt: Einfachere ML-Modelle sind von sich aus opak und werden als interpretierbar bezeichnet. Komplexere KI-Systeme werden zu Blackboxen und erfordern nachträgliche Erklärungen, um sie transparent zu machen (Arrieta et al., 2020; Herm et al., 2023;

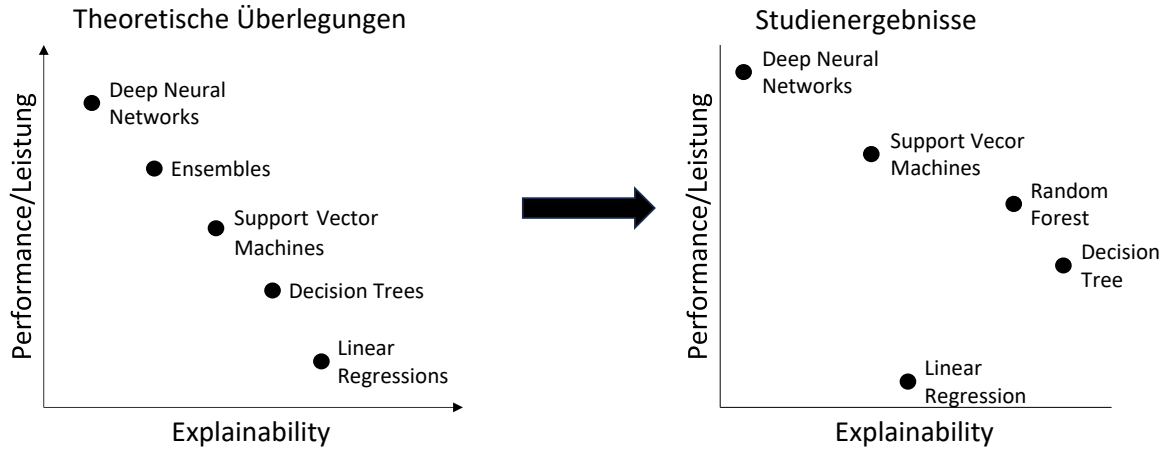
Mohseni et al., 2021; Murdoch et al., 2019; Rudin, 2019). Weshalb komplexere KI-Systeme den opaken, einfacheren Systemen vorgezogen werden, ist ihre genauere Abbildung der komplexen Welt, die zu besseren Vorhersagen oder Ergebnissen führt. Performance und Transparenz wirken demnach gegenläufig, es kommt zu einem Zielkonflikt zwischen möglichst guten und möglichst verständlichen KI-Modellen (Arrieta et al., 2020; Lipton, 2018; Murdoch et al., 2019; Rai, 2020; Rudin, 2019).

Manche Forscher*innen argumentieren, in dieser Abwägung sollte man sich für Erklärbarkeit und Interpretierbarkeit entscheiden, um sicherstellen zu können, dass die Modelle tatsächlich tun, was sie tun sollen (Rudin, 2019). Andere Autor*innen verweisen auf die Möglichkeit, hochkomplexe KI-Algorithmen durch nachträgliche Erklärungen transparent zu gestalten, weshalb „Linear models are not strictly more interpretable than deep neural networks. Despite this claim's enduring popularity, its truth value depends on which notion of interpretability is employed." (Lipton, 2018, S. 42) Zwar bezieht Lipton diese Aussage auf technische Ansätze der Transparenz wie Simulierbarkeit oder Zersetzbarkeit, er stellt aber auch die Frage in den Raum, inwiefern komplexe lineare Modelle, die als einfach geltenden Modelle, tatsächlich nachvollziehbar sind – insbesondere für Endnutzende (siehe auch Gosiewska et al., 2021; Lipton, 2018).

Herm et al. (2023) gingen dieser Frage in ihrer Untersuchung zum Zusammenhang von Explainability und Performance von KI-Systemen nach. Dazu setzten sie mehrere Machine Learning-Modelle ein, beginnend mit einfachen linearen Regressionsmodellen und endend bei Deep Neural Networks, die als komplexeste Klasse angesehen werden (siehe Abbildung 5). Diese Modelle verglichen sie in zwei medizinischen Anwendungen (Herzkrankheit- bzw. Hirntumor-Erkennung in Herz- bzw. Hirn-Scans) sowohl hinsichtlich der Akkuratheit ihrer Klassifizierungen als auch hinsichtlich der wahrgenommenen Güte der Erklärung durch Medizinstudierende. Und tatsächlich zeigen die Ergebnisse entlang der Hypothesen von Herm et al. keinen linearen Zusammenhang von Explainability und Performance (siehe Abbildung 5). Entgegen der Annahme der Autor*innen zeigen die Ergebnisse jedoch: Deep Neural Networks sind die Modelle mit der höchsten Performance und der geringsten Erklärbarkeit. Vergleicht man also KI mit deutlich einfacheren Modellen, bewahrheitet sich das Argument des Tradeoffs zwischen Performance und Erklärbarkeit.

Nachgelagerte Erklärungen ermöglichen es, diese Abwägung zwischen Performance und Erklärbarkeit zu überwinden: „deep neural networks are considered to be black boxes to the end users that are not interpretable by humans. They need to be augmented with XAI to offer any explainability to end users" (Herm et al., 2023, S. 11). Tatsächlich werden KI-Modelle, die durch Ergänzungen aus der „XAI-Toolbox" – in diesem Falle durch SHAP (Lundberg et al., 2020) – erklärbar gemacht werden, von den am wenigsten verständlichen Modellen zu den am besten verständlichen Modellen (Herm et al., 2021).

Abbildung 5: *Theoretischer Tradeoff zwischen Performance und wahrgenommener Erklärbarkeit von verschiedenen Machine Learning-Modellen und seine empirische Überprüfung.*



Anmerkung. Abbildung nach Herm et al., 2023, S. 9, eigene Übersetzung.

Betrachtet man also ausschließlich KI-Modelle, liefern diese im Vergleich zu einfacheren Modellen die bessere Akkuratheit, sind von sich aus nicht erklärbar und bedürfen deshalb zusätzlicher Maßnahmen, um Transparenz herzustellen. So ist es heutzutage „increasingly common for a decision-maker to be presented with a ‚black-box' model along with some measure of its performance – most often accuracy – on held-out data" (Yin et al., 2019, S. 2). In der Anforderungsliste der AI HLEG der Europäischen Kommission ist Akkuratheit einerseits unter dem Konzept „Technische Robustheit und Sicherheit" eine Anforderung an KI. Andererseits wird die Kommunikation über diese Akkuratheit dem Bereich „Transparenz" zugeordnet, weshalb eine Frage der Checkliste für vertrauenswürdige KI im Abschnitt Transparenz lautet: „Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates?" (AI HLEG, 2020, S. 15).

Wie Yin et al. betonen, wird eine solche Angabe der Akkuratheit heutzutage immer häufiger präsentiert (2019). Dies geht insbesondere auf die Tatsache zurück, dass die Ermittlung dieser Performance-Angabe ein Bestandteil der Erstellung der meisten Datenmodelle ist, bei einfachen Regressionen ebenso wie bei komplexen KI-Modellen (Goodfellow et al., 2016; M. Stone, 1974; Y. Xu & Goodacre, 2018). Indem das Modell auf einen dem Modell unbekannten Datensatz angewandt wird, wird die Passung des Modells ermittelt und vergleichbar gemacht (Goodfellow et al., 2016; Molnar, 2019). Dieser Prozess der Kreuzvalidierung nennt sich im Zusammenhang mit Machine Learning auch train-test-split, weil dazu vorhandene Daten in einen Trainings- und einen Testteil aufgeteilt werden. Der Prozess, die genaue Zusammensetzung und Größe des Testdatensets sowie die Messung der Passung sind abhängig von Daten und Anwendung. Entsprechend komplex ist die Wahl der genauen Berechnungsmethode. Je nach Methode unterscheidet sich das Akkuratsheitsmaß. Teilweise besteht es aus einem Wert zwischen 0 (keine Passung) und 1 (komplette Passung), ist eine logarithmische

Angabe oder eine Spanne wie ein Konfidenzintervall (z. B. S. Cramer et al., 2022; Y. Xu & Goodacre, 2018). Meist ist es aber eine Prozentangabe, also "the proportion of examples for which the model produces the correct output" (Goodfellow et al., 2016, S. 101–102).

Die Bewertung der Akkuratheit durch Nutzende von Systemen ist eine davon unabhängige, von weiteren Faktoren abhängige Frage. Ein linearer Zusammenhang zwischen „model performance"/Akkuratheit eines Systems und Vertrauen von Endnutzenden besteht dabei nicht unbedingt, wie auch Lipton betont (2018). Vielmehr zeigen Studien Differenzen zwischen tatsächlicher Akkuratheit und erlebter Akkuratheit (Yin et al., 2019). Und auch Effekte wie Algorithm Aversion demonstrieren, dass Nutzende die tatsächliche Güte eines Systems für ihre Nutzung häufig weniger in Betracht ziehen als andere subjektive Bewertungen. So betonen McNee et al. (2006), mehr als Akkuratheit sei nötig, um gute Entscheidungsunterstützungssysteme zu bauen, beispielsweise die Berücksichtigung von Anforderungen und Erwartungen der Nutzenden an die Systeme.

Zwar lässt sich von den technischen Gegebenheiten der KI-Modell-Erstellung ein Spannungsfeld öffnen zwischen optimierter Modell-Performance und hoher Erklärbarkeit eines Systems. Gleichzeitig ist die Kommunikation dieser Performance, beispielsweise als Akkuratheitsangabe, ein wichtiger Bestandteil algorithmischer Transparenz. Ziel der vorliegenden Arbeit ist es zum einen, die Anforderungen von Nutzenden gegenüber KI-Transparenz zu erheben, wobei auch die Kommunikation von Akkuratheit Betrachtung findet. Zum Zweiten spielen Akkuratheitsangaben als Bestandteil von Transparenz eine Rolle für die Nutzung von KI-Systemen. Welche technischen Ansätze zu transparenter KI über die Angabe von Akkuratheit hinaus bestehen, zeigt das folgende Kapitel auf.

2.2.4. Technische Ansätze zu transparenter KI

Zwei bereits angesprochene technische Konzepte transparenter KI sind Interpretability und Explainability. Wie in Kapitel 2.2.2 dargelegt, lässt sich eine Unterscheidung von Interpretability und Explainability anhand der Komplexität der Algorithmen bzw. KI-Modelle vornehmen: Interpretierbarkeit ist eine passive Eigenschaft und besteht aus einfacheren ML-Modellen, deren Prozesse von sich aus interpretierbar, also nachvollziehbar, sind (Rudin, 2019; Arrieta et al., 2020). Zu diesen einfacheren Modellen, denen grundsätzlich Interpretability zugeschrieben wird, zählen Regressionsmodelle aller Art, Entscheidungsbäume, auf Fuzzy-Rules basierende Systeme oder Kombinationen dieser Modelle (sogenannte Ensembles) (Ali et al., 2023; Arrieta et al., 2020; Gosiewska et al., 2021; Herm et al., 2023; Rudin, 2019).

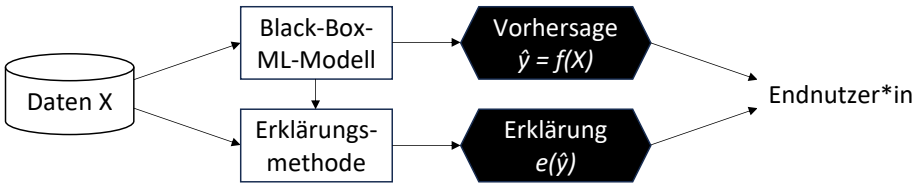
Dagegen bezeichnet Explainability eine herzustellende Eigenschaft, die meist durch nachgelagerte Post-Hoc-Erklärungen erreicht wird für ein System, das andernfalls nicht verständlich wäre (Miller 2018; Rai 2020). Explainability bezieht sich also auf die Darstellung der Funktionsweise komplexer KI-

Systeme, die ansonsten als Blackboxen betrachtet würden (Arrieta et al., 2020; Carvalho et al., 2019; Gunning et al., 2021; Miller, 2019; Mohseni et al., 2021; Rai, 2020). Diese Blackbox-Modelle sind vielschichtige Artificial oder Deep Neural Networks, also KI-Systeme, die aus mehreren Schichten von Entscheidungsknoten und selbstständig gebildeten Abhängigkeiten bestehen. Andere Forscher*innen unterscheiden die Begriffe Explainability und Interpretability allerdings auf andere Weise (z. B. Ali et al., 2023) oder verwenden sie austauschbar (z. B. Molnar, 2019; Murdoch et al., 2019).

Da sich die vorliegende Arbeit auf komplexe KI-Modelle konzentriert, die von sich aus nicht erklärbar sind, soll ein näherer Blick auf Explainability und ihre Methoden folgen. XAI als Forschungslinie in der Informatik, die sich mit der Sichtbarmachung von Blackbox-Modellen befasst, ist seit den 2010er Jahren massiv angewachsen (Arrieta et al., 2020; Gunning et al., 2021; Littman et al., 2021; P. Stone et al., 2016). Die Anzahl der Veröffentlichungen, die im Bereich XAI, Interpretable AI und Explainability eine Literaturübersicht bereitstellen, dabei Konzepte aufarbeiten, Rück- und Ausblick geben und einzelne XAI-Lösungen auflisten und klassifizieren, ist zahlreich (Ali et al., 2023; Angelov et al., 2021; Arrieta et al., 2020; Brasse et al., 2023; Carvalho et al., 2019; A. Das & Rad, 2020; Linardatos et al., 2020; Mohseni et al., 2021; Molnar, 2019; Murdoch et al., 2019; Ras et al., 2022; Zhou et al., 2016). Zusätzlich existieren spezifizierte XAI-Übersichten, wie z. B. für das Gebiet der Sprachverarbeitung (Cambria et al., 2023) oder spezifische Anwendungsdomänen wie den medizinischen Bereich (Dey et al., 2022).

Einerseits bezeichnet XAI die Forschung, die sich mit der Erstellung von Erklärungen für eigentlich nicht transparente Modelle befasst. Andererseits bezeichnet XAI auch das Ergebnis dieser Forschung: Die erklärbare KI selbst. Diese Erklärbarkeit wird, wie zuvor beschrieben, meist durch nachgelagerte Erklärungen erreicht: Post-Hoc. Eine klassische Post-Hoc-Methode ist die „Model Distillation“: Dabei werden kleinere, einfachere und deshalb transparentere Modelle entwickelt, die das Verhalten größerer, komplexerer Modelle nachahmen und dadurch den Entscheidungsprozess der komplexen Modelle nachvollziehbarer machen sollen (Carvalho et al., 2019; Molnar, 2019; Rai, 2020; Ribeiro et al., 2016). Dieser Vorgang lässt sich wie in Abbildung 6 formalisiert darstellen.

Abbildung 6: *Modellhafte Darstellung des Explainable Machine Learning-Prozesses*



Anmerkung. Abbildung nach Carvalho et al., 2019, S. 15, eigene Übersetzung.

Sehr häufig wird XAI anhand von zwei Kategorien klassifiziert: abhängig von Modellgültigkeit und abhängig vom Geltungsbereich der Erklärung. Die Klassifizierung nach Modellgültigkeit unterscheidet

modell-spezifische und modell-agnostische Verfahren (Arrieta et al., 2020; Brasse et al., 2023). Modellspezifische Verfahren sind abhängig von Modell(klasse), also speziell für bestimmte Prozesse entwickelt und nur für diese anwendbar. Modell-agnostische Verfahren hingegen werden zur Anwendung unabhängig von Prozessen oder Modellen für eine generelle Anwendbarkeit entwickelt und analysieren häufig eher Input und/oder Output (Arrieta et al., 2020; Murdoch et al., 2019; Rai, 2020). Ein prominentes agnostisches Erklärmodell ist LIME (Local Interpretable Model-Agnostic Explanations). Es variiert den Input in ein zu erklärendes Modell und analysiert, wann sich die Vorhersage ändert. So kann es Rückschlüsse auf die Entscheidungsprozesse des Modells ziehen und die Faktoren identifizieren, die zu einzelnen Entscheidungen (besonders) beitragen (Ribeiro et al., 2016).

Die Unterscheidung in agnostisch oder spezifisch ist für die Erstellung und Anwendung von XAI relevant, für Nutzende aber können die Ergebnisse identisch aussehen. Anders ist das bei der erklärungsbereichsabhängigen Kategorisierung. Hierbei wird üblicherweise zwischen globalen und lokalen Verfahren unterschieden (Angelov et al., 2021; Herm et al., 2023; Molnar, 2019; Rai, 2020). Globale Explainability liefert Aussagen zu den inneren Prozessen eines Modells, wie es arbeitet und grundsätzlich funktioniert. Damit beantwortet globale Explainability die Frage nach dem „Wie“ (Herm et al., 2023; Molnar, 2019). Lokale oder auch merkmalsabhängige (feature-priented) Explainability bezieht sich auf einzelne Ergebnisse bzw. Vorhersagen des KI-Systems. Es geht darum, warum dieses Ergebnis zustande kam und nicht ein anderes. Lokale Erklärungen beantworten also die Frage nach dem „Warum“ (Herm et al., 2023; Molnar, 2019). Zusätzlich zu den Fragen nach dem Wie und dem Warum ergänzen manche Autor*innen noch weitere Fragestellungen, um Erklärungen zu kategorisieren: „Warum nicht“, um kontrastierende Erklärungen zu geben, „Was wenn“, um Veränderungen sichtbar zu machen, „Wie sonst“, um auf interaktive Art Alternativen (Counterfactuals) zu produzieren oder „Was noch“, um Beispiele ähnlicher Daten zu zeigen (Herm et al., 2023; Mohseni et al., 2021). Die beiden Fragen nach dem Wie und Warum klassifizieren die herkömmlichen XAI-Lösungen der Computerwissenschaften. Auch weitere Erklärungen lassen sich diesen beiden zuordnen.

Im Folgenden werden einige klassische XAI-Ansätze in die beiden Dimensionen global und lokal eingeordnet, um die Kategorien anhand von geläufigen Beispielen zu verdeutlichen (siehe Tabelle 2). Klassische globale XAI-Lösungen sind beispielsweise textliche Erklärungen zum KI-Modell, dazu, wie es trainiert wurde oder funktioniert, ebenso wie die Abbildung eines (vereinfachten) Entscheidungsbaums oder ein Ranking der Eigenschaften, die in einem Modell am meisten Einfluss haben (Global Feature Importance; Caruana et al., 2015; Gurumoorthy et al., 2019). Auch eine Aussage zur Akkuratheit eines Modells wäre eine globale Erklärung (AI HLEG, 2020; Yin et al., 2019). Als lokale

Explainability im Bereich der Bildverarbeitung werden klassische Heatmaps betrachtet, die für einzelne Bilder aufzeigen, welche Pixel in einem Modell maßgeblich für eine Entscheidung waren (Herm et al., 2023; Posada-Moreno et al., 2023; Zintgraf et al., 2017). Die Hervorhebung von Wörtern bei der Texterkennung, textuelle Erklärungen zu den Merkmalen, die zum aktuellen Ergebnis geführt haben, oder Aufstellungen dieser Merkmale (Local Feature Importance) zählen zur lokaler Explainability (Baniecki & Biecek, 2019; Hohman et al., 2019; Ribeiro et al., 2016; Springer & Whittaker, 2018). Ebenso sind Wahrscheinlichkeitsangaben zu einer einzelnen Vorhersage eine lokale Aussage, die als (Un-)Sicherheit des einzelnen Ergebnisses bezeichnet wird (S. Cramer et al., 2022).

***Tabelle 2**: Beispiele von XAI-Lösungen entlang der Kategorien global und lokal sowie nach Inhalten: Funktionalität und Akkuratheitsangabe*

	Funktionalität/Explainability	**Akkuratheitsangabe**
Global	Liste der Einflussfaktoren insgesamt (z. B. Caruana et al., 2015; Gurumoorthy et al., 2019; Ribeiro et al., 2016) Entscheidungsbaum (z. B. Donadello & Dragoni, 2021; Wanner et al., 2020)	Akkuratheit (z. B. Yin et al., 2019)
Lokal	Einflussfaktoren auf das Ergebnis (z. B. Baniecki & Biecek, 2019; Hohman et al., 2019; Lundberg et al., 2020; Ribeiro et al., 2016) Heatmap (z. B. Herm et al., 2023; Posada-Moreno et al., 2023) Eingefärbter Text (z. B. Springer & Whittaker, 2018)	(Un-)sicherheit (z .B. S. Cramer et al., 2022; Mueller et al., 2020)

Über diese strenge Einteilung von lokaler und globaler Explainability hinaus gibt es inzwischen Ansätze der XAI-Forschung, beide zu verknüpfen, um so noch mehr Aussagekraft zu gewinnen (Posada-Moreno et al., 2023). Diese Lösungen sind jedoch noch hochkomplex und richten sich (bisher) ausschließlich an Entwickler*innen.

Die Assessment List for Trustworthy AI der AI HLEG der EU definiert Explainability als „Feature of an AI system that is intelligible to non-experts. An AI system is intelligible if its functionality and operations can be explained non technically to a person not skilled in the art“ (AI HLEG, 2020, S. 26). Damit formuliert die AI HLEG eine Forderung an KI, die in der XAI-Forschung bis in die 2020er Jahre kaum eine Rolle spielte. Denn für lange Zeit hatte sich die Forschung zu Erklärbarkeit und Interpretierbarkeit auf die Perspektive der Entwickler*innen und KI-Expert*innen und nicht der Nutzenden konzentriert (Miller, 2019). Dies zeigte auch eine Analyse von KI-Anwendungen und dem Umgang mit ihren

Stakeholdern (Bhatt et al., 2020). Ziel war es zumeist, KI-Systeme nachvollziehbar zu machen und Prozesse in den Blackboxen zu verstehen, um so die Qualität der Systeme sicherstellen bzw. sie verbessern zu können.

So bestehen die Vorteile von transparenter KI laut Arrieta et al. (2020) in der Sicherstellung von

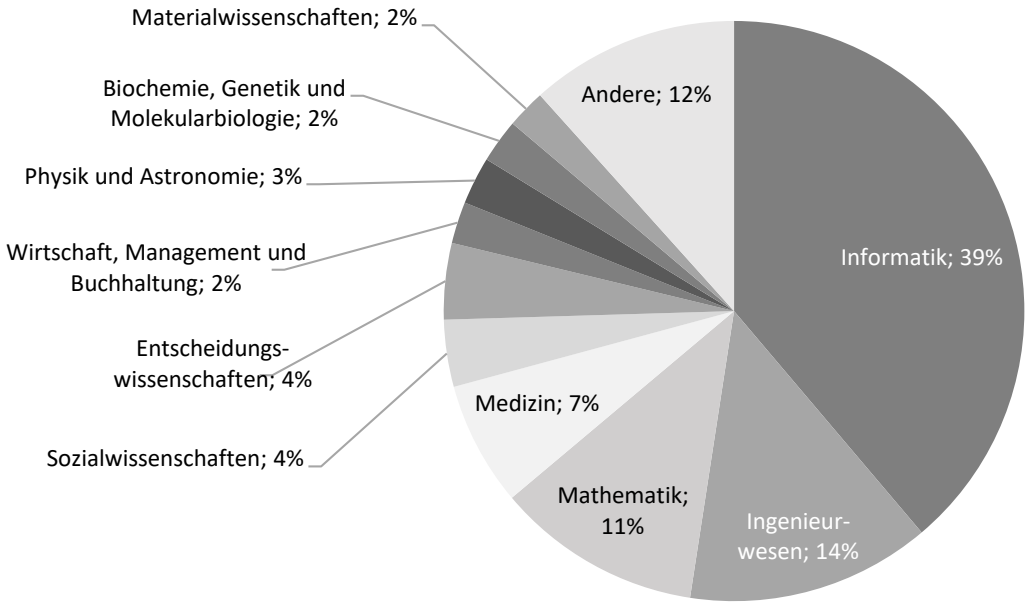
- **Vorurteilsfreiheit**, z. B. im Trainingsdatensatz,
- **Stabilität**, z. B. indem eine schädliche Einflussnahme auf Ergebnisse ausgeschlossen werden kann und
- **Sinnhaftigkeit der Ergebnisse**, da der Zusammenhang der Variablen auf das Ergebnis geprüft werden kann, z. B. wenn eine bestehende, zugrundeliegende Kausalität besteht.

Diese eher technische Sicht – der Bedarf an Transparenz in KI, um sie als Entwickler*innen verbessern zu können – besteht in vielen Forschungsansätzen weiterhin fort. Das Ziel transparenter KI sei es „to provide users with explanations that enable them to understand the system‘s overall strengths and weaknesses; convey an understanding of how it will behave in future/different situations; and perhaps permit users to correct the system‘s mistakes“ (Rai 2020, S. 138). Die Forscher konzentrieren sich auf das System selbst und seine Funktionalität (Larsson und Heintz 2020) und nicht auf das „Hintergrundwissen der Nutzer, ihren Kenntnisstand und die Zeit, die ihnen zur Verfügung steht, um das Proxy-Modell zu verstehen“ (Páez 2019, S. 14-15). Wie sich Abbildung 7 entnehmen lässt, sind mit 64 % der überwiegende Teil der Veröffentlichungen zum Thema Explainability, Interpretability und Transparent AI den Disziplinen Computerwissenschaften, Ingenieurwesen und Mathematik zuzuordnen. Auf die Sozial- und Entscheidungswissenschaften entfallen lediglich 8 %.

Dieser technische Ansatz hat einerseits seine Berechtigung: KI-Systeme müssen akkurat, fair und verlässlich arbeiten. Andererseits spielten in der XAI-Forschung sozialwissenschaftliche Ansätze lange Zeit keine Rolle. Es blieb unklar, wie Explainability verständlich dargestellt werden kann oder wie Visualisierungen wie Heatmaps vom Endnutzenden wahrgenommen werden. Dies konterkariert auch ein häufig geäußertes Argument der Informatik: nämlich, dass „by building more transparent, interpretable, or explainable systems, users will be better equipped to understand and therefore trust the intelligent agents“ (Miller, 2018, S. 3).

Páez hob diesbezüglich 2019 hervor, dass zwar umfassende Erklärungen dazu beitragen könnten, KI-Expert*innen einen Einblick in ein System zu ermöglichen. Dennoch würden diese Anforderungen keine Transparenz und keinen Einblick für Endnutzende bieten. Die Einsicht für diese Lücke wuchs auch durch die Forderungen der EU in den 2020er Jahren und damit einher ging die Erkenntnis, Erklärungen sollten „depend on the context, the severity of the consequences of its decision […] and the relevant stakeholders“ (Felzmann et al., 2020, S. 3348). Páez (2019) plädiert für mehr Verständlichkeit anstelle

***Abbildung 7**: Die Zahl der Veröffentlichungen auf EBSCOhost zu „explainability" oder „interpretability" zusammen mit „ai" oder „artificial intelligence" nach Themenbereichen*



Anmerkung. Abgerufen über scopus, 2017 bis 2023; Stand Juni 2024; Mehrfachzuordnung möglich.

von perfekten Erklärungen. Wichtig seien dazu drei Faktoren: „obtaining the right fit between the interpretative model and the black box model in terms of accuracy and reliability, providing sufficient information about its limitations, and achieving an acceptable degree of comprehensibility for the intended user" (Páez, 2019, S. 14).

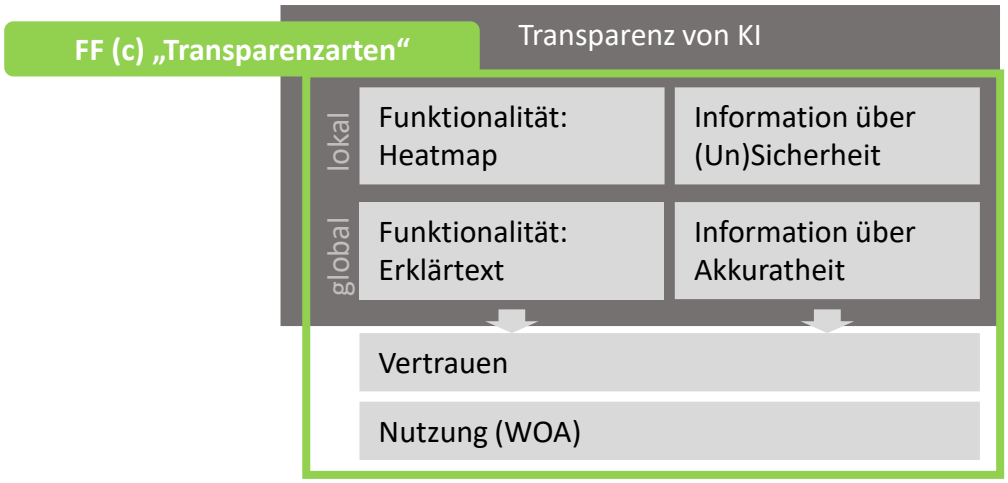
Um die Herausforderungen zu bewältigen und KI-Transparenz für Endnutzende zu erreichen, sind kombinierte Erkenntnisse aus den Sozial- und Computerwissenschaften erforderlich (Larsson & Heintz, 2020). Es gilt, durch Erkenntnisse aus der Sozialwissenschaft das Verständnis von KI-Transparenz zu bereichern, um beispielsweise die Visualisierung interpretierbarer Modelle und deren Wirkung auf Endnutzende zu beleuchten (Murdoch et al. 2019). Es ist also zu klären (siehe Abbildung 8):

> **Forschungsfrage (c) „Transparenzarten": Wie wirken sich verschiedene Arten der KI-Transparenz auf Vertrauen und Nutzung eines Systems aus?**

Dazu wurden ausgehend von den verschiedenen technisch möglichen XAI-Lösungen, die in Tabelle 2 aufgeführt sind, vier verschiedene Transparenzarte ausgewählt, die möglichst unterschiedliche Bereiche abdecken: Sie sind einteilbar in globale und lokale Transparenz sowie Erklärungen zur Funktionalität und Akkuratheitsangaben. Mit einer Heatmap als lokaler Erklärung, textlichen Beschreibungen als globaler Erklärung sowie den beiden lokalen und globalen Akkuratheitsangaben wurden vier sehr weit verbreitete Transparenzarten gewählt. Darüber hinaus sind die beiden Explainability-Arten angelehnt an die Ergebnisse aus Forschungsfrage (b) „Nutzendenanforderungen",

wo generelle Erklärungen zur Vertrauensvermittlung einerseits und lokale Erklärungen andererseits gewünscht wurden (mehr zu den Ergebnissen aus Forschungsfrage (b) finden sich in Kapitel 5.3).

Abbildung 8: *Schematische Darstellung der Forschungsfrage (c) im Rahmen der Gesamtarbeit*



2.2.5. Transparenzverständnis in der vorliegenden Arbeit

Wie die vorangegangenen Ausführungen deutlich machen, ist Transparenz im Zusammenhang mit KI ein weitrechender, auf sehr vielfältige Weise definierter und viele weitere Bestandteile umfassender Containerbegriff. Dabei betreffen die Definitionen KI-Mechanismen und ihre zugrundeliegende Logik bis hin zur Möglichkeit, Systeme zu verbessern und Diskriminierung zu verhindern (Ananny & Crawford, 2018). Durch immer komplexere KI, die zu einer Blackbox wurde, und vor dem Hintergrund politischer Forderungen, z. B. von der EU, wurde Transparenz in KI zu einem „modern, surprisingly complex […] ideal“ (Koivisto, 2016, S. 2).

Die vorliegende Arbeit untersucht die beiden Aspekte von Transparenz, die bereits aus technischer Perspektive beforscht wurden, deren Auswirkungen auf Endnutzende jedoch oft noch unklar bleiben: **Erklärbarkeit und Informationen über die Akkuratheit einer KI**.

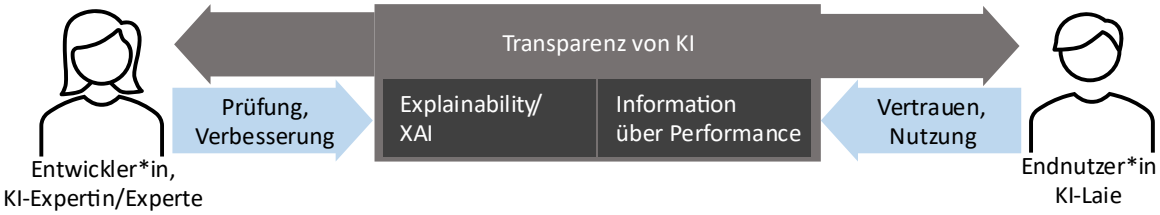
Dabei wurde Explainability lange Zeit als technischer Ansatz verfolgt (Arrieta et al., 2020; Carvalho et al., 2019; Gunning et al., 2021; Miller, 2019; Mohseni et al., 2021; Rai, 2020). Es bestand der Bedarf, KI-Systeme mithilfe von Explanability für Entwickler*innen nachvollziehbar und so verbesserbar zu machen – ein Ansatz, der bis heute im Feld fortbesteht. Gleichzeitig übersieht dieser Ansatz das in der Informatik häufig geäußerte Argument, durch Transparenz Endnutzende in die Lage zu versetzen, den Systemen zu vertrauen (Miller 2018). Denn lange Zeit blieb es in der Forschung zu XAI unklar, wie Explainability verständlich dargestellt werden kann oder wie Visualisierungen wie Heatmaps vom Endnutzenden wahrgenommen werden (Murdoch et al. 2019).

Auch bezüglich der Modell-Leistung lag der Schwerpunkt der Forschung lange Zeit auf der kontinuierlichen Verbesserung der Systeme, also auf der technischen Auseinandersetzung mit

steigender Komplexität und der Verarbeitung immer vielschichtigerer Datensätze. Dies verstärkte einerseits das Spannungsfeld zwischen möglichst guter Modell-Performance und möglichst guter Erklärbarkeit eines Systems, da komplexere Systeme zwar häufig bessere Vorhersagen oder Klassifizierungen liefern, jedoch weniger nachvollziehbar sind. Selbst wenn sich kein linearer Zusammenhang nachweisen lässt, trifft dieser Zusammenhang auf besonders komplexe KI-Systeme, nämlich Deep Neural Networks, zu, die von sich aus nicht erklärbar sind, also Blackboxen darstellen (Herm et al., 2023). Andererseits bildet die Kommunikation dieser Performance, beispielsweise als Akkuratheitsangabe eines KI-Systems, einen wichtigen Bestandteil algorithmischer Transparenz (AI HLEG, 2020). Wie schon bei Explainability befasste sich die Forschung mit der technischen Umsetzung, bevor aus sozialwissenschaftlicher und psychologischer Perspektive die Frage aufkam, wie diese Performance-Informationen eigentlich auf Nutzende von KI wirken.

Diese beiden Perspektiven von Expert*innen einerseits und KI-Laien andererseits auf Explainability bzw. Transparenz im Ganzen verdeutlicht Abbildung 9. Je nach betrachteter Zielgruppe unterscheidet sich die angestrebte Wirkung von KI-Transparenz: Prüfung und Verbesserung der Systeme für KI-Expert*innen, Nutzung und Vertrauen für KI-Laien. Insbesondere der **Forschungsstand zu KI-Laien ist uneinheitlich und lückenhaft**. Die Relevanz einer KI-Nutzung für Laien nimmt aber stetig mit der Zahl der KI-Systeme zu, weshalb sich die vorliegende Arbeit mit dieser Zielgruppe befasst und entsprechend **Nutzung und Vertrauen** in den Blick nimmt. Mehr zu Nutzungsstudien zu transparenter KI findet sich im folgenden Kapitel 2.3.

Abbildung 9: *Perspektiven auf transparente KI für die Nutzendengruppen KI-Expert*innen und KI-Laien*



Die vorliegende Arbeit nutzt die prominente und pragmatische Unterscheidung von KI-Transparenz, die sich auf den Fokus der Erklärung stützt: die Unterscheidung von **globalen und lokalen Verfahren** (Angelov et al., 2021; Herm et al., 2023; Molnar, 2019; Rai, 2020). Globale Explainability beantwortet die Frage nach dem „Wie“, lokale Explainability bezieht sich auf einzelne Ergebnisse und beantwortet die Frage nach dem „Warum“ (Herm et al., 2023; Molnar, 2019). Auch Performance-Informationen lassen sich diesen Kategorien zuordnen: Eine lokale Performance-Angabe betrifft ein einzelnes Ergebnis und wird auch Unsicherheitsangabe genannt. Eine globale Angabe betrifft den gesamten Algorithmus und damit seine Akkuratheit.

Die Forschungsfragen befassen sich mit verschiedenen Detailtiefen und Foki mit den beschriebenen Bestandteilen und Aspekten der Transparenz von KI (siehe auch Abbildung 10 in Kapitel 2.4). Forschungsfrage (a) „Fehlerfall" untersucht den Effekt einer **Akkuratheitsangabe** auf die Nutzung eines Algorithmus nach einem Fehlerfall und widmet sich damit einer globalen Angabe zur KI-Performance in einem sehr konkreten Nutzungsfall. Forschungsfrage (b) „Nutzendenanforderungen" hingegen setzt sich mit einem qualitativen Ansatz mit den Nutzendenanforderungen an KI-Transparenz auseinander. **Transparente KI wird dabei als Ganzes** betrachtet, jedoch in der Analyse die beschriebenen Kategorien, lokal und global, berücksichtigt. Die letzte Forschungsfrage (c) „Transparenzarten" bezieht sich auf **vier unterschiedliche Transparenzarten** und untersucht ihre Wirkung auf Nutzung und Vertrauen. Angelehnt an die Übersicht über verschiedene Transparenzarten (Tabelle 2 in Kapitel 2.2.4) werden vier Arten der Transparenz verglichen: je eine lokale und globale Art der Erklärbarkeit sowie eine lokale Unsicherheits- und eine globale Akkuratheitsangabe.

Auf diese Weise untersucht die vorliegende Arbeit ein sehr breites Transparenzverständnis einerseits, geht aber auch in die Tiefe und konzentriert sich auf einzelne Transparenzarten, um dort Effekte prüfen und Zusammenhänge verdeutlichen zu können. Im Folgenden werden bestehende Arbeiten zum Zusammenhang von transparenter KI und ihrer Nutzung aufgeführt, um die Arbeit in den Rahmen dieser Vorarbeiten einzubetten, bestehende Forschungslücken aufzuzeigen und die verwendete Methodik und Herangehensweise abzuleiten.

2.3. Nutzungsstudien zu transparenter KI

In einer bevölkerungsrepräsentativen Befragung von insgesamt 1.221 Deutschen zu ihren Assoziationen zum Begriff „Algorithmus" nannten sie am häufigsten „Genauigkeit". Gleichzeitig äußerten 79 Prozent der Befragten, sie zögen menschliche Entscheidungen automatisierten vor (Fischer & Petersen, 2018). Diese grundsätzliche Neigung, menschlichen Rat algorithmischem vorzuziehen zeigen einige Studien, andere proklamieren inzwischen aber auch eine sogenannte „Algorithm Appreciation", insbesondere bei objektiven und mit Zahlen verbundenen Aufgaben (Burton et al., 2020; Logg et al., 2019; Morewedge, 2022). Dennoch besteht im Kontext von KI noch eine erhebliche Skepsis (Brauner et al., 2024; Haupt et al., 2024; Vianello et al., 2023).

Das Ziel ist nicht, diese Skepsis auszuräumen oder eine gedankenlose KI-Nutzung zu erreichen. Vielmehr stellt sich die Frage, wie eine sinnvolle, informierte Nutzung gestärkt werden kann, insbesondere da die Zahl der KI-Systeme, auch im privaten Bereich, stetig zunimmt. Als ein häufig diskutiertes Mittel wird hierbei KI-Transparenz genannt, die durch Aufklärung, Information und Nachvollziehbarkeit Vertrauen steigern und eine informierte Nutzung von KI-Systemen erlauben soll. Während die politische Forderung nach Transparenz und Erklärbarkeit von KI zunimmt, zeigen Studien

jedoch ein uneinheitliches Bild von den Auswirkungen von Transparenz. Die folgenden Abschnitte verdeutlichen den aktuellen Forschungsstand zu Nutzungsstudien, die sich mit den Effekten transparenter KI befassen.

Die Studien zur Nutzung von KI unter bestimmten Bedingungen sind zahlreich. Teilweise werden verschiedene Eigenschaften der Algorithmen manipuliert, teilweise die Nutzung von KI und menschlichen Ratgebern verglichen. Unter letztere Kategorie fallen auch die meisten Studien zu Algorithm Aversion, wie in Kapitel 2.1.4 beschrieben.

Zur Identifikation relevanter Studien wurde für diesen Abschnitt eine Analyse mit Google Scholar und ScienceDirect durchgeführt. Es ist zu betonen, dass aufgrund der sehr zahlreichen Ergebnisse kein Review der gesamten Literatur, sondern vielmehr eine Sichtung der Literatur und eine Priorisierung aufgrund der Relevanz (Zitationen) bzw. Neuigkeit (Erscheinungsdatum) vorgenommen wurde. Die dabei identifizierten Veröffentlichungen werden in den folgenden Kapiteln aufgeführt (siehe auch Anhang A).

Zum Ziel dieser Analyse wurden bei einer ersten Suche zu Beginn des Jahres 2021 und bei einer zweiten im Jahr 2023 der Suchterm [[„transparency“ AND „artificial intelligence“] OR [„Explainability“ AND „artificial intelligence“]] AND [„Laboratory study“ OR „Experimental design“ OR „User study”] entsprechend den Vorgaben der Plattform eingegeben. In der ersten Suche wurde keine Zeitbegrenzung, in der zweiten eine Begrenzung auf Studien ab 2020 vorgenommen. Von den in der ersten Suche erlangten Funden wurden zunächst doppelte Eintragungen entfernt und dann die übrigen 2.844 anhand des Titels und ggf. des Abstracts gescannt. Dabei war das Ziel, diejenigen Veröffentlichungen zu identifizieren, in denen tatsächlich Nutzungsstudien stattgefunden und auch Transparenzaspekte untersucht wurden. Eine überwiegende Zahl der Studien beschäftigte sich jedoch mit der reinen Entwicklung von KI oder von explainable AI und ihrer Anwendung aus technischer Entwicklungssicht. Teilweise wurden KI-Systeme mit Nutzenden evaluiert, jedoch ohne Transparenz oder Explainability zu thematisieren. Entsprechende Veröffentlichungen wurden von der weiteren Analyse ausgeschlossen.

Bei der detaillierten Sichtung der als relevant angesehenen Paper wurden zum einen weitere, das Thema nicht in der intendierten Weise behandelnde Veröffentlichungen ausgeschlossen. Gleichzeitig konnten durch Verweise in den Quellenangaben andere, passendere Veröffentlichungen ermittelt und ergänzt werden. Trotzdem wurden auf diese Weise im Jahr 2020 lediglich 31 Studien identifiziert, die den Anforderungen entsprachen: Nutzendenstudien im Bereich transparente KI, in denen Transparenzfaktoren untersucht werden. Bei der gleichen Suche nur zwei Jahre später, die zum Ziel hatte, neu hinzugekommene Studien zu identifizieren, wurden mit dem gleichen Vorgehen 34 weitere

Paper für den viel kürzeren Zeitraum von 2020 bis 2023 ermittelt[2]. Die identifizierten Veröffentlichungen finden sich in Anhang A. In den folgenden Kapiteln ist weitere, nicht in der Übersicht in Anhang A aufgeführte Literatur zitiert, die zum Verständnis der Thematik relevant ist. Dazu gehören beispielsweise Übersichtswerke, die keine eigenen Studien enthalten oder Veröffentlichungen, die dem Verständnis von Effekten transparenter KI dienen, jedoch keine Transparenzeffekte untersuchen.

Die überwiegende Mehrzahl der identifizierten, aber hier nicht aufgeführten Studien überprüften nicht die Effekte von Transparenzarten oder manipulierten Eigenschaften auf die abhängigen Variablen wie Vertrauen oder Nutzung, sondern prüften in erster Linie einen selbst entwickelten Explanability-Algorithmus (z. B. Dragoni et al., 2020; Li et al., 2022; Musto et al., 2019). Einerseits sind diese Untersuchungen Nutzungsstudien zu Transparenz in KI und dienen dem Zweck, nicht nur XAI zu entwickeln, sondern auch ihre Wirkung zu prüfen. Entsprechende Nutzungsstudien sind auch unter den hier berichteten zu finden, beispielsweise Springer (2019) oder Venkatesh, Thong et al. (2016). Andererseits wurden zahlreiche Studien ausgeklammert, deren Fokus nicht auf der Nutzungsanalyse lag, sondern in denen in erster Linie erhoben wurde, wie ein neuer XAI-Ansatz bewertet wurde, ohne tiefere Erkenntnisse zu befördern. Da die Übergänge fließend sind, handelt es sich bei den im Folgenden berichteten Studien letzten Endes um eine subjektive Auswahl und die zusammengestellte Liste kann nicht als vollumfänglich betrachtet werden. Ziel ist es vielmehr, einen Eindruck zu vermitteln, welche Studien wann und wie durchgeführt wurden bzw. welche uneinheitlichen Ergebnisse die Studien erlangten.

Zunächst werden dazu die häufig erhobenen und für diese Arbeit relevanten abhängigen Variablen Nutzung und Vertrauen eingeordnet (2.3.1). Anschließend erfolgt eine Darlegung der Studien, die sich mit der Kommunikation von Akkuratheit und deren Effekten befassen (2.3.2) sowie der Studien, die sich mit den Effekten von klassischer Transparenz befassen (2.3.3).

2.3.1. Abhängige Variablen bei Transparenzstudien

In der Literatur wird Transparenz von KI-Systemen sowohl als Ergebnis als auch als Methode betrachtet (Rader et al., 2018). Während einige Studien verschiedene Aspekte des Systems, seines Interfaces oder hinzugefügter Erklärungen variieren und im Anschluss die wahrgenommene Transparenz evaluieren, betrachten andere den Einfluss von Transparenz auf Variablen wie Akzeptanz, Vertrauen und Nutzung. In der frühen Forschung von XAI war das Hauptargument für die Notwendigkeit algorithmischer Transparenz das zu steigernde Vertrauen und die Akzeptanz. Der Ansatz der vorliegenden Arbeit folgt dieser Argumentation von Transparenz als Mittel, stellt sie aber gleichermaßen in Frage. Denn diese

[2] Bei einem der Paper wurde zwischenzeitlich das Veröffentlichungsdatum auf 2024 geändert, eine weitere relevante Studie aus dem Jahr 2024 wurde nachträglich ergänzt.

Annahme vernachlässigt den multidimensionalen Charakter von Vertrauen, der dazu führt, dass „there is no simple correlation between explanation and trust, and that an adequate analysis of trust requires taking into account contextual factors that can foster or hinder it" (Páez, 2019, S. 9). Diese Aussage spiegelt sich in den zahlreichen uneinheitlichen Ergebnissen der Nutzungsstudien wider.

Als **Ziel von Transparenz werden häufig Vertrauen in und Nutzung eines Systems** genannt. Diese Variablen sind sicherlich zentral, um die Auswirkungen von Transparenz auf Endnutzende zu untersuchen. Gleichzeitig fokussieren Studien auf die Wahrnehmung des Systems oder Transparenzbeurteilungen (z. B. Molina & Sundar, 2022; Z. Zhang et al., 2021). Weitere analysierte abhängige Variablen beinhalten die subjektive Zufriedenheit mit dem System oder dessen Ergebnis (z. B. Alam & Mueller, 2021; Dominguez et al., 2019; Ford et al., 2020) oder die seitens der Nutzenden wahrgenommene Rechenschaftspflicht eines Algorithmus (Rader et al., 2018). Einzelne Studien berücksichtigen physiologische Maße wie die Herzrate als Proxis für Stress und Erregung bei der Nutzung von Algorithmen (Alexander et al., 2018) oder die Pupillenveränderungen als Maß für die kognitive Belastung (Karran et al., 2022). Der Tabelle in Anhang A lassen sich die verschiedenen abhängigen Variablen der Nutzungsstudien entnehmen.

Gleichzeitig wird **Vertrauen** auf vielfältige Weisen operationalisiert (Schmidt et al., 2020): Einerseits werden subjektive Zuschreibungen wie subjektive Akkuratheit oder Kompetenz, Zuverlässigkeit, Integrität oder allgemeines Vertrauen über Skalen erhoben (z. B. Alam & Mueller, 2021; Joinson et al., 2010; Karran et al., 2022; Molina & Sundar, 2022; Ribes et al., 2021; Verberne et al., 2012). Andererseits argumentieren manche Forscher*innen Verhaltensmaße dienten als Operationalisierung von Vertrauen (als „behavioral trust") und untersuchen die tatsächliche Nutzung eines algorithmischen Systems, um darüber Aussagen über Vertrauen zu tätigen (Alexander et al., 2018; B. Berger et al., 2021; Schmidt et al., 2020; Yin et al., 2019; Y. Zhang et al., 2020). Dabei gilt zum einen sicherlich der Zusammenhang von Vertrauen als wichtige Voraussetzung für Technologienutzung (Schmidt et al., 2020). Zum anderen lässt sich aus hohem (kognitivem) Vertrauen nicht automatisch auf hohe Nutzung schließen, wie beispielsweise Daschner & Obermeier (2022) in einer Untersuchung mit algorithmischen Ratgebern zeigen.

Bis ungefähr 2020 untersuchte viel Forschung aus dem Bereich „transparente KI" ihre Wirkung auf Vertrauen und andere subjektive Maße. Häufig wurden einzelne Systeme getestet und eine subjektive Transparenzwahrnehmung in Verbindung gebracht mit Vertrauen in diese Systeme (z. B. Chen & Sundar, 2018; Cheng et al., 2019; Eslami et al., 2018; Venkatesh, Thong, et al., 2016). Die tatsächliche **Nutzung** der Systeme war weniger im Zentrum und selten als abhängige Variable betrachtet, jedoch hat die Zahl der Studien mit tatsächlichem Verhalten als abhängiger Variable in den letzten Jahren deutlich zugenommen (siehe Anhang A). Entlang dieser Studien argumentiert auch die vorliegende

Arbeit, dass die Nutzung als Verhaltensvariable in Zusammenhang mit Vertrauen zu sehen ist, aber Vertrauen weder eine notwendige noch hinreichende Voraussetzung für Nutzung darstellt. Man stelle sich beispielsweise Situationen vor, in denen Nutzende unter Bedingungen wie Zeitdruck und begrenzter Informationslage eine Entscheidung treffen müssen. In diesen Fällen könnte die Nutzung einer beratenden KI hoch ausfallen, auch wenn das Vertrauen in das System eigentlich gering ist.

Die Theorie des geplanten Verhaltens betont die Nutzungsintention als notwendige Voraussetzung für Verhalten, jedoch nicht als zwangsläufige Konsequenz (Ajzen, 1991). Studien berichten von Zusammenhängen von bis zu 75 %, teilweise aber auch deutlich darunter, je nach Fragestellung und Verhaltensweise (Ajzen, 1991, 2014; Ajzen & Fishbein, 1977). Theoretische Nachfragen oder hypothetische Vignettenstudien sind sicherlich hilfreich, um Zusammenhänge zu evaluieren oder zu sondieren. Interviewstudien oder Fokusgruppen wie in Forschungsfrage (b) „Nutzendenanforderungen“, die über hypothetische Nutzungen diskutieren, dienen dem Zweck, tiefere Einblicke in die Annahmen und Motive von Nutzenden zu erhalten. Darüber hinaus stellen jedoch Verhaltensmaße die besten Prädiktoren für zukünftiges Verhalten dar. Entsprechend wird in Forschungsfrage (a) „Fehlerfall“ und (c) „Transparenzarten“ das Nutzungsverhalten als Konsequenz von verschiedenen Transparenzmanipulationen erhoben und geht damit über eine Vielzahl von Studien hinaus, die nur hypothetisches Verhalten oder Vertrauen erheben.

2.3.2. Akkuratheitsangaben und Nutzung von KI-Systemen

Wie bereits dargelegt, nimmt die aktuelle Arbeit ein sehr breites Verständnis von Transparenz an und schließt dort auch – wie beispielsweise von der AI HLEG der Europäischen Kommission vorgegeben – die Kommunikation über KI-Akkuratheit ein. Die Relevanz der Transparenz ergibt sich auch aus dem zu Beginn angesprochenen Effekt des Advice Discounting, der besagt, dass Ratschläge abgewertet werden aufgrund einer vorliegenden Informationsasymmetrie (Yaniv & Kleinberger, 2000). Der zufolge hat ein*e Nutzer*in zwar Informationen über die eigenen Entscheidungsrechtfertigungen, nicht aber über die des Ratgebers. Transparenz, auch in Form von Informationen über Entscheidungssicherheit, könnte helfen, diese Asymmetrie auszugleichen.

Studien dazu, unter welchen Bedingungen Ratschläge von Menschen angenommen werden, zeigen den **großen Einfluss von Sicherheitsangaben, Reputation und Vorerfahrung mit dem Ratgeber** (Bonaccio & Dalal, 2006, 2010; Harvey & Fischer, 1997; Sniezek & Van Swol, 2001; Yaniv & Kleinberger, 2000). Das Selbstvertrauen („confidence“) des Ratgebers, ist einer der wichtigsten Faktoren bei der Wahl, ob und wie sehr ein Ratschlag angenommen wird (Bonaccio & Dalal, 2006; Hütter & Fiedler, 2019; Price & Stone, 2004; Van Swol, 2011; Van Swol & Sniezek, 2005). Operationalisiert wurde dies u. a. mit Angaben auf einer Likert-Skala oder mit Wahrscheinlichkeitswerten (Bonaccio & Dalal, 2010), also Aussagen wie „Der Ratgeber ist sich mit dem Ratschlag zu 80 % sicher“. Daran anschließend stellt

sich die Frage, inwiefern sich diese Aussage über das Selbstvertrauen auf einen algorithmischen Ratschlag übertragen lässt, also auf eine Akkuratheitsangabe wie „Der Algorithmus ist sich mit dem Ergebnis zu 80 % sicher“.

Wie sich immer wieder zeigt, beeinflussen auch bei maschinellem Rat Aussagen über die Akkuratheit eines Systems seine Wahrnehmung und Nutzung. Während Nutzende sich eher **auf ein System mit hoher Sicherheit verlassen, überprüfen sie seine Entscheidungen bei einer niedrigen Sicherheitsaussage öfter, nehmen sie nicht an, vertrauen ihr weniger** (z. B. Antifakos et al., 2005; Ford et al., 2020; Yin et al., 2019; Y. Zhang et al., 2020). Aussagen über (Un-)Sicherheiten eines Systems sind also nicht per se positiv, sondern machen transparent, inwiefern dem System vertraut werden kann. Sie wirken also als Moderatoren (Venkatesh, Thong, et al., 2016). Dies gilt in Anwendungsbereichen mit hohem wie auch niedrigem Risiko (Antifakos et al., 2005; Yin et al., 2019). Bei Sicherheitsangaben eines KI-Systems für Einkommensvorhersagen zeigt sich der klassische Effekt: Nutzende folgen einem System mehr, je höher die angegebene Sicherheit, und mehr als ohne Angabe. Dabei ist weniger relevant, ob Versuchspersonen einer Systemempfehlung blind folgen müssen oder ob sie die Empfehlung vor ihrer Entscheidung sehen (Y. Zhang et al., 2020). Und auch im Vergleich mit einer lokalen Erklärung haben Sicherheitsangaben einen größeren Einfluss auf die KI-Nutzung (Y. Zhang et al., 2020; Z. Zhang et al., 2021). In einer Studie von Pálfi und Kolleginnen zeigte sich kein Zusammenhang von Akkuratheitsangabe und Nutzung bei einem Krebserkennungssystem im Einsatz bei Allgemeinmediziner*innen (2022). Dies steht entgegen vorangehenden Studien, könnte aber an der in der Studie herrschenden sehr hohen Systemakzeptanz und -nutzung liegen.

Die hohe Überzeugungskraft der Sicherheitsangaben birgt jedoch auch Gefahren: Nutzende folgen einem als sicher präsentierten System auch bei falschen Entscheidungen; ebenso erhalten sie eine negative Bewertung eines Systems aufgrund dargestellter geringer Sicherheit selbst dann aufrecht und widersprechen ihm häufiger, wenn es korrekte Ergebnisse liefert (Lim & Dey, 2011). In einem Vergleich von **Erklärung und Sicherheitsangabe zeigte sich ebenso ein möglicher irreführender Effekt** beider Faktoren. Hohe Sicherheitswerte bei dennoch falschen Ergebnissen führten dazu, dass Versuchspersonen eine um 10 % schlechtere Leistung erzielten als ohne Sicherheitsangaben. Besonders in schwierigeren Aufgaben – es ging darum, Texte als positiv oder negativ konnotiert zu klassifizieren – verließen sich Versuchspersonen übermäßig häufig auf den Algorithmus. Jedoch ließen sie sich durch Erklärungen, markierte Worte in den Texten, vorschnell von einer genaueren Analyse und dem eigentlich korrekten Ergebnis abbringen (Schmidt et al., 2020). Bei hoher postulierter Sicherheit, aber falschen Ergebnissen halfen Erklärungen zum Ergebnis dabei, die Wahrnehmung der Systemperformance nicht abstürzen zu lassen (Lim & Dey, 2011).

Erleben Versuchspersonen allerdings ein fehlerhaftes System, nimmt das Vertrauen in dieses System ab, selbst wenn die zuvor dargestellte Akkuratheit hoch ist. Umgekehrt wird durch das Erleben eines korrekten Systems eine zuvor als niedrig dargestellte Akkuratheit nur langsam angepasst (Yin et al., 2019). Die postulierte **Akkuratheit eines Algorithmus beeinflusst maßgeblich die Wahrnehmung des Algorithmus,** bis hin zu dem Punkt, an dem **falsche Ergebnisse nicht mehr hinterfragt werden**. Gleichzeitig überschreibt das Erleben eines Algorithmus als fehlerhaft die Akkuratheitsangaben und lässt sich nur schwer überwinden.

In einem Videospiel mit der Aufgabe, die Welt zu retten, wurden die Nutzung bzw. Über- und Unterinanspruchnahme von Hinweisen eines automatisierten Ratgebers unter verschiedenen Bedingungen erhoben: Unsichere Umgebungen und eine hohe postulierte Verlässlichkeit des automatisierten Ratgebers führten dazu, dass Ratschläge übermäßig in Anspruch genommen oder im umgekehrten Fall zu selten genutzt wurden (Sutherland et al., 2016). Aufgrund der sehr künstlichen Umgebung muss in Frage gestellt werden, ob die Versuchspersonen die automatisierten Ratgeber als solche wahrnahmen. Dennoch bestätigen diese Ergebnisse den **starken Einfluss der KI-Ratgeber besonders bei unsicheren, komplexen Entscheidungen**.

In einer Studie, in der Erklärungen und Sicherheitsratings für Kunst-Empfehlungen eines Kunst-Empfehlungssystems präsentiert wurden, zeigte sich der Effekt der Erklärung sehr viel positiver als der der Sicherheitsangaben. In nachgelagerten Interviews erklärten die Versuchspersonen, die Sicherheitsangaben als nicht zutreffend und mit teilweise unter 50 % als zu negativ wahrgenommen zu haben. Dabei zeigte sich, so die Autor*innen, dass die Sicherheitsratings schlicht falsch verstanden worden waren (H. Cramer et al., 2008). Der Effekt ist also weiterhin anzunehmen – und es stellt sich als umso wichtiger heraus, ein Verständnis der Angaben sicherzustellen.

Wie hoch die Sicherheitsangaben dabei zu sein haben, um zu überzeugen, ist nicht eindeutig. Einige Studien untersuchen mehrere Stufen, häufig zwischen 50 und knapp 100 %. Andere vergleichen nur zwei oder wenige Gruppen. Die Ergebnisse zeigen meist, wie schon beschrieben: je sicherer, desto überzeugender (Antifakos et al., 2005; Ford et al., 2020; Lim & Dey, 2011; Madhavan & Wiegmann, 2007; Yin et al., 2019; z. B. Y. Zhang et al., 2020). Zu niedrige Akkuratheitsangaben von teilweise 80 oder gar 60 % führen dazu, dass Systeme weniger genutzt werden als ohne Angabe (Ford et al., 2020; T. Kim & Song, 2020). In Abhängigkeit von den KI-Nutzenden untersuchte eine Studie zu wirtschaftlichen Prognosen den Einfluss der Höhe der Sicherheitsangaben: Sie manipulierte die Höhe der tatsächlichen KI-Akkuratheit als höher, gleich oder niedriger als die der Nutzenden. Die Studie umfasste mehrere Durchgänge, nach denen die Fehlerraten der Versuchspersonen sowie die des Algorithmus offengelegt wurden. Zeigte der Algorithmus höhere oder die gleiche Akkuratheit wie die Teilnehmenden, nutzten sie ihn weiterhin. Im Falle einer schlechteren Akkuratheit nahm die Nutzung

ab (Daschner & Obermaier, 2022). Die eigene Leistung dient also als Schwelle, die ein ratgebender Algorithmus mindestens zu erreichen hat.

In einer anderen Studie wurde eine Sicherheitsaussage mit sozialen Nutzungshinweisen verglichen: die Aussage, der Algorithmus sei 75 % sicher, mit den Aussagen, 50 bzw. 70 % der anderen Nutzenden würden ihn in Anspruch nehmen. Es zeigt sich: **Algorithmische Hilfe wird beansprucht, wenn andere sie auch nutzen**, unabhängig von der Anzahl der anderen. Die Akkuratheitsaussage hingegen verbessert die Nutzung nicht (Alexander et al., 2018). Allerdings verbesserte sich die Leistung der Nutzenden in der Aufgabe – es galt den Weg durch ein gezeichnetes Labyrinth zu finden – durch den Algorithmus nicht, in einer Bedingung verschlechterte sie sich sogar. Die zögerliche Nutzung in der Bedingung mit Akkuratheitsaussage war also eigentlich folgerichtig.

Ebenso wie ein sozialer Hinweis spielt auch die Präsentation der Akkuratheit eine Rolle: In einer Spracherkennungsanwendung, die gesprochene in geschriebene Wörter umwandelte, wurden durch visuelle Hinweise Angaben zur Sicherheit des Systems gemacht. In der Hälfte der Fälle ergänzte ein virtueller Avatar, der in einer Sprechblase eigentlich nur wiederholte, was bereits zu sehen war, die Darstellung. Der Avatar führte zu signifikant höherem Vertrauen in das System (Weitz et al., 2019). Will man überzeugen, lassen sich neben einer möglichst hohen zu erreichenden Akkuratheit andere Wege wählen, um die Nutzung zu erhöhen. Gleichzeitig sollte es beim **Ziel möglichst korrekter Entscheidungen nicht darum gehen, durch irrelevante Darstellungen die Nutzung in die Höhe zu treiben**, sondern zu erheben, wie Inhalte verständlich vermittelt und eine überlegte Nutzung erreicht werden kann. Deshalb werden in der vorliegenden Studie verschiedene Akkuratheitsangaben eingesetzt und verglichen. Besonders zu betonen ist die speziell für die Studie (a) „Fehlerfall“ durchgeführte Berechnung der Akkuratheitsangaben, mit dem Ziel, die postulierten Sicherheitsangaben der tatsächlichen Wahrscheinlichkeit entsprechen zu lassen (veröffentlicht in Werz et al., 2021). In Studie (c) „Transparenzarten“ werden zwei Akkuratheitsangaben mit zwei Explainability-Arten verglichen, um ihre Wirkungen auf Vertrauen und tatsächliche Nutzung zu erheben. In beiden Studien wurde darauf geachtet, dass die Hinweise der vorgeblichen Algorithmen die Leistung der Versuchspersonen verbesserten. Sich von Akkuratheitsangaben oder Explainability überzeugen zu lassen, wäre in den Experimenten also die richtige Entscheidung gewesen.

2.3.3. Zusammenhang von Transparenz und Nutzung

Die ambivalente Sicht auf Transparenz in den Sozialwissenschaften, die dem eher computerwissenschaftlichen Ansatz der Transparenz als „modern [...] ideal“ (Koivisto 2016, S. 2) gegenübersteht, spiegelt sich in den uneinheitlichen Ergebnissen von Nutzendenstudien wider, die Transparenz untersuchen. So hebt Springer (2019) hervor, „in some settings [...] transparency improves algorithmic perceptions because users may better understand system behavior“ (S. 101),

während in anderen Kontexten „transparency can have other quite paradoxical effects, […] cause users to have worse perceptions of a system, trusting it less“ (Springer, 2019, S. 101). Eine umfassende Übersicht über die analysierten Studien ist Anhang A zu entnehmen.

Erste Studien zu algorithmischen Systemen befassten sich mit den damals innovativen und neuen Empfehlungssystemen für beispielsweise Musik, Kunst oder Filme. Obwohl diese Empfehlungssysteme häufig mithilfe von „Nachbarschaft“, Ähnlichkeit oder anderen regelbasierten Algorithmen realisiert wurden und eigentlich erklärbar waren, stellten die Empfehlungen für die Nutzenden damals häufig Blackboxen dar, wie Herlocker et al. (2000) betonen. Erklärungen, wie die personalisierten Filmempfehlungen zustande kommen, wurden in den Studien beispielsweise über Anzeigen, wie „ähnliche Leute wie du“ den empfohlenen Film bewertet hatten, visualisiert. Empfehlungen mit einer solchen Erklärung wurden Empfehlungen ohne Erklärung vorgezogen, ohne jedoch die Filmauswahl zu vereinfachen (Herlocker et al., 2000). Ebenso zeigte sich ein Zusammenhang zwischen der Bewertung der Erklärungen und der Zufriedenheit mit dem Empfehlungssystem, wiederum ohne Effekt auf Auswahleffektivität (Gedikli et al., 2014). Bei einem Vergleich verschiedener Einflussfaktoren stellte sich die Webseitenqualität als wichtigster Faktor für Vertrauen in das System heraus, während Transparenz nur einer von vielen weiteren Faktoren war (Nilashi et al., 2016).

Während also Erklärungen zu ansonsten nicht nachvollziehbaren Systemen eher **positive Effekte** zeigten, demonstrierten einige Studien zum Facebook-Newsfeed-Algorithmus, die Mitte der 2010er Jahre stattfanden, negative Effekte von Transparenz: In Situationen, in denen Nutzende keine Kenntnis vom Algorithmus hatten, führte die Offenlegung seiner Existenz zunächst zu negativen Emotionen, Überraschung und Ärger (Eslami et al., 2015; Rader et al., 2018). Transparenz hatte hier also negative affektive Auswirkungen auf die Wahrnehmung des Systems. Die in all diesen Studien umgesetzte Transparenz war jedoch sehr einfach. Sie reichte von einfachen Erklärungen der Filmempfehlungen zur Information, wie ein Algorithmus die Nachrichten im Facebook-Newsfeed sortiert.

Untersuchungen mit vielschichtigeren Ansätzen, wie beispielsweise von E-Government-Webseiten und -Services, zeigten komplexere Effekte, die durch Transparenz ausgelöst werden. Bei hoher (bzw. geringer) Vollständigkeit oder Genauigkeit der Information erhöhte (bzw. verringerte) Transparenz das Vertrauen und die Nutzungsintention des Systems (Venkatesh, Thong, et al., 2016). Im Kern zeigt es jedoch, wie schon beim Facebook-Newsfeed: Transparenz steht nicht für sich allein, sondern hängt von den vermittelten Informationen ab. **Transparenz macht Probleme bzw. positive Aspekte sichtbar** und wirkt abhängig davon positiv oder negativ.

Dabei steht die Transparenz nicht nur in Zusammenhang mit den Inhalten, die sie offenlegt, sondern auch mit den Nutzenden und deren Erwartungen an ein System: So zeigten Erklärungen eines Kunst-

Empfehlungssystems bei Versuchspersonen mit erhöhter Expertise einen negativen Effekt auf die wahrgenommene Kompetenz des Systems, da sie durch die Erklärungen das System als zu vereinfacht wahrnahmen (H. Cramer et al., 2008). Erklärungen dazu, wie eingetippte Wörter automatisiert kategorisiert wurden, führten ebenfalls zu Annahmen über die Funktionsweise des Systems, die, obwohl es eigentlich korrekt arbeitete, als nicht optimal empfunden wurden, beispielsweise weil sie von den eigenen Erwartungen abwichen. Entsprechend bevorzugte nach der Nutzung die Hälfte der Versuchspersonen das intransparente System (Springer, 2019).

Weitere Untersuchungen belegen, dass zu viel Information und zu viele Details das Verständnis verringern und folglich auch das Vertrauen in ein System beeinträchtigen (Yu et al., 2017; Zhao et al., 2019). Hingegen erhöhte die Interaktion mit einem Erklärbot und dadurch die Wahl, wie viele Informationen Versuchspersonen erhalten wollten, ihr Vertrauen sowie ihr subjektives und tatsächliches Verständnis eines KI-Systems im Vergleich zu einer statischen Präsentation von Erklärungen oder keinen Erklärungen (Sun & Sundar, 2022). Diese Ergebnisse zeigen: Transparenz verändert – eher als erleichtert – das Verständnis eines Systems, teilweise mit negativem Effekt. Wichtig ist, dass die **Erklärungen zum Bedarf der KI-Nutzenden passen**.

Dieser Bedarf nach KI-Transparenz ist nicht stetig, sondern unterscheidet sich je nach Interaktionssituation. Nutzende erachten Transparenz besonders dann als sinnvoll, wenn das System fehlerhaft oder nicht übereinstimmend mit den eigenen Erwartungen wahrgenommen wurde (Springer, 2019). Bei einem medizinischen Diagnose-System stiegen durch Erklärungen die Zufriedenheit und das Vertrauen von Versuchspersonen nur in den Momenten, in denen ihre Diagnosen unklar oder (noch) falsch waren. Nachdem das System die richtige Diagnose gestellt hatte, war die Zufriedenheit unabhängig von Erklärungsart oder keiner Erklärung gleich hoch (Alam & Mueller, 2021). **Transparenz und Erklärungen werden also besonders bei Unsicherheit als wichtig** erachtet.

Den Effekt der Unsicherheit auf einen dadurch erhöhten Bedarf nach Transparenz zeigen auch folgende Studien zu Transparenz im Fehlerfall: Versuchspersonen misstrauten einem eigentlich funktionierenden System, als sie beobachteten, dass es einen Fehler machte. Folgten jedoch (mögliche) Erklärungen für den Fehler, stieg das Vertrauen wieder (Dzindolet et al., 2003). Auch Lucic et al. (2020) untersuchten die Effekte von Erklärungen zu Fehlern des Systems. Diese führten zwar zu einem erhöhten Verständnis und informierteren Umgang mit dem System, zeigten bei ihnen aber keinen Effekt auf Vertrauen und Akzeptanz. Auch wenn **im Fehlerfall eher Erklärungen verlangt** werden, scheint die Transparenz die Erfahrung nicht ungeschehen zu machen.

Der nachteilige Effekt von Transparenz auf Informiertheit, Verständnis und auf Vertrauen zeigte sich bei Schmidt et al. (2020): In einer Aufgabe zur Textklassifizierung vertrauten die Versuchspersonen bei komplexen Texten übermäßig stark auf den Algorithmus und folgten ihm sogar dann, wenn er Fehler machte. Bei uneinheitlichen Erklärungen – nicht eindeutig zuzuordnenden markierten Wörtern – verhielten sich die Versuchspersonen vermutlich unter der Annahme, der Algorithmus würde irren, gegen seine Empfehlung und damit falsch. Ebenso ließen sie sich von vermeintlich plausiblen Erklärungen vorschnell von falschen Entscheidungen des Algorithmus überzeugen. Ähnliche Ergebnisse zeigten sich auch bei Ford et al., wo Versuchspersonen eine fehlerhafte KI mit bespielbasierten Erklärungen eher als korrekt bezeichneten als ohne Erklärung (Ford et al., 2020). Erklärungen in schlecht performenden Systemen führten zu einem größeren Vertrauensverlust, als wenn keine Erklärungen vorlagen (Z. Zhang et al., 2021). Mehr Transparenz und Erklärungen sind also nicht zwangsläufig besser. Vielmehr müssen sie sinnvoll gestaltet sein, um weder zu viel noch zu wenig, sondern ein richtiges Maß an Vertrauen hervorzurufen.

Darüber hinaus lassen die präsentierten Ergebnisse vermuten, dass **Transparenz bei den Nutzenden nicht zu einer Verringerung, sondern vielmehr zu einer Steigerung des kognitiven Aufwands** führt (Du et al., 2019). In einer Studie über eine visuelle Entscheidungsaufgabe beeinflussten verschiedene Varianten einer visuellen (Heatmap-) Erklärung zwar die kognitive Belastung („cognitive load“), diese zeigte aber keinen Zusammenhang mit der Erwartung, das System produziere gute Ergebnisse (Karran et al., 2022). Wie bereits Herlocker et al. (2000) und Gedikli et al. (2014) zeigten, nehmen Nutzende also höhere kognitive Aufwände in Kauf, um ein KI-System zu verstehen.

Die bisherigen Studien beruhten überwiegend auf der Präsentation von Transparenz als inhaltlicher Information. Dies kann, wie dargelegt, die kognitiven Aufwände erhöhen. Tatsächlich sind Versuchspersonen in manchen Fällen sogar bereit, eine Gebühr zu entrichten, um die Transparenz eines Kreditsystems, das über sie urteilt, zu erhöhen (Peters et al., 2020). Und auch mehr Zeit und Aufwand sind Nutzende bereit zu investieren, um ein System zu verstehen. Häufiger als zusätzliche Erklärungen, die sie hätten auswählen können, probierten sie verschiedene Eingabe-Slider aus, um ihre Auswirkungen auf das Ergebnis zu testen und das System so zu verstehen (Tsai & Brusilovsky, 2019). Transparenz und Verständnis für ein System können also auch durch Interaktion und Ausprobieren hergestellt werden: Dabei eignet sich die sogenannte How-to-Transparenz besonders für Systeme mit weniger Eingabevariablen und einfacheren Zusammenhängen (Mohseni et al., 2021).

Transparenz muss also auf die richtige Weise umgesetzt werden. Diese richtige Weise ist, wie die beschriebenen Studien zeigen, abhängig von Systemeigenschaften und Anwendung der KI – und von den Nutzenden selbst. Auch wenn die Betrachtung der individuellen Einflussfaktoren auf die Mensch-KI-Interaktion nicht im Fokus dieser Arbeit steht, zeigen Studien den großen Einfluss der Vorerfahrung:

Diese überschreibt Erklärungen und **reduziert insbesondere dann Vertrauen und Leistung, wenn die Erklärungen nicht zur Vorerfahrung passen**. Auch die **Einstellung gegenüber Algorithmen und KI spielt eine wichtige Rolle**: Eine positive (vs. negative) Einstellung führt zu einem höheren (vs. geringeren) Vertrauen gegenüber einem System und überschreibt auch verschiedene Transparenzeffekte (Molina & Sundar, 2022; Sundar, 2020). Um Effekte aus dieser Richtung bei den hier durchgeführten quantitativen Studien (a) „Fehlerfall" und (c) „Transparenzarten" auszuschließen, wurde ein Design und eine Aufgabe gewählt, die möglichst wenig Vorerfahrung oder Einstellungen mit sich bringen sollten.

Fragt man Nutzende vor Interaktion mit entweder einem intransparenten oder einem transparenten System, welches sie voraussichtlich bevorzugen werden, wählen sie das transparente System: In Interviews über eine Vogelerkennungs-App, für die verschiedene Transparenz-Arten präsentiert wurden, gaben potentielle Nutzende an, Systeme mit Erklärungen denen ohne zu bevorzugen, mit einer Präferenz für visuelle Erklärungen. Ihre Annahme war, die Erklärungen zu nutzen, um ihr Vertrauen in die App zu steuern und einen sinnvollen Umgang mit ihr zu lernen (S. S. Y. Kim et al., 2023). In dieser qualitativen Studie zeigte sich, eher als generelle Erklärungen bevorzugten die Befragten die Erklärungen, die die Einzelteile, auf denen die Entscheidung beruht, darstellten. Forschungsfrage (b) „Nutzendenanforderungen" nutzt mit ihrem qualitativen Ansatz eine ähnliche Methodik wie Kim et al. (2023) und diskutiert speziell erstellte Frontends von drei KI-Apps. In Studie (b) stehen jedoch mehr die Anforderungen an die Transparenz im Vordergrund als ihre Auswirkungen, auf die die Studie von Kim et al. fokussierte. Zu diesem Ziel stehen drei unterschiedliche KI-Apps zur Diskussion, um so Unterschiede der Anforderungen je nach KI-Eigenschaften ermitteln zu können.

Wie die aufgeführten Beispiele verdeutlichen, ergeben sich aus Aussagen aus Befragungen zur Nutzung und tatsächlichen Nutzungsstudien häufig divergierende Ergebnisse. Um einem System vertrauen zu können, wünschen sich Nutzende Erklärungen, wenn man sie direkt danach fragt. In quantitativen Studien sind die Zusammenhänge nicht so eindeutig positiv. Gleichzeitig ist der Widerspruch nicht so deutlich, wie er scheinen mag: Die bloße Steigerung von Vertrauen sollte nicht das Ziel sein, vielmehr gilt es, Nutzende zu informieren und zur Regulierung ihres Vertrauens zu ermächtigen. Die qualitativ geäußerte Präferenz von Transparenz bedeutet eigentlich die Forderung nach sinnvoller, verständlicher und nutzbarer Transparenz. Entsprechend ist es wichtig, in den Nutzungsstudien zu Transparenz, also Forschungsfrage (a) „Fehlerfall" und (c) „Transparenzarten", zu ermitteln, ob die Nutzung überhaupt zu besseren Ergebnissen führte, eine Steigerung der Nutzung durch Transparenz also sinnvoll war.

Gedikli et al. unterscheiden zwischen objektiver und subjektiver (user-perceived) Transparenz. Erstere „reveals the actual mechanisms of the underlying algorithm. However, there might be a number of

reasons why it might be more appropriate to present more user-oriented 'justifications' […]" (Gedikli et al., 2014, S. 369). Während in vielen XAI-Studien die Endnutzenden, die insbesondere im Alltagskontext häufig KI-Laien darstellen, sehr wenig Beachtung finden, zeigen die Ergebnisse der Literaturübersicht, die Vielfalt der Studien, abhängiger Variablen und Ergebnisse (siehe Anhang A). Um die qualitativ erhobenen Aussagen aus Forschungsfrage (b) „Nutzungsstudien" zu ergänzen, vergleicht die Forschungsfrage (c) „Transparenzarten" in einem experimentellen Setting vier verschiedene Transparenzarten: Dabei wird eine visuelle, lokale Erklärung – wie auch in Kim et al. (2023) – mit einer textlichen Erklärung zum Algorithmus verglichen, also eine lokale mit einer globalen Funktionalitätserklärung, dem Kern von XAI. Zusätzlich werden eine lokale und eine globale Akkuratheitsangabe verglichen. Es ergibt sich also ein 2 (Funktionalität vs. Akkuratheitsangabe) x 2 (lokal vs. global)-Design.

Bisherige Studien zur Frage, welche Art der Erklärung, lokal oder global, für Endnutzende wichtiger ist, zeigen eine **Tendenz zu lokaler Explainability**. Während Rudin (2018) argumentiert, je wichtiger die zu lösende Aufgabe, desto wichtiger seien Nutzenden globale Erklärungen, bestätigt sich diese Annahme in einer empirischen Untersuchung nicht (Wanner et al., 2022). Bei einer medizinischen Diagnose-KI schützten lokale, nicht aber globale Erklärungen zum Diagnoseprozess im Falle einer zunächst falschen Diagnose vor Unzufriedenheit (Alam & Mueller, 2021). In einer weiteren Studie befanden Endnutzende die lokalen Erklärungen als am aussagekräftigsten und mit der höchsten Erklärkraft, während sich die How-, also die globale Erklärung, nicht von der Bedingung ohne Erklärung abgrenzte (Herm et al., 2023). Diese Ergebnisse gehen einher mit den zuvor berichteten Präferenzen der Interviewten in einer Vogelbestimmungsapp, die Erklärungen zu den Einzelheiten der jeweiligen Empfehlungen den generellen Erklärungen bevorzugten (S. S. Y. Kim et al., 2023).

Der Blick in die Details zeigt jedoch auch hier die Abhängigkeit der Erklärung von ihren Effekten: Bei Herm et al. (2023) wurden visuelle Erklärungen genutzt, die lokal sehr aussagekräftig waren, während die globale visuelle Erklärung kaum Information enthielt. Ähnliches zeigte eine Untersuchung mit Wirtschaftsstudierenden, die ein Unterstützungssystem zu Vorhersagen über Produktnachfragen nutzten. Globale Erklärungen halfen, die Prozesse des Algorithmus zu verstehen, jedoch legten sie offen, wie simpel der Algorithmus rechnete, weshalb die Nutzenden den Rat als weniger wertvoll einschätzten, ihn seltener nutzten und schlechter abschnitten, als diejenigen, die keine Erklärung erhielten (Lehmann et al., 2020). Die Studie zeigte also weniger den negativen Effekt von globalen Erklärungen als den negativen Effekt verletzter Erwartung an die Funktionen des Systems (H. Cramer et al., 2008; Springer, 2019). Eine Richtung ist in der Frage, **ob lokale oder globale Erklärungen besser wirken, also nicht abschließend geklärt**. Entsprechend widmet sich Forschungsfrage (c) „Transparenzarten" diesem Thema.

Die hier aufgeführten Studien zeigen zusammenfassend: Wichtiger als Transparenz im kleinsten Detail herzustellen, ist es, sie auf relevante Erklärungen und aussagekräftige Teile zu beschränken. Diesbezüglich fordern Carvalho et al. (2019, S. 27) für die Forschung zu Erklärbarkeit:

> „The […] research field needs to focus more on the comparison of existing explanation methods instead of just creating new ones: only with interpretability assessment, including metrics that help to measure interpretability and context definitions that assist in comparing different use cases, can we know in which direction the explanation methods should aim."

Während in Forschungsfrage (a) „Fehlerfall" dazu eine sehr einfache Art der Transparenz, die Akkuratheitsangabe, für einen spezifischen Nutzungsfall gewählt wurde, dient Forschungsfrage (b) „Nutzendenanforderungen" der breiteren Annäherung an das Thema, welche Transparenzaspekte überhaupt als relevant angesehen werden. Zuletzt sind die Transparenzarten, die in Forschungsfrage (c) „Transparenzarten" verglichen werden, sehr unterschiedlich gestaltet, um verschiedene Aspekte von Transparenz, nämlich Erklärbarkeit sowie Akkuratheitsangaben zu vergleichen, aber gleichzeitig Einflüsse wie Kontext oder Vorwissen konstant zu halten (siehe Kapitel 2.2.4 „Technische Ansätze zu transparenter KI").

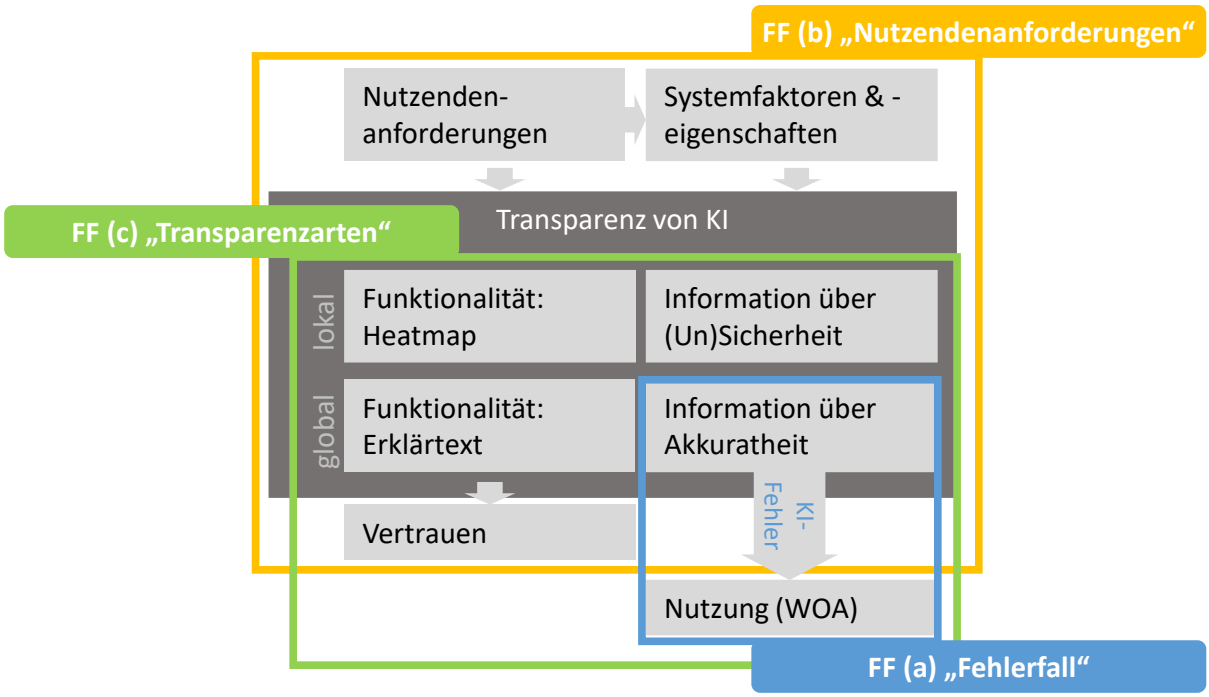
2.4. Zusammenfassend: Fragestellungen

Aufbauend auf den dargelegten Vorarbeiten und theoretischen Überlegungen widmet sich die vorliegende Arbeit der übergeordneten Fragestellung, **wie sich Transparenz von KI-Entscheidungsunterstützungssystemen auf die Nutzung dieser Systeme durch Endnutzende auswirkt**. Dabei wird diese Fragestellung mit einem Mixed Method-Design in drei verschiedenen Detailgraden untersucht:

a. Forschungsfrage „Fehlerfall": Inwieweit führen Angaben von Akkuratheit eines Algorithmus dazu, dass dieser auch nach einem Fehlerfall genutzt wird?
b. Forschungsfrage „Nutzendenanforderungen": Welche Anforderungen an Transparenz in KI bestehen für Laiennutzende und inwiefern unterscheiden sie sich nach Eigenschaften der KI?
c. Forschungsfrage „Transparenzarten": Wie wirken sich verschiedene Arten der KI-Transparenz auf Vertrauen und Nutzung eines Systems aus?

Abbildung 10 stellt dar, wie die drei Forschungsfragen auf dem Feld der Transparenz in KI anzuordnen sind und miteinander in Bezug stehen.

Abbildung 10: *Darstellung des Untersuchungsgegenstands der vorliegenden Arbeit und der in den Forschungsfragen adressierten Aspekte*



3. Methodisches Vorgehen der Arbeit und ihr Beitrag

Um die übergeordnete Fragestellung, **wie sich Transparenz von KI-Entscheidungsunterstützungssystemen auf die Nutzung dieser Systeme durch Endnutzende auswirkt,** anzunähern, wurden drei untergeordnete Forschungsfragen und jeweils drei verschiedene Blickwinkel gewählt. Dieses gewählte Mixed Method-Design dient dazu, ein breites Verständnis zur Fragestellung durch verschiedene Ansätze zu erhalten.

Nach Venkatesh und Kolleginnen (2016) gilt es zu Beginn einer Untersuchung, das Ziel des Mixed Method-Designs festzulegen, wonach sich die vorliegende Arbeit einem „**complementary purpose**" zuordnen lässt, bei dem „complementary views about the same phenomenon" gewonnen werden (S. 38). In einer frühen Kategorisierung von Mixed Method-Designs von Greene et al. (1989), wird Complementarity charakterisiert durch eine Kombination verschiedener Methoden, die sich mit **Dimensionen des gleichen Phänomens** eher als einem einzigen Phänomen befassen (im Gegensatz zu z. B. einer „Triangulation", bei der verschiedene Methoden eingesetzt werden, um ein einzelnes Phänomen sehr detailliert zu untersuchen und die Reliabilität zu erhöhen). Das Phänomen Transparenz in KI wird durch den vorliegenden Ansatz in drei Dimensionen zerlegt: Der Ansatz (a) „Fehlerfall" widmet sich einem technisch sehr einfach umzusetzenden Transparenzaspekt, Akkuratheit, um dessen Auswirkungen in einem (häufig replizierten) Fall der Mensch-KI-Interaktion quantitativ, in kontrollierten Bedingungen und durch Manipulation der Transparenz genau zu untersuchen. Der Ansatz (b) „Nutzendenanforderungen" ist sehr viel breiter angesetzt und untersucht qualitativ in Fokusgruppen, welche Anforderungen an Transparenz KI-Laien äußern, um so die Anforderungen abhängig von Systemgegebenheiten ableiten zu können. Hingegen vergleicht Ansatz (c) „Transparenzarten" quantitativ vier aus der Literatur abgeleitete Transparenzarten in ihrem Effekt auf die Nutzung von Laien. Der Ansatz (c) ist damit in Bezug auf die **Spezifität** zwischen Ansatz (a), der als äußerst spezifisch betrachtet werden kann, und dem weniger spezifischen Ansatz (b) einzuordnen.

Die Forschungsfragen wurden **unabhängig** voneinander in drei Phasen („Multistrand Design") und **gleichzeitig** (concurrent) untersucht (Greene et al., 1989; Venkatesh, Brown, et al., 2016). Sie bauen also nicht aufeinander auf, sondern beleuchten parallel drei Aspekte der übergeordneten Fragestellung. Gleichzeitig ist zu betonen, dass die drei Untersuchungen nicht zeitlich parallel stattfanden, sondern zunächst Studie (a), dann (b) und dann (c) durchgeführt wurden.

Der Kategorisierung von Venkatesh und Kolleginnen folgend ist zu berichten, dass **zwei qualitative** und **eine quantitative** Studie durchgeführt wurden, gleichzeitig aber die Methoden **gleichwertig** sind und keine Methode bzw. kein Ergebnis das andere dominiert. Die Teilnahme an den drei Studien war unabhängig in dem Sinne, dass die Teilnehmenden in Studie (a), (b) und (c) unterschiedlich waren und

nicht mehrfach teilgenommen haben (es wurde in der Rekrutierung darauf hingewiesen, Teilnehmende von einer der anderen Studien mögen bitte nicht erneut teilnehmen). Es handelt sich also um „**Probability Sampling**“ mit parallelen Stichproben (Venkatesh, Brown, et al., 2016). Während die Studien (a) „Fehlerfall“ und (c) „Transparenzarten“ einem deduktiven, hypothesentestenden Ansatz folgen, zeichnet sich Studie (b) „Nutzendenanforderungen“ durch einen überwiegend induktiven, explorativen Ansatz aus.

Mit dieser Herangehensweise widmet sich die vorliegende Arbeit der **Forschungslücke**, die zwischen technisch detaillierter Forschung und der Nutzung dieser technischen Erkenntnisse für die Anwendung für Endnutzende besteht. Ihr Beitrag besteht in der Auseinandersetzung damit, wie sich die bisher sehr technisch behandelte Thematik der Transparenz in KI für Endnutzende übersetzen lässt. Durch die breite Herangehensweise bezieht die Arbeit sich dabei nicht nur auf spezifische Fragestellungen nach einzelnen Zusammenhängen oder Phänomenen, wie z. B. der Algorithm Aversion, sondern schafft mit drei Ansätzen einen Überblick über das Feld der KI-Transparenz und ihre verschiedenen Facetten. So sollen nicht nur Einzelergebnisse produziert werden, die lediglich ein weiteres Mosaiksteinchen zum ambivalenten Bild von KI-Transparenz und ihrer Auswirkung auf die Nutzung beitragen. Vielmehr sollen die Ergebnisse am Ende ein **breites Bild von KI-Transparenz** ergeben, das die Sicht von Laien-Nutzenden darlegt, in verschiedenen Kontexten, mit ihren Abhängigkeiten und Effekten auf Vertrauen und Nutzung. Ziel ist es, auf Grundlage der Erkenntnisse Handlungsempfehlungen für Entwickler*innen und die Politik formulieren zu können, die diese bei der Umsetzung von KI-Transparenz für Endnutzende unterstützen.

Die drei Ansätze werden im Folgenden nacheinander, der zeitlichen Durchführung folgend, berichtet. Jeder Studie ist eine kurze, zusammenfassende theoretische Einordnung vorangestellt, gefolgt von Methodik, Ergebnissen und Diskussion dieses Ansatzes. Um das Gesamtbild von KI-Transparenz zusammenzusetzen, werden am Ende der Arbeit die Erkenntnisse der drei Studien zusammengeführt. Dies geschieht insbesondere in Kapitel 7 „Einordnung der Ergebnisse und Diskussion“ mit einem Überblick über die Studien, den zu ziehenden gemeinsamen Erkenntnissen und der ausführlichen Ableitung von Implikationen für Praxis und Forschung. Am Ende folgt Kapitel 8 mit einer Zusammenfassung der gesamten Arbeit und schließlich Kapitel 9 mit Fazit und Ausblick.

4. Effekt der Akkuratheitsinformation auf die Nutzung nach einem Fehlerfall (Forschungsfrage a)

Bezüglich der Nutzung von KI-Systemen treten zwei vermeintlich gegenläufige Effekte auf: Overtrust und Algorithm Aversion. Beide zeigen sich in Studien und in der Praxis und wirken sich negativ auf Entscheidungsprozesse aus. Ersteres bezeichnet übermäßiges Vertrauen in KI bzw. ein algorithmisches System, das in der Folge zu wenig kritisch genutzt wird und beispielsweise Fehler unüberlegt übernommen werden (Madhavan & Wiegmann, 2007; Springer et al., 2018). Im Gegensatz dazu bezeichnet Algorithm Aversion die übermäßige Ablehnung algorithmischer Systeme nach einem Fehlerfall, selbst wenn das System insgesamt zu besseren Entscheidungen verhelfen würde (Dietvorst et al., 2014; Filiz et al., 2023). Da beide Effekte rationaler, bestmöglicher Entscheidungsfindung im Weg stehen, sind Maßnahmen zur Regulierung vonnöten.

Beide Effekte fußen, so die Annahme, in der falschen Annahme, Algorithmen irrten sich nie. Im Folgenden wird sich auf den Ablehnungseffekt, Algorithm Aversion, konzentriert und untersucht, ob er sich reduzieren lässt, wenn Transparenz darüber hergestellt wird, dass jeder Algorithmus eine gewisse Unsicherheit beinhaltet. Die dazu durchgeführte quantitative Studie wird im Folgenden berichtet.

4.1. Theoretische Einordnung und Hypothesen

Der Effekt der Algorithm Aversion bezeichnet den Nutzungseinbruch und Vertrauensverlust in ein Entscheidungsunterstützungssystem, nachdem Nutzende Fehler des Systems beobachteten. Ausführlich eingeführt und besprochen wird der theoretische Hintergrund dieses Effekts und seiner Auswirkungen in Kapitel 2.1.4. Das Problematische an Algorithm Aversion: Da Algorithmen bzw. KI irgendwann zwangläufig Fehler begehen, scheint er unvermeidlich (Prahl & Swol, 2017).

Ein Grund für Algorithm Aversion liegt darin, dass algorithmische Empfehlungen anders bewertet werden als menschliche (Burton et al., 2020; Logg et al., 2019; Morewedge, 2022; Renier et al., 2021). Aufgrund zu hoher Erwartungen an Algorithmen verletzen aufkommende Fehler die herrschende Perfektheitsannahme und führen deshalb zu großer Ablehnung (Daschner & Obermaier, 2022; Dietvorst et al., 2014; Dzindolet et al., 2003; Madhavan & Wiegmann, 2007; Renier et al., 2021).

Die Annahme, Algorithmen arbeiteten perfekt, zeigt sich auch dann, wenn ein einzelner Fehler eines KI-Systems von den Nutzenden als Zeichen dafür interpretiert wird, das System sei defekt (Reich et al., 2022). Entsprechend sind Nutzende eher gewillt, das System weiterhin zu nutzen, wenn sie die Möglichkeit erhalten, die Entscheidungen des Algorithmus nach einem Fehler – und sei es nur minimal – anzupassen (Dietvorst et al., 2018). Wirkungsvoll zur Überwindung von Algorithm Aversion ist

außerdem, den Nutzenden zu zeigen, das System lerne durch Fehler hinzu (Reich et al., 2022). Gleichzeitig stellt sich die Frage, ob nicht schon allein die Aufklärung darüber, dass einzelne Fehler vorkommen können, ja gar zu erwarten sind und keine Anzeichen für einen fehlerhaften Algorithmus darstellen, dazu beitragen könnte, Algorithm Aversion zu überwinden. Die vorliegende Arbeit argumentiert, schon die Information, KI und Algorithmen hätten zwangsläufig bestimmte Fehlerwahrscheinlichkeiten, könnte diesen schädlichen Perfektheitsanspruch ins rechte Licht rücken und so einen Ansatz zur Überwindung von Algorithm Aversion bieten. Daraus ergibt sich die Forschungsfrage:

> **FF (a) „Fehlerfall“: Inwieweit führen Angaben von Akkuratheit eines Algorithmus dazu, dass dieser auch nach einem Fehlerfall genutzt wird?**

Da es sich bei Algorithm Aversion um einen Effekt handelt, der nicht rational begründbar ist, wird in den meisten Studien zur Algorithm Aversion keine Verhaltensintention, sondern die tatsächliche Nutzung von Algorithmen erhoben. Denn auch wenn berichtetes Vertrauen oder Verhaltensintention wichtige Voraussetzungen für eine Technologienutzung sind, führen sie nicht automatisch zu tatsächlichem Nutzungsverhalten (Ajzen, 2014; Daschner & Obermaier, 2022; Schmidt et al., 2020). In Experimenten zu Algorithm Aversion wird häufig die Nutzung von algorithmischen Ratschlägen vor und nach einem Fehler miteinander oder mit der Nutzung von menschlichen Ratschlägen verglichen (Burton et al., 2020; Logg et al., 2019; Mahmud et al., 2022; Önkal et al., 2009). Als Algorithm Aversion werden in diesen Studien sowohl die signifikante Verringerung der Nutzung nach einem Fehler im Vergleich zur Nutzung vorher als auch eine signifikant niedrigere Nutzung algorithmischer im Vergleich zu menschlichen Empfehlungen nach einem Fehler bezeichnet.

Das in solchen Untersuchungen häufig eingesetzte Experimentaldesign ist das eines Judge-Advisor-Systems (JAS; Sniezek & Buckley, 1995). Ein JAS bildet eine Interaktion ab, in der die Person, die eine Entscheidung trifft, als Judge und die Quelle eines Ratschlags zu dieser Entscheidung als Advisor bezeichnet werden. In den Studien zur Algorithm Aversion handelt es sich bei den Advisor üblicherweise um verschieden gestaltete Algorithmen. Teilweise werden auch algorithmische und menschliche Advisor verglichen (z. B. Dietvorst et al., 2014; Logg et al., 2019; Prahl & Swol, 2017). Das JAS ermöglicht die Untersuchung von Faktoren, die beeinflussen, unter welchen Bedingungen und wie sehr der Judge die Entscheidungsempfehlungen des Advisors in die eigene Entscheidung integriert. Im Experimentaldesign trifft bei einer Entscheidungsaufgabe zunächst der Judge eine Entscheidung. Anschließend folgt die Empfehlung des Advisors, woraufhin der Judge seine Entscheidung anpassen kann und damit eine endgültige Entscheidung abgibt (Bonaccio & Dalal, 2006; Sniezek & Buckley, 1995). Anhand dieser wechselseitigen Interaktion lässt sich durch die Differenz zwischen erster und finaler Entscheidung der Einfluss der Empfehlung des Advisors auf die Entscheidung des Judges

quantifizieren. Dieser Messwert wird, seit ihn Dawes und Corrigan (1974) zum ersten Mal nutzten, als Weight of Advice (WOA) bezeichnet (z. B. Bonaccio & Dalal, 2006; Daschner & Obermaier, 2022; Logg et al., 2019; Önkal et al., 2009).

In diesem Experimentalparadigma lassen sich die Bedingungen, unter denen Entscheidungen getroffen werden, ebenso wie die Advisor selbst variieren: Der Vergleich von menschlichen mit algorithmischen Advisor wurde bereits angesprochen. Für die vorliegende Fragestellung hingegen werden verschieden gestaltete algorithmische Advisor verglichen, die sich hinsichtlich ihrer Akkuratheitsangaben unterscheiden. Dazu werden, aufbauend auf den vorangegangenen Darstellungen, die folgenden Hypothesen aufgestellt:

> **H1:** Die Nutzung des Algorithmus (WOA) nach Fehlern sinkt stärker, wenn zuvor keine Akkuratheitsangaben gegeben wurden als mit Akkuratheitsangaben.
>
> **H2:** Im Vergleich zu Durchgängen ohne Fehlererfahrung sinkt die Nutzung eines Algorithmus (WOA) nachdem Fehler von diesem erlebt wurden (Algorithm Aversion).
>
> **H3:** Nach einem Fehler ist die Nutzung des Algorithmus (WOA) höher, wenn der Algorithmus Akkuratheitsangaben aufweist als ohne solche Angaben.

Um diese Hypothesen zu prüfen, wurde ein Experiment aufgesetzt, in dem die Akkuratheitsangaben bzw. ihr Vorhandensein manipuliert wurden. Dazu wurde die positive Formulierung von Akkuratheit gewählt („xx % korrekt“), da ein Vergleich mit negativ formulierten Fehlerwahrscheinlichkeiten eine geringere Akzeptanz gezeigt hatte (T. Kim & Song, 2020). In der Literatur wurden darüber hinaus Zusammenhänge zwischen der Einstellung einer Person gegenüber Algorithmen und ihren mathematischen Fähigkeiten (Logg et al., 2019) oder auch ihrem Bildungsniveau (Thurman et al., 2019) gefunden. Da die Einstellung einer Person gegenüber Algorithmen auch Auswirkungen auf die Nutzung der Algorithmen hat, wird in der vorliegenden Studie zudem explorativ untersucht, ob weitere Persönlichkeitseigenschaften Auswirkungen auf den Algorithm Aversion-Effekt zeigen. Es wird ein Zusammenhang zwischen der allgemeinen Einstellung einer Person gegenüber Algorithmen, ihrer Kontrollüberzeugungen im Umgang mit Technik, Risikobereitschaft, Kognitionsbedürfnis, Entscheidungsfreude und weiteren Einzelfaktoren untersucht. Die Details des Experimentaldesigns und die Rahmenbedingung zur Erhebung sowie zur Auswertung werden im Folgenden näher erläutert.

4.2. Methode

Die Untersuchung, ob sich durch Transparenz über die Algorithmusakkuratheit die Nutzung des Algorithmus auch nach einem Fehlerfall erhalten und Algorithm Aversion abschwächen lässt, fand im Rahmen eines Online-Experiments statt. Dabei hatten die Versuchspersonen die Aufgabe, das Gewicht

von auf Fotos abgebildetem Gemüse bzw. Obst zu schätzen (im Folgenden nur noch Gemüse genannt). Nach Abgabe einer initialen Gewichtschätzung wurde ihnen die Einschätzung eines zuvor beschriebenen algorithmischen Systems präsentiert. Bei diesem als KI dargestellten System handelte es sich um ein „Wizard-of-Oz-System", bei dem der angebliche Algorithmus in Wahrheit fest programmiert war und keine Gewichtschätzung durchführte. Mehr Informationen zum Ablauf und zum verwendeten Material folgen in den nächsten Kapiteln.

Das Studiendesign wurde zum ersten Mal in Werz et al. (2020) beschrieben und vorgestellt. Die Datenerhebung fand u. a. im Rahmen der Masterarbeit von Jacqueline Engels statt, die ihre Abschlussarbeit in Betreuung der Autorin durchführte. Die Studie wurde vor der Durchführung und Analyse präregistriert[3] (siehe auch Anhang B).

4.2.1. Stichprobe

Am Online-Experiment nahmen insgesamt 257 Personen teil, wobei nicht vollständig abgeschlossene Durchführungen schon zu Beginn ausgeschlossen wurden. Weitere 88 Datensätze mussten ausgeschlossen werden, weil eine als Manipulationscheck eingefügte Kontrollfrage fehlerhaft beantwortet wurde. Die Stichprobe bestand schließlich aus n = 169 Personen zwischen 18 und 68 Jahren (M = 29,43, SD = 10,95). Von diesen waren 100 Teilnehmerinnen weiblich. Weitere Daten zur Stichprobebeschreibung stellt Tabelle 3 dar.

***Tabelle 3**: Demographische Daten der Stichprobe zur Forschungsfrage (a)*

	Häufigkeit	Prozent
Arbeitsverhältnis		
Student/in	77	45,6%
Angestellte/r	76	45%
Freiberufler/in	12	7,1%
Arbeitslos	1	0,6%
Rentner/in	3	1,8%
Höchster Bildungsabschluss		
Hauptschulabschluss	1	0,6%
Realschulabschluss	7	4,1%
Abitur	41	24,3%
Lehre/Berufsausbildung	23	13,6%
Universitäts- oder Fachschulabschluss	97	57,4%
Muttersprache		
Deutsch	163	96,4%
Nicht Deutsch	6	3,6%
Routine mit Computern		
Gar nicht	2	1,2%
Ein wenig	45	26,6%
Sehr	122	72,2%

[3] https://aspredicted.org/5zb9y.pdf

Die Rekrutierung erfolgte über verschiedene E-Mail-Verteiler der Psychologiestudierenden der RWTH Aachen, über den Verteiler der Mitarbeitenden des IMA der RWTH Aachen sowie über persönliches Kontaktieren im Bekanntenkreis. Teilnahmevoraussetzung war ein Mindestalter von 18 Jahren. Als Aufwandsentschädigung erhielten die Psychologiestudierenden eine halbe Versuchspersonenstunde. Unter allen anderen Teilnehmenden wurden drei 10 €-Gutscheine für Amazon verlost.

4.2.2. Versuchsablauf

Die Online-Studie wurde mittels der Online-Plattform SoSci Survey realisiert und den Teilnehmenden zwischen dem 01.04.2021 und dem 18.04.2021 zur Verfügung gestellt. Die Dauer der Teilnahme an der Online-Studie betrug ca. 15 bis 20 Minuten.

Zu Beginn des Experiments erfolgte die Begrüßung der Versuchspersonen. Nach Sichtung und Einwilligung der Informationen zum Datenschutz (siehe Anhang C) startete der inhaltliche Teil. Den gesamten Ablauf der Studie stellt Abbildung 11 im oberen Prozessdiagramm dar. Zunächst füllten die Versuchspersonen die Skala zur Erhebung der Einstellung gegenüber Algorithmen aus (Attitude Towards Algorithm-Scale, ATAS; Bock & Rosenthal-von der Pütten, 2023). Anschließend wurde ihnen erklärt, die Studie untersuche einen Algorithmus, der der kontaktlosen Gewichtsbestimmung von Gemüse und Obst dient und auf Grundlage einzelner Bilder dieses Gewicht mittels Bilderkennung sowie Volumen- und Dichteberechnungen bestimme. An der Stelle wurden durch drei verschiedene Akkuratheitsangaben zum Algorithmus die drei unterschiedlichen Zwischensubjektfaktor-Bedingungen eingeführt. Anschließend erfolge die randomisierte Zuteilung der Versuchspersonen zu einer der drei Bedingungen:

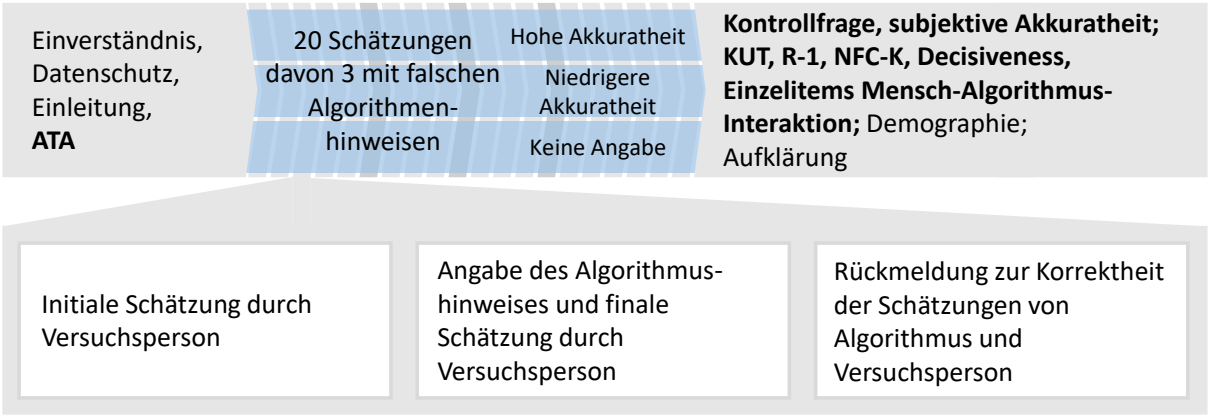
- **Hohe Akkuratheitsbedingung (HA)** mit der Information, der Algorithmus weise eine Genauigkeit von **90,1 %** auf
- **Niedrige Akkuratheitsbedingung (NA)** mit dem Hinweis, der Algorithmus weise eine Genauigkeit von **78,9 %** auf
- Bedingung ohne Genauigkeitsangabe der **Kontrollgruppe (KG): keine Information** über die Genauigkeit des Algorithmus

Die darauffolgenden Instruktionen waren für alle Versuchspersonen identisch. Sie erhielten die Aufgabe, anhand von Bildern das Gewicht von Gemüse zu schätzen. Nach dieser ersten eigenen Angabe erhielten sie die Schätzung des Algorithmus. Daran anschließend hatten sie die Möglichkeit, ihre initiale Einschätzung anzupassen und so ihre finale Schätzung abzugeben. Abbildung 11 stellt in ihrem unteren Teil den Ablauf eines einzelnen Trials dar. Abbildung 12 zeigt die Schätzaufgabe. Nach jeder abgeschlossenen Schätzaufgabe wurde den Versuchspersonen Rückmeldung dazu gegeben, welche Schätzungen korrekt gewesen waren: ihre finale und/oder ihre erste Schätzung sowie die des

Algorithmus (eine Schätzung galt als korrekt, wenn sie sich innerhalb von +/- 20 % des wahren Gewichts befand; siehe Abbildung 11 bzw. Abbildung 13). Die Versuchspersonen durchliefen diese Interaktion in 20 randomisierten Durchgängen, in denen sie jeweils ein Bild aus einem Pool von 37 Bildern zur Schätzung erhielten.

Von den insgesamt 20 Durchgängen gab der Algorithmus in drei Durchgängen (im 6., 10. und 14.) eine fehlerhafte Empfehlung. Bei den Empfehlungen der anderen Durchgänge lag der Algorithmus richtig. Um ein Diskrepanzerleben zwischen der angegebenen Genauigkeit des Algorithmus und der beobachteten Leistung zu vermeiden, wurden die Werte der algorithmischen Akkuratheit (90,1% und 78,9%) sowie die Anzahl der richtigen und falschen Schätzungen des Algorithmus so gewählt, dass sie den angegebenen Akkuratheitsangaben entsprachen (Werz et al., 2021).

Abbildung 11: *Ablauf des Experiments (a). Der obere Teil zeigt den gesamten Ablauf: drei dunkelgraue Pfeile stehen für drei falsche Trials. In fetter Schrift sind Skalen/Items dargestellt, die in 4.2.3.2 näher erläutert werden. Der Detailblick darunter zeigt den Ablauf eines einzelnen Trials.*



Nach der Schätzaufgabe folgte die Kontrollfrage zum Manipulationscheck der Akkuratheitsbedingungen. Dazu sollten die Versuchspersonen anhand eines Drop-Down-Menüs angeben, in welchem Bereich die Sicherheit des algorithmischen Systems laut den Anweisungen gelegen hatte (Antwortoptionen: „es gab keine Angabe zur Sicherheit“, „55-64 %“, „65-74 %“, „75-84 %“, „über 85 %“). Personen, die hierbei falsche Angaben machten, wurden bei der Auswertung ausgeschlossen, da für sie von keiner erfolgten Manipulation ausgegangen werden konnte. Zudem sollten die Versuchspersonen mittels eines Schiebereglers angeben, wie sie die tatsächliche Trefferquote des Algorithmus einschätzten (von „0% (nie richtig)“ bis „100% (immer richtig)“).

Im Anschluss wurden die Versuchspersonen gebeten, verschiedene Items zur Mensch-Algorithmus-Interaktion zu beantworten (siehe Kapitel 4.2.3.3), die Einschätzungen über den Algorithmus enthielten und die Motivation der Algorithmusnutzung abfragten. Zusätzlich beschrieben die Versuchspersonen mittels eines offenen Textfeldes kurz ihre Vorgehensweise in der Zusammenarbeit mit dem Algorithmus. Ziel dieser Items sowie der offenen Abfrage war es, die gewählte Strategie in der

Zusammenarbeit mit dem Algorithmus zu erfassen. Danach sollten sie den Fragebogen zur Kontrollüberzeugung im Umgang mit Technik (KUT; Beier, 1999, 2004), die Kurzskala zur Risikobereitschaft (R-1; Beierlein et al., 2014), die Kurzskala des Need for Cognition (NFC-K; Beißert et al., 2015) sowie die Subskala Decisiveness (Roets & Van Hiel, 2007) ausfüllen (siehe Kapitel 4.2.3.3).

Anschließend wurden demografische Daten (Alter, Geschlecht, höchster Bildungsabschluss, Arbeitsverhältnis, Muttersprache, Routine im Umgang mit Computern) abgefragt. Die Versuchspersonen wurden außerdem aufgefordert, in einem offenen Textfeld Algorithmen zu benennen, die sie in ihrem beruflichen oder privaten Alltag verwenden, um so Rückschlüsse auf ihre bisherigen Erfahrungen mit KI treffen zu können. Auf der darauffolgenden Seite wurde den Versuchspersonen für ihre Teilnahme an der Studie gedankt und sie wurden über das eigentliche Ziel und die Manipulation der Studie aufgeklärt. Abschließend konnten die Versuchspersonen ihren E-Mail-Kontakt zur Incentivierung bzw. Teilnahme an der Verlosung sowie für den Erhalt weiterer Informationen zur Studie angeben.

4.2.3. Material

Im Folgenden werden das verwendete Material für die Schätzaufgabe, die Erhebung der abhängigen Variablen und die Manipulation der Bedingungen und abschließend die weiteren Erhebungsinstrumente im Detail vorgestellt.

4.2.3.1. Schätzaufgabe und abhängige Variable

Zur Untersuchung der Frage, ob die Angabe der algorithmischen Akkuratheit die Algorithm Aversion reduzieren kann, kam ein JAS zum Einsatz. Dieses wurde als Wizard-of-Oz-Experiment umgesetzt. Das bedeutet, die Versuchspersonen wurden Glauben gemacht, sie interagierten mit einem automatisierten/KI-basierten System, das in Wahrheit aber keines war (Dahlbäck et al., 1993; Gu et al., 2024). Im vorliegenden Fall handelte es sich nicht um ein echtes Vorhersagemodell zur Bildanalyse, sondern um manuell und in Abhängigkeit vom wahren Gewicht definierte Gewichtangaben. Konzept und Design der Studie wurden erstmalig in Werz et al. (2020) dargelegt.

Bei der Entscheidungsaufgabe handelte es sich um eine Schätzaufgabe, in der das Gewicht von auf Fotos abgebildetem Gemüse geschätzt werden musste (siehe Abbildung 12). Diese Aufgabe wurde gewählt, da sie drei Bedingungen erfüllt, die für die vorliegende Untersuchung als notwendig erachtet wurden:

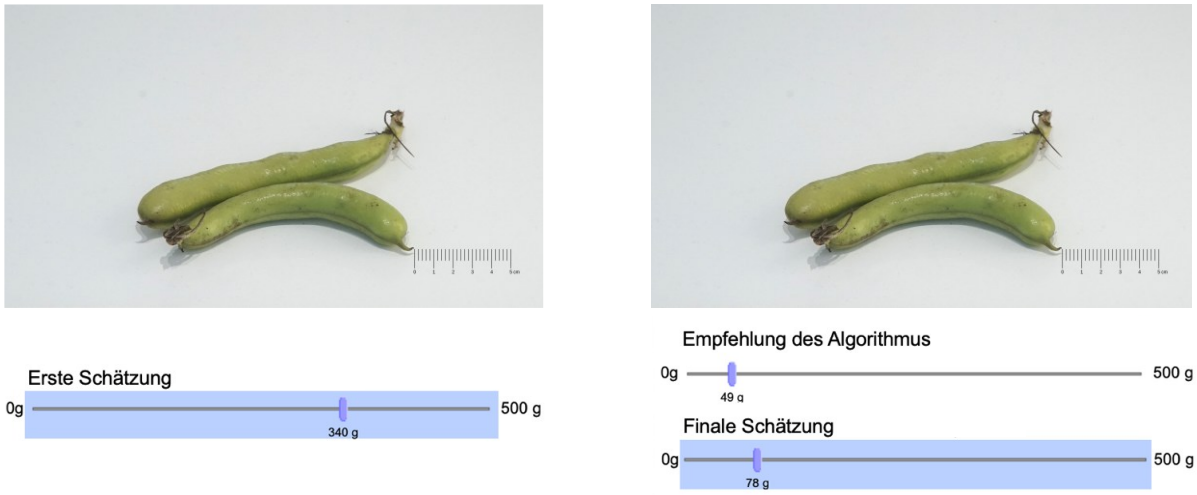
1. Die Aufgabe muss objektiv richtig oder falsch gelöst werden können (z. B. keine Vorhersagen in die Zukunft).
2. Die Aufgabe sollte Antworten auf einer Skala, nicht nur binäre Lösungen (z. B. Ja/Nein), ermöglichen.

3. Der Einfluss von Vorerfahrung und Expertise oder eines besonders leidenschaftlichen Themas (z. B. Geldanlage, geschichtliche Ereignisse oder Sport) sollte vermieden werden, um Verzerrungen bei einzelnen Versuchspersonen zu vermeiden.

In der einzelnen Schätzaufgabe wurde den Versuchspersonen randomisiert eines von 37 Bildern mit darauf abgebildetem Gemüse präsentiert. In der unteren rechten Ecke jedes Bildes war jeweils ein Größenmaßstab von 5 cm abgebildet. Das Material für diese Aufgabe wurde selbst erstellt.

Mittels eines Schiebereglers (von 0 g bis 500 g) sollten die Versuchspersonen zunächst das Gewicht des abgebildeten Gemüses schätzen, also als „Judges“ im JAS tätig werden. Im Anschluss wurde ihnen das angeblich vom Algorithmus ermittelte Gewicht als Empfehlung präsentiert – der Hinweis des „Advisors“. Daran anschließend konnten die Versuchspersonen ihre initiale Entscheidung mittels des Schiebereglers anpassen und so ihre finale Schätzung abgeben. In Abbildung 12 sind die Abbildungen einer einzelnen Schätzaufgabe grafisch dargestellt.

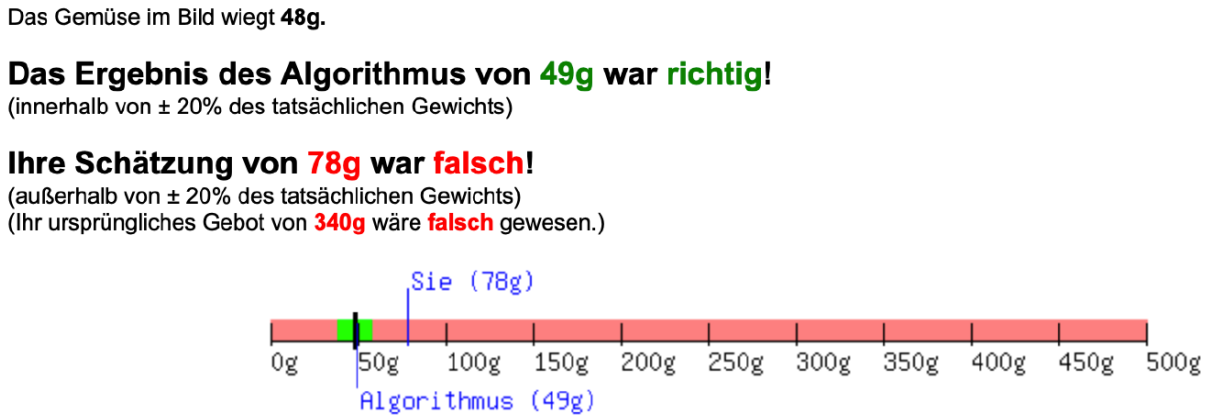
Abbildung 12: *Darstellung der beiden Abbildungen der Schätzaufgabe. Zunächst war der linke Teil zur Abgabe der Schätzung durch die Versuchspersonen zu sehen und anschließend der rechte, der die Empfehlung des Algorithmus und die finale Schätzungseingabe enthält.*



Durch einen solchen Aufbau der Entscheidungsaufgabe als JAS lässt sich mit dem WOA ein Verhaltensmaß für die Nutzung des Advisors erstellen. Der WOA bildet ab, in welchem Ausmaß der Ratschlag des Advisors berücksichtigt wird. Dazu relativiert er den Anteil der Schätzungsanpassung an der initialen Schätzung und besteht üblicherweise aus einer Zahl zwischen 0 und 1 (wobei höhere oder niedrigere Zahlen vorkommen können; Dawes & Corrigan, 1974; Logg et al., 2019). Ein WOA von 0 bedeutet, die Versuchsperson ignorierte die Empfehlung des Algorithmus und nahm keine Anpassung an den Algorithmus vor. Bei einem WOA von 1 folgte sie der Empfehlung vollständig und übernahm sie für die finale Entscheidung. Als abhängige Variable war von Interesse, ob sich Effekte der Bedingungen und der Fehlererfahrung auf die WOAs zeigen.

Die Fehlererfahrung bezeichnet die Erfahrung, dass der Algorithmus einen Fehler macht, also eine falsche Schätzung abgibt. Um diese zu ermöglichen, folgte im Anschluss an die finale Gewichtschätzung die Information über das tatsächliche Gewicht des präsentierten Gemüses. Dabei wurde auch das Ergebnis der ersten und der finalen sowie der Schätzung des Algorithmus angezeigt. Die Darstellung von wahren bzw. falschen Schätzungen wurde farblich markiert (siehe Abbildung 13). Ein Ergebnis galt als richtig, wenn es maximal 20 % vom tatsächlichen Gewicht abwich. Da der Algorithmus in allen Bedingungen drei Mal falsche Schätzungen abgab, durchliefen die Versuchspersonen drei algorithmische Fehlererfahrungen.

Abbildung 13: *Beispiel einer Ergebnispräsentation, wie sie nach jedem Trial erschien. Sie enthält das wahre Gewicht des zuvor abgebildeten Gemüses, das Ergebnis des Algorithmushinweises sowie der eigenen Schätzungen. In der Grafik grün markiert ist der als korrekt gewertete Bereich.*



Die Gestaltung der Between-Subject-Bedingungen der Akkuratheitsangabe sowie die der Fehlererfahrung legt das folgende Kapitel dar.

4.2.3.2. Bedingungen

Zunächst wurde entschieden, die beiden Akkuratheitslevel deutlich über 70 % zu wählen, um nicht von vornherein Ablehnung zu provozieren (Daschner & Obermaier, 2022; Ford et al., 2020; T. Kim & Song, 2020). Außerdem wurden zwei Akkuratheitslevel untersucht, um potenzielle Effekte der Höhe auszuschließen oder zu überprüfen.

Durch diese Ansprüche ergab sich die Herausforderung, ein realistisches, statistisch korrektes Verhältnis zwischen postulierter und tatsächlich erlebter Akkuratheit zu gewährleisten, bei einer gleichbleibenden Anzahl an Trials und algorithmischen Fehlern in allen Bedingungen. Mögliche Diskrepanzen könnten, wie andere Studien zeigen, zu Vertrauensverlusten führen (Ford et al., 2020; Yin et al., 2019). Gleichzeitig galt es jedoch die drei verschiedenen Bedingungen (hohe, niedrige und keine Akkuratheitsangabe) in einem ansonsten möglichst identischen Setting zu vergleichen. Die dazu angestellten Berechnungen wurden in Werz et al. (2021) detailliert dargelegt. So wurden zwei Akkuratheitslevel ermittelt, 90,1 % und 78,9 %, bei denen das Vorkommen von drei Fehlern bei 20

Trials dem angegebenen Akkuratheitslevel entspricht sowie mit den für jedes einzelne Bild festgelegten Abweichungen vom tatsächlichen Gewicht übereinstimmt.

Diese beiden Akkuratheitslevel bildeten die zwei Bedingungen für hohe und niedrigere Akkuratheitsangabe (HA bzw. NA). Zusätzlich kam eine dritte Bedingung hinzu: keine Akkuratheitsangabe bzw. die Kontrollbedingung (KG). Einer dieser drei Zwischen-Subjekt-Bedingungen wurden die Versuchspersonen zu Beginn des Experiments ohne ihr Wissen zufällig zugeteilt. Am Ende enthielten die Gruppen n_{HG} = 59, n_{NG} = 58, n_{KG} = 52 Teilnehmende.

Zusätzlich zu dieser Zwischen-Subjekt-Bedingung war eine weitere, für alle Versuchspersonen erhobene Bedingung die algorithmische Fehlererfahrung. Diese Innersubjektvariable wies zwei Stufen auf: vor und nach der Fehlererfahrung. Für jede Versuchsperson wurde dazu die Algorithmusnutzung (WOA) vor und die nach den algorithmischen Fehlern erhoben. Die Berechnung dieser Variable wird in Kapitel 4.2.4 erläutert.

4.2.3.3. Weitere Erhebungsinstrumente

Neben den bereits beschriebenen Variablen wurden zusätzlich Skalen eingesetzt, um explorativ den Einfluss verschiedener Einstellungs- und Persönlichkeitsmerkmale auf Algorithm Aversion zu prüfen. Zunächst wurden zwei Skalen ergänzt, die die Einstellung der Versuchspersonen zu Technik bzw. Algorithmen erhoben und damit die soziotechnische Interaktion abbilden sollten.

Attitudes Towards Algorithms-Skala (ATAS). Die ATAS misst die allgemeine Einstellung einer Person gegenüber Algorithmen anhand einer 7-stufigen Antwortskala (1 = „Stimme überhaupt nicht zu“ bis 7 = „Stimme voll und ganz zu“) mit insgesamt 17 Items (Bock & Rosenthal-von der Pütten, 2023). Die Items werden den drei Subskalen Objektivität, Ethik und Leistung zugeordnet. Eine Übersicht der Items ist in Anhang D angefügt. Beispielitems sind „Algorithmen behandeln alle Menschen gleich“ für die Subskala Objektivität (O; 6 Items), „Menschen könnten sich durch Algorithmen fremdbestimmen lassen“ für die Subskala Ethik (E; 7 Items) und „Algorithmen können Daten schneller analysieren als ein Mensch“ für die Subskala Leistung (L; 4 Items). Die ATAS wurde von Nikolai Bock am Lehrstuhl für Technik und Individuum der RWTH Aachen für Forschungen zur Mensch-Algorithmus-Interaktion entwickelt und wird aktuell validiert (Bock & Rosenthal-von der Pütten, 2023). In der vorliegenden Untersuchung wiesen die Subskalen überwiegend annehmbare interne Konsistenzen auf (Cronbachs α_O = ,84; Cronbachs α_E: ,74; Cronbachs α_L = ,69), wobei der Wert der Leistungsskala knapp unter 0,7 lag. Die Reliabilität der Gesamtskala war mit einem Cronbachs α = ,78 akzeptabel.

Kontrollüberzeugungen im Umgang mit Technik (KUT). Der KUT misst die Kontrollüberzeugungen und Selbstwirksamkeitsüberzeugungen einer Person, die sie beim Problemlösen im Umgang mit technischen Systemen hat und die wiederum das Erleben und Verhalten dieser Person in der Mensch-

Technik-Interaktion beeinflussen (Beier, 1999, 2004). Er umfasst acht Items, die auf einer 6-stufigen Antwortskala (1 = „Stimme gar nicht zu“ bis 6 = „Stimme voll zu“) beantwortet werden. Ein Beispielitem ist „Ich kann ziemlich viele technische Probleme, mit denen ich konfrontiert bin, allein lösen“. Der vollständige KUT findet sich in Anhang E. Die vorliegende interne Konsistenz war mit einem Cronbachs α = ,91 als sehr hoch einzuordnen.

Darüber hinaus wurde eine Skala ergänzt, die die Art und Weise quantifiziert, wie Individuen Probleme und Unsicherheiten angehen. So wurde erhoben, inwiefern das Kognitionsbedürfnis einer Person ihre Ausprägung der Algorithm Aversion beeinflusst.

Need for Cognition-Kurzversion (NFC-K). Die Skala NFC-K misst die Ausprägung des Konstrukts Kognitionsbedürfnis einer Person anhand einer 7-stufigen Antwortskala (1 = „Trifft überhaupt nicht zu“ bis 7 = „Trifft ganz genau zu“; Beißert et al., 2015). Die Kurzskala basiert auf der Originalskala von Cacioppo und Petty (1982) und besteht aus insgesamt vier Items, die die zwei Facetten „Engagement“ und „Freude“ des Konstrukts widerspiegeln. Eine Übersicht der Items ist in Anhang F dargestellt. Ein Beispielitem lautet: „Ich habe es gern, wenn mein Leben voller kniffliger Aufgaben ist, die ich lösen muss“. Die NFC-K ist aufgrund der geringen Item-Anzahl ökonomisch und gilt trotz der geringen Reliabilität (Cronbachs Alpha von ,51 bis ,54) als objektives und valides Messinstrument für das Kognitionsbedürfnis (Beißert et al., 2015). Die Reliabilität war in der vorliegenden Erhebung mit einer internen Konsistenz von Cronbachs α = ,64 sogar etwas höher.

Die nachfolgenden Skalen zur Risikobereitschaft bzw. Entscheidungsfreude wurden ergänzt, da untersucht werden sollte, wie Versuchspersonen mit der Unsicherheit der Schätzaufgabe umgehen. Da sowohl die Nutzung eines Algorithmus als auch seine Nichtnutzung als Risiko angesehen werden könnten, galt zu prüfen, ob sich ein Zusammenhang mit Algorithm Aversion zeigen würde. Auch andere Studien zu Algorithm Aversion ergänzten Skalen dieser Art (z. B. Herm et al., 2023).

Kurzskala Risikobereitschaft-1 (R-1). Die R-1 misst die selbsteingeschätzte allgemeine Risikobereitschaft einer Person anhand einer 7-stufigen Antwortskala (1 = „Gar nicht risikobereit“ bis 7 = „Sehr risikobereit“) mit einem Item: „Wie schätzen Sie sich persönlich ein: Wie risikobereit sind Sie im Allgemeinen?“ (Beierlein et al., 2014). Die R-1 wird im sozioökonomischen Panel einer bevölkerungsrepräsentativen Umfrage verwendet und gilt aufgrund der Testökonomie, der Reliabilität und der Validität als geeignetes Messinstrument für die allgemeine Risikobereitschaft (Beierlein et al., 2014).

Subskala Decisiveness. Die Skala Decisiveness ist eine Subskala des Fragebogens Need for Closure und misst die Ausprägung des Konstrukts Entscheidungsfreude einer Person (Roets & Van Hiel, 2007, 2011). Hierbei zielt sie auf interindividuelle Unterschiede im Bedürfnis nach schnellen und eindeutigen

Entscheidungen ab. In der vorliegenden Studie wurde das Konstrukt anhand einer 7-stufigen Antwortskala (1 = „Trifft überhaupt nicht zu“ bis 7 = „Trifft ganz genau zu“) erhoben. Da zum Zeitpunkt der Erhebung noch keine deutsche Übersetzung der Subskala existierte, wurden die sechs englischsprachigen Originalitems zunächst vom Englischen ins Deutsche und anschließend von zwei unabhängigen Personen vom Deutschen ins Englische rückübersetzt. Es wurde dann abgeglichen, ob die Übersetzungen noch die ursprüngliche Bedeutung hatten. Das Prinzip der Rückübersetzung entspricht der gängigen Vorgehensweise bei der Übersetzung von Fragebogenitems (Brislin, 1970). Eine Übersicht der englischsprachigen Originalitems und der ins Deutsche übersetzten Items findet sich in Anhang G. Ein Beispielitem ist „Ich würde schnell ungeduldig und gereizt werden, wenn ich nicht direkt eine Lösung zu einem Problem fände“. Die eindimensionale Subskala weist mit einer akzeptablen Reliabilität (Cronbachs α= .73) gute psychometrische Eigenschaften auf (Roets & Van Hiel, 2007). Die interne Konsistenz war im vorliegenden Fall etwas geringer, jedoch noch fast im akzeptablen Bereich: Cronbachs α = ,69.

Einzelitems zur Mensch-Algorithmus-Interaktion. Zur Erfassung weiterer Faktoren, die die menschliche Interaktion mit dem Algorithmus beeinflussen könnten, wurden neun Einzelitems erstellt, die auf einer 7-stufigen Antwortskala (1 = „Stimme gar nicht zu“ bis 7 = „Stimme voll zu“) erfasst wurden. Die Items bezogen sich beispielsweise auf das Vertrauen in den Algorithmus, die Zufriedenheit mit diesem und die Wahrnehmung der Notwendigkeit des Algorithmus. Zudem wurden bekannte Motivationsfaktoren in Online-Experimenten wie Spaß, der Wunsch, sich zu vergleichen, sowie der Wunsch, etwas über sich selbst zu erfahren, erhoben (Jun et al., 2017). Eine Übersicht der Items stellt Tabelle 4 dar. Da die Items nicht als Skala, sondern als Einzelitems ausgewertet wurden, wurden für diese auch keine Reliabilitätsmaße erhoben.

Tabelle 4: *Übersicht über selbst erstellte Items zur Erfassung der Mensch-Algorithmus-Interaktion*

Nr.	Item
1	Bei der Abgabe meiner Gewichtsschätzungen war ich mir sicher.
2	Ich hatte Vertrauen in das algorithmische System.
3	Ich war mit den Empfehlungen des Algorithmus zufrieden.
4	Ich habe die Empfehlungen des Algorithmus als hilfreich wahrgenommen.
5	Ich habe versucht besser zu sein als der Algorithmus.
6	Die Schätzaufgaben haben mir Spaß gemacht.
7	Ich fand es interessant herauszufinden, wie meine Leistung in den Schätzungen war.
8	Ich wollte die Aufgabe so schnell wie möglich hinter mich bringen.
9	Ich habe versucht ein möglichst richtiges Ergebnis zu erzielen.

4.2.4. Auswertung

Eine a-priori-Poweranalyse mit G*Power (Faul et al., 2007) für eine messwiederholte ANOVA mit drei Zwischen- und zwei Innensubjektfaktoren ergab, der Stichprobenumfang müsste bei einer Power von

95% für einen kleinen Effekt von f = 0,16 (Cohen, 1992) bei mindestens 156 Versuchspersonen liegen. Da die Stichprobe zuletzt aus 169 Versuchspersonen bestand, würde ein kleiner Effekt demnach bei Vorhandensein mindestens mit einer Wahrscheinlichkeit von 95% gefunden. Die Auswertung der Daten erfolgte über die Software SPSS (IBM SPSS Statistics 26). Für alle statistischen Berechnungen wurde ein α-Fehler-Niveau von ,05 genutzt.

Zur Überprüfung der Hypothesen stand in allen Analysen die Nutzung des Algorithmus als abhängige Variable im Zentrum. Diese Nutzung wurde pro Durchgang anhand des WOA quantifiziert. Dieser wurde anhand der folgenden Formel berechnet (Dawes & Corrigan, 1974; Dietvorst et al., 2014):

$$\text{WOA} = \frac{\text{Finale Schätzung} - \text{Initiale Schätzung}}{\text{Algorithmusangabe} - \text{Initiale Schätzung}}$$

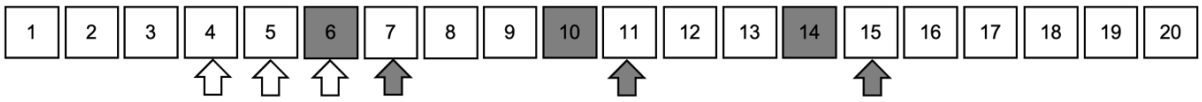
Üblicherweise zeigen sich als WOA Werte zwischen 1 und 0. Dabei gilt: Je höher der Wert, desto mehr wird der Algorithmus genutzt. Ein WOA größer 1 bedeutet, die ursprüngliche Entscheidung wird nicht nur in Richtung der Empfehlung des Algorithmus korrigiert, sondern geht über diese hinaus. Ein Wert kleiner 0 bedeutet, die finale Schätzung wurde nach der Empfehlung des Algorithmus noch weiter von dieser entfernt.

Da eine präzise Nutzung des Schiebereglers zur Einstellung einer genauen Zahl schwierig war, wurde bei Werten von unter 0 oder über 1 davon ausgegangen, diese seien auf die unpräzise Verwendung des Schiebereglers und nicht auf eine willentliche Eingabe zurückzuführen. Um dieser Annahme gerecht zu werden, wurden die WOAs korrigiert: Werte höher als 1 wurden als 1, Werte unter 0 als 0 festgelegt. Von 1014 Fällen von Interesse wurde die Korrektur 44 Mal vorgenommen. Diese Umkodierung von Ausreißerwerten, zum ersten Mal vom Statistiker Charles Winsor beschrieben, wird Winsorisierung genannt (engl.: winsorizing; Frieman et al., 2018) und beschreibt die Beschneidung von Werten auf einen festgelegten Kernbereich. Dieses Vorgehen wurde bereits in der Präregistrierung dieser Studie festgelegt und damit so verfahren wie bei anderen Algorithm Aversion-Studien (Daschner & Obermaier, 2022; Logg et al., 2019). Erst im Anschluss an die Winsorisierung wurden die im Folgenden beschriebenen Mittelwerte berechnet. Wenn nicht anders angemerkt, beruhen die in Abschnitt 4.3 aufgeführten Ergebnisse auf winsorisierten WOAs. Jedoch erfolgten auch zusätzlich Analysen mit nicht-winsorisierten Werten (siehe Abschnitt 4.3.2).

Wie bereits berichtet war die Bedingung „Fehlererfahrung“ zweistufig: mit und ohne Fehlererfahrung. Zur Messung der Nutzung des Algorithmus ohne Fehlererfahrung wurden WOAs zu Beginn des Experiments gewählt, in denen die Versuchspersonen noch keinen Algorithmusfehler erlebt hatten. Diese erfolgten erst im 6., 10., und 14. Durchgang. Um den Versuchspersonen die Möglichkeit zu geben, sich mit dem Versuchsdesign vertraut zu machen, wurden die ersten drei Trials als Probedurchgänge betrachtet. Zur Berechnung des WOA ohne Fehlererfahrung dienten die Durchgänge

4, 5 und 6. Der Wert für die Variable „Nutzung vor Fehlererfahrung“ entsprach dem Mittelwert dieser drei Trials. Die Nutzung des Algorithmus mit Fehlererfahrung wurde durch Mittelung der WOAs aus den Durchgängen 7, 11 und 15 unmittelbar nach algorithmischen Fehlern erhoben. Durch dieses Vorgehen wurde die gleiche Zahl von Trials für beide Bedingungen berücksichtigt. Abbildung 14 stellt die Verrechnung der Trials grafisch dar.

Abbildung 14: *Darstellung der Mittelung der WOAs. Die Rechtecke repräsentieren 20 Durchgänge. Grau gekennzeichnete Rechtecke stellen Durchgänge dar, in denen der Algorithmus eine fehlerhafte Empfehlung gab. Mit weißen Pfeilen gekennzeichnete Durchgänge wurden zum WOA ohne Fehlererfahrung gemittelt, mit grauen Pfeilen gekennzeichnete Durchgänge zum WOA mit Fehlererfahrung.*



Zunächst wurde zur Testung der Hypothesen eine 2x3 gemischtfaktorielle Varianzanalyse (ANOVA) berechnet mit dem Innersubjektfaktor Fehlererfahrung (Faktorstufen: ohne Fehlererfahrung, mit Fehlererfahrung) und dem Zwischensubjektfaktor Genauigkeitsangabe des Algorithmus (Faktorstufen: HA bzw. NA sowie KG). Hypothese 1 postulierte, dass die Nutzung des Algorithmus nach algorithmischen Fehlern weniger stark abnimmt, wenn die Akkuratheit des Algorithmus bekannt ist (HA und NA), im Vergleich dazu, wenn keine Information zur algorithmischen Akkuratheit gegeben ist (KG). Hierfür war also die Interaktion von Akkuratheitsinformation und Fehlererfahrung von Interesse. Zur Beantwortung von Hypothese 2, die im Kern die Replikation des Algorithm Aversion-Effekts, die Abnahme der Algorithmusnutzung nach Fehlererfahrung, postulierte, galt es, den Haupteffekt der Fehlererfahrung zu prüfen. Für alle ANOVAs wurde als Maß zur Effektstärke das η^2p verwendet, das ebenfalls über die Software SPSS (IBM SPSS Statistics 26) ermittelt wurde.

Zuletzt wurde zur Untersuchung der dritten Hypothese, wonach nach einem Fehler die Algorithmusnutzung in Bedingungen mit Akkuratheitsangabe höher ist im Vergleich zu der ohne Akkuratheitsangabe, ein t-Test für unabhängige Stichproben berechnet. Die Genauigkeitsangabe des Algorithmus diente als zweistufige unabhängige Gruppierungsvariable (HA und NA gemittelt ergeben die G-Bedingung als Bedingung mit Akkuratheitsangabe vs. KG als Bedingung ohne Akkuratheitsangabe). Die Nutzung des Algorithmus nach algorithmischen Fehlern stellte die abhängige Variable dar. Die Effektstärke des t-Tests für unabhängige Stichproben wurde mittels Cohens d bestimmt und nach Lenhard und Lenhard (2017) berechnet.

4.3. Ergebnisse

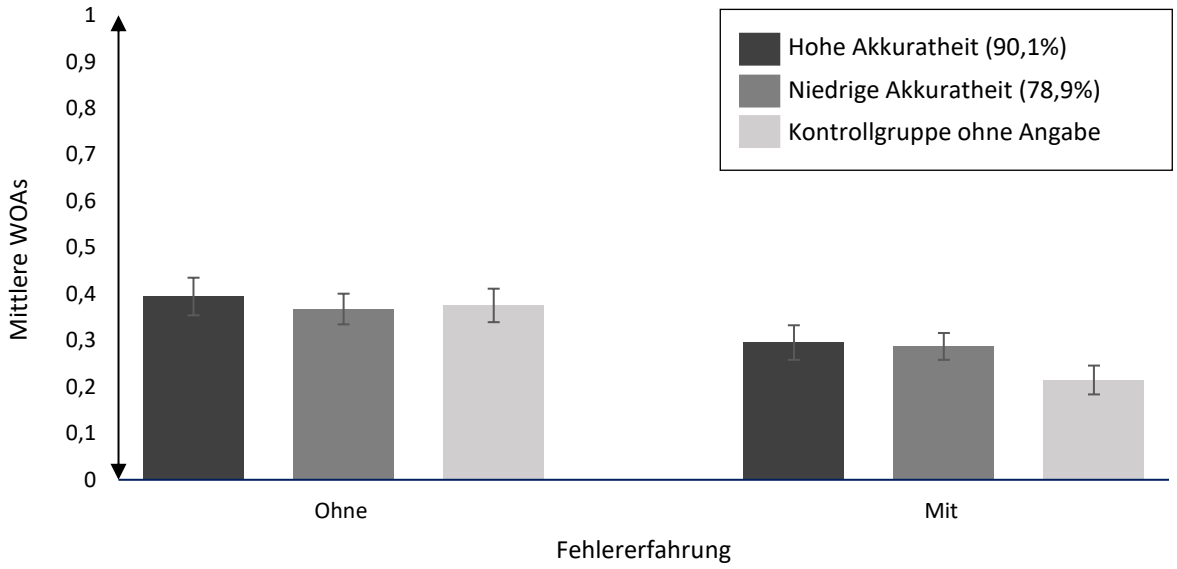
In diesem Abschnitt werden die Ergebnisse aus der ersten Erhebungsrunde berichtet. Dabei erfolgt zunächst der Bericht der Hypothesentestung, der nach dem präregistrierten Vorgehen durchgeführt wurde. Anschließend werden eine ergänzende Analyse zur Hypothesentestung berichtet, die über die

Präregistrierung hinausgeht (Kapitel 4.3.2), sowie Analysen weiterer erhobener Konstrukte (Kapitel 4.3.3 und 4.3.4).

4.3.1. Hypothesentestung

Zur Untersuchung, inwiefern die Angabe einer algorithmischen Akkuratheit dazu führt, dass das System auch nach einem Fehlerfall genutzt wird, wurde zunächst in einer 2×3-gemischtfaktoriellen ANOVA die Hypothese 1 untersucht. Die grafische Darstellung der Ergebnisse der 2×3-gemischtfaktoriellen ANOVA kann Abbildung 15 entnommen werden.

***Abbildung 15**: Gemittelte WOAs aus den Durchgängen 4, 5 und 6 (ohne Fehlererfahrung) und den Durchgängen 7, 11 und 15 (mit Fehlererfahrung) für die unterschiedlichen Bedingungen. Je größer der mittlere WOA, desto höher die Nutzung des Algorithmus.*



Anmerkung. Die Whisker repräsentieren +/- einen Standardfehler.

Entgegen der Hypothese zeigte sich in der ANOVA kein signifikanter Interaktionseffekt der Faktoren Fehlererfahrung und Bedingung ($F(2,166) = 1{,}34$, $p = ,265$, $\eta^2p = 0{,}16$). Dies lässt sich auch deskriptiv nachvollziehen: Obwohl die Abnahme in der Bedingung ohne Genauigkeitsangabe (KG) deskriptiv tatsächlich am höchsten war (ohne Fehlererfahrung: $M = 0{,}38$, $SD = 0{,}26$; mit Fehlererfahrung: $M = 0{,}21$, SD = 0,22), nahm die Nutzung des Algorithmus nach algorithmischen Fehlern auch in der HA-Bedingung (ohne Fehlererfahrung: $M = 0{,}39$, $SD = 0{,}31$; mit Fehlererfahrung: $M = 0{,}30$, $SD = 0{,}29$) sowie in der NA-Bedingung (ohne Fehlererfahrung: $M = 0{,}37$, $SD = 0{,}25$; mit Fehlererfahrung: $M = 0{,}29$, $SD = 0{,}22$) ab. Die Hypothese 1 ist also abzulehnen.

Jedoch lässt sich Hypothese 2, der zufolge die Nutzung des Algorithmus im Fehlerfall abnimmt, bestätigen. Dies zeigte der entsprechende Haupteffekt für den Faktor Fehlererfahrung ($F(1,166) = 30{,}44$, $p < ,001$, $\eta^2p = 0{,}16$). Für den Faktor der Bedingung ergab sich kein signifikanter Haupteffekt ($F(2,166) = 0{,}69$, $p = ,502$, $\eta^2p = 0{,}008$).

Die dritte Hypothese wurde mittels eines t-Tests für unabhängige Stichproben untersucht und ließ sich bestätigen. Zu ihrer Berechnung wurden die Nutzungswerte aus den beiden Genauigkeitsbedingungen HA und NA als G-Bedingung zusammengefasst und mit der KG-Bedingung verglichen. Der t-Test für unabhängige Stichproben mit einseitiger Testung wies einen signifikanten Unterschied auf (t(167) = 1,87, p = ,032, d = 0,31). Die Nutzung des Algorithmus nach algorithmischen Fehlern war höher in der G-Bedingung (M = 0,29, SD = 0,25) als in der KG-Bedingung (M = 0,21, SD = 0,22).

4.3.2. Hypothesentestung ohne Winsorisierung

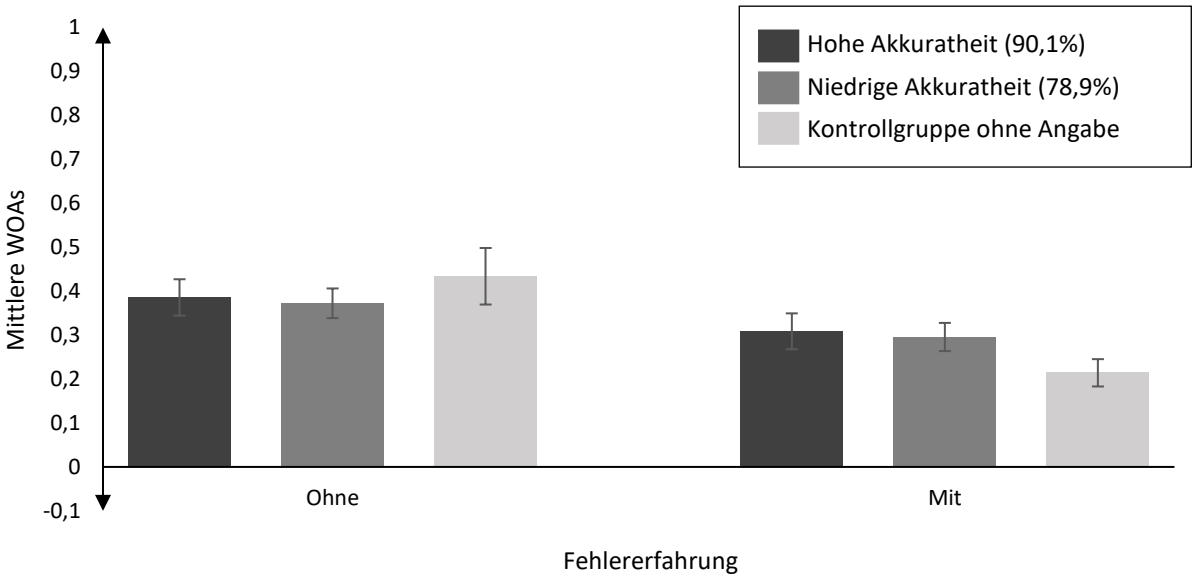
Die zuvor dargestellten Ergebnisse beruhen, wie im Methodenteil beschrieben und in der Präregistrierung angegeben, auf der Annahme, WOA-Werte über 1 oder unter 0 basierten auf einer unpräzisen Verwendung des Schiebereglers und nicht auf dem Willen der Versuchspersonen. Dementsprechend wurden die WOAs korrigiert, sodass Werte über 1 auf 1 bzw. Werte unter 0 auf 0 winsorisiert, also zurückkodiert, wurden. Diese Winsorisierung der WOAs auf maximal 1 bzw. minimal 0 kann durchaus kritisiert werden. Denn durch sie wird die tatsächlich vorhandene Varianz der Werte künstlich eingeschränkt. Nicht zuletzt kann die Annahme, Werte größer als 1 bzw. kleiner als 0 wären aufgrund des Schiebereglers und nicht willentlich herbeigeführt, nicht geprüft werden. Möglicherweise sind auch ungewöhnliche Werte von den Versuchspersonen gewünscht. Eine nachträgliche Veränderung der Daten ist also kritisch zu sehen, weshalb im Folgenden eine zusätzliche Analyse ohne winsorisierte WOAs erfolgte. Die WOAs wurden dazu also nicht verändert und aus den originalen Werten die Mittelwerte ohne Fehlererfahrung bzw. mit Fehlererfahrung berechnet.

Wie schon zuvor wurde auch mit diesen Werten zunächst eine 2x3-gemischtfaktoriellen ANOVA berechnet. Die dabei genutzten, unkorrigierten Werte können Abbildung 16 entnommen werden.

Die 2×3-gemischtfaktorielle ANOVA ergab einen signifikanten Interaktionseffekt der Faktoren Fehlererfahrung und Bedingung (F(2,166) = 3,07, p = ,049, η^2p = 0,04). Die Abnahme der Nutzung nach algorithmischen Fehlern war am höchsten in der KG-Bedingung (ohne Fehlererfahrung: M = 0,43, SD = 0,46; mit Fehlererfahrung: M = 0,21, SD = 0,22), gefolgt von der HA-Bedingung (ohne Fehlererfahrung: M = 0,39, SD = 0,32; mit Fehlererfahrung: M = 0,31, SD = 0,31) und der NA-Bedingung (ohne Fehlererfahrung: M = 0,37, SD = 0,26; mit Fehlererfahrung: M = 0,30, SD = 0,24). Hypothese 1 ließe sich mit den nicht winsorisierten WOAs also bestätigen.

Darüber hinaus ergab sich auch ein signifikanter Haupteffekt für den Faktor Fehlererfahrung (F(1,166) = 21,89, p < ,001, η^2p = 0,12). Wie schon zuvor kann Hypothese 2 also bestätigt werden. Die Nutzung des Algorithmus nach algorithmischen Fehlern (M = 0,27, SD = 0,26) nahm im Vergleich zur Nutzung vor algorithmischen Fehlern ab (M = 0,4, SD = 0,35). Für den Faktor der Bedingung gab es keinen signifikanten Haupteffekt (F(2,166) = 0,11, p = ,895, η^2p = 0,001).

Abbildung 16: *Gemittelte WOAs aus den Durchgängen 4, 5 und 6 (ohne Fehlererfahrung) und den Durchgängen 7, 11 und 15 (mit Fehlererfahrung) für die unterschiedlichen Bedingungen ohne Korrektur der WOA-Werte durch Winsorisierung. Je größer der mittlere WOA, desto höher die Nutzung des Algorithmus.*



Anmerkung. Die Whisker repräsentieren +/- einen Standardfehler.

Der t-Test für unabhängige Stichproben mit einseitiger Testung zeigte, dass sich die Nutzung des Algorithmus nach algorithmischen Fehlern signifikant unterschied wenn Akkuratheitsangaben gemacht wurden im Vergleich zur Kontrollbedingung (KG) ohne diese Angabe ($t(167) = 2{,}00$, $p = {,}024$, $d = 0{,}33$). Die Nutzung des Algorithmus nach algorithmischen Fehlern war höher in der Bedingung mit Akkuratheitsangabe ($M = 0{,}3$, $SD = 0{,}28$) als in der KG-Bedingung ($M = 0{,}21$, $SD = 0{,}22$). Dies bestätigt erneut Hypothese 3.

4.3.3. Weitere Einflussfaktoren

Zur Untersuchung möglicher Einflussfaktoren auf den Effekt der Algorithm Aversion wurden die allgemeine Einstellung gegenüber Algorithmen, Kontrollüberzeugungen im Umgang mit Technik, Risikobereitschaft, Kognitionsbedürfnis und Entscheidungsfreude gemessen. Um Zusammenhänge zwischen diesen Konstrukten und Algorithm Aversion zu untersuchen, wurde zunächst Algorithm Aversion operationalisiert durch die Differenz des WOA ohne Fehlererfahrung und dem mit Fehlererfahrung. Je höher der Wert dieser Differenz, desto größer ist die Differenz der Nutzung des Algorithmus vor und nach algorithmischen Fehlern und somit der Algorithm Aversion. Zwischen den möglichen Einflussfaktoren und der so erstellten Variable zur Quantifizierung der Algorithm Aversion wurden Spearman-Korrelationen berechnet, deren Ergebnisse in Tabelle 5 dargestellt sind. Um dem präregistrierten Vorgehen zu folgen, wurden zur Berechnung die auf minimal 0 und maximal 1 winsorisierten WOA-Werte genutzt. Jedoch zeigte eine Analyse der nicht-winsorisierten Werte keine anderen auftretenden Signifikanzen als die im Folgenden berichteten (siehe Anhang H).

Tabelle 5: *Ergebnisse der Spearman-Korrelationen zwischen Algorithm Aversion und möglichen Einflussfaktoren*

	Algorithm Aversion	
	Korrelations-koeffizient	Signifikanz-niveau
Allgemeine Einstellung gegenüber Algorithmen	ρ = ,018	*p* = ,817
Kontrollüberzeugungen im Umgang mit Technik	ρ = -,113	*p* = ,144
Risikobereitschaft	ρ = -,077	*p* = ,320
Kognitionsbedürfnis	ρ = -,197*	*p* = ,010
Entscheidungsfreude	ρ = -,024	*p* = ,760

Anmerkung. Werte, die mit * gekennzeichnet sind, sind auf einem Fehlerniveau von ,05 signifikant.

Es zeigte sich ein signifikanter Zusammenhang lediglich zwischen Kognitionsbedürfnis und Algorithm Aversion: Je höher das Kognitionsbedürfnis, desto geringer die Algorithm Aversion. Zwischen den anderen möglichen Einflussfaktoren und Algorithm Aversion gab es keinen signifikanten Zusammenhang. Es ist zu betonen, dass es sich hierbei nur um den Effekt der Algorithm Aversion handelt, also die Abnahme der Algorithmusnutzung nach dem Fehler. *Ob* der Algorithmus genutzt wurde, wurde hiermit nicht erhoben.

Zudem wurde untersucht, ob weitere Faktoren, die die menschliche Interaktion mit dem Algorithmus beeinflussen könnten, in Zusammenhang mit Algorithm Aversion stehen. Dazu wurden Spearman-Korrelationen zwischen den Ergebnissen der Einzelitems zur Mensch-Algorithmus-Interaktion und der Variable Algorithm Aversion berechnet. Die Ergebnisse sind in Tabelle 6 dargestellt. Nur das Item „Ich habe versucht besser zu sein als der Algorithmus“ zeigte einen signifikanten Zusammenhang zu Algorithm Aversion. Je mehr man der Aussage zustimmte, desto geringer war die Abnahme der Algorithmusnutzung nach dem Fehlererleben. Das gleiche Muster zeigte sich bei den nicht-winsorisierten Werten, deren Analyse in Anhang I einzusehen ist.

Tabelle 6: *Ergebnisse der Spearman-Korrelation zwischen Algorithm Aversion und den Items zur Mensch-Algorithmus-Interaktion*

	Algorithm Aversion	
	Korrelations-koeffizienten	Signifikanz-niveau
Bei der Abgabe meiner Gewichtsschätzungen war ich mir sicher.	ρ = ,065	*p* = ,400
Ich hatte Vertrauen in das algorithmische System.	ρ = ,084	*p* = ,279
Ich war mit den Empfehlungen des Algorithmus zufrieden.	ρ = ,029	*p* = ,704
Ich habe die Empfehlungen des Algorithmus als hilfreich wahrgenommen.	ρ = -,138	*p* = ,074
Ich habe versucht besser zu sein als der Algorithmus.	ρ = -,183*	*p* = ,017
Die Schätzaufgaben haben mir Spaß gemacht.	ρ = -,061	*p* = ,429
Ich fand es interessant herauszufinden, wie meine Leistung in den Schätzungen war.	ρ = -,053	*p* = ,492
Ich wollte die Aufgabe so schnell wie möglich hinter mich bringen.	ρ = -,005	*p* = ,947
Ich habe versucht ein möglichst richtiges Ergebnis zu erzielen.	ρ = ,018	*p* = ,819

Anmerkung. Werte, die mit * gekennzeichnet sind, sind auf einem Fehlerniveau von ,05 signifikant.

Zudem wurde explorativ untersucht, ob eine der demografischen Variablen einen Einfluss auf den Algorithm-Aversion-Effekt hat. Hierfür wurden gemischtfaktorielle ANOVAs mit dem Innersubjektfaktor Fehlererfahrung (Faktorstufen: ohne Fehlererfahrung, mit Fehlererfahrung) und jeweils dem Zwischensubjektfaktor Geschlecht (Faktorstufen: weiblich, männlich), Arbeitsverhältnis (Faktorstufen: Student/in, Angestellte/r, Freiberufler/in, Arbeitslose/r, Rentner/in), höchster Bildungsabschluss (Faktorstufen: Hauptschulabschluss, Realschulabschluss, Abitur, Lehre/Berufsausbildung, Universitäts- oder Fachhochschulabschluss) sowie Routine mit Computern (Faktorstufen: gar nicht, ein wenig, sehr) berechnet. Keine der unabhängigen Faktoren hatte einen signifikanten Effekt auf die Nutzung des Algorithmus. Die Ergebnisse sind in Anhang J dargestellt.

4.3.4. Nutzung des Algorithmus

Als Weiteres galt es zu ermitteln, ob die Nutzung des Algorithmus überhaupt zu signifikant besseren Ergebnissen geführt hatte bzw. hätte als die selbstständige Beantwortung der Schätzaufgaben ohne Algorithmushinweis. Dazu wurde die mittlere Abweichung aller initialen Schätzungen vom tatsächlichen Gewicht und die durchschnittliche algorithmische Abweichung vom tatsächlichen Gewicht berechnet. Anschließend wurde anhand der mittleren Abweichung der initialen Schätzungen (M = 51,51, SD = 14,71) und der algorithmischen Schätzungen (M = 20,53, SD = 6,02) ein t-Test für gepaarte Stichproben gerechnet, $t(168) = 26{,}725$, $p < {,}001$. Das Ergebnis zeigte deutlich, dass die Abweichung der initialen Schätzungen signifikant größer ist als die der algorithmischen Schätzungen. Dem Algorithmus zu folgen war/wäre also in jedem Fall besser, als den eigenen initialen Einschätzungen zu folgen.

Zuletzt wurden die Items zur Interaktionsstrategie, bei der die Teilnehmenden Aussagen zur Mensch-Algorithmus-Interaktion bewerteten, ausgewertet (siehe Tabelle 7). Es zeigt sich überwiegend eine positive Annahme und Einstellung gegenüber dem Algorithmus: Den Aussagen zu Vertrauen, Zufriedenheit und dazu, wie hilfreich der Algorithmus wahrgenommen wurde, wurde im Durchschnitt zugestimmt. Gleichzeitig lässt sich aus der Zustimmung zum Spaß an den Schätzaufgaben bzw. der Ablehnung, die Aufgaben möglichst schnell erledigen zu wollen, von einer Motivation der Teilnehmenden ausgehen. Gleichzeitig sind solche Fragen angesichts einer offensichtlichen Erwünschtheit mit Vorsicht zu genießen. Die sehr hohe Zustimmung zur Aussage, ein möglichst richtiges Ergebnis erzielen zu wollen, ist vor dem Hintergrund der tatsächlichen Algorithmusnutzung und der im vorherigen Abschnitt gezeigten Nützlichkeit desselben zur Erlangung möglichst guter Ergebnisse bemerkenswert hoch.

Tabelle 7: *Zustimmung auf einer 7-stufigen Likert-Skala zu den Items der Mensch-Algorithmus-Interaktion*

	M	*SD*
Bei der Abgabe meiner Gewichtsschätzungen war ich mir sicher.	3,08	1,20
Ich hatte Vertrauen in das algorithmische System.	4,24	0,91
Ich war mit den Empfehlungen des Algorithmus zufrieden.	4,50	0,80
Ich habe die Empfehlungen des Algorithmus als hilfreich wahrgenommen.	4,62	1,02
Ich habe versucht besser zu sein als der Algorithmus.	4,52	1,39
Die Schätzaufgaben haben mir Spaß gemacht.	4,60	1,22
Ich fand es interessant herauszufinden, wie meine Leistung in den Schätzungen war.	4,95	1,07
Ich wollte die Aufgabe so schnell wie möglich hinter mich bringen.	2,82	1,14
Ich habe versucht ein möglichst richtiges Ergebnis zu erzielen.	5,21	0,80

4.4. Diskussion

Zur Untersuchung der Frage, inwieweit Angaben von Akkuratheit eines Algorithmus dazu führen, dass dieser auch nach einem Fehlerfall genutzt wird, wurde ein Online-Experiment vorgestellt. Darin erhielten die Versuchspersonen in einer Schätzaufgabe Unterstützung durch ein vermeintliches autonomes, algorithmisches System und konnten entscheiden, wie sehr sie den Hinweisen des Algorithmus folgten. So ließ sich die Algorithmusnutzung abhängig von drei Bedingungen – einer ohne und zwei mit den Hinweisen, der Algorithmus sei zu 78,9 bzw. 90,1 % akkurat – vergleichen mit den Fällen, in denen die Versuchspersonen den Algorithmus unvoreingenommen nutzten, und mit denen, in denen er zuvor einen Fehler gemacht hatte. Mithilfe dieses 3x2-Designs wurde der Einfluss von Akkuratheitsangaben auf den bereits häufig nachgewiesenen Effekt der Algorithm Aversion untersucht.

Im folgenden Abschnitt werden die zuvor berichteten Ergebnisse in den Kontext der theoretischen Vorarbeiten, der aktuellen Forschungslandschaft und ihrer eigenen Erhebung gesetzt. Zunächst erfolgt im Rahmen der Einordnung der Hypothesen die Diskussion, wie Akkuratheitsinformation auf Algorithm Aversion wirkt. Anschließend werden die Ergebnisse im Licht weiterer Forschung besprochen und zuletzt die zu beachtenden Limitationen diskutiert.

4.4.1. Wirkung der Akkuratheitsinformation auf Algorithm Aversion

Tatsächlich bestätigen die Ergebnisse der Untersuchung mit 169 Versuchspersonen erneut den stabilen Effekt der Algorithm Aversion. Wie Hypothese 2 annahm, **zeigt sich der Effekt der Abnahme der Algorithmusnutzung nach einem Fehlerfall robust und mit großem Effekt**. Damit reiht sich die Studie in eine Vielzahl vorangehender Forschungsarbeiten ein (Burton et al., 2020; Dietvorst et al., 2014; Filiz et al., 2023). Eine Maßnahme zur Überwindung von Algorithm Aversion ist besonders deshalb vonnöten, da auch in der vorliegenden Untersuchung eine Nutzung des Algorithmus zu besseren Schätzungen führte. Die Abnahme der Nutzung nach einem Fehlererlebnis verschlechtert also die Urteile der Nutzenden.

Eine Maßnahme zur Überwindung postulierten die Hypothesen 1 und 3 mit dem Einfluss der Akkuratheitsinformation. Es zeigte sich, die **Nutzung eines Algorithmus nach einem Fehlerfall ist höher, wenn zuvor Akkuratheitsinformation über ihn vermittelt wird als ohne solche Information** (Hypothese 3 bestätigt). Die in Bezug auf Algorithm Aversion zentrale Hypothese, die zur Interaktion von Akkuratheitsbedingung und Fehlererfahrung (Hypothese 1), kann allerdings nur eingeschränkt bestätigt werden. In einer Analyse ohne Vorverarbeitung der Daten zeigte sich eine signifikante Interaktion mit sehr kleinem Effekt. Diese ist in erster Linie auf die Bedingung ohne Akkuratheitsinformation zurückzuführen, bei der sich nach den Fehlerfällen die mittlere Nutzung halbierte (und auch deutlich an Varianz einbüßte). Diese Interaktion scheint in starkem Maße auf einzelnen Werten außerhalb des üblichen Nutzungsbereichs zu beruhen, da bei der zunächst durchgeführten ANOVAS mit Beschränkung der Nutzungswerte auf die Werte zwischen 0 und 1 die Interaktion nicht signifikant wurde. Die Richtung war allerdings auch hier deskriptiv angedeutet. Über den bestehenden Forschungsstand hinaus zeigt die Studie also den positiven Effekt von Akkuratheitsangaben: Nach einem Fehler werden Algorithmen, die eine Akkuratheitsinformation offenlegen, mehr genutzt als solche ohne Informationen. Gleichzeitig scheint seine Wirkung auf Algorithm Aversion höchstens klein auszufallen, also nicht für die Überwindung des Effekts auszureichen.

Insbesondere vor dem Hintergrund, dass die Nutzung des Algorithmus in der vorliegenden Studie trotz seiner Fehler die Schätzungen der Versuchspersonen deutlich verbesserte bzw. verbessert hätte, ist selbst der geringe Einfluss durch Akkuratheitsinformation wichtig. Das vorliegende Ergebnis bestärkt damit die Annahme, algorithmische Unterstützung führt zu überwiegend besseren Ergebnissen (Dawes & Corrigan, 1974; Dietvorst et al., 2018; Grove et al., 2000; Logg et al., 2019) und dass Algorithm Aversion demnach eine Gefährdung guter Entscheidungsfindung darstellt. Die Wirkung anderer Maßnahmen, wie die Möglichkeit, das Algorithmusergebnis zu beeinflussen oder zu erleben, wie ein System nach einem Fehler dazulernt, wurde in vorangegangenen Studien gezeigt (Dietvorst et al., 2018; Gubaydullina et al., 2021; Reich et al., 2022). Auch wenn diese Maßnahmen möglicherweise stärkere Effekte erzielen, ist die hier propagierte sehr einfach umzusetzen. Die Studie leistet also einen wichtigen Beitrag dazu, Algorithm Aversion zu überwinden: Informationen über die Akkuratheit eines Algorithmus entstehen bei der Erstellung von KI-Systemen (Yin et al., 2019). Ihre Offenlegung reicht nicht aus, um Algorithm Aversion zu überwinden und doch zeigt sich der positive Effekt auf die Nutzung nach einem Fehlerfall.

4.4.2. Weitere Einflussvariablen auf Algorithm Aversion

Vorangehende Studien zeigten Zusammenhänge zwischen der Einstellung einer Person gegenüber Algorithmen und ihren mathematischen Fähigkeiten (Logg, 2017) oder ihrem Bildungsniveau (Thurman

et al., 2019). In anderen Untersuchungen zeigte sich die Expertise in der Anwendungsdomäne als Einflussfaktor auf die Nutzung eines Algorithmus (Logg et al., 2019). In der vorliegenden Studie trat der Effekt der **Algorithm Aversion unabhängig von demographischen Variablen, der Vorerfahrung mit Computern oder der Einstellung gegenüber Algorithmen** auf. Dieses Ergebnis verdeutlicht: Während eine Expertise in der Anwendungsdomäne den Effekt zu verstärken zeigt, wirkt sich Erfahrung mit Computern nicht auf Algorithm Aversion aus. Erfahrung mit KI und Wissen über sie – Algorithm Literacy – werden sogar mit einer Abschwächung des Effekts verbunden (Burton et al., 2020).

Um zu überprüfen, inwiefern weitere Persönlichkeitseigenschaften und Einstellungen einen Zusammenhang mit der Stärke des Algorithm Aversion-Effekts haben, wurden zusätzlich zu demographischen Variablen fünf weitere Konstrukte erhoben: allgemeine Einstellung gegenüber Algorithmen, Kontrollüberzeugung im Umgang mit Technik, Risikobereitschaft, Entscheidungsfreude und Kognitionsbedürfnis. Von diesen zeigte lediglich das Kognitionsbedürfnis einen Zusammenhang mit der Stärke der Algorithm Aversion.

Das Kognitionsbedürfnis („Need for cognition“) ist ein Persönlichkeitsmerkmal, das beschreibt, wie hoch das Engagement und die Freude an Denkaufgaben sind (Cacioppo & Petty, 1982). Personen mit hoher Ausprägung akzeptieren Empfehlungen von Expertensystemen nicht sofort, sondern prüfen zuerst ihre Sinnhaftigkeit (Bless et al., 1994; Dijkstra, 1999). Der in dieser Studie gefundene Zusammenhang zwischen **hohem Kognitionsbedürfnis und niedriger Algorithm Aversion** kann demnach darauf zurückzuführen sein, dass Personen mit hohem Kognitionsbedürfnis den Empfehlungen des Algorithmus nicht blind vertrauen, sondern diese hinterfragen. Möglicherweise hinterfragen sie auch ihre eigenen Fähigkeiten stärker und erkennen deshalb den Vorteil, den sie durch den Algorithmus erhalten, deutlicher. Wenn der Algorithmus eine fehlerhafte Empfehlung gibt, sind sie weniger überrascht und enttäuscht und dadurch weniger anfällig für Algorithm Aversion.

Dieser Zusammenhang von Kognitionsbedürfnis und Algorithm Aversion bettet sich in bestehende sozial- und medienpsychologische Modelle ein, die eine duale Informationsverarbeitung propagieren. Das **Elaboration Likelihood Modell** (ELM) geht von zwei Arten der Informationsverarbeitung aus, abhängig von kognitiven Ressourcen (Petty & Cacioppo, 1986):

- die **zentrale** Route, die sich durch eine aufwendigere Informationsanalyse und gerichtete Aufmerksamkeit auszeichnet und
- die **periphere** Route, die sich durch die oberflächliche und weniger nachhaltige Auseinandersetzung mit Informationen auszeichnet.

Menschen mit einem hohen Kognitionsbedürfnis verarbeiten eher über die zentrale Route, ebenso wie Menschen, denen das Thema wichtig ist, die Zeit haben oder weniger abgelenkt sind. In ihrer

Veröffentlichung zu diesem Thema stellen Petty und Cacioppo (1986) einen Zusammenhang des ELM mit Biases und Heuristiken her, deren Auftreten ebenso durch die zwei Routen erklärt werden kann (Chaiken, 1980). Anstelle rationaler Informationsverarbeitung nutzen Menschen kognitive „Abkürzungen“, um ihre Aufmerksamkeit und kognitiven Ressourcen sparsam einzusetzen. Dies bildet das Konzept der „begrenzten Rationalität“ ab („bounded rationality“; Kahneman et al., 1982). Tatsächlich sind Nutzende, die den Vorschlag eines Algorithmus ohne Nachdenken übernehmen, weniger angestrengt als diejenigen, die reflektieren und ihn als falsch identifizieren (Dijkstra, 1999). Angesichts des hier gefundenen Zusammenhangs und des auf den ersten Blick „irrationalen“ Charakters von Algorithm Aversion, scheint es plausibel anzunehmen, der Effekt zeige sich besonders, wenn weniger kognitive Ressourcen aufgebracht, Informationen weniger rational und bewusst, sondern mehr auf der peripheren Route oder oberflächlich verarbeitet werden.

Im Rahmen der explorativen Analysen wurden zudem weitere die menschliche Interaktion mit dem Algorithmus beeinflussenden Faktoren und deren Zusammenhang zu Algorithm Aversion untersucht. Dabei zeigte lediglich das Item „Ich habe versucht besser zu sein als der Algorithmus“ einen signifikanten Zusammenhang zu Algorithm Aversion. Je mehr man der Aussage zustimmte, desto geringer war der Algorithm-Aversion-Effekt. Dieses Ergebnis ist vermutlich in erster Linie ein methodisches Artefakt: Personen, die versuchten, besser als der Algorithmus zu sein, nutzten mutmaßlich die Empfehlungen des Algorithmus nicht. Stattdessen blieben sie bei ihren ursprünglichen Schätzungen. Da sie ihre Schätzungen nicht in Richtung der algorithmischen Empfehlung korrigierten, lag ihr WOA stets bei 0. Entsprechend zeigten sie keinen Algorithm-Aversion-Effekt, da das WOA kontinuierlich niedrig blieb und nicht in Durchgängen nach algorithmischen Fehlern abnahm.

4.4.3. Limitationen

Bei der Interpretation der Ergebnisse sind verschiedene Einschränkungen zu beachten, die sich auf die Generalisierbarkeit und Übertragbarkeit der Ergebnisse auswirken. Dabei gilt es, sowohl das Studiendesign als auch seine Durchführung, ebenso wie inhaltliche Beschränkungen oder die Stichprobe kritisch zu hinterfragen.

Ein augenscheinlicher Kritikpunkt am Studiendesign wurde bereits im Rahmen der Auswertung besprochen: die vorgenommene Winsorisierung der WOAs auf Werte zwischen 0 und 1. Einerseits ergeben diese Werte auf den ersten Blick mehr Sinn, da sie die Annäherung an den Algorithmus von keiner Annäherung (0) bis zur kompletten Übernahme (1) abbilden. Werte darüber – Veränderung der ersten Einschätzung noch über Algorithmushinweis hinaus – oder darunter – Veränderung der ersten Einschätzung weg vom Algorithmushinweis – scheinen weniger rational. Andererseits treten eben auch andere Werte auf. Alle Annahmen, warum diese Werte vorkommen, sind nicht prüfbar. Aus diesem Grund wurden im weiteren Verlauf der Studie sowohl winsorisierte als auch originale WOAs

angegeben und diskutiert. Die Ergebnisse sind zwar ähnlich, aber unterscheiden sich in der entscheidenden Interaktion. Gleichzeitig erlaubte der Vergleich die Feststellung, dass der signifikante Interaktionseffekt von Akkuratheitsangabe und Fehlererfahrung auf WOAs über 1 zurückzugehen scheint, die in der Kontrollbedingung vor dem Fehler vorhanden waren und nach dem Fehler zurückgingen. Diese Schlussfolgerung – und eine entsprechende Berücksichtigung in der Ergebnisinterpretation – ist möglich, da ANOVAs anfällig auf die Extremwerte reagieren. Zukünftige Studien sollten diese oder ähnliche Einschränkungen durch die Bedienbarkeit der Studie von vornherein ausschließen. Beim Einsatz eines Schiebereglers sollten zusätzliche Optionen bereitgestellt werden, um Schätzungen einzugeben und Empfehlungen des Algorithmus bequem übernehmen zu können.

Als zweiten möglichen Kritikpunkt am Studiendesign ist die Auswahl der beiden Akkuratheitsangaben zu nennen. Was genau definiert für die Nutzenden hohe, was niedrige Akkuratheit in Bezug auf Algorithmen, was auf KI? In der Studie wurden zwei Werte verwendet: 78,9 und 90,1 %. Ob diese Werte aber tatsächlich als hohe bzw. niedrigere Akkuratheit wahrgenommen wurden oder wie die Angaben auf die Nutzung wirkten, wurde nicht erforscht. Vielleicht wurden beide Werte als niedrig angesehen oder eine Angabe von 99 % hätte die Nutzung deutlich erhöht. Gleichzeitig ging es in der Studie aber nicht um eine Nutzungserhöhung per se, sondern um die Effekte der Akkuratheitsinformation auf eine Nutzung nach tatsächlichen Fehlern. Um einen Effekt der Akkuratheitsinformation auch ohne Fehlereinfluss erkennen zu können, wurden nicht nur eine, sondern zwei Akkuratheitsbedingungen gewählt und diese mit der Bedingung ohne Angabe verglichen. Da kein Haupteffekt für die Bedingung auftrat und die Nutzungswerte der beiden Akkuratheitsbedingungen beinahe identisch waren, kann ein entsprechender Effekt tatsächlich ausgeschlossen werden. Gleichzeitig wären in zukünftiger Forschung eine weitere Ausdifferenzierung und Untersuchung dieser Werte denkbar.

Als dritter Kritikpunkt ist die niedrige Relevanz der Aufgabe zu nennen: Weder ist bei der Gewichtschätzung von Gemüse mit besonderen Auswirkungen im Fehlerfall zu rechnen, noch gab es große externe Anreize für die Versuchspersonen, möglichst gute Schätzungen zu erlangen. Daraus resultiert die Frage, welchen Einfluss eine höhere Relevanz auf die verschiedenen Bestandteile des Experiments gehabt hätte: Steigt bei höherer Relevanz die Nutzung insgesamt – oder nimmt sie ab? Steigt der Effekt der Algorithm Aversion – oder nimmt er ab? Und verändert sich der Einfluss der Akkuratheitsinformation bei steigender Relevanz auf die Nutzung insgesamt wie auch auf die Nutzung nach dem Fehlerfall?

Vorangehende Forschung ist hier uneinheitlich: KI-Nutzungszahlen nehmen bei höherer Relevanz sowohl zu (Logg et al., 2019) als auch ab (Longoni et al., 2019). Vor dem Hintergrund des ELM könnte argumentiert werden, mit zunehmender Relevanz nehme die Verarbeitung der Information zu (Petty

& Cacioppo, 1986) und entsprechend die Algorithm Aversion ab. Gleichzeitig könnte aber auch die Enttäuschung bei einem schwerwiegenden Fehler größer und die Ablehnung entsprechend stärker ausfallen, was Filiz et al. als „tragedy of algorithm aversion“ bezeichnen (Filiz et al., 2023, S. 1). Beide Richtungen sind plausibel. Bisherige Studien deuten eher eine Verstärkung der Algorithm Aversion bei erhöhter Aufgabenrelevanz an (Filiz et al., 2023; Longoni et al., 2020). Dass der Effekt trotz der geringen Relevanz in der vorliegenden Studie auftrat, bestärkt ihn also eher.

Bezüglich der Auswirkungen der Relevanz auf die Akkuratheitsinformation sind wiederum verschiedene Zusammenhänge denkbar. Laut ELM nimmt die Verarbeitung von Information bei steigender Relevanz zu (Petty & Cacioppo, 1986). Der in der Studie gefundene Zusammenhang des Kognitionsbedürfnisses mit geringerer Algorithm Aversion deutet in die gleiche Richtung. Gleichzeitig beeinflussen Faktoren wie die Emotionalität des Themas, eigene Vorannahmen zur Eignung von KI für die Aufgabe, eine vorhandene Expertise zum Thema oder denkbare Kontextfaktoren (z. B. Leistungsdruck, Überwachung, soziale Kontrolle) darüber hinaus die individuelle Relevanzeinschätzung einer Aufgabe. Aufgrund der Vielzahl der Einflussfaktoren hatte die vorliegende Studie zum Ziel, den Einfluss von Vorerfahrung, Einstellung und Emotionalität zu minimieren und eine möglichst neutrale Aufgabe zu wählen. Gleichzeitig wären zukünftige Studien mit größeren Anreizen, eine möglichst richtige Entscheidung zu treffen, in Bereichen mit größerer Relevanz (z. B. Medizin oder Finanzen) oder in anwendungsnäheren Kontexten (z. B. zur Entscheidungsunterstützung im Arbeitskontext) sinnvoll und würden das Bild hinsichtlich der aufgeworfenen Fragen ergänzen.

Zuletzt ist die Zusammensetzung der Stichprobe zu nennen, die, auch wenn sie eine breite Altersstruktur aufwies, doch überwiegend aus dem universitären Kontext stammte. Damit ist sie jünger, besser ausgebildet und sicherlich im Umgang mit Algorithmen bzw. KI geübter als die deutsche Allgemeinbevölkerung. Für Menschen mit hoher „Algorithm Literacy“, wie sie in der Stichprobe vertreten sind, ist die Information, ein Algorithmus mache Fehler, keine allzu große Überraschung. In Bezug auf eine Überwindung von Algorithm Aversion, so könnte man argumentieren, profitierten diese Versuchspersonen nicht so sehr von der Akkuratheitsinformation wie KI-Nutzende, die weniger Vorwissen mitbringen. Gleichzeitig zeigt der starke Effekt der Algorithm Aversion, der auch bei dieser Stichprobe auftrat, das Potential, das – selbst bei dieser Stichprobe – bei seiner Überwindung vorhanden wäre. Insgesamt war die Nutzung mit mittleren WOAs von 0,3 bis 0,4 erschreckend gering, angesichts der Tatsache, wie viel besser die Versuchspersonen gewesen wären, wenn sie dem Algorithmus gefolgt wären. Das bedeutet, selbst in dieser überdurchschnittlich gut ausgebildeten Stichprobe waren die Nutzenden nicht in der Lage zu reflektieren, wie sinnvoll eine Algorithmusnutzung trotz seiner Fehler gewesen wäre.

Insgesamt ergeben sich aus den Limitationen konkrete Ansätze für zukünftige Studien – die Nutzung des Algorithmus noch einfacher zu gestalten – wie auch verschiedene Forschungsansätze und -fragen, derer sich nachfolgende Studien annehmen sollten. Aufbauend auf den Ergebnissen, ihrer Interpretation und unter Berücksichtigung der gerade diskutierten Limitationen lassen sich aus der Studie einige Implikationen für die Praxis ableiten, die im Folgenden dargelegt werden.

4.5. Implikationen aus Studie (a) „Fehlerfall“

Um die Implikationen bestmöglich umsetzbar zu gestalten, wird im Folgenden unterschieden zwischen den Zielgruppen Praxis und Forschung. Zunächst werden praktische Implikationen abgeleitet, die sich insbesondere an Entwickler*innen und Designer*innen von KI richten. Anschließend werden die noch offenen oder aus den Ergebnissen resultierenden Anknüpfungspunkte für weitere Forschung erläutert.

4.5.1. Implikationen für die Praxis

Zunächst ist im Sinne einer Transparenz, die Nutzende bestmöglich befähigt, gute Entscheidungen zu treffen, festzuhalten, dass Akkuratheitsinformationen in der vorliegenden Studie, wenn auch in geringem Maße, zu einer höheren Nutzung von Algorithmen nach einem Fehlerfall führten als ohne eine solche Information. In diesem Sinne sollten entsprechende Informationen also geliefert werden. Auch die AI HLEG fordert im Rahmen von Transparenz „open communication about the limitations” von KI (2020, S. 14). Da sie leicht zu ermitteln sind, sollte es zum Standard von KI-Anwendungen gehören, diese Informationen offenzulegen.

Eine zu niedrige Akkuratheit hält Nutzende davon ab, ein System zu verwenden – so zeigen frühere Forschungsarbeiten (Antifakos et al., 2005; Ford et al., 2020; Yin et al., 2019). Jedoch sollte die Fragestellung nicht lauten, wie niedrig die Akkuratheit sein darf, um eine möglichst hohe Nutzung zu erreichen. Im Sinne eines bestmöglichen Ergebnisses sollte die Abwägung zur Nutzung davon abhängen, wie hoch ein*e Nutzer*in die eigene Akkuratheit einschätzt. Wenn jemand über keinerlei Erfahrung oder Wissen in einem Gebiet verfügt, könnte ein Algorithmus selbst mit niedriger Akkuratheit hilfreich sein. Gleichzeitig steigt mit der Offenlegung der Akkuratheit auch die Anforderung an Entwickler*innen oder KI-Anbieter, schlechte Performance zu verbessern, da ihr System andernfalls gar nicht genutzt wird. Ziel einer den Menschen unterstützenden KI sollte sein, mit der Offenlegung der algorithmischen Akkuratheit die Informiertheit der Nutzenden zu steigern und ihnen die Entscheidung zu überlassen, wie sehr sie sich auf ein System verlassen möchten.

Implikation 1 für Entwickler*innen von KI:

Ermöglichen Sie die Einschätzung der Güte ihrer KI, indem Sie ihre Akkuratheit offenlegen. Bei guter Akkuratheit fördert dies die Nutzung. Bei schlechter Akkuratheit sollten Sie nachbessern.

Vor diesem Hintergrund sei auf Burton et al. (2020) verwiesen, die in ihrem Review zu Algorithm Aversion „Algorithm Literacy“ als wichtige Maßnahme propagieren. Diese umfasst das Wissen „how to interact with algorithmic tools, how to interpret statistical outputs, and how to appreciate the utility of decision aids“ (Burton et al., 2020, S. 223). Laut Burton et al. stellt Algorithm Literacy einen wichtigen Schritt zur Überwindung von Algorithm Aversion dar, indem sie insbesondere die falschen Erwartungen an KI-Systeme reguliere.

Gleichzeitig kritisieren Burton et al. (2020) impliziere die Forderung nach Algorithm Literacy, die Verantwortung liege bei den Nutzenden sich „literate“ zu machen. Im Rahmen dieser Arbeit jedoch wird die Anforderung an KI-Entwickler*innen oder -Anbieter deutlich, überhaupt ein Verständnis des Systems zu ermöglichen, also Transparenz herzustellen. Für die Regulierung falscher Erwartungen an KI scheint die Darstellung der Algorithmusakkuratheit, wie in der vorliegenden Studie geschehen, allein nicht auszureichen. Ein schützender Effekt gegen Algorithm Aversion fiel sehr gering aus. Eine umfassende Algorithm Literacy bedarf aber möglicherweise mehr: Neben der Einschätzung der Leistung der KI gehört dazu auch die Fähigkeit zu beurteilen, wie gut die eigene Leistung im Vergleich ist. Erste Studien zeigen, dies hilft den Nutzenden abzuleiten, wie sehr der eigenen und wie sehr der KI-Lösung vertraut werden sollte (Daschner & Obermaier, 2022).

Implikation 2 für Entwickler*innen von KI:

Ermöglichen Sie die Einschätzung der Akkuratheitswerte vor dem Hintergrund der eigenen Fähigkeiten. Dies könnte geschehen, indem beispielsweise die Leistung der Nutzenden ohne KI ebenso wie die mit KI kommuniziert werden.

Insbesondere bei selbstlernenden KI-Systemen wäre zusätzlich Information darüber, wie genau das System arbeitet und wie es lernt, wenn ein Fehler passiert, wichtige weitere Schritte bei der Ermöglichung von „Algorithm Literacy“ und der Überwindung von Algorithm Aversion.

Implikation 3 für Entwickler*innen von KI:

Für den Fehlerfall sollten Nutzende die Möglichkeit haben, Feedback zu geben. Im besten Falle kommuniziert das System außerdem, wie es aus dem Fehler lernt und sich für zukünftige Entscheidungen verbessert.

Darüber hinaus folgt aus dem Verständnis von Algorithm Aversion als einem Effekt der menschlichen Informationsverarbeitung der Anspruch, Mensch-KI-Interaktion ganzheitlich zu betrachten und bei ihrer Gestaltung kognitive Prozesse zu berücksichtigen. Das bedeutet, die Entwicklung guter KI darf sich nicht ausschließlich auf technische Fragestellungen beziehen, also lediglich untersuchen: Was ist mit KI möglich? Vielmehr muss ein „human centered“ Verständnis etabliert, also Technologie als

Werkzeug angesehen und für die Nutzenden entwickelt und gestaltet werden. Die zu stellende Frage muss dann lauten: Wie ist KI nützlich? Das dahinterliegende Menschenbild basiert auf der Theorie der optimalen, menschlichen Intuition, die Gerd Gigerenzer als prominentester Vertreter verfolgt. Anstatt menschliche Rationalität als begrenzt zu verstehen und reparieren zu wollen, geht es darum, menschliches Entscheiden als bestmöglich ausgebildet zu betrachten. KI als technische Unterstützung gilt es so zu entwickeln, dass sie den Entscheidungsprozess bestmöglich ergänzt („augmented intelligence“ ist hierfür das Stichwort; Sadiku & Musa, 2021).

Implikation 4 für Entwickler*innen von KI:

Verstehen Sie KI als technische Unterstützung für den menschlichen Entscheidungsprozess. Die KI und ihre Transparenz müssen sich dem Menschen anpassen, nicht umgekehrt.

4.5.2. Anschließende Forschungsfragen

Aus technischer Sicht ergibt sich in Bezug auf die Vielzahl der Aufgaben, in denen KI inzwischen unterstützt, der Bedarf nach weiteren Studien zur Wirkung von Akkuratheitsangaben. Eine Befragung von Amerikaner*innen zeigte, dass 75 % den Ergebnissen von ChatGPT vertrauen (Grigutyté, 2023). Gleichzeitig ist es zunehmend schwierig, die Akkuratheit dieser immer komplexeren Systeme zur Bearbeitung immer umfassenderer Fragestellungen zu quantifizieren. Das erste Problem dabei ergibt sich durch die Komplexität der Fragestellungen, für die diese KI-Systeme eingesetzt werden: Worauf soll sich ein Test zur Akkuratheit genau beziehen? Darauf, wie gut die Fragestellung verstanden wurde, wie gut einzelne Unter- und Teilaspekte gelöst werden konnten oder wie gut die KI zur Lösung überhaupt geeignet ist? Das zweite Problem beruht auf der Tatsache, dass nicht alle Aussagen geprüft werden können. Man denke beispielsweise an Prognosen oder subjektive Abwägungen. Zugespitzt wird die Problematik, wenn KI-Systeme selbst zum Fakten-Check eingesetzt werden. Dass eine Lüge nur lange genug wiederholt werden muss, bis sie als Meinung stehen gelassen wird, zeigt das Beispiel der gestohlenen Wahl von Donald Trump („YouTube will no longer take down false claims about U.S. elections“: Bond, 2023; Nix & Ellison, 2023). Wie verarbeitet ein KI-System solche Informationen in der Folge? Hier wird das Thema der Akkuratheit von KI ethisch aufgeladen. Doch auch schon ohne diese ethische Komponente gilt von technischer Seite die Frage zu beantworten, **anhand welcher Maße die Akkuratheit immer komplexerer KI gemessen werden kann**.

In Kooperation mit sozialwissenschaftlicher Forschung gilt es anschließend zu untersuchen, wie Akkuratheit kommuniziert werden sollte und wie sich verschiedene Akkuratheitsangaben auf die Nutzung – vor und nach einem Fehler – auswirken. Der Entscheidung der Nutzenden, ob eine Akkuratheit hoch genug ist oder als zu gering eingeschätzt werden sollte, geht voraus die korrekte Wahrnehmung und Interpretation dieser Information. **Wie genau kann und sollte die Kommunikation**

von Akkuratheit gestaltet sein? Daraus ergeben sich im Detail weitere Fragestellungen: Wie genau werden unterschiedliche Arten von Akkuratheitsangaben interpretiert (z. B. prozentuale Angaben, Visualisierungen, absolute Zahlen, die Gegenüberstellung der eigenen und der algorithmischen Akkuratheit)? Wie ändert sich ein mentales Modell, das Nutzende von einem KI-Systems haben, wenn ihnen verschiedene Arten der Akkuratheitsinformation präsentiert werden, wie wenn es einen Fehler macht? Wie sollten die Informationen gestaltet sein, um möglichst wenig fehlzuleiten, im Fehlerfall aufzuklären und gutes Entscheiden zu ermöglichen?

Darüber hinaus ergeben sich durch den Ansatz Algorithm Aversion als einen Effekt der Informationsverarbeitung beim Entscheiden anzusehen (Kahneman et al., 1982; Petty & Cacioppo, 1986) weitere Möglichkeiten zu seiner Überwindung. Die Annahme ist, Algorithm Aversion stellt einen **„irrationalen Effekt" dar, der sich besonders dann zeigt, wenn kognitive Ressourcen knapp sind** und Informationen oberflächlich verarbeitet werden. Diese Erklärung sollte geprüft werden, indem z. B. Anreize bewusst variiert oder die kognitive Verarbeitung von Informationen bei der Algorithmusnutzung beeinflusst werden. Ein entsprechender Zusammenhang würde darauf hindeuten, dass bei der Nutzung von Algorithmen besonders auf eine vertiefte Verarbeitung geachtet werden sollte und Zeitdruck oder Informationsüberfluss vermieden werden sollten.

Zuletzt macht es angesichts der Relevanz von Algorithm Aversion und der in der Studie sehr niedrigen Nutzungszahlen sicherlich Sinn, sich mit der Frage der Interaktionsstrategien mit KI auseinanderzusetzen. Die Analyse zeigte: Das beste Ergebnis erzielten die Versuchspersonen, die sich möglichst viel auf den Algorithmus verließen. Auch bewerteten sie die Zusammenarbeit mit dem Algorithmus überwiegend positiv: Zustimmung bei Vertrauen und Zufriedenheit und er wurde im Durchschnitt als eher hilfreich wahrgenommen. Ebenso herrschte in der Abfrage der Strategien sehr hohe Zustimmung zur Aussage, man habe versucht, ein möglichst gutes Ergebnis zu erzielen, und die überwiegend richtigen Ergebnisse des Algorithmus wurden kommuniziert. Trotz alledem fiel die Nutzung so zurückhaltend aus. Es stellt sich also die Frage, **ob die Kommunikation des Nutzens des Algorithmus deutlicher geschehen muss und, wenn ja, wie sie erfolgreich gestaltet werden kann**. Außerdem sollte bezüglich der Interaktionsstrategien besser verstanden werden, wie diese auf die tatsächliche Nutzung einzahlen, welche hilfreich für eine gelingende Mensch-KI-Interaktion sind und wie diese gefördert werden können. Nur wenn diese Interaktion besser verstanden wird, können KI-Systeme entwickelt werden, die das menschliche Entscheiden sinnvoll ergänzen.

Zusammenfassend ergeben sich damit die folgenden Fragen für zukünftige Forschung:

- Wie können aus technischer Sicht immer komplexere KI-Systeme in ihrer Akkuratheit bewertet werden?

- Wie kann eine erfolgreiche Kommunikation von (immer komplexeren) KI-Systemen aussehen: Wie werden verschiedene Akkuratheitsangaben interpretiert und was leitet sich daraus für die Kommunikation von Akkuratheit ab?
- Wie kann eine Anpassung bestehender mentaler Modelle von KI durch Akkuratheitsangaben (im Fehlerfall) gefördert werden, die möglichst gutes Entscheiden gewährleistet?
- Inwiefern hängt Algorithm Aversion mit einer niedrigen kognitiven Verarbeitung von Information ab? Lässt sich der Effekt durch Manipulation dieser Verarbeitungsleistung, z. B. durch Anreize oder Primes, reduzieren?
- Wie kann Nutzenden der Vorteil einer Algorithmusnutzung verdeutlicht werden? Welche Strategien sind für eine erfolgreiche Zusammenarbeit besonders sinnvoll und wie lassen sie sich fördern?

4.6. Zwischenfazit zur Studie (a) „Fehlerfall“

Die vorliegende Studie untersuchte drei Hypothesen. Es zeigte sich erneut, wie robust der Effekt der Algorithm Aversion ist (Hypothese 2). Außerdem bestätigte sich die Annahme, nach einem Fehler würden Algorithmen mit Akkuratheitsinformation häufiger genutzt als die ohne diese Information (Hypothese 3). Die zentrale Hypothese 1, nach der sich der Nutzungseinbruch nach einem Fehler durch Akkuratheitsinformationen verringert, ließ sich nur eingeschränkt bestätigen. Es ergibt sich also, wenn überhaupt, nur ein geringer Einfluss der Akkuratheitsbedingung bei der Überwindung von Algorithm Aversion.

Damit ergibt sich für die **Forschungsfrage (a) – Inwieweit führen Angaben von Akkuratheit eines Algorithmus dazu, dass dieser auch nach einem Fehlerfall genutzt wird?** – ein gemischtes Bild. Zwar nehmen Nutzende Algorithmen mit Akkuratheitsangaben nach Fehlern mehr in Anspruch als solche ohne Akkuratheitsangaben. Gleichzeitig verringert sich die Nutzung im Vergleich zu der vor einem Fehler nicht (in großem Maße). Insgesamt überraschen die sehr niedrigen Nutzungszahlen des Algorithmus, der eigentlich überwiegend korrekte Schätzungen lieferte und damit die Leistung der Versuchspersonen verbessert hat bzw. hätte. Für Entwickler*innen folgt die Empfehlung, Akkuratheitsangaben bereitzustellen, diese aber möglichst in Relation zur Leistung der Nutzenden und damit für diese einschätzbar zu gestalten. Durch den gezeigten Einfluss des Kognitionsbedürfnisses auf Algorithm Aversion ergibt sich außerdem ein Ansatzpunkt für weitere Forschung. Auch gilt es, das Konzept der „Algorithm Literacy“ weiterzuverfolgen, das einerseits als Maßnahme zur Überwindung von Algorithm Aversion identifiziert wurde und andererseits Nutzende erst in die Lage versetzt, die Vor- und Nachteile einer Nutzung zu reflektieren.

Wie die Studie zeigt, benötigen Nutzende für eine überlegte und zum bestmöglichen Ergebnis führende Nutzung eines Algorithmus mehr Informationen als seine Akkuratheit. Und es ist klar: Die Offenlegung von Akkuratheit ist nur ein Aspekt von Transparenz (AI HLEG, 2020). Doch welche Anforderungen an Transparenz von KI-Unterstützungssystemen haben nicht-technische Endnutzende? Welche Bestandteile von KI-Transparenz ergeben sich für diese Zielgruppe? Ist die Kommunikation der Akkuratheit Teil davon und welche weiteren lassen sich ermitteln? Auch die Rolle von Transparenz im Fehlerfall gilt es weiter zu untersuchen, um Schlussfolgerungen und Empfehlungen für die Umsetzung von KI abzuleiten, die eine den Nutzenden dienliche KI-Transparenz beinhaltet. Wie die vorliegende Studie zeigte, ist die Offenlegung algorithmischer Akkuratheit nur ein Element von KI-Transparenz, die sehr viel mehr umfasst.

5. Anforderungen von Endnutzenden an KI-Transparenz (Forschungsfrage b)

Um ein umfassendes Verständnis der Effekte transparenter KI auf Nutzung und Vertrauen in KI-Systeme zu erlangen, wurde eine qualitative Fokusgruppenstudie durchgeführt. Diese Studie behandelt das vorliegende Kapitel. Eine erste zusammenfassende Veröffentlichung zu dieser Untersuchung wurde durch die Autorin in Werz et al. (2024) getätigt.

5.1. Theoretische Einordnung und Einführung der Methodik

Neue Gesetze verlangen zunehmend Transparenz von digitalen Medien, Plattformen und KI (z. B. die DSGVO der EU) und beziehen sich dabei auf spezifische Funktionsweisen, wie beispielsweise das Recht auf Information über die eigenen Daten. Gleichzeitig wird unter dem Begriff Explainability die Frage verfolgt, was technisch machbar ist: Wie können KI-Prozesse und -Ergebnisse für Entwickler*innen transparent und nachvollziehbar gemacht werden und inwiefern ist es technisch möglich, verschiedene Arten von KI nachvollziehbar zu machen? Da die KI selbst (immer) häufig(er) eine Blackbox darstellt, kann sie nur durch zusätzlich eingesetzte Methoden im Nachgang erklärbar gemacht werden (Arrieta et al., 2020; Mohseni et al., 2021). Eine prominente Unterscheidung dieser Erklärungen ist die in globale und lokale Transparenz. Globale Transparenz bezeichnet dabei eine Erklärung der gesamten KI und ihrer Prozesse und beantwortet die Frage nach dem Wie. Lokale Transparenz hingegen erklärt das Zustandekommen eines einzelnen Ergebnisses und stellt die Frage nach dem Warum (Ali et al., 2023; Mohseni et al., 2021; Wanner et al., 2020). Ausführlich wird der Begriff Transparenz, sein technisches Verständnis und verschiedene Kategorisierungen in Kapitel 2.2 dargelegt.

Zusammengenommen bezeichnet Transparenz die Möglichkeit, nachvollziehen zu können, wie ein System funktioniert und warum es bestimmte Ergebnisse produziert – auf eine Weise, die als verständlich und ausreichend informativ wahrgenommen wird. In Bezug auf Endnutzende, die oftmals kein oder wenig technisches Vorwissen aufweisen als private Nutzende von KI, ist jedoch häufig noch unklar, was für sie Transparenz bedeutet (Mahmud et al., 2022; Molina & Sundar, 2022; Páez, 2019) – und ob es sich in Wie und Warum oder in andere Anteile zerlegen lässt. Diese Frage, was Endnutzende als zum Verständnis notwendig erachten und außerdem ausreichend informativ in Bezug auf KI und ihre Nutzung, wird im Folgenden untersucht. Die leitende Fragestellung b lautet:

> **FF b „Nutzendenanforderungen“: Welche Anforderungen an Transparenz in KI bestehen für Laiennutzende und inwiefern unterscheiden sie sich nach Eigenschaften der KI?**

Mithilfe von Fokusgruppen sollen deshalb die nutzerseitigen Anforderungen an KI-Transparenz erhoben werden. Ausgehend von der übergeordneten Fragestellung b wurden dazu zwei zentrale

Leitfragen abgeleitet, die mithilfe der Fokusgruppen beantwortet wurden und die Inhaltsanalyse nach Mayring leiteten. Die erste der beiden lautet:

FF b (1): Welche Anforderungen an Transparenz in KI-Apps haben Laiennutzende?

Dieser Fragestellung zugrunde liegen die theoretischen und empirischen Funde zu lokaler und globaler Transparenz, die im Bereich Explainability unterschieden werden (Ali et al., 2023; Herm et al., 2023; siehe Kapitel 2.6). Angelehnt daran gilt zu ermitteln, welche dieser Transparenzarten für Laien besonders zentral angesehen werden und ob ihre Anforderungen an Transparenz darüber hinausgehen.

In der vorliegenden Studie liegt der Fokus weder auf Unterschieden zwischen Nutzenden noch auf den Einflussfaktoren aus der Umwelt. Vielmehr geht es darum, ausgehend von KI-Systemen die Anforderungen der Nutzenden an Transparenz zu erheben. Wie die große Diversität der Ergebnisse der Nutzungsstudien von KI zeigt (siehe auch Kapitel 2.3), kommen Kontext und Eigenschaften eines Systems bei Fragen von Vertrauen und Nutzung elementare Bedeutung zu (z. B. Laato et al., 2022; Mohseni et al., 2021): Auf welche Weise interagieren Nutzende mit dem System, welche Relevanz hat es für das Leben der Nutzenden? Wird es eher zur Beschleunigung von Prozessen oder zu einer möglichst detaillierten Ausführung eingesetzt, dient es der Unterhaltung oder unterstützt es im Alltag? (Ehsan et al., 2024; Förster et al., 2020; Sieger et al., 2022; Sun & Sundar, 2022; Tabrez et al., 2022, etc., siehe Kapitel 2.3) Diese Systemeigenschaften und Anwendungsgebiete bestimmen maßgeblich die Wahrnehmung der Systeme durch die Nutzenden. Damit wirken sie sich mutmaßlich auch auf deren Anforderungen an die Systeme aus. Um diesen Einfluss auf die Anforderungen an KI-Transparenz zu erheben, lautet die zweite Fragestellung:

FF b (2): Was erwarten Nutzende abhängig von gegebenen Systemeigenschaften?

Um diese Systemeigenschaften zu untersuchen, wurden drei beispielhafte KI-Apps erstellt, die sich im Aufgabenkontext, in ihrer Bedienbarkeit und ihrer Autonomie unterscheiden (siehe Kapitel 5.2.4.). Ziel war es, möglichst verschiedene Eigenschaften in den drei KI-Apps umzusetzen. Die Relevanz der Aufgabe sowie möglicher Fehler zeigen sich in diversen Studien als ein zentrales Kriterium für KI-Akzeptanz und die Wirkung von Transparenzmaßnahmen (Förster et al., 2020; Wanner et al., 2020). So wurden für die Apps drei Kontexte gewählt, die sich durch möglichst unterschiedliche Relevanz auszeichnen: eine Finanzberatungs-App (Anlageberatung), eine Freizeit-App mit gesundheitlichen Implikationen (Pilz-Identifikation) und eine App zur Unterhaltung (Musikauswahl). Die Bedienbarkeit war mit textlicher Eingabe, Bild- und Stimmeingabe für die drei Apps unterschiedlich.

Ein Pre-Test der Apps, der in Abschnitt 5.2.4 beschrieben wird, stellte sicher, dass die Unterschiede in den Apps auch wie intendiert wahrgenommen wurden. Dazu wurden verschiedene wahrgenommene

Nutzen der App (Unterhaltung, Zeitersparnis, kognitive Entlastung etc.) sowie wichtige Funktionen der App quantitativ abgefragt. Ebenso wurde sichergestellt, dass die Apps unterschiedlich relevant bzw. gleich verständlich und ansprechend wahrgenommen wurden.

Wie die beiden Leitfragen in eine qualitative Untersuchung umgesetzt wurden, welche konkreten Fragen in den Fokusgruppen diskutiert, wie die KI-Apps gestaltet und gepretestet und wie die Ergebnisse ausgewertete wurden, legt das folgende Kapitel dar.

5.2. Methode

Zur Untersuchung der dargelegten Fragestellungen wurde ein qualitativer Ansatz, bestehend aus Fokusgruppenbefragungen und einer anschließenden qualitativen Inhaltsanalyse, gewählt, um so Zusammenhänge und Hintergründe möglichst breit erheben zu können. Ziel war eine induktive Annäherung an das Thema, um ein breites Bild zu erlangen sowie mögliche Anknüpfungspunkte für weitere Forschung zu ermitteln. Im Folgenden werden das methodische Vorgehen bei diesem Ansatz eingeordnet, die Abläufe und verwendeten Materialien der Fokusgruppen detailliert beschrieben sowie das zur Auswertung genutzte Verfahren dargelegt.

5.2.1. Projekt und Forschungskontext

Die Fokusgruppen waren Teil des Projekts TAIGERS (Transparency in Artificial Intelligence: Considering Explainability, User and System Factors), gefördert im Rahmen des Exploratory Research Space (ERS) als Open Seed Fund[4] der RWTH Aachen University.

Das Projekt wurde in Zusammenarbeit mit dem Human-Computer Interaction Center (HCIC) der RWTH Aachen University durchgeführt und untersuchte System- und Nutzendenfaktoren, die das Vertrauen und die Nutzung von transparenter KI beeinflussen. Das HCIC konzentrierte sich innerhalb des Projektes auf Nutzendenfaktoren, während am IMA und dabei federführend von der Autorin dieser Arbeit die Systemeigenschaften untersucht wurden. Entsprechend wurden die Fokusgruppen mit ihrem Fokus auf Systemeigenschaften von der Autorin dieser Arbeit konzipiert und durchgeführt.

5.2.2. Aufbau und Durchführung der Fokusgruppen

Die Fokusgruppen dauerten ca. 2,5 Stunden und fanden online mithilfe von Zoom statt, wobei Miro als digitales Kollaborationstool eingesetzt wurde, um Inhalte zu präsentieren, zentrale Punkte der Diskussion zu protokollieren und visuell durch den Workshop zu leiten. Die Fokusgruppen fanden am 27.09.21, 14.10.21 und 28.10.21 statt. Die Fokusgruppen wurden per Videomitschnitt aufgezeichnet

[4] Gefördert vom Bundesministerium für Bildung und Forschung (BMBF) und dem Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (MKW) im Rahmen der Exzellenzstrategie von Bund und Ländern

und anonym transkribiert (siehe Anhang K)[5]. Jede Fokusgruppe bestand aus fünf zentralen Bausteinen (siehe Abbildung 17):

(1) Begrüßung und Einführung in das Projekt sowie in das Kooperationstool Miro
(2) Kurzvorstellung durch die Teilnehmenden und Abfrage der Erfahrung mit KI
(3) World-Café zur Besprechung dreier KI-Apps
(4) Gruppendiskussion
(5) Bewertung von KI-Statements

Im Folgenden werden diese einzelnen Bausteine näher beschrieben.

(1) Begrüßung und Einführung in das Projekt sowie in das Kooperationstool Miro

Der erste Baustein diente der Einführung in das Forschungsprojekt TAIGERS sowie in die Thematik der Fokusgruppe. Dabei wurde auch die Relevanz der Fokusgruppen erläutert. Anschließend erfolgte eine kurze Einführung in das Kollaborationstool Miro, um die Teilnehmenden auf die Nutzung vorzubereiten. Notizen auf dem Board wurden jedoch ausschließlich von den Moderator*innen der Fokusgruppen getätigt. Den Teilnehmenden diente Miro lediglich zur Präsentation der Inhalte, die sie über einen geteilten Bildschirm wahrnahmen.

(2) Kurzvorstellung durch die Teilnehmenden und Abfrage ihrer Erfahrung mit KI

Innerhalb ihrer Kurzvorstellung sollten die Teilnehmenden erklären, welche Berührungspunkte oder Vorerfahrungen sie bisher mit KI hatten, sowie einen Smiley auswählen, der ihre Einstellung gegenüber KI beschreibt. Dabei diente die Vorstellungsrunde dazu, den persönlichen Bezug der Teilnehmer*innen zum Thema zu fördern und die Vorerfahrungen der Teilnehmenden zu ermitteln.

(3) World-Café zur Besprechung dreier KI-Apps

Die Diskussion der Fokusgruppen fand mithilfe der World-Café-Methode statt. Die World-Café-Methode beschreibt eine Form der Gruppendiskussion, bei der die Teilnehmenden auf eine beliebige Tisch-Anzahl aufgeteilt werden und eine festgelegte Zeit über ein vorgegebenes „Tisch-Thema“ diskutieren. Nach Ablauf der Zeit wechseln die Teilnehmenden den Tisch und diskutieren ein neues Thema. Pro Tisch ist ein*e Moderator*in vorgesehen, welche*r die Diskussion leitet (Löhr et al., 2020). Die Methode ermöglicht es allen Teilnehmenden, alle Themen zu diskutieren. Möglich ist eine Variation der Fragestellungen über die Runden hinweg.

Im Rahmen des World-Cafés fand die Umsetzung der beiden zuvor hergeleiteten Leitfragen statt. Diese Leitfragen wurden in drei Fragestellungen expliziert, die in den drei Runden durch die Fokusgruppen

[5] Die Transkripte sind aufgrund ihres Umfanges online abgelegt und einsehbar unter https://osf.io/5tpdv/files

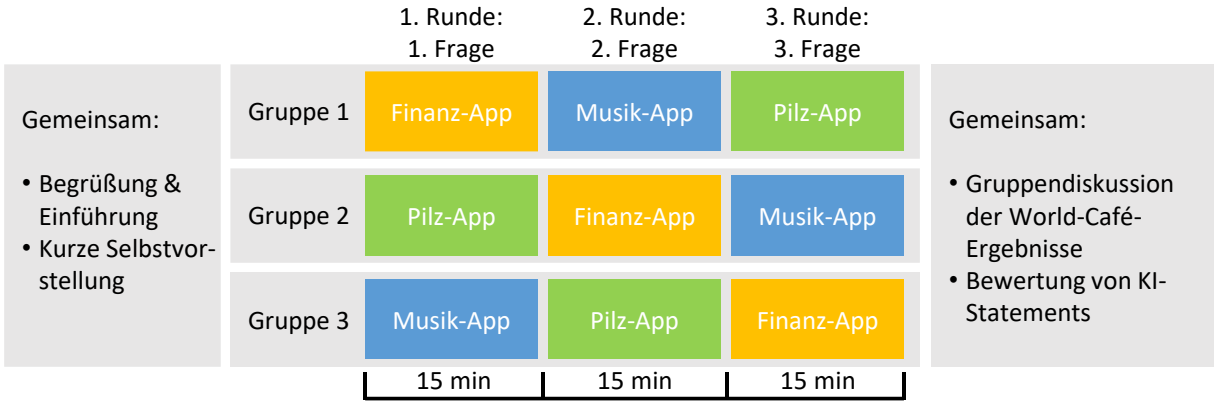
diskutiert wurden. Um jedoch zu vermeiden, dass Transparenz eingefordert wird, ohne ein genaueres Verständnis von ihr zu haben, wurde der Begriff „Transparenz“ in den Fragestellungen vermieden. Vielmehr sollte durch indirekte Fragen das Thema eingegrenzt und Implikationen von Transparenz ermittelt werden.

Zur Veranschaulichung der drei KI-Apps dienten eine Kurzbeschreibung sowie ein Screenshot der Apps (siehe Abbildung 18). Entsprechend der World-Café-Methode wurden drei virtuelle Tische gebildet, also drei Gruppen. Jede Gruppe diskutierte pro Durchgang die ihr vorliegende App unter der in dieser Runde geltenden Fragestellung:

1. Runde: Was muss die App erklären? In welchem Teil der App benötigst du welche Informationen?
2. Runde: Unter welchen Voraussetzungen würdest du die App nutzen?
3. Runde: Wie gehst du damit um, wenn du merkst, dass die App falsch liegt?

Die Fokusgruppen von 12 bzw. sieben Teilnehmenden wurden so aufgeteilt, dass jede „Tisch-Gruppe“ aus drei bis vier Teilnehmenden bestand. Eine Runde dauerte dabei 15 Minuten, dann wechselten die App, die Frage und die Moderation. Daraus resultierte eine Dauer der World-Café-Diskussion von 45 Minuten (siehe Abbildung 17).

Abbildung 17: *Ablauf der Fokusgruppen-Workshops mit detaillierter Darstellung der World-Café-Phase, die zwischen zwei gemeinsamen Diskussionen stattfand. Die Moderation blieb für die jeweiligen Apps konstant.*

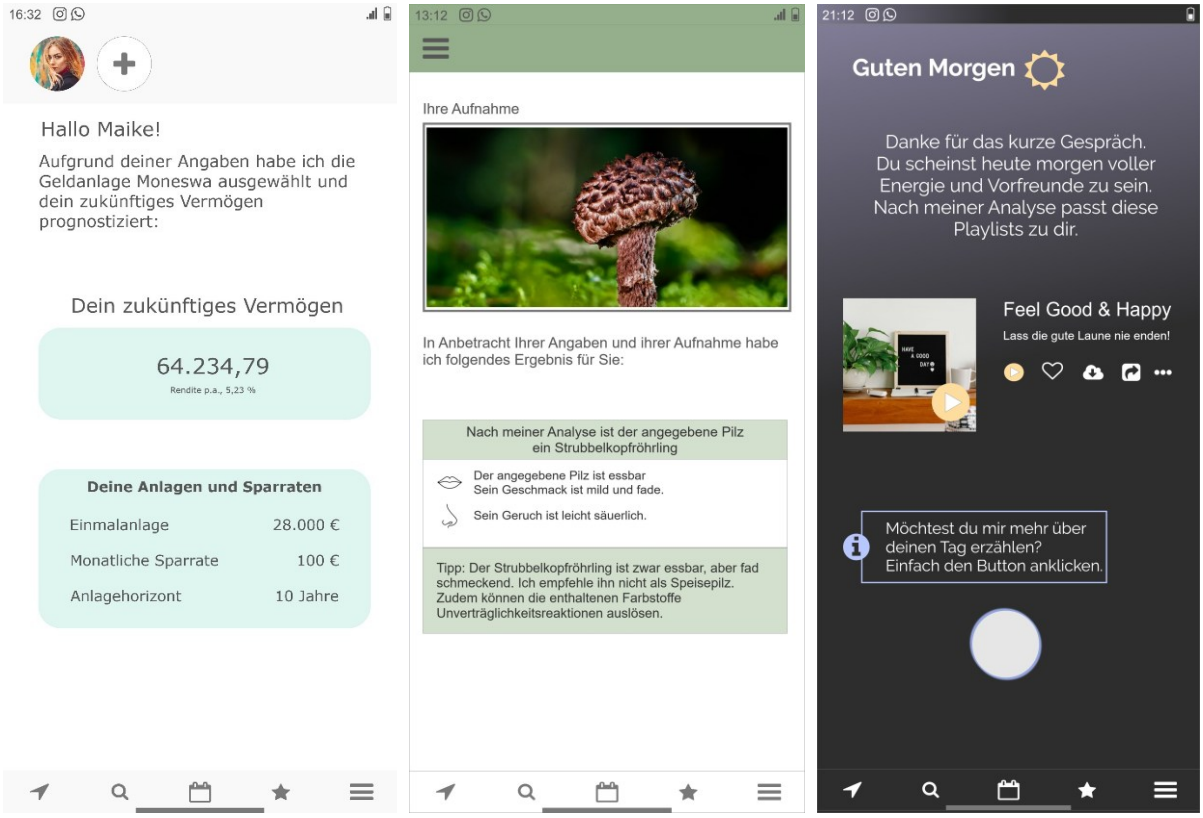


Ziel der drei verwendeten, fiktiven Apps war es, möglichst verschiedene Systemeigenschaften abzudecken (siehe Kapitel 5.1). Die drei Apps – eine Geldanlage-Beratungsapp, eine Pilzidentifikationsapp und eine Musikauswahlapp – stammen aus drei verschiedenen Anwendungs- und Anforderungsbereichen. Sie unterschieden sich nach Bedienbarkeit bzw. Eingabe der Daten: schriftlich, visuell-bildlich oder per Stimme. Durch die unterschiedlichen Kontexte unterschieden sie sich in den Anforderungen, die an sie gestellt wurden hinsichtlich ihrer Funktion: auf Sicherheit und Präzision in der Finanz-App, auf Gültigkeit und Präzision in der Pilz-App sowie Vereinfachung und Zugänglichkeit in der Musik-App. Zudem standen die Finanz- und die Pilz-App für eine hohe Relevanz,

da Fehler besonders schwerwiegend wären. Die Musik-App stand für eine niedrige Relevanz. Die Apps wurden hinsichtlich der Wahrnehmung dieser Systemeigenschaften im Vorfeld gepretestet (siehe Abschnitt 5.2.4).

Die fiktiven Screenshots wurden mithilfe der Software AXURE erstellt. Darüber hinaus dienten die Apps *Oskar*, *Spotify* sowie *Pilzführer Lite*[6] als Design-Vorlage bei der Ausarbeitung der Screenshots. Als Fotos dienten frei verfügbare Fotos von den Webseiten pexels.com sowie pixabay.com. Abbildung 18 stellt die drei Apps dar.

Abbildung 18: *Die drei abgebildeten Screenshots der KI-Apps dienten in den Fokusgruppen in der World-Café-Phase als Grundlage zur Diskussion. Links dargestellt ist die App zur Beratung bei Finanzanlagen (Finanz-App), in der Mitte die App zur Identifikation von Pilzen (Pilz-App), rechts die App zur autonomen Erstellung von Musik-Playlists (Musik-App).*



(4) Gruppendiskussion

Im Anschluss an die World-Café-Phase, die als zentrale Phase der Fokusgruppe bezeichnet werden kann, wurden deren Ergebnisse in einer Gruppendiskussion zusammengetragen. Dabei fassten zunächst die drei Moderatorinnen der „Tische“ die vorangegangenen Diskussionen zusammen, um im Anschluss nochmals mit allen Teilnehmenden gemeinsam Auffälligkeiten und weitere Kommentare zu diskutieren. Als Leitfragen dienten dabei die folgenden drei Fragen:

[6] https://www.oskar.de/, https://www.spotify.com/de/free/, https://apps.apple.com/ch/app/pilzführer-lite-pilze/id689909338

- Welche Aspekte sind bei jeder Anwendung gleich?
- Welche Informationen benötigt ihr von allen Anwendungen?
- Was muss jede Anwendung erklären?

Durch die gemeinsame Diskussion dieser Fragen sollten einerseits die Unterschiede und Gemeinsamkeiten der Apps diskutiert und darauf aufbauend die übergeordneten Aspekte von Transparenz aus Nutzendensicht reflektiert werden.

(5) Bewertung von KI-Statements

Im letzten Teil der Fokusgruppe wurde der Aspekt Transparenz explizit angesprochen und kritisch reflektiert. Dies geschah anhand von drei nacheinander präsentierten KI-Thesen. Die Thesen waren provokant formuliert, um die Diskussion anzuregen und die Teilnehmenden zu einer Positionierung zu bewegen:

1. Auch wenn sich die KI mir erklärt – ich glaube nicht, dass sie mir wirklich alles offenlegt.
2. Ich erwarte von einer KI, dass sie fehlerfrei funktioniert. Wie sie auf ihr Ergebnis kommt, ist mir egal.
3. Nur wenn die KI schwerwiegende Entscheidungen trifft, will ich wissen, wie sie funktioniert.

Die Diskussion dieser Thesen stellte den Abschluss des Fokusgruppen-Workshops dar. Anschließend wurde den Teilnehmenden für ihre Teilnahme gedankt und sie entlassen.

5.2.3. Stichprobe

An den drei Fokusgruppenterminen nahmen insgesamt *n* = 26 teil, jeweils 12, 7 und 7 Personen. Davon waren *n* = 15 weiblich. Zur Teilnahme an einer Fokusgruppe wurde kein spezielles Vorwissen benötigt. Vielmehr wurde explizit darauf hingewiesen, dass KI-Expert*innen und Entwickler*innen bitte nicht teilnehmen sollen. Die Einladungsmail findet sich im Anhang L und wurde im erweiterten Familien- und Freundeskreis der Autorin und von Kolleg*innen sowie am HCIC- und IMA-Lehrstuhl versendet. In einer kurzen schriftlichen Abfrage gaben *n* = 3 Teilnehmer*innen als höchsten Bildungsabschluss einen Masterabschluss und *n* = 18 einen Bachelor, Diplom oder Magister an. Die verbleibenden *n* = 5 hatten Abitur bzw. eine Berufsbildung. Die Hälfte der Teilnehmenden waren Studierende, die andere Hälfte Angestellte. Auf einer Skala von 1 (*gar nicht routiniert*) bis 5 (*sehr routiniert*) gaben die Teilnehmenden eine sehr hohe Routine im Umgang mit Computern an (*Median* = 5 mit 59 %, die verbleibenden 41 % gaben 4 bzw. 3 an). Das Wissen im Bereich KI wurde auf einer Skala von 1 (*überhaupt kein Wissen*) bis 5 (*sehr viel Wissen*) mit einem Median von 3 bewertet. Dabei gaben 80 % der Befragten Stufe 2 oder 3 an, 15 % Stufe 4 und eine Versuchsperson äußerte, sehr viel Wissen im Bereich KI zu haben. Damit ist

anzunehmen, dass die Stichprobe einen etwas höheren Bildungsstand und etwas mehr Wissen über KI hat als die Allgemeinbevölkerung in Deutschland.

5.2.4. Pre-Test des Materials

Zur Evaluation der verwendeten KI-App-Darstellungen wurde das Material im Vorfeld und unabhängig von den Fokusgruppen gepretestet. Dazu wurden die drei Apps nacheinander und randomisiert in einer Online-Umfrage präsentiert und dann von den *n* = 15 speziell für den Pre-Test rekrutierten Teilnehmenden hinsichtlich verschiedener Variablen bewertet (Abbildung 18 stellte die drei Apps dar). Aufgrund der geringen Stichprobengröße wurden für die vergleichenden Analysen nonparametrische Verfahren gewählt.

Für die Variablen „Eingabefunktion“, „Nutzen der App“ und „Wichtigste Funktionen der App“ wurden verschiedene Auswahlmöglichkeiten geboten (siehe Fragebogen in Anhang M). Da es sich um messwiederholte Gruppen mit binären Antwortoptionen (gewählt vs. nicht gewählt) handelte, wurde zum statistischen Vergleich der drei Apps ein Cochran-Q-Test durchgeführt. Außerdem bewerteten die Teilnehmenden die drei Apps hinsichtlich der vier Variablen „Relevanz des Themas“, „Relevanz von Fehlern“, „wie ansprechend“ sowie „wie verständlich ist die App“ auf einer 7-stufigen Likert-Skala. Diese Bewertungen wurden mittels des Friedmann-Tests für Messwiederholungen mit mehr als zwei Gruppen verglichen.

Die intendierte Eingabeoption für die Apps wurde verstanden: Bei der Musik-App wurde überwiegend korrekt „Sprache“ (86,7 %, mit *n* = 1 bzw. 6,7 % je „manuelle Auswahl“ bzw. „Bilderkennung“), bei der Pilz-App ausschließlich „Bilderkennung“ (100 %) und bei der Finanz-App „manuelle Auswahl“ (60 %) wie auch „Schreiben“ (40 %) korrekt erkannt.

Bezüglich der Frage, welchen Nutzen die Apps erfüllen sollten, trat überwiegend der intendierte Unterschied auf. Dies war der Fall bei Kostenersparnis, wo sich die Finanz-App von den beiden anderen unterschied, dem zugeschriebenen Nutzen zur Unterhaltung, wo sich die Musik-App von den beiden anderen unterschied sowie den individuellen Fehlern, wo sich nur Pilz- und Musik-App voneinander unterschieden. Erwartet worden war auch ein Unterschied zwischen Musik- und Finanz-App (siehe Tabelle 8 und Tabelle 9).

Tabelle 8: *Intendierter Unterschied zwischen den Apps Finanz-App (Finanz), Pilz-App (Pilz) und Musik-App (Musik) und tatsächlich durch die Pre-Test-Teilnehmenden wahrgenommener Unterschied in der Frage „Welchen Nutzen erfüllt die App?“*

Nutzen der App	**Intendierter Unterschied**	**Wahrgenommener Unterschied**
Kostenersparnis	Finanz > Musik = Pilz	Finanz > Musik = Pilz
Unterhaltung	Musik > Finanz = Pilz	Musik > Finanz = Pilz
Weniger individuelle Fehler	Finanz = Pilz > Musik	Pilz = Finanz, Pilz > Musik = Finanz

Tabelle 9: *Auswertung der Mehrfachauswahl zur Frage „Welchen Nutzen erfüllt die App?“ im Pre-Test. Anzahl der Nennungen und relative Häufigkeit pro App. Vergleich der Nennungen für die drei Apps durch einen Cochran-Q-Test mit paarweisen Vergleichen*

	Pilz-App		Music-App		Finanz-App		Vergleich
Nutzen	n	%	n	%	n	%	Cochran-Q-Test
Kostenersparnis	/[a]	/[a]	/[a]	/[a]	6[b]	42,90 %[b]	**χ2 (2) = 12,00, p = ,002**
kognitive Entlastung	8	57,10 %	7	46,70 %	8	57,10 %	χ2 (2) = 0,18, p = ,913
Unterhaltung	3[a]	21,40 %[a]	13[b]	86,70 %[b]	1[a]	7,10 %[a]	**χ2 (2) = 17,71, p < ,001**
Zeitersparnis	4	28,60 %	6	40,00 %	8	57,10 %	χ2 (2) = 3,42, p = ,180
weniger individuelle Fehler	9[a]	64,30 %[a]	/[b]	/[b]	4[ab]	28,60 %[ab]	**χ2 (2) = 13,56, p = ,001**

Anmerkung. n = 15; signifikante Ergebnisse sind fett markiert und Unterschiede der Häufigkeiten mit unterschiedlichen Exponenten dargestellt.

Zur Frage „Welche Funktionen sind Ihnen bei der App wichtig?“ ergab sich ein Unterschied zwischen den Apps bei „Schnelligkeit“, „Sicherheit“ (entgegen Erwartungen kein Unterscheid zwischen Musik- und Pilz-App), „Gültigkeit“ und „Akkuratheit“ (entgegen den Erwartungen kein Unterschied zwischen Finanz- und Musik-App). Auch war der Unterschied in der Erwartung an eine „einfache Bedienung“ zwischen Musik- und den anderen Apps nicht signifikant, was daran lag, dass es insgesamt oft ausgewählt wurde (siehe Tabelle 10 und Tabelle 11).

Tabelle 10: *Intendierter Unterschied zwischen den Apps Finanz-App (Finanz), Pilz-App (Pilz) und Musik-App (Musik) und tatsächlich durch die Pre-Test-Teilnehmenden wahrgenommener Unterschied in der Frage „Welche Funktionen sind Ihnen bei der App wichtig?“*

Funktionen der App	**Intendierter Unterschied**	**Wahrgenommener Unterschied**
Schnelligkeit	Musik > Finanz = Pilz	Musik > Finanz = Pilz Musik = Pilz
Sicherheit	Finanz > Pilz > Musik	Pilz = Finanz > Musik Musik = Pilz
Gültigkeit	Finanz = Pilz > Musik	Finanz = Pilz > Musik
Akkuratheit	Finanz = Pilz > Musik	Finanz = Pilz > Musik Finanz = Musik
Einfache Bedienung	Musik > Finanz = Pilz	Musik = Finanz = Pilz

Tabelle 11: *Auswertung der Mehrfachauswahl zur Frage „Welche Funktionen sind Ihnen bei der App wichtig?" im Pre-Test. Vergleich der Nennungen für die drei Apps durch einen Cochran-Q Test mit paarweisen Vergleichen.*

	Pilz-App		Music-App		Finanz-App		Vergleich
Funktion	*n*	%	*n*	%	*n*	%	Cochran-Q
Schnelligkeit	5^{ab}	33,3 %^{ab}	9^{a}	60,00 %^{a}	1^{b}	6,78 %^{b}	**$\chi 2$ (2) = 9,60, *p* = ,008**
Verfügbarkeit	3	20,00 %	3	20,00 %	4	26,70 %	$\chi 2$ (2) = 0,33, *p* = ,846
Sicherheit	9^{ab}	60,00 %^{ab}	2^{a}	13,33 %^{a}	13^{b}	86,78 %^{b}	**$\chi 2$ (2) = 13,29, *p* = ,001**
breites Aufgabenspektrum	/	/	1	6,70%	/	/	$\chi 2$ (2) = 2,00, *p* = ,368
Gültigkeit	8^{a}	53,3 %^{a}	/^{b}	/^{b}	9^{a}	60,00 %^{a}	**$\chi 2$ (2) = 12,17, *p* = ,002**
Komplexität	/	/	/	/	2	13,30 %	$\chi 2$ (2) = 4,00, *p* = ,135
Akkuaratheit	13^{a}	86,78 %^{a}	5^{b}	33,33 %^{b}	10^{ab}	66,78 %^{ab}	**$\chi 2$ (2) = 7,54, *p* = ,023**
Zugänglichkeit	2	13,30 %	4	26,70 %	4	26,70%	$\chi 2$ (2) = 1,60, *p* = ,449
einfache Bedienungen	9	60,00 %	15	100,00 %	7	46,70%	$\chi 2$ (2) = 10,40, *p* = ,006

Anmerkung. n = 15; signifikante Ergebnisse sind fett markiert und Unterschiede der Häufigkeiten mit unterschiedlichen Exponenten dargestellt.

Bezüglich der Relevanz der in der App dargestellten Thematik zeigte der Friedman-Test keine signifikanten Unterschiede der Ränge, die sich deskriptiv jedoch wie erwartet darstellten. Der erwartete Unterschied bezüglich der Relevanz möglicher Fehler zeigte sich. Bezüglich der Wahrnehmung der Apps, wie ansprechend und verständlich sie wahrgenommen wurden, ergab sich wie intendiert kein Unterschied (siehe Tabelle 12).

Trotz geringerer Abweichungen zeigte sich insgesamt das erwartete App-Profil: Die Musik-App dient der Unterhaltung, entsprechend weniger wichtig ist Genauigkeit (Sicherheit, Gültigkeit, Akkuratheit, geringe Fehler-Relevanz), sondern eher Schnelligkeit und einfache Bedienung. Die Pilz-App soll individuelle Fehler reduzieren und entlasten, Fehler wären schwerwiegend, deshalb sollte sie besonders akkurat arbeiten und eine höhe Gültigkeit aufweisen. Die Finanz-App hat einen breiten Nutzungsfall: Kostenersparnis, aber auch Entlastung in einer komplexen Materie mit hoher Relevanz.

Tabelle 12: *Bewertung der Items zur Wahrnehmung der Apps im Pre-Test. Deskriptive Daten pro App sowie Vergleich der Bewertung der drei Apps anhand des Friedmann-Tests*

	Musik-App			Pilz-App			Finanz-App			Vergleich
	M	*SD*	*Med*	*M*	*SD*	*Med*	*M*	*SD*	*Med*	Friedman-Test
Relevanz	3,20	1,42	3	3,80	1,82	4	4,13	1,60	5	$\chi 2$ (2) = 2,70, *p* = ,259
Fehler-Relevanz	2,40	1,24	2	5,73	1,49	6	6,20	0,86	6	**$\chi 2$ (2) = 18,47, *p* < ,001**
Ansprechend	3,40	1,40	3	3,60	1,45	4	3,33	1,40	3	$\chi 2$ (2) = 0,28, *p* = ,872
Verständlichkeit	5,33	1,63	6	5,80	1,15	6	5,20	1,08	5	$\chi 2$ (2) = 1,68, *p* = ,433

Anmerkung. n = 15; Med = Median; 7-stufigen Likert-Skala (1 = nicht relevant/ansprechend/…, 7 = sehr relevant/…); das signifikante Ergebnis ist fett markiert.

Fehler wären besonders schwerwiegend. Deshalb spielen Fragen der Bedienung oder Schnelligkeit eine untergeordnete Rolle, vielmehr sollte sie sicher und akkurat funktionieren.

Da eine Überarbeitung der Apps die gesamten Bewertungen verändert und damit den Ablauf der Studie deutlich in die Länge gezogen hätte, wurden die wenigen Abweichungen als akzeptabel angenommen und die KI-Apps für den Einsatz in der Fokusgruppendiskussion festgelegt. Insbesondere das Ergebnis, dass die Darstellung KI-Apps als ähnlich ansprechend und insgesamt als sehr verständlich bewertet wurde, war eine wichtige Voraussetzung für ihren Einsatz. Wie im Anschluss an die Durchführung die Ergebnisse der Fokusgruppen ausgewertet wurden, wird im nachfolgenden Kapitel erläutert.

5.2.5. Auswertung

Zur Auswertung der Fokusgruppen wurden alle Videoaufnahmen transkribiert und dabei anonymisiert (siehe Anhang K). Das eingesetzte Miro-Board wurde nicht in die Auswertung einbezogen. Die anschließende Auswertung der Transkripte wurde anhand der zusammenfassenden qualitativen Inhaltsanalyse nach Mayring (2002, 2010) durchgeführt. Diese Methode ist besonders passend für die vorliegende Untersuchung, da sie „für solche Fragestellungen besonders geeignet [ist], bei denen das Vorwissen gering ist und die Exploration im Vordergrund steht“ (Kuckartz, 2010, S. 96). Eine Technik der zusammenfassenden Inhaltsanalyse ist die induktive Kategorienbildung (Mayring, 2010). „Eine induktive Kategoriendefinition […] leitet die Kategorien direkt aus dem Material in einem Verallgemeinerungsprozess ab, ohne sich auf vorab formulierte Theorienkonzepte zu beziehen“ (Mayring, 2010, S. 85). Mayring zufolge ist dieses Vorgehen besonders für „Welche“-Fragestellungen geeignet, die im hier gegebenen Fall auch vorliegen: (1) Welche Anforderungen an Transparenz in KI-Apps haben Laiennutzende? sowie (2) Welche Anforderungen lassen sich abhängig von Systemeigenschaften identifizieren?

Nach einem induktiven Prozess wurde eine deduktive Phase angeschlossen, in der einzelne Kategorien hinzugefügt und nach Passung gefüllt wurden. Dieser Prozess der individuellen Anpassung des Vorgehens an Fragestellung und Auswertungseinheit bezieht sich auf Mayrings Beschreibung der Inhaltsanalyse als eine offene Technik: „Sie muss an den konkreten Gegenstand, das Material angepasst sein und auf die spezifische Fragestellung hin konstruiert werden“ (Mayring, 2010, S. 49). Diese deduktiv ergänzten Kategorien waren auf oberster Ebene *Transparenz* sowie die beiden Kategorien in Explainability: *lokal* und *global*.

Zur Beantwortung der Leitfragen wurden die Diskussionsinhalte zunächst gruppiert und jeweils ein Kategoriensystem pro KI-App entworfen. Die Auswertungseinheiten stellten also die Fokusgruppenworkshops zu den drei KI-Apps dar. Als Kodiereinheiten dienten einzelne Wörter, als

Kontexteinheiten die gesamte Aussage der Teilnehmerin/des Teilnehmers. Nach der Erstellung einzelner Kategoriensysteme für die drei Apps wurden sie in einem gesamten Kategoriensystem zusammengeführt. Ziel war es, überlappende, ähnliche und unterschiedliche Kategorien zu identifizieren, um so auf Systemeigenschaften abhängig von den Apps schließen zu können. Zur Präzisierung erfolge eine Überarbeitung der bestehenden Kategorien: Ähnliche Kategorien wurden vereinheitlicht bzw. Unterschiede verdeutlicht. Die hierdurch entstandenen Kategorien wurden in mehreren Schritten weiter abstrahiert und in Oberkategorien zusammengefasst. Die Verfeinerung, Schärfung und Zusammenfassung der Kategorien fanden in mehreren Schleifen statt. Um die Reliabilität der Auswertung zu erhöhen, führten zwei Auswerterinnen die Analyse durch. Während eine Auswerterin alle Transkripte analysierte, nahm eine zweite Auswerterin stichprobenartig (ca. 30 % der) Analysen vor. Die Kategorien ähnelten sich in einem Großteil der Fälle, abweichende Fälle wurden diskutiert. Außerdem konnten so Unklarheiten, mehrdeutige Aussagen oder uneindeutige Kategorien in der Diskussion gelöst werden. Ebenso führten beide Auswerterinnen die letzte Schleife zur Festlegung der Kategorien parallel durch, um sie dann in einer Abschlussdiskussion zusammenzuführen.

Nach dieser induktiven Phase der Kategorienbildung wurden die Kategorien mit der existierenden Forschungsliteratur verglichen, einzelne Kategorien ergänzt und andere in Oberkategorien zusammengefasst. Beim induktiven Vorgehen nach Mayring wird, wie zuvor beschrieben, zunächst jede Aussage nur einer Kategorie zugeordnet. Durch das ergänzte deduktive Verfahren ergaben sich jedoch Fälle, in denen Aussagen mehreren Kategorien zugeordnet wurden.

Als letzter Schritt wurden die getätigten Aussagen pro Kategorie für eine quantitative Darstellung gezählt. Dazu wurden ein bis zwei Aussagen mit einem x, drei bis fünf Aussagen mit xx und sechs oder mehr Aussagen mit xxx markiert. Das Ziel dieses Vorgehens bestand darin, einen umfassenden Einblick über die Bedeutung der Themen für jede KI-App zu erhalten (siehe Tabelle 13, Tabelle 14 und Tabelle 15). Die genauere Vorstellung des Kategoriensystems, die Abgrenzung der Kategorien sowie die Beantwortung der beiden Leitfragen erfolgen im nachfolgenden Kapitel.

5.3. Ergebnisse

Um die übergeordnete Forschungsfrage (b) **Welche Anforderungen an Transparenz in KI bestehen für Laiennutzende und inwiefern unterscheiden sie sich nach Eigenschaften der KI?** zu beantworten, wurden für die Auswertung der Fokusgruppen zwei Leitfragen formuliert: (1) Welche Anforderungen an Transparenz in KI-Apps haben Laiennutzende? und (2) Welche Anforderungen lassen sich abhängig von Systemeigenschaften identifizieren? (siehe Kapitel 5.1) Dazu wird im Folgenden zunächst auf das entstandene Kategoriensystem entlang der drei Hauptkategorien eingegangen, das damit die

Ergebnisse zu Leitfrage 1 darstellt (Kapitel 5.3.1). Anschließend erfolgt ein Bericht hinsichtlich der Unterschiede, die sich zu den Systemeigenschaften identifizieren lassen, die Ergebnisse zu Leitfrage 2 (5.3.2). Die erstellten Kategoriennamen werden kursiv dargestellt, um sie einerseits von den Zitaten in Anführungszeichen abzugrenzen und andererseits den Satzfluss zu erhalten. Aufgrund der Detailtiefe der Ergebnisse schließt Kapitel 5.3.3 mit einer Zusammenfassung der wichtigsten Ergebnisse.

5.3.1. Kategoriensystem

Im Folgenden wird zunächst das entstandene Kategoriensystem vorgestellt und damit die Ergebnisse hinsichtlich der ersten formulierten Leitfrage dargelegt:

FF b (1): Welche Anforderungen an Transparenz in KI-Apps haben Laiennutzende?

Auf oberster Ebene werden dazu drei Kategorien unterschieden: *Systemanforderungen (*Tabelle 13*)*, *Transparenz* (Tabelle 14*)* und *Nutzendenfaktoren (*Tabelle 15*)*. In *Systemanforderungen* finden sich die Aussagen, die Anforderungen an das technische System zuzuordnen sind, wie beispielsweise Informationen über die Funktionen der App, Ansprüche an die Leistung des Systems oder gewünschte Funktionen. Die *Nutzendenfaktoren* setzen sich aus Nutzungsursachen und individuellen Faktoren wie Vorerfahrung und Bedürfnis nach Kontrolle zusammen. Es ist zu betonen, dass die Nutzendenfaktoren nicht im Fokus der Untersuchung standen, diese Punkte aber in den Diskussionen zur Sprache kamen und entsprechend kategorisiert wurden.

Wie zuvor im Rahmen der Auswertung beschrieben (Kapitel 5.2.5), wurden nach einem induktiven Vorgehen einzelne Kategorien deduktiv hinzugefügt. Zum einen wurde die Oberkategorie *Transparenz* ergänzt und insbesondere aus der Kategorie Systemanforderungen diejenigen Kategorien herausgelöst, die sich dieser speziellen Kategorie zuordnen lassen. Zusätzlich wurden der Oberkategorie Transparenz die der Literatur entstammenden Subkategorien *lokal* und *global* hinzugefügt. Zu den zu *Transparenz* zugeordneten Subkategorien 1 zählen *Datenschutz*, *Hintergrundinformationen*, z. B. zu den *Urhebern* des Systems, ebenso wie *Akkuratheitsinformationen* und Fragen der *Sicherheit* und der *Rechenschaft*spflicht (siehe Tabelle 14). Während sich die Subkategorien der *Systemanforderungen* vor allem auf funktionale Aussagen beschränken, wurde *Transparenz* weit gefasst. Da als Kontexteinheiten immer ganze Aussagen von Teilnehmenden herangezogen wurden, kam es durch das deduktive Hinzufügen von Kategorien dazu, dass beispielsweise eine Aussage zu globaler Explainability als Voraussetzung für Vertrauen zunächst der Kategorie *Nutzendenfaktoren – Vertrauen – Bedingungen* und dann zusätzlich der Kategorie *Transparenz – Explainability – global* zugeordnet wurde.

5.3.1.1. Die Kategorie Systemanforderungen

Die Tabelle 13 stellt die Subkategorien der Kategorie Systemanforderungen dar, die im weiteren Verlauf näher erläutert und mit Zitaten belegt werden.

Tabelle 13: *Subkategorien der Oberkategorie Systemanforderungen. Die rechten drei Spalten bilden ab, wie häufig das Thema einer Kategorie abhängig von der genannten KI-App angesprochen wurde (für mehr Details zur Auswertung siehe Kapitel 5.2.5 bzw. Kapitel 5.3.2 für die Ergebnisse).*

Subkategorie 1	Subkategorie 2	Subkategorie 3	Musik-App	Pilz-App	Finanz-App
Benötigte Informationen	Zusätzliche Informationen	keine		x	
		anfallende Kosten	x		
		Bewertung durch andere	x		x
		Seriosität		xx	x
	über Funktionen der App	Anweisungen für Nutzung	x	xx	
		allgemeine Funktionsweise	xxx		xxx
Performance-Ansprüche	besser als andere Apps		x		
	besser als ich		x		x
	Unter Bedingungen		x		
	grundsätzlich gut		x	xx	
Konsequenzen von Fehlern	Löschen / Nutzungsende		xxx	xxx	x
	Freunde warnen			x	
	Feedback geben	Öffentlich		x	
		an App/KI	xxx	xx	
		Entwickler*in/ Expert*in		xxx	
	App-Wechsel		xx		
	Weiter nutzen		xxx	x	
	Vorsichtig weiternutzen			x	
	selbst verantwortlich			xxx	x
	Weitere Information für Fehlerfall			x	
	Keine Konsequenz		xx	x	x
Zusätzlich gewünschte Funktionen	Kompatibilität mit anderen Apps		x		x
	Auswahloptionen	Nutzende letzte Entscheidung	x	x	xxx
		Personalisierung der App	xxx		x
		App lernt selbstständig	x		
		Mehrere Ergebnisvorschläge		x	x
		Erklärungen zum Ergebnis		xxx	xxx
	Risikomanagement	Warnungen/ Empfehlungen	x	xx	x
		Einschränkungen			x
	weitere Interaktionsoptionen	auditiv	x		
		textuell	x		
		anklicken	x		

Zu den *Systemanforderungen* zählen zunächst Aussagen über *benötigte Informationen*. Bei der Musik-App weist die Frage nach anfallenden Kosten auf die Vorerfahrung mit ähnlichen Apps hin:

> „Was ist eigentlich mit Kosten? Die Musik hat ja nun bestimmte Rechte und die Künstler wollen auch leben. Also so eine Information, woher sie streamt und ob Kosten anfallen, wäre ja nicht verkehrt." (FG2_1Musik, 128-130)

Gleichzeitig wird der *Bewertung durch andere* große Bedeutung beigemessen bei der Frage nach der Qualität einer KI. Wobei einerseits die Bewertung der gesamten App berücksichtigt wird –

> „Bevor ich sie runterlade, gucke ich ja trotzdem immer im Store, wie viele Downloads hat es schon. Was hat das für eine Bewertung oder so? Also ich lade das ja nicht random runter, sondern schon mit ein paar Blicken vorher." (FG2_2Musik, 33-35)

– und andererseits bei der Finanz-App Erfahrungsberichte für einzelne Produkte gewünscht werden. Anforderungen zur *Seriosität* der KI-Apps bezogen sich auf die Seriosität der zugrundeliegenden Daten im Falle der Pilz-App sowie des gesamten Erscheinungsbilds, das im Falle der Finanz-App für eine*n Teilnehmende*n als nicht seriös wahrgenommen wurde.

Weitere *Informationen über die Funktionen der App* wurden bezüglich der auf den präsentierten Screenshots nicht sichtbaren Funktionen gefordert: zum einen *Anweisungen für die Nutzung* – „[…] in welchem Winkel man den [Pilz] fotografieren muss oder einfach von unten, von oben, von der Seite" (FG2_1Pilz, 67-68) – und zum anderen zur *allgemeinen Funktionsweise* der Apps:

> „Ich finde es auch irgendwie ein bisschen uneinsichtig. Das sagt ja, das verwaltet selbstständig dein Geld und das finde ich halt so ein bisschen, wie macht das das? Also hat das dann Zugriff auf mein Bankkonto?" (FG3_2Finanz, 142-144)

> „Wenn die mir eine Playlist aussucht, basierend auf meiner jetzigen Stimmung, ob das quasi Playlists sind, die meine jetzige Stimmung verstärken oder verändern sollen. Und inwiefern, also was so wirklich die Intention der Playlist dann ist im Sinne von Emotionsänderung oder Verstärkung der Emotionen." (FG3_Musik, 85-89)

Darüber hinaus äußerten die Teilnehmenden verschiedene *Ansprüche an die Performance* des Systems. Aussagen zur Musik-App ließen sich in die Anforderungen *besser als vergleichbare Apps*, *als man selbst* oder *grundsätzlich gut* unterteilen. Den Anspruch, die App solle besser sein *als man selbst*, hatte ein*e Teilnehmer*in auch an die Finanz-App, während mehrere auch von der Pilz-App eine *grundsätzlich gut*e Leistung erwarteten.

Als *Konsequenz von auftretenden Fehlern* äußerte ein Großteil der Befragten in allen drei Gruppen die Absicht, die App nicht weiterzunutzen und sie zu *löschen*. Ein *App-Wechsel* wurde bei der Musik-App

in Betracht gezogen. Für diese App kam für einige auch trotz Fehlern eine *Weiternutzung* in Frage, auch aufgrund von Verständnis für Anfangsschwierigkeiten:

> „Also wenn da jetzt in der Playlist ein zwei Lieder drin sind, die gar nicht zu meiner Stimmung passen, werde ich wahrscheinlich die Möglichkeit haben zu sagen, gefällt mir, gefällt mir nicht und die App wird wahrscheinlich mit, je länger ich die nutze, wird die dazu lernen, was ich mag und was nicht. Denke ich mal. Also die Fehler werden vielleicht am Anfang dann…, dass die Auswahl nicht so ganz stimmt." (FG2_2Musik, 111-116)

Teilweise wurde auch das fehlerhafte Verhalten bei den Nutzenden selbst vermutet und bei der Pilz-App eine *vorsichtige Weiternutzung* in Betracht gezogen.

Besonders bei der Pilz-App zeigte sich eine Spannung zwischen dem Wunsch andere warnen zu wollen, das System zu verbessern und sich selbst schützen zu wollen. Eine *Warnung an Freunde* oder durch *öffentliches Feedback,* sprich über eine App-Bewertung, dient dem ersteren Anliegen. Die Hoffnung bei der Rückmeldung an *Entwickler*innen oder Expert*innen* oder aber an die *App/KI* ist zumeist, so die KI zu verbessern und zukünftige Fehler zu vermeiden. Andererseits äußerten Viele, für die Nutzung der App *selbst verantwortlich* zu sein:

> „[…] der Hersteller sollte sich im besten Falle auf jeden Fall irgendwie absichern und sagen, hier, wenn ich jetzt einen Pilz falsch erkenne und du isst, weil du doof bist, einen Fliegenpilz und stirbst dran, selber schuld, so ein bisschen. […] sowas gibt es ja auch mit Blumen oder so, also das kenne ich jetzt aus der eigenen Erfahrung, dass das schon relativ gut funktioniert, und deswegen würde ich mal sagen, ich glaube, die Technik kann das schon. Aber wenn halt echt mal was passiert, ja, wie T8 eben schon gesagt hat, auf eigene Verantwortung auf jeden Fall." (FG1_Pilz, 188-198)

Bei Pilz- und Finanz-App sehen einige die eigene Verantwortung besonders dann gegeben, wenn zuvor auf mögliche Fehler hingewiesen wurde, z. B. durch Wahrscheinlichkeiten oder Warnungen:

> „Wenn aber diese Sachen [Wahrscheinlichkeiten und Referenzbilder] gegeben wären, dann würde ich sagen: ja ok, dann liegt die Verantwortung halt bei mir und dann kann ich sie als Hilfe dazu immer noch nutzen, auch wenn sie falsch gelegen hat." (FG1_Pilz, 99-101)

> „[…] solange die App jetzt in einem realistischen Rahmen scheitert und sagt, ja, okay, es ist, natürlich, es ist nicht optimal gelaufen, aber es ist in dem Rahmen nicht optimal gelaufen, vor dem wir euch vorher gewarnt haben. Dann ist es ja im Grunde der eigene, die eigene Entscheidung gewesen, dieses Risiko zu akzeptieren." (FG1_Start und Finanz, 761-765)

Aussagen, laut denen aus einem Scheitern der KI *keine Konsequenzen* gezogen werden würden, beziehen sich auf die geringen Auswirkungen eines solchen Fehlers bei der Musik-App und der Pilz-App, bei der ein Fehler „nur halb so schlimm [ist] wie wenn das jetzt mit dem eigenen Geld passiert oder etwas wirklich Wichtigem“ (FG1_Pilz, 412-413). Der Grund für eine gleichgültige Reaktion angesichts eines KI-Fehlers bei der Finanz-App liegt für die/den Teilnehmer*in in der Natur von Finanzanlagen: „dann kann man halt nichts machen. Es gibt halt manchmal Situationen, da ist es egal, welche App du benutzt, es gehen alle runter“ (FG1_Start und Finanz, 787-789).

Eine weitere Subkategorie 1 der *Systemanforderungen* stellt *zusätzliche gewünschte Funktionen* an die Apps dar. Dazu gehörte die *Kompatibilität mit anderen Apps*, ebenso wie verschiedene, weitere *Auswahloptionen* im Nutzungsprozess. Als zusätzliche Funktion äußerte ein Großteil der Teilnehmenden den Punkt, als *Nutzende die letzte Entscheidung*sinstanz sein zu wollen. Diese Subkategorie 3 war besonders bei der Finanz-App zentral:

> „Aber trotzdem will ich nicht, dass die App das dann komplett allein macht und will trotzdem bei jedem Schritt gefragt werden. Aber ich fände es cool, wenn die App mir dann so eine Push-Benachrichtigung gibt: Normalerweise, wenn sie jetzt online wären, würden sie das und das tun, wollen sie das wirklich tun?“ (FG3_Finanz, 308-311)

Aber auch bei der Pilz-App und der Musik-App war es einigen Teilnehmenden wichtig, Eingabevariablen kontrollieren und die letzte Entscheidung selbst treffen zu können. Weitere Anmerkungen bezogen sich bei der Musik-App auf die noch genauere *Personalisierung der App* oder ihre Fähigkeiten *selbstständig dazuzulernen*.

Bei Pilz- und Finanz-App bestand darüber hinaus der Wunsch als Ergebnis nicht nur eines, sondern *mehrere Vorschläge* angezeigt zu bekommen, aus denen man im Falle der Pilz-App den passenden Pilz heraussuchen bzw. im Falle der Finanz-App die passende Anlage wählen könnte. Hier besteht eine enge Verwandtschaft zu *Explanability*, insbesondere *lokal*er Art, also einem *Transparenz*-Aspekt. Die hier zugeordneten Aussagen waren solche ohne den Wunsch nach Erklärung, während Aussagen, die Fragen nach dem Warum oder weiteren Ergänzungen enthielten, der Oberkategorie Transparenz zugeordnet wurden. Eine Aussage der Subkategorie 3 *Mehrere Ergebnisvorschläge* lautet beispielsweise:

> „Und ich finde bei so einer App hätte man noch viel, viel mehr Infos geben können, um irgendwie annähernd eine Entscheidung treffen zu können, ob man den Pilz mitnimmt oder nicht. Zum Beispiel könnte die KI einem noch die Suchergebnisse, die fünf Suchergebnisse, die danach am ehesten darauf zutreffen würden, mitanzeigen, dass man da nochmal abwägen

> kann, okay, sieht der vielleicht doch dem ähnlicher oder wie weit darf der Pilz schon vergammelt sein, also wenn man das noch irgendwie erkennt.“ (FG2_1Pilz,24-30)

Zusätzlich wurden *weitere Erklärungen zum einzelnen Ergebnis* gewünscht. Auch diese Subkategorie 3 weist eine Nähe zur Transparenz-Subkategorie *Explainability – lokal* auf. Während es sich bei lokaler Explainability jedoch um Erklärungen zur Funktionalität handelt, bezieht sich die Subkategorie 3 *Erklärungen zum einzelnen Ergebnis* auf Informationen, die zusätzlich zu dem Ergebnis gewünscht werden. Besonders bei der Pilz-App hatten die Teilnehmenden viele Anmerkungen in diese Richtung: „Oder noch mal ein Vergleichsbild aus einem Lexikon, das man dazu angezeigt bekommt zu seinem eigenen oder so.“ (FG2_1Pilz, 52-53)

> „Vielleicht auch, dass man dazu schreibt, noch mal so eine schriftliche Beschreibung des Pilzes. Also dass man eben nicht nur anhand eben eines Fotos oder so, wie das jetzt gerade quasi dargestellt ist, hat, sondern dass man noch mal am Pilz selber kontrollieren kann, der hat jetzt aber auf jeden Fall von unten Flecken oder so. […] Ja, also, dass du dann halt noch mal drunter schauen kannst und das eben vernünftig kontrollieren kannst.“ (FG1_Pilz, 221-227)

In der Subkategorie 2 *Risikomanagement* diskutierten die Teilnehmenden einerseits die Anforderung von einer „mitdenkenden“ App *Warnungen/Empfehlungen* erhalten zu wollen, z. B. zum Geltungsbereich der Pilz-App, die nur in deutschen Wäldern, nicht in tropischen, anzuwenden sei oder bei besonders risikobehaftetem Anlageverhalten. Ein*e Teilnehmer*in wünschte sich, das Risiko durch *Einschränkungen* bei der Finanz-App reduzieren zu können.

Die letzte Subkategorie 2 der *Systemanforderungen* betrifft Anmerkungen zu *weiteren Interaktionsoptionen*. Die Sprachinteraktion mit der Musik-App wurde bezüglich vieler Faktoren diskutiert (beispielsweise auch zum Thema Datenschutz oder mögliche Fehler), weshalb sich für diese App der Wunsch nach weiteren Interaktionsoptionen ergab, seien sie *auditiv*, *textuell* oder durch *Anklicken*.

5.3.1.2. Die Kategorie Transparenz

Wie zuvor beschrieben, ist die Oberkategorie *Transparenz* eine deduktiv erstellte Kategorie. Das bedeutet, nach dem induktiven Vorgehen wurde sie ergänzt und Kategorien aus der Oberkategorie *Systemanforderungen* übertragen, um die Transparenzaspekte herausgelöst und genauer betrachten zu können. Tabelle 14 stellt die Subkategorien der Oberkategorie Transparenz dar.

Tabelle 14: *Subkategorien der Oberkategorie Transparenz. Die rechten drei Spalten bilden ab, wie häufig das Thema einer Kategorie abhängig von der genannten KI-App angesprochen wurde (für mehr Details zur Auswertung siehe Kapitel 5.2.5 bzw. Kapitel 5.3.2 für die Ergebnisse).*

Subkategorie 1	Subkategorie 2	Subkategorie 3	Musik-App	Pilz-App	Finanz-App
Datenschutz	keine Sorge				x
	Sorge	Rückverfolgbarkeit/ Anonymität	x		
		Datenverkauf an Dritte	x		
		Sammeln kritischer Daten	xxx		x
	Information	keine Information	xx	xx	
		Kontrolle über geteilte Daten	x		
		Verarbeitung der Daten	xxx	x	
		Sicherung der Daten			x
		Art der gespeicherten Daten	xxx		xxx
Hintergrund-informationen	Profit				xxx
	Ethische Bedenken		x		xx
	Urheber			xxx	xx
Sicherheit des Systems	Sicherheitsgarantie				xxx
	Audits			x	x
Explainability	global	Grundsätzliche Erklärungen	x	x	x
		Trainingsdaten	xx	xxx	x
	lokal	Warum-Erklärung	xxx	xx	x
		Warum nicht-Erklärung		x	
	Darstellung		x		
	besonders im Fehlerfall				x
	Zweifel an Funktionalität			x	
	kein Interesse		xx		x
Accountability/ Rechenschaft					xxx
Akkuratheits-information				xxx	xx

Die erste Subkategorie, die in Bezug auf *Transparenz* identifiziert wurde, und das besonders prominent bei der Musik-App, ist die Subkategorie 1 *Datenschutz*. Eine Person sprach an, sich *keine Sorgen* zu machen, da wir im Internet sowieso ständig unsere Daten angeben. Gründe für Sorgen rund um den Datenschutz hingegen betrafen die *Rückverfolgbarkeit/Anonymität* eigener Daten, ihr *Verkauf an Dritte* sowie generell das *Sammeln kritischer Daten*. Letzterer Punkt umfasste Aussagen dazu, wie Daten gespeichert werden, welche Daten gesammelt werden und:

> „Auch wann. Also viele Apps sagen ja immer, es ist nur aktiv, wenn die App aktiv ist, aber im Endeffekt nehmen die einen die ganze Zeit auf oder sowas. Das da quasi so eine Erklärung

> kommt, wann man aufgenommen wird. Also wahrscheinlich während dieses Gesprächs und dann auch nur dann. Das fände ich ganz beruhigend zu wissen.“ (FG1_Musik, 54-58)

Bezüglich der *Informationen*, die die Teilnehmenden im Bereich Datenschutz forderten, gaben einige an, *keine Informationen* zu wünschen. Die Argumentation ähnelte der aus vorherigen Zitaten, mit der begründet wurde, sich keine Sorgen um den Datenschutz zu machen: Die eigenen Daten seien sowieso schon online, bei Nutzung eines Dienstes nehme man eine Speicherung in Kauf oder – bei der Pilz-App – die Daten seien nicht sensibel.

Wünsche nach Informationen umfassten bei der Musik-App die Möglichkeit, die Weitergabe der eigenen *Daten kontrollieren* und z. B. löschen zu können, sowie bei der Finanz-App Informationen darüber, wie Zugang oder Daten *gesichert* sind, z. B. gegen Zugriffe Dritter. Darüber hinaus waren die *Art der gespeicherten Daten* sowie ihre *Verarbeitung* Themen von Interesse bei der Musik-App:

> „Also ich glaube gerade bei Spracherkennung bin ich auch immer sehr…, beziehungsweise gerade, wenn es auch um so Stimmungsdaten geht und was an meinem Tag passiert ist. Das sind ja schon sehr, sehr persönliche Sachen und da würde ich wirklich einfach gerne wissen, was damit passiert.“ (FG1_Musik, 204-208)

> „Ja, für mich sind so Datenschutz-technische Informationen wichtig. […] Die klickt man aber eigentlich nur relativ schnell weg, wenn das so 50.000 Sätze sind oder Zeilen. Aber dass da ganz am Anfang vielleicht kurz erklärt wird, ok, das nehmen wir von dir auf, sei es deine Stimme, sei es dein Standort, sehr wahrscheinlich auch, was auch immer. Also dass man da kurz noch einen ganz kurzen Überblick bekommt, was überhaupt aufgenommen wird und wie das vielleicht verarbeitet wird.“ (FG3_Musik, 18-24)

Und auch bei der Finanz-App stellten die Teilnehmer*innen die Frage nach der *Art der gespeicherten Daten*: „Ich persönlich stelle mir so ein bisschen die Frage, welche Informationen soll ich überhaupt preisgeben?“ (FG1_Start und Finanz, 535-536)

Während der *Datenschutz* besonders prominent in den Gruppendiskussionen zur Musik-App besprochen wurde, fanden sich sehr viel weniger Aussagen zu dieser Subkategorie 1 in den beiden Diskussionen zur Pilz- und Finanz-App. Diese Verteilung drehte sich in der nächsten Subkategorie 1, die als *Hintergrundinformationen* zusammengefasst wurde, um. In deren drei Subkategorien 2 *Profit*, *ethische Bedenken* und *Urheber* wurden insbesondere bei der Finanz-App zahlreiche Informationen gefordert. Das Thema *Profit* wurde ausschließlich bei der Finanz-App diskutiert: „Ja, also bei mir ist auf jeden Fall auch die Skepsis, […] dass ich nicht weiß, welche Interessen dieser Algorithmus hier bedient.“ (FG3_Finanz, 259-261)

> „Also das Wichtigste wäre für mich, wer profitiert. Sind es meine Interessen, die da im Endeffekt im Vordergrund stehen oder sind es doch die Interessen von irgendwelchen Daten, irgendwelche Datenanalysen? Machen die Profit durch meine Abschlüsse oder so? Das wäre mir wichtig." (FG3_Finanz, 173-177)

Auftretende *ethischen Bedenken* betrafen verschiedene Aspekte bei der Musik- und der Finanz-App. Bei letzterer ging es in einigen Aussagen in erster Linie um die Übereinstimmung mit einem eigenen Anspruch, den zu erfüllen der App nicht zugetraut wurde:

> „Ich finde da aber auch besonders kritisch noch mal, dass vermutlich die Strategie, mit der man das meiste Vermögen macht, nicht unbedingt der ethischen oder nachhaltigsten Strategie entspricht und das wäre für mich ein sehr entscheidender Faktor, dass ich eben bei Anlage-Themen eben ethische Entscheidungen treffen wollen würde und ich gehe nicht davon aus, dass die App das gewährleistet." (FG3_Finanz, 42-47)

Hingegen betraf eine Sorge bei der Musik-App die grundsätzliche Beeinflussung und Steuerung der Stimmung, was als ethisch problematisch wahrgenommen wurde.

Ähnlich wie die Profit-Diskussion drehte sich die Frage nach dem *Urheber* darum, wer denn hinter der App steht, also „wer die programmiert hat, von welcher Firma, die App entwickelt wurde" (FG3_Finanz, 34), oder zu beantworten „[…], wenn in die eine KI haben, die dir Geld generiert: Wieso bieten die die dann an? Was ist so wirklich der Hintergrund dieser Firma?" (FG1_Start und Finanz, 430-431) Diese Skepsis herrschte besonders bei der Finanz-App. Doch auch bei der Pilz-App kam wiederholt die Frage nach den Urhebern auf, die insbesondere als Qualifizierung der KI-Empfehlungen von Interesse war: „Oder dass irgendwo in der App steht, wer die entwickelt hat. Also auf welcher Wissensdatenbank die Infos basieren." (FG2_1Pilz, 59-60)

> „Na ja, bei der App steht jetzt zum Beispiel unten drin ‚Ich empfehle ihn nicht als Speisepilz'. Da wäre für dann schon die Frage: Wer ist denn dieser Ich? Ist das der Herr Müller von nebenan, der gerade mal eine App entwickelt hat und aus dem Lexikon irgendwelche Pilze eingescannt hat? Oder ist das vielleicht die Deutsche Gesellschaft für Pilzforschung? Weiß ich nicht. Was das denn schon eher irgendwo realistischer macht, dass das stimmen könnte, was da steht." (FG2_2Pilz, 42-47)

Auch die Frage nach der *Sicherheit des Systems*, sei es durch *Sicherheitsgarantien* oder *Audits*, wurde bei der Finanz-App sehr häufig diskutiert und bei der Musik-App gar nicht. *Sicherheitsgarantien* kamen nur bei der Finanz-App zur Sprache: „Ich würde es auch, glaube ich, nur nutzen, wenn ich so eine Art Versicherung habe, dass ich nur […] so oder so viel verlieren könnte. Ich meine, bei Investitionen, weiß man ja, glaube ich, nie so genau." (FG1_Start und Finanz, 529-532)

> „Also ich fände interessant […], wenn eine Garantie bestünde. Das heißt, ich gehe nach so und so viel Jahren nicht im Minus da aus der Nummer raus, sondern es ist ein festgelegter Gewinn, den die App dann generieren muss, sonst wird das von dem Betreiber abgedeckt irgendwie. Dann würde ich es eventuell in Betracht ziehen." (FG2_Finanz, 50-54)

In der Subkategorie 2 *Audits* finden sich Aussagen zur Pilz-App – „Also es sollte schon irgendwie immer, wenn man sowas anbietet, denke ich mir, irgendwie geprüft sein, dass es korrekt ist" (FG2_1Pilz, 84-85) – und zur Finanz-App:

> „Das heißt, eine App, die ich vielleicht nutzen würde, die müsste dann irgendwie vielleicht von, keine Ahnung, der Stiftung Warentest kommen oder so was. Halt so einer glaubhaft unabhängigen Institution, die kein Gewinninteresse an meiner Anlagenentscheidung hat." (FG3_Finanz, 267-270)

Eine weitere sehr umfangreiche Kategorie stellt die Subkategorie 1 *Explainability* dar. Ihre zwei Subkategorien 2 *global* und *lokal* wurden nachträglich deduktiv ergänzt.

Die Subkategorie 2 *globale* Explainability enthält zwei Subkategorien 3: *Grundsätzliche Erklärungen* und *Trainingsdaten*. Es finden sich in beiden Kategorien Aussagen von allen drei Apps, wobei nur wenige zur Finanz-App – „Mehr Informationen über die App. Wie die funktioniert, was die für einen Algorithmus hat" (FG3_Finanz, 33-34) – und mehrere zur Pilz-App getätigt wurden. So lassen sich den *grundsätzlichen Erklärungen* bei der Pilz-App folgende Aussagen zuordnen:

> „Dass [die KI] das dann auch sagt: hör mal, ich bin mir nicht sicher. Pilz A ist essbar, Pilz B nicht. Lass es lieber. So ungefähr. Also dass man da auf jeden Fall irgendwie transparent das erklärt und nicht einfach, weiß ich nicht, weil… Ja, wie rechnet der das überhaupt? Mit einer Wahrscheinlichkeit? Und wenn es dann 51% zu 49% ist, nimmt der dann den mit 51%? Also das wäre eben auch so die Frage: Wie funktioniert das Ganze überhaupt, dann in dem Moment?" (FG1_Pilz, 244-250)

> „Die Frage ist ja auch, auf was die KI reagiert, was wir mit der KI anstellen. Also ob unsere Beteiligung an der App auch in diesem Lernen irgendwie reinspielt oder ob das ein geschlossenes System ist, und wir greifen daraus Informationen ab. Weil das würde für mich das Ganze auch noch mal unsicher machen, wenn ich wüsste, ok, Menschen wie ich ziehen halt im Wald und sagen oh ne, das sieht anders aus. […] Das heißt, das wäre vielleicht schon auch irgendwie interessant zu wissen, wer oder wie das Ergebnis zustande kommt." (FG3_1Pilz, 178-185)

Die *Trainingsdaten* umfassten hingegen spezifischere Aussagen nach dem Ursprung der Daten bzw. nach der „Information, […] basierend auf welchen Infos die App klassifiziert, oder einfach ihre

Informationen über die Pilze herbekommt“ (FG3_2Pilz, 76-77) oder inwiefern „[…] unter langwierigen oder mehrfachen Testszenarien irgendwie im Real-Einsatz getestet [wurde] oder irgendwie so was“ (FG2_1Pilz, 154-155).

Lokale Explainability umfasst Transparenz-Aspekte, die den Prozess hinter der Ergebnisfindung betreffen. Die in *Systemanforderungen* aufgeführten Subkategorien 3 *Mehrere Ergebnisvorschläge* und *Erklärungen zum Ergebnis* in *Auswahloptionen* beschränken sich auf vordergründige Informationen: Es geht darum, mehr zu diesem Ergebnis bzw. mehrere Alternativen zu erfahren (siehe Kapitel 5.3.1.1). Hier in *Transparenz* liegt der Fokus auf Erklärungen zum Zustandekommen der Ergebnisse. Die *lokale* Explainability ist in zwei Subkategorien 3, *Warum-Erklärungen* und *Warum-nicht-Erklärungen*, geteilt. Es geht also um Erklärungen, die entweder erläutern, warum dieses Ergebnis zustande kam oder warum es nicht zustande kam. Letztere Kategorie ergibt sich nur mit einer Aussage aus der Pilz-App-Diskussion, die eine automatische Kontrastierung mit giftigen Pilzen wünscht. Aussagen zu *Warum-Erklärungen* hingegen finden sich bei allen drei Apps: „[…] also aus welcher Quelle die Informationen zum Pilz kamen oder auch wie sozusagen die Bewertung des Pilzes zustande kam.“ (FG1_Pilz, 370-371)

> „Und dann finde ich es auch relativ interessant zu sehen, wenn die schon nicht den Code öffentlich machen wollen, was ja ein Nicht-ITler dann eh nicht verstehen kann, dass man dem Nutzer wirklich dann sagt, ok, anhand der Schlagwörter, anhand deiner Stimmfarbe, an deiner Tonlage, haben wir herausgefunden das und das.“ (FG3_Musik, 148-152)

> „Ich glaube, wenn sie mir auf Nachfrage, das muss nicht alles auf der Start-Homepage sein für Leute, die da einfach gar kein Bock drauf haben, aber sie sollte mir sagen können, warum, auf welchen Entscheidungen sie welche Anlage-Entscheidungen getroffen hat und wie sich dann auch dieses Ziel, diese 5% Rendite, ergeben.“ (FG1_Start und Finanz, 678-681)

In dieser letzten Aussage steckt schon zu einem Teil die nächste Subkategorie 2, die aber in erster Linie in Bezug auf die Musik-App zum Vorschein trat: die *Darstellung* der Explainability. Dabei wurde der Vorschlag gemacht, Informationen zur Funktionsweise der App einerseits bei einer Einführungstour durch die App zu erhalten oder andererseits „[…] in den Benutzerrichtlinien oder in den Einstellungen. Gibt so einen Unterpunkt, wo man auf Info gehen kann, und dann steht da, das ist der Algorithmus, mit dem die KI Entscheidungen trifft.“ (FG3_Musik, 133-135).

Der Wunsch, *besonders im Fehlerfall* eine Erklärung zu erhalten, wurde in der Finanz-App angesprochen: „[…] Also dann muss ja irgendwas schiefgelaufen sein und dann wüsste ich schon ganz gerne, was ist da schiefgelaufen und kann sowas nochmal passieren […]“ (FG1_Start und Finanz, 771-773). Das Motiv für Transparenz stellt hier einen Schutz vor (weiteren) Fehlern dar. Ein weiteres Motiv

für Transparenz zeigt sich in der Subkategorie 2 *Zweifel an Funktionalität*, der zufolge Transparenz dem grundsätzlichen Zweifel gegenüber dem System entgegenwirken soll.

Die letzte Subkategorie 2 in Explainability ist zusammengefasst unter *Kein Interesse* und umfasst Aussagen zur Musik- sowie Finanz-App. Denjenigen, die sich in dieser Kategorie äußern, steht die Funktionalität an erste Stelle, während der technische Weg zu dieser Funktionalität irrelevant ist: „[…] also eigentlich bin ich so ein Typ Mensch, dem relativ egal ist, wie genau die Technik funktioniert, Hauptsache, sie funktioniert." (FG1_Start und Finanz, 353-355)

> „Aber jetzt bei der Musik App, ehrlich gesagt, ist mir das da jetzt nicht so wichtig, dass ich weiß, wie genau das funktioniert. Also so lange das tatsächlich dann funktioniert, […] ist es mir relativ egal, wie genau die Technik dahinter jetzt ist oder wie genau, ja, die App jetzt herausfindet, welche Musik ich gerne hören möchte." (FG1_Abschluss, 142-146)

Dem entgegen äußert ein*e Teilnehmer*in das geringe Interesse an Erklärungen aus der Erwartung heraus, diese wegen eines grundsätzlich fehlenden Zugangs zum Thema nicht zu verstehen:

> „Ich glaube, ich würde, wenn es um Finanzen geht, ein persönliches Gespräch immer bevorzugen, weil man überhaupt gar keinen Plan hat, wie das programmiert ist, welcher Algorithmus dahintersteckt. […] Also, finde ich, da hast du ein Programm und du hast den Code noch nicht mal, und selbst wenn du den Code hättest, würdest du es nicht verstehen." (FG3_Finanz, 119-125)

Neben *Datenschutz*, *Hintergrundinformation*, *Sicherheit* und *Explainability* ergaben sich zwei weitere Subkategorien 1 in der Oberkategorie Transparenz. Zum einen *Accountability/Rechenschaft* und zuletzt *Akkuratheitsinformation*. Beide sind auch Bestandteile allgemeiner Transparenzdefinitionen, wie in den Kapiteln 2.2.1 und 2.2.2 dargelegt. Die *Rechenschaftspflicht* wurde lediglich in Bezug auf die Finanz-App, hier aber vielfach, angesprochen: „Und zweite Frage wäre halt, wer haftet am Ende dafür, falls die KI komplett, kompletten Müll verursacht […]?" (FG1_Start und Finanz, 365-366)

> „Und eine Person kann man auch noch zur Rechenschaft ziehen, wenn die einen jetzt betrügt, dann kann man sagen, hier XY hat das und das gesagt, stimmt das? Aber die App kann ja nicht zur Rechenschaft gezogen werden. Da ist dann die Frage: Wer wird dann danach zur Rechenschaft gezogen? Ist das der Programmierer? Ist das die Firma? Keine Ahnung. Wer auch immer." (FG3_Finanz, 128-133)

Die Forderung nach *Akkuratheitsinformationen* kam einige Male bei der Finanz-App und ausführlich bei der Pilz-App zur Sprache. Diese Subkategorie 1 weist Parallelen mit der Subkategorie 2 *selbst verantwortlich* im Fehlerfall auf, in der häufig angesprochen wurde, bei einer vorangehenden Warnung zur App-Leistung selbst für mögliche Fehler verantwortlich zu sein. Als eine Maßnahme konzentriert

sich die Subkategorie 1 *Akkuratheitsinformationen* auf die Warnung selbst. Dabei geht es um „[…] so eine Risikoeinschätzung […], also wie viel Prozent Wahrscheinlichkeit wird diese Vorhersage verwirklicht oder, dass du also so eine Einschätzung [bekommst] zwischen Gewinn und Verlust“ (FG1_Start und Finanz, 798-800). Es geht darum, zu erfahren: „Wie sicher ist das denn eigentlich, dass er einen Pilz erkennt? Theoretisch müsste da noch irgendwie so ein Indexmarker dann sagen, ok, ich habe den zu 80 % identifiziert oder ich bin mir totsicher, dass der Pilz essbar ist“ (FG2_2Pilz, 52-55) und „man nicht nur das eine Ergebnis kriegt, sondern dass die KI eben auch zugibt, dass sie es nicht genau weiß, und sagt, ja, es könnte das oder das oder das sein.“ (FG2_1Pilz, 104-106)

5.3.1.3. Die Kategorie Nutzendenfaktoren

Da die *Nutzendenfaktoren* nicht im Fokus der Analyse standen, ergaben sich für diese Oberkategorie die wenigsten Inhalte. Tabelle 15 stellt die Subkategorien der Oberkategorie Nutzendenfaktoren dar.

Tabelle 15: *Subkategorien der Oberkategorie Nutzendenfaktoren. Die rechten drei Spalten bilden ab, wie häufig das Thema einer Kategorie abhängig von der genannten KI-App angesprochen wurde (für mehr Details zur Auswertung siehe Kapitel 5.2.5 bzw. Kapitel 5.3.2 für die Ergebnisse).*

Subkategorie 1	Subkategorie 2	Subkategorie 3	Musik-App	Pilz-App	Finanz-App
Nutzen	Kein Nutzen		xxx	xxx	xxx
	individueller Nutzen	Neues entdecken	xx		
		Unterstützung im Alltag		xx	x
		Neugierde/testen	x	x	x
		lebensrettend		x	
		unter Bedingungen		x	xx
		zeitsparend	x		
Individuelle Faktoren	Kontrolle	Wunsch nach Kontrolle	xxx	xxx	xxx
		keine Kontrolle nötig	x		
	Vorerfahrung	positiv		x	
		neutral			x
		negativ		x	xxx
	Risikowahrneh-mung	Hohes persönliches Risiko	x	xxx	xx
		Niedriges persönliches Risiko	x		
	Vertrauen	Kein Vertrauen		xxx	xxx
		situationsabhängig		xxx	xxx
		Vertrauen vorhanden		x	

Die Oberkategorie teilt sich auf in *Nutzen* und *individuelle Faktoren*. In *Nutzen* wurden Aussagen zu Nutzungsgründen zugeordnet. Über alle drei Apps hinweg, gab es Diskussionsteilnehmende, die die jegliche Nutzung der App ablehnten (*kein Nutzen*). Dabei ging es um die Art der Interaktion – „[…] Wobei ich es mir komisch vorstelle mich mit meinem Handy zu unterhalten“ (FG2_2Musik, 47-48) – bis hin zu einer grundsätzlichen Skepsis gegenüber der Funktionalität:

> „[…] ich würde die App wahrscheinlich überhaupt nicht nutzen, weil die viel zu lange brauchen würde für mich, um auf mich individuell zu reagieren. Daher wäre mir auch egal wie schnell

> oder wie die App das zusammenstellt, weil es wahrscheinlich nicht funktionieren würde.“ (FG2_2Musik, 142-145)

Bei der Pilz-App wurde eher der Nutzungsfall selbst in Frage gestellt: „For me I think, I would not use this App because I´m not really going for mushrooms […]“ (FG2_2Pilz, 17-18).

> „[…] Also ich würde es nicht machen, wenn ich gar keine Ahnung von Pilzen habe. Und auch, also der App quasi blind vertrauen, wäre eine schlechte Idee. Wenn ich Pilz Experte bin, dann ist das vielleicht ein cooles Gimmick dabei zu haben, aber dann kennt man sich ja eigentlich auch aus.“ (FG3_2Pilz, 141-144)

Die Ablehnung bei der Finanz-App war teilweise sehr vehement und grundsätzlich: „Also ich würde sie unter keinen Umständen benutzen, […]. Es wirkt alles so ein bisschen […] unseriös“ (FG3_Finanz, 16-20) und „[…] ich würde niemals irgendwas mit Geld einfach über eine KI-App oder irgendwie so was machen.“ (FG2_Finanz, 47-48)

> „Ja, wenn alles so einfach wäre, was mit Geld vermehren zu tun hat, dann wären alle Leute reich, dann würde jeder die paar Pfennig, die er übrig hat, irgendwo reintun und würde sichergehen, dass es funktioniert, ne. Also das ist ein illusorisches Denken, was da ist, das geht gar nicht. Also von daher, so was ist Quatsch. Totaler Quatsch.“ (FG3_Finanz, 292-295)

Der Grund durch die Nutzung *Neues entdecken* zu wollen, war einer, der nur in Bezug auf die Musik-App zur Sprache kam. Hingegen sprachen einzelne Teilnehmende das Motiv, *Unterstützung im Alltag* erhalten zu wollen, bei einzelnen Pilz- und Finanz-App an. Der Beweggrund die App aus *Neugierde auszuprobieren* und zu testen, wurde für alle drei Apps geäußert. Den *lebensrettend*en Aspekt, im Falle einer Vergiftung zu wissen, was man gegessen habe, sprach ein*e Teilnehmer*in an. Ein*e weitere*r betonte, durch die Musik-App *Zeit sparen* zu wollen. Die Nutzung unter Vorbehalt und nur *unter Bedingungen* wurde insbesondere für die Finanz-App geäußert. Diese Bedingungen betrafen zum einen bereits zuvor angesprochene finanzielle Beschränkungen sowie eine Versicherung.

Die Subkategorien 2, die sich bei *individuellen Faktoren* ergaben, waren *Kontrolle*, *Vorerfahrung*, *Risikowahrnehmung* und *Vertrauen*. Zu *Kontrolle* zählte zum einen die Subkategorie 3 *Wunsch nach Kontrolle*. Diese Subkategorie weist Parallelen mit den Subkategorien 3 *Nutzende letzte Entscheidung* in den *Systemanforderungen* auf. Während es dort im Kern um das System ging, wiesen Aussagen hier explizit Bezug zu individuellen Meinungen oder Einstellungen auf. Die Subkategorie 3 *Wunsch nach Kontrolle* umfasst zahlreiche Aussagen zu allen drei Apps, die den Kern enthalten, die Kontrolle nicht abgeben, sondern selbst behalten zu wollen: „[es] wäre nicht so meine App, weil ich möchte selber entscheiden, welche Musik ich höre, und ich will nicht, dass das eine App macht irgendwie.“ (FG3_Musik, 272-273)

> „[…] wenn ich mir nicht sicher bin, finde ich, kann man ja auch immer noch mal selbst googeln, so dass ich das Gefühl hätte, dass das so ein nützliches Tool zur Unterstützung ist, aber ich im Endeffekt dann noch immer meine eigenen Entscheidungen treffe.“ (FG3_2Pilz, 43-46)

> „Habe ich das auch richtig verstanden, dass ich da gar kein Mitspracherecht mehr hätte? Weil da steht, es ermittelt basierend auf deinen Antworten und Marktprognosen, eine für dich passende Anlagestrategie und verwaltet selbstständig dein Geld. […] Weil dann käme das sowieso überhaupt nicht in Frage.“ (FG3_Finanz, 271-275)

Eine einzelne Aussage befand, bei der Musik-App sei *keine Kontrolle nötig*.

Die Subkategorie 2 *Vorerfahrung* befasst sich mit Aussagen, wonach mit ähnlichen Apps bereits *positive*, *neutrale* und *negative* Erfahrungen gemacht wurden. Entsprechend der Erfahrungen fühlten sich die Personen einer Nutzung der Pilz- und Finanz-App zugeneigt oder bei schlechten Erfahrungen abgeschreckt.

Die *Risikowahrnehmung* der drei Apps unterschied sich je nach App. Bei der Musik-App wurde nur von einer Person ein *hohes persönliches Risiko* festgestellt, während bei der Pilz- und der Finanz-App viele entsprechende Aussagen gefunden wurden. Diese reichten von geplünderten Konten oder einem insolventen Anbieter bei der Finanz-App bis zu einer Verfälschung des eigenen Risikoempfindens und der Gefahr für das eigene Leben bei der Pilz-App. Nur zwei Aussagen bei der Musik-App betrafen ein *niedriges persönliches Risiko*.

Die Subkategorie 2 *Vertrauen* enthält ausschließlich Aussagen zu Pilz- und Finanz-App. Ein großer Teil der Aussagen findet sich in der Subkategorie 3 *Kein Vertrauen* und verdeutlicht die große Skepsis gegenüber den beiden Apps. Aussagen zu *vorhandenem Vertrauen* beschränkten sich auf die Pilz-App. Mehrere Aussagen zu Pilz- und Finanz-App ließen sich *situationsabhängigem* Vertrauen zuordnen, wobei sich dort die Erkenntnisse aus den vorherigen Kategorien widerspiegeln: Informationen über Urheber, mehrere Ergebnisse zum eigenen Vergleich, ganz allgemein mehr Informationen oder die Einschätzung von Dritten über Zertifikate oder Versicherungen – das waren die vertrauenserweckenden Maßnahmen, die hier erneut angesprochen wurden.

5.3.2. Unterschiede nach Systemeigenschaften

Neben der Erhebung der Transparenzanforderungen, die von Laiennutzenden an KI gestellt wurden, bezieht sich die zweite Leitfrage auf die Anforderungen an Transparenz, abhängig von den Systemeigenschaften. Nachdem bei der Beschreibung des Kategoriensystems im vorigen Kapitel die Ergebnisse entlang der Kategorien des Systems – in Zeilen – beschrieben wurden, folgt nun eine Betrachtung der Spalten und damit die Beantwortung der zweiten Leitfrage (siehe Tabelle 13 für die

Oberkategorie *Systemanforderungen*, Tabelle 14 für die Oberkategorie *Transparenz* und Tabelle 15 für die Oberkategorie *Nutzendenfaktoren):*

FF b (2): Was erwarten Nutzende abhängig von gegebenen Systemeigenschaften?

Der Fokus lag darauf, zu analysieren, welche Themen in den Apps unterschiedlich stark präsent waren. Neben der reinen Anzahl der Aussagen pro Kategorie wird im Folgenden auch deren inhaltliche Ausrichtung betrachtet.

Bei den Anforderungen an *weitere Informationen* allgemein oder *über Funktionen der App* kamen bei allen Apps Kommentare und Fragen auf, wobei sich wiederholt und im Vergleich mit den anderen Apps die Vorerfahrung mit der Musik-App ähnlichen Systemen zeigte. Die gewünschten zusätzlichen Informationen bezogen sich beispielsweise auf das Bezahlmodell. Die Performance-Ansprüche wurden in Abgrenzung mit bestehenden Apps – besser als diese – geäußert. Obwohl einige Aussagen zu Performance-Ansprüchen bei der Musik-App getätigt wurden, wurden auch bei den anderen beiden Apps hier Ansprüche gestellt (besser als ich bei der Finanz- und insgesamt gut bei der Pilz-App). Der Einfluss der Vorerfahrung wurde besonders deutlich in der Subkategorie 1 *Zusätzlich gewünschte Funktionen (Systemanforderungen*). Die *Kompatibilität mit anderen Apps* ebenso wie die *Personalisierung der App* wurde fast ausschließlich bei der Musik-App diskutiert. Die präzisen Vorstellungen und Beschreibungen, teilweise auch in Abgrenzung zu bestehenden Systemen, machten bei der Musik-App den großen Einfluss bestehender Systeme deutlich: „Wenn du die App runterlädst, musst du so ein Profil ausfüllen. Dann kannst du anklicken, welche Musikstile du magst. Das wäre relativ simpel." (FG2_2Musik, 155-156)

Hingegen wurden *Mehrere Ergebnisvorschläge* und *Erklärungen zum Ergebnis* ausschließlich bei Pilz- und Finanz-App gefordert. Das Motiv bei der Pilz-App war in erster Linie der Wunsch, die Ergebnisse selbst nochmal prüfen zu wollen. Hier wurde bei der einem Empfehlungssystem sehr ähnlichen Pilz-App auch die Vorerfahrung mit ähnlichen Systemen deutlich:

> „Das mit dem Vergleichsfoto ist tatsächlich eine gute Idee, weil wenn man dann noch mal ein Foto von Google hat zum Beispiel, wo man sowieso wahrscheinlich nachgucken würde, wenn man wissen will, was das für ein Pilz ist, dann wird man auf jeden Fall gut abgesichert." (FG3_2Pilz, 52-55)

Von der Finanz-App hingegen wünschten die Teilnehmenden in erster Linie mehr Information zum Ergebnis, da die App an der Stelle nur vermittelt: „[...] dieses Unternehmen da vorstellen vielleicht. Also wo ist der Sinn von dem Unternehmen? Quasi, dass ich nicht selber noch mal googeln muss, was ist das da, was der mir vorschlägt?" (FG1_Start und Finanz, 569-572) Diese Aussage und das

Selbstverständnis, nochmal selbst zu googeln, zeigten den hohen Bedarf, die Ergebnisse prüfen zu wollen.

Was *Konsequenzen aus Fehlern* anging, bestanden diese über alle Apps hinweg häufig in *Löschung* bzw. Nutzungsabbruch. Eine *Weiternutzung* war fast nur bei der Musik-App, derjenigen mit der geringsten Fehlerrelevanz, ein Thema: Einige Teilnehmende äußerten analog zum Pre-Test, ein Fehler der Musik-App wäre nicht so folgenreich und entsprechend nicht besonders schlimm: „Wenn das [Musik-Ding] falsch liegt, suche ich mir selbst über einen anderen Kanal was Anderes raus. Das finde ich jetzt nicht so schlimm wie bei den anderen beiden." (FG1_Musik, 391-392) Hingegen kam das Thema, die Verantwortung für die Nutzung bei Fehlern selbst zu tragen, nur bei der Pilz- und Finanz-App auf.

Diese eigene Verantwortung wurde besonders dann betont, wenn die App vorher Warnung ausgesprochen hätte. Hier lässt sich ein Bezug zur Subkategorie 1 *Akkuratheitsinformationen* herstellen, die auch ausschließlich bei Pilz- und Finanz-App Aussagen gefordert wurden.

Die *finale Entscheidung* zu treffen (in *Systemanforderungen*) bzw. *Kontrolle* zu behalten (in *Nutzendenfaktoren*) war ein zentrales, vielleicht das wichtigste Bedürfnis der Teilnehmenden über alle drei Apps hinweg. Dabei war das Bedürfnis einerseits bei den beiden risikoreicheren Apps ausgeprägter: So bestanden die Teilnehmenden darauf, wenn sie der Finanz-App überhaupt Entscheidungsoptionen zugestanden, mitentscheiden zu können und die Anlage nicht einfach der KI zu überlassen:

> „Bevor die KI irgendetwas, eine Überweisung oder so ab einem gewissen Wert, tätigt, dass man eine Push-Notification bekommt, wo man dann halt Approve oder Disapprove drücken kann und das halt für einen selbst so angenehm wie möglich zu machen, um halt auch noch Kontrolle darüber zu haben, falls die KI mal einen Fehler macht." (FG1_Start und Finanz, 414-418)

Bezüglich der Pilz-App konnten sich zwar viele vorstellen, sie auszuprobieren, aber würden ihr selten als einziger Quelle vertrauen, sondern zusätzliche Quellen zurate ziehen:

> „[…] wenn ich mir nicht sicher bin, finde ich, kann man ja auch immer noch mal selbst googeln so, dass ich das Gefühl hätte, dass das so ein nützliches Tool zur Unterstützung ist, aber ich im Endeffekt dann noch immer meine eigenen Entscheidungen treffe." (FG3_2Pilz, 43-46)

Während die Fokusgruppenteilnehmenden bei der App zur Unterhaltung, der Musik-App, andererseits häufig grundsätzlich bereit waren, sie auszuprobieren – siehe auch die Subkategorie 3 *Neugierde/testen* in *individueller Nutzen* –, bestanden sie trotzdem auf der Funktion, in die Musikauswahl eingreifen zu können. Im Kern waren sie also auch hier nicht bereit, die Kontrolle abzugeben: „Wenn die KI mir vorschreibt, welche Playlist ich zu hören habe, das würde ich sofort

ablehnen. Aber wenn ich quasi skippen kann in der Playlist, entscheide ich wieder selber, was ich höre“ (FG2_Abschluss, 116-118).

Ein weiterer Unterschied für die drei Apps ergab sich in der Subkategorie2 *weitere Interaktionsoptionen*, wobei Bedenken und Skepsis gegenüber der Interaktion durch Sprache deutlich wurden: Eine Änderung bzw. *weitere Interaktionsoptionen* wurden ausschließlich für die Musik-App gefordert. Dazu passt der Befund aus der Transparenz-Subkategorie 1 *Datenschutz*: Dieses Thema kam bei der Musik-App besonders häufig zur Sprache, sei es als *Sorge* über die *Sammlung kritischer Daten* oder als Forderung nach *Informationen*, wie die Daten geschützt werden. Die Begründung lautete häufig, die Sprachinteraktionen würden als besonders sensible Daten angesehen, die außerdem missbrauchsanfällig seien.

> „[…] Also viele Apps sagen ja immer, es ist nur aktiv, wenn die App aktiv ist, aber im Endeffekt nehmen die einen die ganze Zeit auf oder sowas. Dass da quasi so eine Erklärung kommt, wann man aufgenommen wird. […]“ (FG1_Musik, 54-57)

Bei der Pilz-App hingegen herrschten wenig Sorgen über Datensicherheit, ein Interesse an Datenverarbeitung hatte eher andere Gründe: „Und vielleicht auch die Information, was mit meinem Bild passiert, ob das irgendwie in so eine Datenbank aufgenommen wird, das die Erkennung dann wahrscheinlich auch noch mal optimiert.“ (FG1_Pilz, 381-383)

Bei der Finanz-App wollten einige Teilnehmende wissen, wie ihre Daten verarbeitet werden und ob sie die App auch nutzen können, ohne sensible Daten – genannt wird zum Bankkonto oder zum Arbeitgeber – preiszugeben. Eine größere Sorge galt hier allerdings der Sicherheit eigener Daten und dabei insbesondere vor dem Zugriff auf das eigene Konto:

> „[…] wie ist das alles abgesichert, was ist, wenn das irgendwie gehackt wird oder so was? Haben die dann irgendwie Zugriff auf mein Bankkonto oder, keine Ahnung, was für Risiken gibt es? Also wenn das gehackt wird, ist das immer so was, ist dann das Geld futsch? […] Ist jetzt halt sehr intransparent, finde ich. Vor allem bei so einem sensiblen Thema wie Geld.“ (FG3_Finanz, 150-155)

Im Gegensatz dazu kam die Aussage, man mache sich *keine Sorgen* oder wünsche *keine Informationen* zum Datenschutz bei allen Apps vor. Als Begründung diente häufig das Argument, man teile sowieso sehr viele Daten.

Während in der Oberkategorie Transparenz der Datenschutz also einen besonderen Schwerpunkt bei der Sprach-App zur Unterhaltung erfährt, zeigte sich bei den beiden Apps mit höherer Relevanz des Themas bzw. von Fehlern ein höherer Anspruch an *Hintergrundinformationen*. Dabei waren besonders die Informationen im Fokus, die potenziell vertrauenserweckend sein könnten oder gar Sicherheiten

durch Dritte versprechen. Anforderungen zum *Risikomanagement* wurden fast ausschließlich bei Pilz- und Finanz-App diskutiert. Hintergrundinformationen zum *Urheber* wurden ausschließlich bei diesen beiden gefordert. So wünschte ein*e Teilnehmer*in Informationen zur Urheberschaft bzw. Datengrundlage und fragte sich

> „[…] waren da Pilzexperten mit involviert und können validieren, dass die Informationen, die über die verschiedenen Pilzarten rausgegeben werden, so stimmen? Oder ja, wie seriös sind die Daten und die Informationen, die über einen Pilz weitergegeben werden." (FG3_2Pilz, 81-84)

Auch Anforderungen zur *Sicherheit des Systems*, sei es durch *Sicherheitsgarantien* oder durch *Audits* durch unabhängige Institutionen, wurden nur bei der Pilz- und Finanz-App angesprochen. Im Finanzbereich wurden dabei Banken einerseits und unabhängige Institute andererseits genannt.

> „Es hilft natürlich, wenn das Ganze auf einer Firma basiert, die es schon etwas länger gibt, die vorher schon… zum Beispiel eine Bank, die es schon lange gibt. […] Dann würde ich halt sagen, okay, die werden vermutlich nicht ihr Image riskieren, indem sie versuchen, mich zu scamen." (FG1_Start und Finanz, 722-726)

> „Das heißt, eine App, die ich vielleicht nutzen würde, die müsste dann irgendwie vielleicht von, keine Ahnung, Stiftung Warentest kommen oder so was. Halt so einer glaubhaft unabhängigen Institution, die kein Gewinninteresse an meiner Anlagenentscheidung hat." (FG3_Finanz, 267-270)

Zuletzt kamen auch geforderte Informationen zur Rechenschaftspflicht lediglich aber wiederholt bei der Finanz-App zum Tragen, also derjenigen App, bei der Fehler als besonders fatal wahrgenommen wurden.

Hingegen zeigten sich bei der Musik-App keine Wünsche über Informationen zur Urheberschaft. Vertrauen war dort weniger durch einen externen, vertrauenswürdigen Akteur herzustellen, vielmehr wollten die Teilnehmenden selbst die Vorschläge der App testen und beurteilen. Allerdings wurden von einigen Teilnehmenden auch bei der Musik-App Fragen nach dem Geschäftsmodell erhoben, allerdings standen diese weniger im Zusammenhang mit gewünschter Transparenz des Systems, sondern schienen eher von Erfahrungen mit ähnlichen Systemen wie Spotify herzurühren:

> „Was ist eigentlich mit Kosten? Ne, die Musik hat ja nun bestimmte Rechte und die Künstler wollen auch leben. Also so eine Information woher sie streamt und ob Kosten anfallen, wäre ja nicht verkehrt." (FG2_1Musik, 128-130)

Die Transparenz-Subkategorien innerhalb von *Explainability* zeigen hingegen eine gleichmäßige Streuung über alle Apps und es ließen sich keine Systemabhängigkeiten identifizieren. Selbst die Aussage, man habe *kein Interesse* an Erklärungen, wurde, zumeist mit der Begründung das technische sowieso nicht zu verstehen, bei der risikobehafteten Finanz-App ebenso getätigt wie der zur Unterhaltung dienlichen Musik-App.

In den *Nutzendenfaktoren* fand sich die grundsätzliche Ablehnung über alle Apps hinweg in der Subkategorie 2 *Kein Nutzen*. Auch die Aussagen zum *Nutzen* der Apps waren über alle drei Apps gestreut, wobei die Subkategorien 3 *Neues entdecken* ebenso wie *Neugierde* einen Überhang bei der Musik-App aufwiesen und die Nutzung *unter Bedingungen* besonders bei der Finanz-App betont wurde. Darüber hinaus wurde erneut deutlich, dass in allen Apps ein ausgeprägter *Wunsch nach Kontrolle* herrscht.

Interessanterweise und entgegen dem Eindruck, der gewonnen wird, wenn man die zuvor getätigten Aussagen inhaltlich bewertet, wurden zur Musik-App keine expliziten Aussagen zu Vorerfahrung gemacht. Bei der Finanz-App bezieht sich der größte Anteil an Aussagen zu *negativer Vorerfahrung* nicht auf technische Unterstützungsdienste in Finanzfragen, sondern umfasst Aussagen über Banken, Trading und Anlagen allgemein. Darüber lässt sich sicherlich ein Teil der grundsätzlichen Ablehnung gegenüber der Finanz-App erklären und es verdeutlicht den großen Einfluss, den die Vorerfahrung hat: Sie rührte nicht nur von der Nutzung ähnlicher technischer Systeme, sondern beruhte auf systemunabhängigen Domänen- oder Alltagserfahrungen.

Auch die Wahrnehmung eines *hohen persönlichen Risikos* bei der Pilz- und der Finanz-App bzw. eines *niedrigen persönlichen Risikos* bei der Musik-App, spiegelte die Bewertungen aus dem Pre-Test wider. Aussagen zum Vertrauen beziehen sich ausschließlich auf Pilz- und Finanz-App. Viele davon äußern nur eingeschränktes bzw. gar *kein Vertrauen* in die entsprechenden Apps.

5.3.3. Zusammenfassung der Ergebnisse

Aufgrund der zahlreichen und detailliert aufgeführten Ergebnisse werden diese im Folgenden kurz zusammengefasst. Hinsichtlich der Oberkategorie der **Systemanforderungen** fanden sich Informationen über die Funktionen der App, Ansprüche an die Leistung des Systems oder gewünschte Funktionen. Es zeigten sich die folgenden Ergebnisse:

Zusammenfassung zur Oberkategorie Systemanforderungen:

1. Weitere benötigte Informationen betreffen in erster Linie Informationen über die App, wie anfallende Kosten oder Bewertung durch andere, sowie zu Funktionen der App, z. B. durch weitere Anweisungen.

2. Bei den Performance-Ansprüchen lassen sich die Anforderungen unterscheiden: Die Anwendung soll besser sein als andere Apps, als man selbst und grundsätzlich gut.
3. Als Konsequenzen von Fehlern wurden am häufigsten das Löschen der App/das Nutzungsende genannt. Auch kam das Bedürfnis auf, Feedback zu geben in der App oder direkt an Entwickler*innen. Teilweise kam ein App-Wechsel, besonders bei der Musik-App auch eine Weiternutzung in Betracht. Viele Teilnehmende sahen insbesondere bei der Pilz-App die Verantwortung für den Fehler bei sich selbst.
4. Zusätzliche Funktionen, die sich Teilnehmende von den Apps wünschen, betrachten insbesondere, als Nutzende die letzte Entscheidung behalten zu wollen. Eine weitere Personalisierung war besonders bei der Musik-App, weitere Erklärungen zum Ergebnis wurden bei Pilz- und Finanz-App gewünscht.

Bezüglich der nachträglich ergänzten Oberkategorie der **Transparenz** kamen besonders häufig Anforderungen an den Datenschutz zur Sprache. Außerdem zeigten sich Anforderungen hinsichtlich Hintergrundinformationen, zu Explainability sowie zu Akkuratheitsinformationen.

Zusammenfassung zur Oberkategorie Transparenz:

1. Anforderungen zum Datenschutz waren besonders prominent bei der Musik-App. Es zeigten sich Sorgen zur Anonymität und zur Rückverfolgbarkeit der Daten. Informationen wurden gefordert zur Verarbeitung und zur Art der gespeicherten Daten – Letzteres auch bei der Finanz-App.
2. Hintergrundinformationen zu Profit und ethischen Bedenken waren besonders bei der Finanz-App gefragt. Dabei schien die Frage nach dem Urheber wichtig für das Vertrauen in die App. Dies zeigte sich auch bei der Pilz-App.
3. Sicherheiten des Systems durch Garantien/Audits wurden besonders zur Finanz-App gefordert.
4. Bei globaler Explainability waren grundsätzliche Erklärungen von Interesse, sehr häufig angesprochen dabei Trainingsdaten. Bei lokaler Explainability wurden Warum-Erklärungen zu den Ergebnissen bei allen drei Apps genannt. Ziel war Transparenz im Fehlerfall und um Zweifel auszuräumen.
5. Teilweise jedoch bestand auch kein Interesse an Erklärungen mit der Begründung, diese seien zu komplex, nicht relevant oder nicht zu verstehen.
6. Akkuratheitsinformationen wurden in Bezug auf Pilz- und Finanz-App gefordert.

Zuletzt setzt sich die Oberkategorie Nutzendenfaktoren aus Nutzungsursachen und individuellen Faktoren wie Vorerfahrung und Bedürfnis nach Kontrolle zusammen.

Zusammenfassung der Oberkategorie Nutzendenfaktoren:

1. Bezüglich allen drei Apps äußerten Teilnehmende, keinen Nutzen in der Anwendung zu sehen.

2. Neugierde und Austesten wurden als Beweggründe für alle drei Apps geäußert. Eine Nutzung unter Bedingungen wurde besonders bei der Finanz-App betont.
3. Der Wunsch nach Kontrolle war ein zentraler Punkt über alle drei Apps hinweg. Dass keine Kontrolle nötig sei, wurde nur einmal bei der Musik-App erwähnt.
4. Der Einfluss der Vorerfahrung, besonders negativer, zeigte sich ausgeprägt bei der Finanz-App, wobei es sich nicht nur um KI-Angebote, sondern auch um Vorerfahrungen mit Banken und Anlagen handelte.
5. Hohes persönliches Risiko wurde besonders bei der Pilz- und Finanz-App wahrgenommen, ebenso wie kein Vertrauen oder nur bedingtes Vertrauen ausgesprochen.

Für die Beantwortung der zweiten Leitfrage wurden die Unterschiede zwischen den Apps analysiert, um daraufhin Schlüsse für **Anforderungen hinsichtlich der Systemeigenschaften** zu ziehen. Dabei zeigten sich die folgenden Kernergebnisse:

1. Allgemeine Anforderungen an Informationen und Funktionen variieren je nach Vorerfahrung mit ähnlichen Systemen.
2. Der Einfluss der Vorerfahrung zeigt sich bei der Musik-App durch hohe Erwartungen an Personalisierung und Kompatibilität mit anderen Apps. Gleichzeitig wurde eine eher geringe Fehlerrelevanz festgestellt.
3. Bei der Pilz-App zeigt sich der Einfluss der Vorerfahrung im Wunsch nach mehreren Ergebnisvorschlägen und Erklärungen, ebenso wie in der hohen Relevanz von Akkuratheitsinformationen.
4. Bei der Finanz-App zeigten sich hoher Informationsbedarf zu Ergebnissen und viele Sicherheitsbedenken, außerdem eine hohe wahrgenommene Fehlerrelevanz und persönliches Risiko mit einem großen Einfluss negativer vorheriger Erfahrungen.
5. Das Bedürfnis nach Kontrolle und die finale Entscheidung selbst fällen zu wollen, zeigte sich bei allen Apps und war bei risikoreich wahrgenommener KI nochmals stärker.
6. Datenschutzbedenken kamen in vielerlei Hinsicht bei der Musik-App zur Sprache, hauptsächlich aufgrund ihres Bestandteils der Sprachinteraktion, weniger bei der Pilz-App.
7. Hintergrundinformationen waren gewünscht zur Steigerung von Vertrauen, besonders bei hoher Fehlerrelevanz, z. B. durch Informationen zu Urhebern und externe Sicherheiten (Audits, Lizenzen). Eine Bewertung durch Testen oder durch Dritte kam bei der Musik-App zum Tragen.
8. Geringer Nutzen wurde als Grund für eine Ablehnung der Anwendung bei allen drei Apps genannt. Eine Ablehnung aufgrund mangelnden Vertrauens nur bei Pilz- und Finanz-App.

5.4. Diskussion

Mit den Fokusgruppendiskussionen galt es die Frage (b) zu untersuchen, **welche Anforderungen an Transparenz in KI bestehen für Laiennutzende und inwiefern unterscheiden sie sich nach Eigenschaften der KI.** Während im technischen Kontext Klassifikationen von transparenter KI beispielsweise in globale und lokale Transparenz bestehen, ist in Bezug auf Endnutzende mit wenig technischem Vorwissen jedoch häufig unklar, was für sie Transparenz bedeutet. Aus der übergeordneten Forschungsfrage wurden zwei Leitfragen abgeleitet, die insbesondere Auswertung und Bericht der Ergebnisse strukturierten:

FF b (1): Welche Anforderungen an Transparenz in KI-Apps haben Laiennutzende?

FF b (2): Was erwarten Nutzende abhängig von gegebenen Systemeigenschaften?

Um das Transparenzverständnis im Hinblick auf verschiedene KI-Arten zu untersuchen, wurden drei beispielhafte Apps mit mehreren Fokusgruppen und drei Fragen diskutiert. Die Fragen bezogen sich in drei Runden darauf, (1) was die App erklären sollte, (2) unter welchen Voraussetzungen eine Nutzung in Frage käme und (3) was im Fehlerfall geschehen würde. Sie sprachen also nicht explizit die Frage nach der Transparenz an, vielmehr war das Ziel, durch indirekte Diskussionen über die App ein Transparenzverständnis abzuleiten.

Aus den Ergebnissen lassen sich zentrale Einflussfaktoren identifizieren, die maßgeblich das Transparenzverständnis beeinflussen. Diese werden im Folgenden hergeleitet (Kapitel 5.4.1), bevor das aus den Ergebnissen ableitbare Verständnis von Transparenz zusammengefasst und diskutiert wird. Dieses umfasst die technischen Kategorien von lokaler und globaler Explainability (Kapitel 5.4.2) sowie den Aspekt der Kontrolle (Kapitel 5.4.3), gefolgt von Hintergrundinformationen zu Urheber und Zertifikaten (Kapitel 5.4.4) sowie zu Privatsphäre und Datenschutz (Kapitel 5.4.5). Die Ergebnisse lassen sich in verschiedene bestehende Theorien einbetten, was in Kapitel 5.4.6 beschrieben wird. Am Ende folgen die zu beachtenden Limitationen bezüglich des Vorgehens und der Ergebnisinterpretation.

5.4.1. Einflussfaktoren auf Transparenzanforderungen

Der Einfluss der Vorerfahrung manifestiert sich an verschiedensten Stellen und zieht sich durch alle Fokusgruppendiskussionen. Obwohl die Teilnehmenden Skepsis gegenüber allen drei Apps äußerten, zeigte sich der negative Einfluss der Vorerfahrung besonders bei der Finanz-App. Diese wurde teilweise rigoros abgelehnt, mit der Begründung, grundsätzlich kein Geld anzulegen oder aufgrund schlechter Erfahrungen bei dem Thema niemandem zu trauen. Die Vorerfahrungen im Anwendungsbereich – Geldanlage in dem Fall – zeigen sich hier deutlich. Hingegen waren einige Diskussionsteilnehmende bereit und interessiert, die Pilz-App auszuprobieren, oft mit dem Verweis auf bereit bekannte ähnliche Anwendungen. Bei schlechten Erfahrungen mit ähnlichen Apps zur Pflanzenbestimmung überwog die

Skepsis. Bei der Musik-App war die Bereitschaft, das System zu testen, am höchsten. Gleichzeitig wurde auch am meisten auf die Konkurrenz durch ähnliche Systeme wie Spotify hingewiesen, weshalb der tatsächliche Nutzen dieser App wiederum in Frage gestellt wurde. Ebenfalls entsprechend der Erfahrungen, die mit verfügbaren Musikdiensten gemacht wurden, kamen bei der Musik-App wiederholt Fragen nach den Kosten der App auf.

Die Vorerfahrung mit dem Anwendungsbereich ebenso wie die Bekanntheit von als ähnlich wahrgenommenen Apps spielen bei den Ansprüchen in zweierlei Hinsicht eine große Rolle: Vorerfahrung prägt die Einstellung gegenüber den Apps, wobei Vorerfahrungen hier sehr breit zu verstehen sind. Sie beschränken sich nicht nur auf Vorerfahrungen mit technischen Systemen, sondern umfassen auch Menschen, wie beispielsweise den Bankberater, oder negative Einstellungen gegenüber der Thematik. Gleichzeitig prägt Vorerfahrung mit ähnlichen Apps die Anforderungen an neue Apps. Dies zeigt sich an Vorschlägen, wie Ergebnisse präsentiert werden sollen – Vergleichsbilder mit Akkuratheitsangabe bei einem Bilderkennungssystem oder der Like/Not-like-Button zur Rückmeldung und zum Trainieren der Musikempfehlungs-KI. Die konzeptionelle Unterscheidung von Entscheidungsunterstützungssystemen und Empfehlungssystemen spielt für Laien keine große Rolle: Während Letztere mehrere Optionen aufbereiten und den Vergleich sichtbar machen, bereiten Erstere eine Entscheidung vor und liefern nur ein Ergebnis (Mohseni et al., 2021). Auch wenn sich die vorliegende Arbeit eher den Entscheidungsunterstützungssystemen widmet, sind die Vorerfahrungen mit anderen Systemen bedeutend und fließen in die Erwartungen an (neuere) KI-Systeme ein. Die Ergebnisse spiegeln Untersuchungen zu Vertrauen und Automation bzw. KI wider, in denen der große Einfluss der Vorerfahrung deutlich wird (Bedué & Fritzsche, 2022; Hoff & Bashir, 2015). Studien zeigen außerdem: Wenn die durch Transparenzmaßnahmen erfolgenden Erklärungen nicht zur Vorerfahrung passen, beeinflusst das die Wahrnehmung des Systems und überschreibt sogar Transparenzeffekte (Molina & Sundar, 2022; Sundar, 2020). Sowohl **mögliche Vorerfahrungen mit ähnlichen Systemen als auch vorherige Erfahrungen und Einstellungen im Hinblick auf den Anwendungsbereich** sollten daher berücksichtigt werden bei der Entwicklung von KI-Systemen. Sie beeinflussen die Wahrnehmung des Systems und damit auch die Anforderungen an Funktionalitäten und seine Transparenz. In einer Transparenzmatrix, die zur Anwendung der Ergebnisse entwickelt wurde und im Implikationen-Kapitel (5.5.1) näher vorgestellt wird, sind daher negative wie positive Vorerfahrungen als Einflussfaktoren auf die KI-Transparenz enthalten.

Während bei der Finanz-App die im Hintergrund laufenden Prozesse für Skepsis sorgten und viel über Geld, Banken und Anlagen diskutiert wurde, spielten die Prozesse zur Musikauswahl eine geringere Rolle bei der Musik-App. Dieser Aspekt wurde möglicherweise als weniger problematisch angesehen. Diskutiert wurde hier vielmehr die Eingabemethode: das Gespräch. Da die Inhalte als privat und die

Stimme als sensibel wahrgenommen wurden, war das meistbesprochene Thema bei der Musik-App Datenschutz und Privatsphäre. Bei der Pilz-App hingegen war eines der wichtigsten Themen die Darstellung der Ergebnisse. Es zeigt sich: **Transparenz betrifft weniger ganze Systeme, sondern viel häufiger einzelne Teilbereiche**. Während bei Explainability üblicherweise Datenverarbeitung und Systemprozesse zur Ergebnisermittlung im Mittelpunkt stehen und meist bei der Ergebnisdarstellung präsentiert werden, **umfasst Transparenz für Laien eine Vielzahl weiterer und in der Gewichtung andere Aspekte.** Diese umfassen z. B. die Interaktionsweise, die Datenspeicherung oder auch Hintergrundinformationen zu Entwickler*innen oder zum Geschäftsmodell.

Zusätzlich zur Vorerfahrung hatte das Anwendungsgebiet der KI auch Auswirkungen darauf, inwiefern sich die Teilnehmenden als Expert*innen fühlten bzw. die Expertise ausgelagert war. Dies zeigte sich beispielsweise bei der Pilz-App: Da sich die meisten Diskutant*innen nicht mit dem Thema auskannten, hatten sie einerseits das Bedürfnis, die Expert*innen hinter dem System kennenzulernen. Andererseits wollten sie sich zumindest ein wenig zu Expert*innen entwickeln, indem sie mehr Informationen erhalten, um ihre Entscheidung fundierter treffen zu können. Inwiefern sich dies in der Praxis umsetzen lässt, ist in Frage zu stellen. Bei der Musik-App sahen sich die Nutzenden selbst als Expert*innen für ihren Musikgeschmack und können deshalb die Qualität der App selbst – und aufgrund der geringeren Fehlerrelevanz auch risikoärmer – prüfen. Das Wissen um die Expertise der Nutzenden zeigt sich auch, wenn Diskussionsteilnehmende zur Beurteilung der Musik-App Bewertungen durch andere Nutzende heranziehen möchten. Im Gegensatz dazu wurden bei den beiden anderen Apps die ausgelagerten Expert*innen hinterfragt: Wer steht als Pflanzenkenner*in hinter der Pilz-App, welche Bank setzt die Finanz-App um bzw. – da einige den Banken nicht trauten – welche unabhängige Institution zertifiziert die KI? **Je nach Anwendungsbereich unterscheidet sich, wo die Expertise für die KI liegt und abhängig davon ändert sich die Frage, bei welchen Aspekten Vertrauen und damit Transparenz nötig sind**. Das Thema, die Expertise für KI bei Dritten zu verorten und damit Vertrauen auszulagern, wird in Kapitel 5.4.4 Urheber und Zertifikate weiter diskutiert.

Zuletzt hat die wahrgenommene Fehlerrelevanz einen großen Einfluss auf den Anspruch an KI-Transparenz. Wie bereits im Pre-Test festgestellt wurde, unterscheiden sich die Apps in Bezug darauf, wie schwerwiegend ein möglicher Fehler wäre. Dies spiegelten auch entsprechende Aussagen in den Diskussionen wider: Fehler in der Musik-App wurden als weniger folgenschwer wahrgenommen als in der Finanz-App. Die Pilz-App war für die einen weniger gefährlich, während die anderen in einem Fehler von ihr Gefahr für Leib und Leben sahen. Dieser Fehlerrelevanz folgt eine entsprechende Sensibilität im Umgang mit der App – und daran anschließend ein erhöhter Transparenzanspruch. Der Bedarf nach Hintergrundinformationen zum Urheber der App oder vertrauensbildenden Zertifikaten kam nur bei den Apps mit hoher Fehlerrelevanz auf. Auch Fragen zur Rechenschaftspflicht und der

erhöhte Bedarf nach Hintergrundinformationen zeigten sich bei KI mit hoher Fehlerrelevanz sehr deutlich. **Insgesamt sind Skepsis und Vorbehalte und damit die Ansprüche an die Apps mit hoher Fehlerrelevanz höher**. Vorausgreifend auf die Transparenzmatrix in den praktischen Implikationen (Kapitel 5.5.1) liegt in diesen Erläuterungen der Grund, weshalb der Effekt der Fehlerrelevanz auf die Bedarfe nach Sicherheit über Dritte bzw. über Informationen zum Urheber dort als „sehr hoch" angegeben wird. Gleichzeitig ist bei hoher Fehlerrelevanz der Bedarf nach Erklärungen allgemein vorhanden, ebenso wie der nach lokaler Transparenz – jedoch nicht in selbem Maße und deshalb lediglich „hoch" (Kapitel 5.5.1).

Ähnlich wie bei gesteigerter Fehlerrelevanz wird die Transparenz von Apps bewertet, die als ethisch kritisch wahrgenommen werden. Allerdings ist diese Wahrnehmung, mehr noch als die Fehlerrelevanz, von hoher Subjektivität gekennzeichnet: Die einen hatten ethische Bedenken bei der Auswahl von Anlagen durch die Finanz-App, andere sahen in der Beeinflussung der Stimmung durch die Musik-App eine ethische Herausforderung. Die von diesen Personen in der Folge getätigten Aussagen zeigen: **Mehr noch als bei der Fehlerrelevanz steigt durch ethische Vorbehalte die Skepsis gegenüber KI und damit der Anspruch an Transparenz in fast jedem Aspekt**.

Auch bei Apps mit geringer Fehlerrelevanz kommen Themen wie Datenschutz zum Tragen und das Bedürfnis nach bestimmten Transparenzaspekten, z. B. nach Kontrolle (siehe auch Kapitel 5.4.3), besteht. Die Bereitschaft, die App einfach mal auszuprobieren oder einen Fehler zu verzeihen, ist dennoch höher. Auf den erhöhten Bedarf nach vertrauensbildenden Maßnahmen bei den als kritisch wahrgenommenen Apps wird in Kapitel 5.4.4 weiter eingegangen.

5.4.2. Globale und lokale Transparenz

Forderungen nach globaler und lokaler Explainability zeigen sich bei allen Apps. Gleichzeitig ist ihre Ausprägung aber nicht größer als die gegenüber anderen Anforderungen, sondern eben nur ein Teil des Transparenzverständnisses.

Globale Transparenz bezeichnet die Antwort auf die Frage nach „Wie funktioniert das System generell", betrifft also grundsätzliche Funktionsweisen und Prozesse der KI. In den Diskussionen dazu lassen sich zwei maßgebliche Strömungen erkennen: einerseits ein gewisses Interesse an diesen Prozessen, andererseits die Bemerkung, man würde die zugrundeliegenden Funktionsweisen nicht verstehen, was entweder das Interesse an globaler Transparenz einschränkt oder ein gänzliches Nichtinteresse begründet. Obwohl Endnutzende also einsehen, dass ihnen zu viele Details keinen Nutzen bringen, äußern sie doch teilweise ein Interesse an den Entscheidungswegen der KI. Hervozuheben ist dabei die Ergänzung einiger Nutzender, die Information könnte optional abrufbar sein für diejenigen, die sich dafür interessieren. Dies spiegelt die Ergebnisse anderer Studien wider, die

zeigen: Zu viel Transparenz kann überfordern und negativ wirken. **Vielmehr gilt es, eine schrittweise dem Wissens- und Interessenstand der Nutzenden angepasste Transparenz herzustellen**.

Lokale Transparenz, also Erklärungen, wie die einzelnen Ergebnisse zustande kommen, sind in allen Apps von Interesse – auch bei der Musik-App. Es scheint also nicht (nur) darum zu gehen, durch lokale Explainability einzelner Ergebnisse besonders risikoreiche Entscheidungen absichern zu wollen, sondern um ein eher generelles Informationsbedürfnis. Gerade bei der Musik-App zeigt sich der Einfluss des Neuheitsaspekts: Über die Funktion der Musikauswahl hinaus – die ja bekannt ist von anderen Musikdienstleistern – zielen die Fragen auf die Analyse der Sprache, der Identifikation der Stimmung und deren Verknüpfung mit der Musik, also den Funktionsweisen, die bei der Musik-App über bisher bekanntes hinausgehen. **Das Bedürfnis nach Transparenz zeigt demzufolge eine Abhängigkeit von Neuem bzw. könnte sich abschwächen, wenn Gewöhnung mit Prozessen oder Systemen eintritt.**

Andererseits verschwimmen die Grenzen von globaler und lokaler Transparenz häufig. Laien führen diese technische, theoretische Unterscheidung nicht in dem Maße aus, wie es die Definitionen vorsehen. Eine Umsetzung derselben ist weniger zielführend, als sich an Aspekten der Transparenz zu orientieren, die über diese klassische Explainability hinausgehen. Auf diese Aspekte wird im Folgenden weiter eingegangen.

5.4.3. Kontrolle

Wie zuvor beschrieben, war das Bedürfnis nach Kontrolle über das finale Ergebnis der App zu entscheiden, eines der wichtigsten Merkmale über alle drei Apps hinweg. Dabei zeigte sich das Bedürfnis bei den beiden Apps mit hoher Fehlerrelevanz ausgeprägter. Allerdings bestanden die Fokusgruppenteilnehmenden bei der App zur Unterhaltung, der Musik-App, trotzdem darauf, in die Musikauswahl eingreifen zu können, trotz der grundsätzlichen Bereitschaft, die App auszuprobieren. Im Kern waren sie also auch hier nicht bereit, die Kontrolle abzugeben, auch wenn einige Teilnehmende auf die geringe Fehlerrelevanz bei der Musik-App hinwiesen. Während also die Tragweite der Entscheidung bzw. möglicher Fehler eine Rolle spielt bei der Entscheidung, ob Nutzende bereit sind, ein System zu nutzen, waren sich alle Teilnehmenden **über die Apps hinweg einig, dass sie stets die Kontrolle wahren und keine Entscheidungen abgeben** möchten.

Noch im Pre-Test zeigte sich, der Nutzen aller drei Apps besteht unter anderem in kognitiver Entlastung: Acht bzw. sieben der 15 Pre-Test-Teilnehmenden gaben dies an. In den Fokusgruppen jedoch kam nun ein anderes Muster zum Vorschein: Die **Nutzenden sind nicht bereit oder willens, sich komplett entlasten zu lassen, sondern wollen – je nach Fehlerrelevanz mehr oder weniger – Aufwand für das finale Ergebnis betreiben**. Dies spiegelt sich in den Ergebnissen aus der Literatur wider, die zeigten, dass Transparenzbedingungen einen erhöhten kognitiven Aufwand erfordern, transparente

Darstellungen von den Versuchspersonen aber trotzdem denen ohne Transparenz vorgezogen werden (Du et al., 2019; Gedikli et al., 2014; Herlocker et al., 2000). Durch Interaktion mit dem Algorithmus ließ sich außerdem das Vertrauen in das System erhöhen (Molina & Sundar, 2022). Für die Umsetzung von Transparenz bedeutet dies: Zusätzlicher Aufwand ist gerechtfertigt, wenn er am Ende ein Verständnis für die relevanten Fragen ermöglicht.

Ein weiterer Punkt betrifft das Gefühl, durch die App Aufgaben abgenommen zu bekommen, die eigentlich Freude bereiten. Ein*e Teilnehmer*in gab an, sich gern mit dem Thema Anlagen auseinanderzusetzen. Manche Personen möchten die Befriedigung, sich in ein Thema einzuarbeiten und eine begründete Entscheidung zu treffen, vielleicht gar nicht von einer KI abgenommen bekommen. In der abschließenden Diskussion über alle Apps äußerte ein*e Teilnehmer*in zusammenfassend: „KI ist ein nützliches Werkzeug, das darf mich im Alltag gerne unterstützen, aber darf mir nicht zu viele Entscheidungen abnehmen, die ich im Endeffekt selber treffen kann oder ablehnen möchte“ (FG2_Abschluss, 99-101).

Aus diesem Satz jedoch resultiert die enorme Anforderung, die an Transparenz gestellt wird. Denn **um sinnvoll Kontrolle ausüben zu können, müssen Nutzende über die Hintergründe und Funktionsweisen von KI Bescheid wissen**. Nicht umsonst geht die Forderung nach Transparenz häufig mit der nach Entscheidungsautonomie einher und umgekehrt (z. B. AI HLEG, 2019).

5.4.4. Urheber und Zertifikate

Ein weiterer, vielbesprochener Punkt waren Informationen zu Urheber oder der Datengrundlage der Apps – jedoch ausschließlich bei den beiden risikobehafteten Apps, bei denen die Nutzenden keine Domainenexpert*innen waren. Es scheint, dass die Diskussionsteilnehmer*innen die Strategie verfolgten, **durch die Information, von wem die App entwickelt wurde, auf die Qualität der App schließen zu können**. Bei der Finanz-App zeigte sich dies im Wunsch nach Transparenz zum dahinterstehenden Geschäftsmodell bzw. denjenigen, die ein dahinterstehendes Gewinninteresse haben könnten. Das Vertrauen in die Systeme kann, so die Teilnehmenden, durch eine vertrauenswürdige Einrichtung erlangt werden. Bekannte Institutionen wie Banken, Wissensträger oder zertifizierende Einrichtungen lieferten mit ihrem Namen also die Möglichkeit, für ein neues, noch wenig bekanntes System als Vertrauensträger einzustehen. Dies entspricht den Ergebnissen der Befragung von Baldauf et al. (2020) zu medizinischen Diagnose-Systemen, bei denen Zertifikate und Empfehlungen der eigenen Ärztin/des eigenen Arztes oder von Expert*innen als besonders vertrauensbildende Maßnahmen beschrieben wurden.

Dieser Mechanismus, **Vertrauen durch Dritte** zu etablieren oder – anders gesprochen – die Unsicherheit in Bezug auf die Vertrauenswürdigkeit eines Gegenübers durch Dritte zu reduzieren, wird auch als **Transfer** bezeichnet (Corves & Schön, 2020; McKnight & Carter, 2009; Stewart, 2003). In Bezug

auf Internet und Verlinkungen beschreibt es das Phänomen, dass Vertrauen in eine Entität bei einer wahrgenommenen Verknüpfung auf eine andere Entität übergehen kann (Stewart, 2003). Sind keine anderen Informationen über das Vertrauensobjekt zu erhalten, kann als letzte Stufe bei der Vertrauensbildung ein Transfer von bekannten Quellen, z. B. von ausführenden Personen oder Institutionen, für Vertrauen sorgen (Corves & Schön, 2020). Schon im Jahr 2010 prognostizieren Bierhoff und Rohmann die zunehmende Bedeutung von Reputation und Zertifizierung für Systemvertrauen. Sie steige insbesondere angesichts der wachsenden und immer weniger kontrollierbaren Anzahl an (technischen) Einheiten, denen Nutzende Vertrauen entgegenbringen müssen (2010). Auch eine aktuelle qualitative Studie bestätigt den Bedarf an – unter anderem – Zertifizierungen und Standards aufseiten von Experteninstitutionen einerseits und Regulierungen aufseiten von Regierungen andererseits, um KI-Vertrauen zu etablieren (Bedué & Fritzsche, 2022).

Zu betonen ist der je nach Anwendungsbereich unterschiedliche Aspekt, für den die Teilnehmenden vertrauensbildende Maßnahmen einforderten. Bei den risikoreicheren Apps wurden eher holistische Ansätze verlangt: Die Entwickler*innen hinter der gesamten KI sollten qualifiziert sein, die Anbieter der KI vertrauenswürdig, ein unabhängiges Institut sollte die App geprüft haben. Bei der weniger risikoreichen KI, der Musik-App, hingegen, deren Qualität Nutzende anhand der eigenen Expertise für ihren Musikgeschmack selbst ermitteln können, wurden Sicherheiten nur bezüglich einzelner Aspekte – Datenschutz in diesem Fall – relevant. Dieser erhöhte Bedarf an **vertrauensbildenden Maßnahmen besonders bei risikoreicher KI** findet sich wieder in den aktuell entwickelten Gesetzen der EU im AI Act. Bei risikoreicher und hochrisikoreicher KI – Letztere betrifft Domänen, die auch ohne KI gesetzlich geregelt sind wie Medizin oder Recht – bestehen höhere Anforderungen, die unter anderem durch Zertifizierungen gewährleistet werden sollen (Verordnung über künstliche Intelligenz, 2024). Darüber hinaus bestehen spezielle Regulationen für besondere Unteraspekte von KI bzw. digitalen Angeboten allgemein: Prominentes Beispiel ist die DSGVO, in der Maßnahmen und Rechte zum Datenschutz und zur Privatsphäre verankert sind (DSGVO: Datenschutz-Grundverordnung, 2016).

Zusammengefasst lässt sich also der **Bedarf nach vertrauensbildenden Informationen** feststellen, insbesondere in risikobehafteten KI-Anwendungen: Diese **Hintergrundinformationen umfassen zum einen solche zu Urheberschaft und genutzten Quellen und zum anderen zum Interesse dieser Urheber oder Anbieter der KI, sei es an entstehenden Gewinnen oder der Datennutzung**. Damit sind diese Hintergrundinformationen ein über das technische Verständnis von Transparenz hinausreichender Aspekt, der bei Laien zum Tragen kommt. Während einige Teilnehmende explizit kommentierten, kein Interesse an technischen Erklärungen zu haben, sind diese Hintergrundinformationen intuitiver verständlich und ähneln denen, nach denen wir auch menschliche Ratgeber beurteilen. Die Frage nach der Expertise einer unterstützenden Person, also worauf baut sie

ihr Wissen, wie hat sie es erlangt oder bewiesen, ist ein Aspekt, den Nutzende bei der Überlegung, ob sie angebotene Unterstützung in Anspruch nehmen, berücksichtigen (Bonaccio & Dalal, 2006). In Bezug auf KI **entsteht Vertrauen durch bekannte, vertrauensvolle (Urheber- und Anbieter-) Namen, aber auch durch unabhängige Zertifikate und gesetzliche Regulierungen**, die eingehalten werden müssen und Standards gewährleisten. In der in den Implikationen aufgeführten Transparenzmatrix (Kapitel 5.5.1) wird dementsprechend bei hoher Fehlerrelevanz der Zusammenhang von „Sicherheit über Dritte“ oder „Informationen zum Urheber“ als „sehr hoch“, Informationen zu globaler oder lokaler Transparenz lediglich als „hoch, nach Interesse“ gekennzeichnet.

5.4.5. Privatsphäre und Datenschutz

Das meistbesprochene Thema bei der Musik-App war – in allen Gruppen und sehr viel häufiger als bei den anderen beiden Apps – das Thema Datenschutz. Informationen zum eigenen Befinden, zum Tag und generell zu Sprachdaten wurden als sehr sensibel wahrgenommen, weshalb ein großer Bedarf nach Transparenz über Datenverarbeitung und -nutzung aufkam. Im Gegensatz dazu fielen die datenschutzrechtlichen Sorgen bei den beiden anderen im Ergebnis risikobehafteteren Apps geringer aus.

Einerseits ist damit Datenschutz ein Aspekt, der über das technische Transparenzverständnis hinaus bei Laien eine Rolle spielt. Andererseits zeigt sich der Aspekt der Datensicherheit als ein Unterbereich, der sich nur sensiblen Datenbelangen zeigt. Die Einschätzung davon, was sensible Datenbelange sind, scheint sehr subjektiv zu sein und ist wiederum durch den Faktor Vorerfahrung beeinflusst. So steht im Unterhaltungskontext, in dem KI schnell und ohne Aufwand funktionieren soll, einerseits Sicherheit nicht im Fokus, ebenso wenig wie Akkuratheit. Allerdings existieren bereits vergleichbare Anwendungen, die eine sprachgesteuerte Interaktion ermöglichen, wie z. B. Siri und Alexa. In deren Zusammenhang wird regelmäßig von Datenschutzverletzungen und Datenmissbrauch berichtet (M. R. Das, 2024). Die Wahrnehmung von Sprachdaten als besonders privat und sensibel könnte teilweise auch von solchem Vorwissen beeinflusst sein. Nichtsdestotrotz zeigt sich an diesem Aspekt besonders deutlich, dass **selbst in KI, die weniger risikobehaftet wahrgenommen wird, Transparenz eine Rolle spielt, wenn auch möglicherweise weniger in Bezug auf das gesamte System und eher für einzelne Unteraspekte**. Dies entspricht der Vertrauenskonzeption nach Bierhoff und Rohmann (2010), nach der Vertrauen keinen dichotomen Zustand darstellt, sondern aus mehreren Schritten besteht und sich auf bestimmte Situationen – bzw. bestimmte Teilsysteme – bezieht. Diese Unteraspekte zu identifizieren und transparent zu gestalten, ist also selbst in Systemen mit geringer Fehlerrelevanz von Bedeutung. Entsprechende Zusammenhänge zwischen als sensibel wahrgenommenen Daten und den Auswirkungen auf Transparenzanforderungen finden sich in der Transparenzmatrix (Kapitel 5.5.1). Dort werden, wenn ein System als „sensible Daten“ verarbeitend wahrgenommen wird,

Anforderungen an Sicherheit über Dritte als „hoch“ und der Bedarf nach Informationen zum Datenschutz als „sehr hoch“ dargestellt.

Zuletzt sei auf einen besonderen Effekt in der Forschung zu Privatsphäre hingewiesen: Das sogenannte Privacy-Paradox steht für das Phänomen, wonach Nutzende in Befragungen Privatsphäre und Datenschutz als relevante Aspekte nennen, die sie in ihrer Nutzung berücksichtigen würden. Untersucht man aber die tatsächliche Nutzung, zeigen sie sorgenfreies Verhalten, ohne Rücksicht auf Datenschutz (Barth & de Jong, 2017). Besonders bei qualitativen Erhebungen, wie der vorliegenden, sind Aussagen über Datenschutz als Voraussetzung für die Nutzung von KI also mit Vorsicht zu interpretieren. Inwiefern sich dieser Effekt des Datenschutzes auch auf die weiteren hier ermittelten Transparenzaspekte übertragen lässt, gilt es zu diskutieren (siehe Kapitel 5.4.7). Jedoch lassen sich, wie der folgende Abschnitt zeigt, die vorliegenden Ergebnisse in etablierte Interkations-, Verhaltens- und Nutzungstheorien einreihen, was wiederum die Gültigkeit der Ergebnisse bestärkt.

5.4.6. Einbettung in bestehende Theorien

Die Ergebnisse aus den Fokusgruppen lassen sich vor dem Hintergrund verschiedener Theorien aus den Bereichen der Kommunikationswissenschaften und der Psychologie einordnen. Eine klassische und inzwischen immer wieder überarbeitete Theorie aus den Kommunikationswissenschaften zu initialen Interaktionen ist die „**Uncertainty Reduction Theory**“ (URT; C. R. Berger & Calabrese, 1975). Diese Theorie, ursprünglich für menschliche Interaktionen entwickelt, besagt im Kern, das Interesse bei initialen Interaktionen bestehe in erster Linie darin, Informationen zu sammeln, die die eigene Unsicherheit über das gezeigte und zukünftige Verhalten des Gegenübers verringern. Wenn die nötigen Informationen gefunden werden, reduzieren sie die Unsicherheit und das wahrgenommene Interaktionsrisiko. Die Informationssuche zur Reduzierung der Unsicherheit kann, wenn keine verbale Interaktion möglich ist, auch mittelbar geschehen, z. B. durch Beobachtung oder durch Informationen anderer (Venkatesh, Thong, et al., 2016).

In seiner „**Motivation to Reduce Uncertainty-Theorie**“ (MRU), die auf der URT aufbaut, ergänzt Kramer die motivationale Komponente (Kramer, 1999). Sie überwindet verschiedene Kritikpunkte an der URT: Zum einen bedeutet mehr Information nicht automatisch weniger Unsicherheit. Zum anderen hat die Toleranz einer Person gegenüber Unsicherheit einen großen Einfluss darauf, ob Unsicherheitsreduzierung betrieben wird oder nicht. Die MRU berücksichtigt also situationale und individuelle Unterschiede. Noviz*innen und Expert*innen unterscheiden sich in ihrer Unsicherheit und demnach der Motivation, diese zu verringern. Ebenso führt ein Aufwand, der als besonders hoch wahrgenommen wird, möglicherweise dazu, dass die Kosten zur Unsicherheitsreduktion zu hoch erscheinen und Unsicherheiten akzeptiert werden.

Sowohl die Implikationen durch die URT als auch die Erweiterung in der MRU lassen sich auf die hier gefundenen Ergebnisse zu Transparenz in KI übertragen. Zunächst ist der Wunsch nach zusätzlicher Information per se eine Maßnahme zur Unsicherheitsreduktion. Dass hierfür Hintergrundinformationen zum Urheber, aber auch Informationen von Dritten herangezogen werden, entspricht der klassischen URT. Bei zufriedenstellender Information kann das Interaktionsrisiko als ausreichend gesenkt wahrgenommen werden und (weitere) Interaktion, also Nutzung, folgen. Hingegen reduzieren Erklärungen, die sehr technisch ausfallen, nicht verstanden werden oder für andere Bereiche geliefert werden als die als notwendig wahrgenommenen, keine Unsicherheit und erhöhen auch nicht die Nutzung. Gleichzeitig wird, entsprechend der MRU, nur begrenzt viel Aufwand in Kauf genommen, um Unsicherheit zu reduzieren. Dies ist beispielsweise der Fall, wenn die Interaktion als wenig risikobehaftet wahrgenommen wird: Unsicherheiten müssen nur in Bezug auf die Aspekte reduziert werden, die ausreichend relevant scheinen. Zuletzt macht es einen Unterschied, ob Laien oder Expert*innen ein System nutzen möchten: Die in der Studie befragten Laien zeigten Unsicherheiten in anderen Bereichen als technische Expert*innen und besaßen eine andere (und auch untereinander unterschiedliche) Motivation, diese zu reduzieren.

Darüber hinaus helfen Modelle, die die Vertrauensbildung oder -zusammensetzung erklären, weitere Ergebnisse dieser Arbeit einzuordnen. Im **„Integrativen Modell für organisationales Vertrauen“** („Integrative Model of organisational Trust“) von Mayer, Davis und Schoorman wirken drei Faktoren wahrgenommener Vertrauenswürdigkeit auf das Vertrauen: Fähigkeit, Wohlwollen und Integrität (1995). Tatsächlich spiegeln sich diese Faktoren in den Fragen wider, die die Diskussionsteilnehmer*innen stellten: Wie gut ist das System in Bezug auf spezielle Fragestellungen, handelt es in meinem Interesse und, zuletzt, ist das Dargestellte auch die Wahrheit? In Bezug auf KI scheinen sich diese Mechanismen der Vertrauensbildung ebenso zu zeigen wie in menschlicher Interaktion.

Die sehr prominenten **„Technology Acceptance-Modelle“** (TAM 1, 2 und 3) beinhalten als Haupteinflussfaktoren auf Technologieakzeptanz den „Ease of Use“ und die „Perceived Usefulnes“ eines technologischen Systems. Tatsächlich schien die „Perceived Usefulnes“ bei der Musik-App als geringer angesehen und könnte deshalb zu Zurückhaltung geführt haben. Allerdings waren Vorbehalte gegenüber den beiden risikobehafteteren Apps anderer Natur, sodass weder „Ease of Use“ (was mit dem vorliegenden Ansatz nicht untersucht wurde) noch „Perceived Usefulness“ eine besonders große Rolle spielen. Es ist davon auszugehen, dass die beiden Faktoren in der tatsächlichen Nutzung größere Relevanz erlangen. Ergänzend sei die Differenzierung des Modells von der „Intention to Use“ und der tatsächlichen „Usage“ erwähnt. Dies leistet der Annahme Folge, dass eine Intention – oder eine

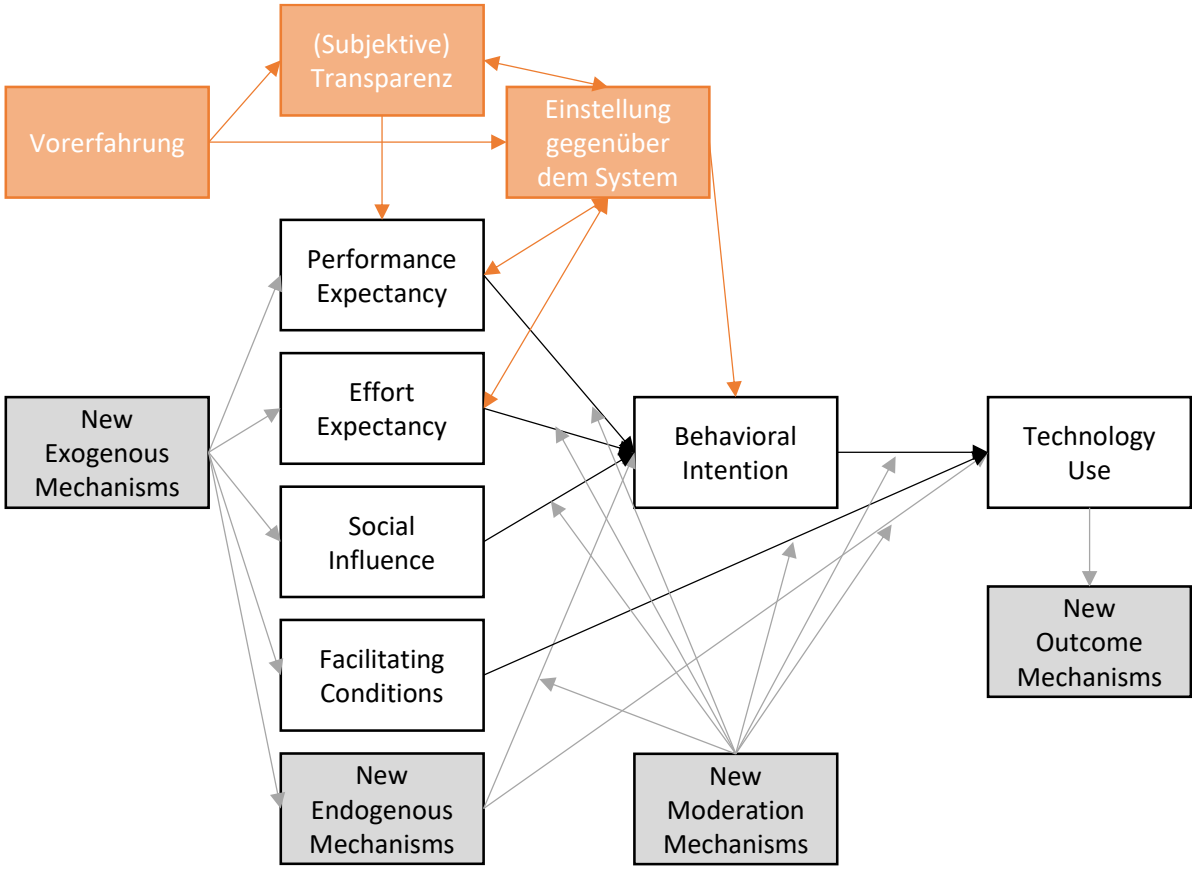
entsprechende Aussage in einer qualitativen Studie – noch keine Nutzung darstellt, sondern nur als einflussnehmende Variable betrachtet werden kann.

Als erweiterte Theory zum TAM ist die „**Unified Theory of Acceptance and Use of Technology**“ und das in diesem Zusammenhang erstellte UTAUT-Modell (siehe Abbildung 19) ein prominentes Modell für Technologieakzeptanz und -nutzung (Venkatesh et al., 2003). Als vereinendes, umfangreiches Nachfolgemodell der Technology Acceptance beinhaltet es jedoch keine Faktoren, die die Transparenz des Systems oder zumindest eine Einstellung gegenüber dem System abbilden. Auch in einer Erweiterung des Modells von Dwivedi et al. findet sich keine Transparenz, jedoch etwas allgemeiner die Einstellung gegenüber dem System (2019). In ihrer Studie zeigte sich, alle vier Hauptvariablen – „Performance Expectancy“, „Effort Expectancy“, aber auch „Social Influence“ und „Facilitating Conditions“ – werden von der Einstellung gegenüber dem System beeinflusst. Die Einstellung selbst wirkt wiederum stark auf die tatsächliche Nutzung. Eine Ergänzung der Variable „Einstellung gegenüber dem System“, die aufgrund der in der vorliegenden Studie Ergebnisse vorgenommen wird, würde Wechselwirkungen zwischen „Performance Expectancy“ und „Efford Expectancy“ postulieren sowie den starken Einfluss der Einstellung auf (zumindest) die Nutzungsintention (siehe rote Ergänzungen in Abbildung 19).

Bezüglich des Einflusses von Transparenz legen die Ergebnisse der aktuellen Studie nahe, dass sie die Einstellung gegenüber dem System und darüber hinaus auch seine zu erwartende Leistung („Performance Expectancy“) beeinflussen. Zusätzlich wird die wahrgenommene Transparenz aber auch von der Einstellung gegenüber einem System und von Vorerfahrung beeinflusst. In einem Review von Venkatesh et al. (2016), in dem das UTAUT-Modell um weitere Faktoren ergänzt wird, stellen die Autor*innen recht allgemein weitere sogenannte endogene Faktoren neben die bestehenden vier Hauptfaktoren, die die Verhaltensintention beeinflussen (siehe Abbildung 19). Zu diesen endogenen Faktoren zählen sie beispielsweise Vertrauen. Ein auf die Hauptfaktoren wirkender exogener Faktor – eine weitere ergänzte Kategorie – wäre das wahrgenommene Risiko. Die gefundenen Ergebnisse deuten an, dass es sich bei Transparenz um einen weiteren endogenen – dem System immanenten – Faktor handelt, der die Verhaltensintention beeinflusst. Darüber hinaus ist der Einfluss der Transparenz auf das Vertrauen in ein System ebenso wie auf das wahrgenommene Risiko zu postulieren. Auf Grundlage der hier präsentierten Ergebnisse kann also davon ausgegangen werden, dass die in Abbildung 19 vorgenommenen roten Ergänzungen „Vorerfahrung“, „(subjektive) Transparenz“ und „Einstellung gegenüber dem System“ weitere Beziehungen zu Elementen wie „Social Influence“ oder „Technology Use“ aufweisen sowie zu aktuell nicht explizit enthaltenen Faktoren wie „Vertrauen in das System“ oder „wahrgenommenes Risiko“. Da insbesondere in KI-Systemen Transparenz und Vorerfahrung wichtige Einflussfaktoren darstellen, stellen eine Ergänzung und

quantitative Prüfung der Faktoren im UTAUT einen wichtigen Schritt dar, um das Modell für die Nutzung immer komplexerer KI-Technologien anzupassen.

Abbildung 19: *Modell der UTAUT in Schwarz-Weiß. Graue Boxen und Pfeile stellen Erweiterungen des klassischen Modells dar; die Ergänzung roter Boxen und Pfeile erfolgte auf Basis der vorliegenden Ergebnisse.*



Anmerkung. Abbildung der schwarz-weißen und grauen Boxen nach Venkatesh et al., 2016, S. 335.

5.4.7. Limitationen

Bei Konzeption, Durchführung und Auswertung der Studie ergaben sich verschiedene die Gültigkeit der Ergebnisse einschränkende Faktoren, auf die im Folgenden eingegangen werden soll. In methodischer Hinsicht lassen sich die Zielgruppe, die Auswertung, die indirekte Fragemethodik und der qualitative Ansatz aufführen.

Bei der untersuchten Zielgruppe ist, da sie aus dem Umfeld des IMA und des TAIGERS-Projekts rekrutiert wurde, ein überdurchschnittliches technisches Wissen anzunehmen, auch wenn bei der Rekrutierung selbst darauf geachtet wurde, keine KI-Expert*innen oder -Entwickler*innen einzuladen. Ebenso ist eine höhere akademische Nähe als in der Allgemeinbevölkerung vorhanden. Dies zeigt sich auch in der mittleren Selbsteinschätzung zum Wissen über KI, das in der Gruppe vorherrschte: Es gab niemanden, der sich kein Wissen zuschrieb, jedoch wählten 80 % der Teilnehmenden eine mittlere Stufe des Wissens. Bei der befragten Gruppe herrschte teilweise eine sehr große Skepsis gegenüber

KI. Deshalb könnte man annehmen, sie weise erhöhte eine Sensibilität und deshalb eine besonders vorsichtige Annäherung an das Thema auf – es wurde also eine überdurchschnittlich sensibilisierte Gruppe befragt. Dem entgegen könnte man auch argumentieren, mit der Nähe zu KI/Technik steige die Gewöhnung und Begeisterung gegenüber KI, und deshalb waren in der befragten Gruppe überdurchschnittlich positive Stimmen vertreten. In einer Untersuchung von Schoeffer et al. (2022) zeigte sich der Effekt von erhöhter AI Literacy auf ein erhöhtes Vertrauen in und Präferenz für ein KI-System mit Transparenz. In der vorliegenden Studie lässt die Breite der getätigten Aussagen keine einzelne Richtung erkennen. Inhaltlich zeigen die Aussagen nur selten ein besonders hohes KI-Verständnis, vielmehr betreffen sie Einstellungen, Meinungen und Anmerkungen aus KI-Privatgebrauch und -nutzung. In einer zukünftigen Untersuchung sollte dieser Zusammenhang untersucht und das Vorwissen der Versuchspersonen ausführlicher berücksichtigt werden.

Aufgrund der qualitativen Methodik spielt die Subjektivität der Ergebnisse im Vergleich zu quantitativen Studien eine wichtige Rolle. Als eine Maßnahme kann die Auswertung mit zwei Auswerterinnen betrachtet werden, die sich regelmäßig austauschten. Dennoch ist es möglich, dass die vorliegenden Kategorien bei anderen auswertenden Personen nicht identisch identifiziert worden wären. Deshalb wurden, um die Intersubjektivität zu erhöhen, beim Bericht der Ergebnisse möglichst viele Zitate aufgeführt, um einen guten Einblick in die Daten zu gewährleisten. Dies ermöglicht Leser*innen ein eigenes Bild darüber, inwiefern die gezogenen Schlüsse stichhaltig sind. Auch die Schlussfolgerungen und Implikationen wurden mit einer ausführlichen Diskussion durchgeführt, um so Intersubjektivität zu gewährleisten. Die kritische Reflexion der Stichprobe und ihrer Zusammensetzung dient darüber hinaus dem Zweck, die Reichweite der Ergebnisse zu verdeutlichen. Zuletzt deutet die im Rahmen der Diskussion vorgenommene Einbettung der Ergebnisse in bestehende Theorien auf eine Gültigkeit der Ergebnisse über die gezogene Stichprobe hinaus hin.

Ein weiterer möglicher Kritikpunkt sind die indirekte Fragemethodik und ihre Auswirkungen auf das Ergebnis: Möglicherweise ist das nun resultierende, sehr breite und viel umfassende Verständnis von Transparenz auch auf diese indirekte Fragemethodik zurückzuführen. Wäre die Gruppe explizit nach Transparenz gefragt worden, hätte sie möglicherweise ein engeres Verständnis geäußert, das sehr viel weniger Bestandteile – wie z. B. Datenschutz oder Hintergrundinformationen – umfasst. Man könnte auch argumentieren, durch eine konkrete Frage wären gänzlich andere Bestandteile in der Diskussion über die Apps zum Vorschein gekommen. Dem lässt sich entgegnen: Auch in der auf die Apps folgenden Diskussionsphase, in welcher konkrete Transparenzaussagen besprochen wurden, kamen keine inhaltlich neuen Punkte, sondern vielmehr die zuvor besprochenen Punkte erneut zur Sprache.

Zuletzt stellt der qualitative Ansatz selbst Vor- und Nachteil für die Validität der Ergebnisse dar. Ein qualitatives Vorgehen erlaubt ein sehr viel vertiefteres Eintauchen, Nachfragen und Analysieren einer

Thematik. Gleichzeitig muss es immer der Kritik standhalten, Ergebnisse zu produzieren, die weniger allgemeingültig und nur für die untersuchten Gruppen oder Gegenstände gültig sind. In der vorliegenden Studie stellt sich, ergänzend zur mehr als für Laien üblichen technischen Nähe der Stichprobe, die Frage, inwiefern die *Aussagen* über Verhalten, Einstellungen und Intentionen von Diskussionsteilnehmer*innen mit ihrem tatsächlichen Verhalten zusammenhängen. Am Beispiel der Musik-App lässt sich das Paradoxon aufzeigen, das möglicherweise auch im Bereich Transparenz gilt: Während, wie die Aussagen vermuten lassen, Spotify einen von vielen genutzten Services darstellt, war Datenschutz eines der im Zusammenhang mit der Musik-App meistdiskutierten Themen. Das sogenannte Privacy-Paradox bezeichnet die hier schlummernde Diskrepanz (Barth & de Jong, 2017): Direkt darauf angesprochen, äußert ein Großteil der Nutzenden, sich wegen des Datenschutzes zu sorgen und bei der Nutzung von Onlinesystemen auf die Einhaltung von Privatsphäremaßnahmen zu achten. In ihrem tatsächlichen Verhalten zeigen sie dann aber Gegenteiliges, nutzen unsichere Systeme und geben ihre Daten ohne große Hemmnisse preis. Es stellt sich in Bezug auf Transparenz-Anforderungen die Frage, inwiefern mündlich getätigten Vorbehalten tatsächliches Verhalten folgt – oder auch hier ein Paradox gilt. Vor diesem Hintergrund ist auf den Dreiklang der vorliegenden Arbeit zu verweisen: Insbesondere die dritte Studie, die im Folgenden vorgestellt wird, zielt auf eine quantitative Untersuchung der Wirkung von Explainability bzw. Akkuratheitsinformationen auf die tatsächliche Nutzung eines KI-Systems. Damit sollen die in dieser Studie qualitativ erhobenen Ergebnisse durch eine weitere Perspektive ergänzt werden.

Auf **inhaltlicher Ebene** sind der Einfluss individueller Faktoren sowie die Spezifität der Ergebnisse als Limitationen zu nennen:

Eine Einschränkung der vorliegenden Studie stellt die Tatsache dar, dass individuelle Faktoren nicht explizit betrachtet wurden, also z. B. Voreinstellung gegenüber KI, Wissen über KI und demographische Variablen. Da der Einfluss der Vorerfahrung sogar ohne explizite Betrachtung deutlich wurde, zeigt sich jedoch die große Relevanz dieser Faktoren. Darüber hinaus scheint es sich bei der Vorerfahrung um einen besonders wichtigen Aspekt zu handeln, der selbst bei nicht spezieller Betrachtung hervortritt. Weitere individuelle Einflussfaktoren sollen in anschließenden Studien, möglicherweise mit quantitativem, vergleichendem Design, untersucht werden.

Ebenso lässt sich die Frage nach der Gültigkeit der Ergebnisse auch hinsichtlich der verwendeten Apps stellen. Der zu Beginn dargelegte Forschungsstand zeigt die enorme Spezifität der Ergebnisse für einzelne Systeme. So stellt sich auch in dieser Studie der Einfluss von Fehlerrelevanz und des Anwendungsbereichs des Systems als besonders groß heraus. Gleichzeitig ist aufgrund der hohen Spezifität bisheriger Ergebnisse unklar, ob es möglich ist, Systemvariablen zu abstrahieren und die Ergebnisse über die aktuell diskutierten Apps oder zumindest über ähnliche Systeme hinaus zu

generalisieren. Selbst wenn die Pilz-App gesundheitliche Fragestellungen behandelt, sind ihre Implikationen nicht ohne Weiteres auf beispielsweise einen Krankheitssymptom-Checker übertragbar. Gleichzeitig ist doch denkbar, Ergebnisse, wie den Wunsch nach alternativen KI-Vorschlägen und Akkuratheitsinformationen oder die Möglichkeit, die Vorschläge anhand mehr Informationen selbst zu prüfen, auf ein solches System zu übertragen. Insbesondere die weniger spezifischen Ergebnisse, z. B. zum Einfluss der Vorerfahrung, der Relevanz von Hintergrundinformationen oder der Bedeutung der wahrgenommenen Kontrolle, lassen größere Gültigkeit annehmen.

Unter Berücksichtigung der methodischen und inhaltlichen Limitation wurde diese Übertragung der Gültigkeit abhängig von Systemvariablen im folgenden Kapitel, dem Kapitel 5.5, aufgegriffen und ausgearbeitet. So ergeben sich Implikationen für KI-Systeme abhängig von ihren Gegebenheiten, die bei ihrer Entwicklung beachtet werden sollten, um anhand von Transparenzaspekten – im weiten Sinne – Vertrauen und Nutzung zu stärken.

5.5. Implikationen aus Studie (b) „Nutzendenanforderungen“

Wie Ergebnisse und Diskussion zeigen, stellt sich KI-Transparenz für Endnutzende als sehr viel breiteres Konzept dar, als das oft technisch betrachtete. Der Bedarf nach Hintergrundinformationen, also über Urheber und Geschäftsmodell, über Zertifikate oder Bewertungen durch Dritte, ist zur Vertrauensbildung relevant (Kapitel 5.4.4) und mindestens so groß wie der nach globaler oder lokaler Transparenz (Kapitel 5.4.2). Letztere scheint vor allem dann relevant zu sein, wenn es darum geht, zu beobachten, wie eine KI einen Fehler macht und diesen nachvollziehen zu wollen.

Als wichtige Faktoren, die die Wahrnehmung und Anforderungen an Transparenz maßgeblich formen (Kapitel 5.4.1), ist erstens die wahrgenommene Fehlerrelevanz zu nennen. Haben die KI-Ergebnisse großen Einfluss auf mein Leben, meine Finanzen oder Gesundheit und ist es demnach besonders folgenreich, wenn die KI einen Fehler begeht, steigt der Bedarf nach Transparenz in fast jeder Hinsicht. Auf sehr ähnliche Weise wirkt die Wahrnehmung, die KI betreffe eine ethisch sensible Thematik. Zweitens haben Vorerfahrungen und Einstellungen einen maßgeblichen Einfluss darauf, wie KI-Systeme wahrgenommen werden, welche Aspekte Skepsis hervorrufen und welche Form von Transparenz gefordert wird.

Darüber hinaus ist Kontrolle ein zentraler Einflussfaktor auf die Nutzung von KI (Kapitel 5.4.3). Auch wenn sie keinen klassischen Transparenzaspekt darstellt, besteht sie aus dem Bedürfnis, eigene Entscheidungen treffen zu wollen. Um dazu in der Lage zu sein, ist es nötig, einschätzen zu können, wo Vertrauen und wo Kontrolle bei der Interaktion mit KI sinnvoll sind. Das Bedürfnis nach Kontrolle begründet also das Bedürfnis nach Transparenz.

Die aus diesen Erkenntnissen folgenden Implikationen werden im Folgenden getrennt berichtet für Praxis und Forschung. Für den Einsatz in der Praxis wurde eine Matrix entwickelt, aus der sich aus den gegebenen Systemeigenschaften eines Systems Implikationen zur Umsetzung seiner Transparenz ableiten lassen (siehe Tabelle 16). Aufbauend auf Ergebnissen und Diskussion sowie der dort vorgenommenen Einbettung der Ergebnisse in bestehende Theorien (Kapitel 5.4.6) werden in Kapitel 5.5.2 Anknüpfungspunkte für die Forschung aufgezeigt.

5.5.1. Transparenzmatrix für die Praxis

Auf Grundlage der Ergebnisse lassen sich gegebene Systemeigenschaften und daraus resultierende Implikationen für die Transparenz des Systems in Beziehung setzen. Daraus ergibt sich eine Matrix, die Tabelle 16 darstellt. Sie ist geeignet für Entwickler*innen, Designer*innen und Anbieter*innen eines KI-Systems und kann während seiner Entwicklung zurate gezogen werden oder bereits entwickelte Dienste können mit ihrer Hilfe analysiert werden. In beiden Fällen hat die Transparenzmatrix zum Ziel, anhand der gegebenen oder zukünftigen Eigenschaften eines Systems aufzuzeigen, welche Arten und Aspekte von Transparenz nötig sind oder wären.

5.5.1.1. Anwendung der Transparenzmatrix

Die erste Zeile der Matrix enthält die aus den Ergebnissen der Fokusgruppen relevanten Systemeigenschaften (siehe Tabelle 16). Diese Faktoren aufgreifend lässt sich eine Liste erstellen, aus der die auf eine gegebene KI zutreffenden Faktoren auszuwählen sind. Die neun Faktoren sowie die für ihre Auswahl zu stellenden Fragen sind:

- ☐ **Hohe Fehlerrelevanz:** Werden Fehler der KI als schwerwiegend angesehen?
- ☐ **Dient der Unterhaltung:** Dient die KI in erster Linie der Unterhaltung?
- ☐ **Ethische Thematik:** Sind ethische Bedenken in Bezug auf die KI bekannt?
- ☐ **Sensible Daten:** Nutzt die KI sensible Daten (persönliche Daten, die Stimme etc.)?
- ☐ **Negative Vorerfahrung/Einstellung:** Sind in Bezug auf die KI (oder ähnliche Systeme) negative Vorerfahrungen gemacht worden bzw. negative Einstellungen vorhanden?
- ☐ **Fremdes System/keine Vorerfahrung:** Ist das System fremd und wurden noch keine Vorerfahrungen (mit ähnlichen Systemen) gemacht?
- ☐ **Positive Vorerfahrung mit System:** Wurden in Bezug auf die KI (oder ähnliche Systeme) bereits positive Vorerfahrungen gemacht?
- ☐ **Mehrere Ergebnisse zur Auswahl:** Stellt das System mehrere Ergebnisse bereit, aus denen die Nutzenden dann eines auswählen?
- ☐ **Nutzung kostet Geld:** Kostet die Nutzung/die KI Geld?

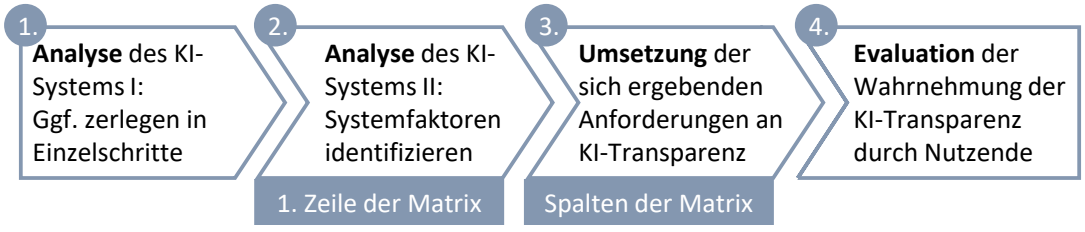
Dabei ist wichtig zu verstehen, dass es sich bei den subjektiven Systemeigenschaften um durch die **Nutzenden wahrgenommene Faktoren** handelt. Das gleiche System kann von verschiedenen

Nutzenden unterschiedlich angesehen werden: Beispielsweise beurteilten einige Fokusgruppenteilnehmende die Musik-App aufgrund der Interaktion mit der eigenen Stimme als ethisch bedenklich. Zur Identifikation der relevanten Systemeigenschaften ist also der Einbezug der Nutzenden zu empfehlen.

Das Beispiel der Musik-App verdeutlicht außerdem, dass es insbesondere bei umfangreicheren KI-Systemen sinnvoll ist, diese vor der Analyse **in ihre einzelnen Prozessschritte zu untergliedern**. Die Musik-App beispielsweise besteht aus den Prozessschritten (a) Eingabe der Daten per Stimme und (b) Vorschlag einer Playliste. Der letzte Schritt, die Nutzung, könnte man als dritten Schritt hinzufügen oder als außerhalb des Systems stattfindend festlegen. Ethische Bedenken werden im Hinblick auf die Eingabe der Daten per Stimme relevant. Hingegen wird beim Vorschlag der Playliste relevant, ob nur eine oder mehrere angezeigt werden.

Das empfohlene Vorgehen beim Einsatz der Matrix besteht also aus vier Schritten (siehe Abbildung 20). Im ersten Schritt wird, falls nötig, die KI in ihre Prozessschritte zerlegt, um sie einzeln analysieren zu können. Im zweiten Analyseschritt wird mithilfe von Nutzenden identifiziert, welche subjektiven Systemeigenschaften aus obiger Liste bzw. der ersten Zeile der Matrix für ein gegebenes System als zutreffend wahrgenommen werden. Anschließend kann in der Matrix abgelesen werden, welche Anforderungen an Transparenz sich aus den Systemeigenschaften ergeben. Im Sinne einer nutzendenzentrierten Entwicklung ist in einem vierten Schritt eine Evaluation des Ergebnisses empfohlen, um zu prüfen, ob die Systemeigenschaften korrekt identifiziert wurden und die Transparenzmaßnahmen das Verständnis des Systems fördern und die Nutzung unterstützen.

Abbildung 20: *Vier Schritte beim Einsatz der Transparenzmatrix (siehe Tabelle 16). In Schritt 2 wird die erste Zeile der Transparenzmatrix genutzt, in Schritt werden 3 die Implikationen aus den Spalten der Matrix abgelesen.*



Zur Ermittlung des Ergebnisses offenbaren die Systemeigenschaften Zusammenhänge mit verschiedenen Implikationen, die in Bezug auf KI-Transparenz zum Tragen kommen (siehe Tabelle 16). Die 13 Punkte dieser Implikationen umfassen die Folgenden, die mit kurzen Erklärungen oder Anregungen zur Umsetzung versehen sind:

→ **Bedarf nach Kontrollgefühl:** Nutzende wollen die letzte Entscheidung haben bzw. die Möglichkeit, Einfluss auf System und/oder Ergebnisse zu nehmen.

- → **Bedarf nach Erklärungen:** Grundsätzlich Erklärungen zum System und seiner Funktionsweise
- → **Lokale Transparenz:** Erklärung, warum ein einzelnes Ergebnis zustande kommt
- → **Globale Transparenz:** Erklärung, wie das System als Ganzes funktioniert
- → **Weitere Informationen zum Ergebnis:** Über die Anzeige des Ergebnisses hinaus wünschen Nutzende mehr Informationen zu diesem Ergebnis (z. B. weitere Fotos, Links).
- → **Mehrere Ergebnisalternativen:** Anstelle eines einzelnen Vorschlags wünschen Nutzende mehrere Vorschläge, aus denen sie einen auswählen.
- → **Sicherheiten über Dritte:** Zertifikate oder soziale Informationen und Bewertungen anderer
- → **Rechenschaftspflichten:** Übernahme von Verantwortung im Falle eines Fehlers
- → **Info über Urheber:** Wer hat die KI entwickelt, was sind Motive, Gewinninteressen etc.?
- → **Infos zum Datenschutz:** Wie werden Daten verarbeitet, wo gespeichert, wann gelöscht?
- → **Akkuratheitsangaben:** Information, wie sicher sich die KI mit dem Ergebnis ist
- → **Akkuratheitsanspruch:** Erwartung hoher Akkuratheit, möglichst weniger Fehler
- → **Zugangsschwelle:** Keine Transparenzanforderung, allerdings lässt sich aus Systemeigenschaften ableiten, als wie hoch die Zugangsschwelle wahrgenommen wird.

Zwei Anmerkungen sind zum Aufbau der Matrix zu tätigen: Zum einen sollen die leeren Felder in der Matrix erklärt werden. Diese stehen für keinen (in der Studie) gefundenen Zusammenhang. Sind sie leer, besteht an der Stelle also kein Zusammenhang zwischen Systemeigenschaft und Transparenzanforderung. Das bedeutet aber nicht, dass die Anforderung nicht durch eine andere zutreffende Systemeigenschaft relevant werden kann.

Zum anderen stellen Vorerfahrungen keine Systemvariablen im eigentlichen Sinne dar. Gleichzeitig zeigte sich ihr Einfluss als enorm wichtig für die Transparenzanforderungen, weshalb sie in einer Transparenzmatrix auch enthalten sein müssen. Außerdem sind Vorerfahrungen doch abhängig vom System. So hatten ein Großteil der Diskussionsteilnehmenden Vorerfahrungen mit Spotify, die ihre Wahrnehmung und Bewertung der Musik-App beeinflussten. Bei KI-Systemen mit prominenten Vorgängern oder Alternativen wären also entsprechende Vorerfahrungen zu berücksichtigen. Ebenfalls waren die insgesamt eher negativen Vorerfahrungen bzw. Voreinstellungen mit Geldanlagen ein Grund für die große Ablehnung der Finanz-App. In Anwendungsbereichen, in denen mit großer Skepsis gerechnet werden muss, oder bei Systemen für Zielgruppen, die eher skeptisch auftreten – man denke beispielsweise an Betriebsräte – gilt es also, bei der Entwicklung von KI-Systemen diese Variable zu berücksichtigen. Um präzise zu sein, wurde die Benennung der Kategorie „**Systemeigenschaften“ um „Interaktionseigenschaften“ ergänzt** (siehe Tabelle 16).

Zusammenfassend lässt sich also als Implikation für Entwickler*innen ableiten:

Implikation 5[7] für Entwickler*innen von KI:

Nutzen Sie die Transparenzmatrix (Tabelle 16), indem Sie den vier Schritten folgen: 1. Ggf. zerlegen in Einzelsysteme, 2. Systemeigenschaften identifizieren, 3. Anforderungen ableiten und umsetzen, 4. Evaluation der KI-Transparenz (Abbildung 20). Beziehen Sie Nutzende mit ein.

Außerdem soll noch eine dritte Implikation ergänzt werden, die sich in den Fokusgruppendiskussionen zeigte und die über die Analyse von Systemeigenschaften hinausgeht: Gruppen von Laiennutzenden sind in ihrer Zusammensetzung sehr heterogen. Manche haben ein Interesse daran, wie die Systeme funktionieren, mit denen sie interagieren. Andere wollen nur, dass die KI funktioniert. Warum ein bestimmtes Ergebnis für sie erfolgt, ist ihnen egal. Um den unterschiedlichen Bedürfnissen von Nutzenden gerecht zu werden, ist also Individualisierung ein wichtiges Thema. Die Gestaltung von Erklärungen, die personalisiert für einzelne Nutzende gestaltet sind, wäre zukünftig sicherlich ein Gewinn, jedoch sehr komplex in der Gestaltung. Für eine einfachere Umsetzung reicht es aus, Transparenz in verschiedenen Schritten optional und auswählbar bereitzustellen. Idealerweise lässt sich durch anklickbare Buttons oder Ein- und Ausblenden von Erklärungen eine KI-Transparenz umsetzen, die verschiedenen Bedürfnissen nach Informationstiefe gerecht wird.

Implikation 6 für Entwickler*innen von KI:

Gestalten Sie Transparenz in verschiedenen Detailtiefen, von einfach bis komplex. Zeigen Sie nicht zu viel Information auf einmal, sondern ermöglichen Sie eine schrittweise Vertiefung.

5.5.1.2. Beispiel einer Anwendung

Zur Veranschaulichung, wie die Transparenzmatrix genutzt werden kann, stelle man sich vor, es soll ein neues KI-System zum Erstellen der Steuererklärung durch Abfotografieren der entsprechenden Unterlagen eingeführt werden. Das System besteht also (a) aus der Eingabe der Daten und (b) aus der Bereitstellung der automatisch ausgefüllten Steuererklärung. Je nach Umfang und Expertise ist für den Großteil einer befragten Nutzendengruppe die **Fehlerrelevanz** des Ergebnisses hoch, für Einzelne geringer. Die App dient nicht der Unterhaltung. Auch ergab die Vorbefragung, dass beide Prozessschritte nicht als **ethisch** problematisch wahrgenommen werden, jedoch sind die verarbeiteten Daten hoch**sensibel**. Die Nutzenden haben positive **Vorerfahrungen** mit der Eingabe per Foto und viele kennen bereits ähnliche Systeme zur Unterstützung bei der Steuererklärung. Auch sind die **Einstellungen** zu einem KI-System zur Steuererklärung nicht schlecht: Das Thema wird zwar überwiegend als nervig wahrgenommen, aber die Hoffnungen auf Unterstützung sind groß. Einige

[7] Die Implikationen 1 bis 4 finden sich bei Forschungsfrage (a) „Fehlerfall“ in Kapitel 4.5.1.

haben gute Erfahrungen mit manuellen Steuerprogrammen gemacht. In diesem Fall sind nicht **verschiedene Ergebnisse** möglich, sondern das System soll die eine, optimale Lösung erstellen. Die Nutzung des Systems wird **Geld kosten**.

Dies als Grundlage nehmend, lässt sich in den Spalten ablesen, welche Implikationen für die KI-Transparenz des Systems entstehen, wenn bei einem System hohe Fehlerrelevanz, sensible Daten, neutrale bis gute Erfahrungen und Kosten des Systems zutreffen. Es ergeben sich hohe Anforderungen an die Möglichkeit, **Kontrolle** auszuüben: Bevor die Erklärung abgeschickt wird, sollte eine manuelle Überarbeitung/Korrektur möglich sein. Auch besteht Bedarf nach Erklärungen insgesamt und sowohl nach **lokaler** als auch nach **globaler** Art. Um das Vertrauen zu stärken, sollte offengelegt werden, wer die **Urheber der App** sind und was sie als Expert*innen auszeichnet. Ebenso ist es zentral, **Datenschutz** zu gewährleisten und transparent zu machen, wie und wo die Daten verarbeitet werden. Aufgrund der (überwiegend) hohen Fehlerrelevanz und der Kosten besteht ein hoher **Akkuratheitsanspruch**. **Akkuratheitsangaben** sind, auch weil es sich um ein neues System handelt, ebenfalls förderlich. Ebenso ist die **Zugangsschwelle** erhöht, aber bei ausreichend Transparenz und aufgrund der positiven Vorerfahrungen besteht die Annahme, dass das System angenommen wird. Bei allen Transparenz-Angaben gilt: Im besten Fall werden diese **nach Interesse abrufbar** gegliedert, z. B., indem man über einen zusätzlichen Infobutton erfahren kann, weshalb was wie ausgefüllt wurde und bei einem weiteren Klick mehr Details erhält.

Tabelle 16: *Transparenzmatrix: Implikationen für Transparenz eines KI-Systems in Abhängigkeit von subjektiv wahrgenommenen System- und Interaktionsfaktoren*

Implikationen für KI-Transparenz	Subjektive System- und Interaktionseigenschaften								
	Hohe Fehler-relevanz	**Dient der Unterhaltung**	**Ethische Thematik**	**Sensible Daten**	**Negative Vor-erfahrung/ -einstellung**	**Fremdes System/keine Vorerfahrung**	**Positive Vorerfahrung mit System**	**Mehrere Ergebnisse möglich**	**Nutzung kostet Geld**
Bedarf nach Kontrollgefühl	sehr hoch	hoch	sehr hoch	hoch	sehr hoch	hoch			sehr hoch
Bedarf nach Erklärungen allgemein	hoch	nach Interesse	sehr hoch		hoch, nach Interesse	hoch	weniger relevant		hoch
Lokale Transparenz	hoch	weniger relevant	sehr hoch					hoch	
Globale Transparenz	nach Interesse	weniger relevant	hoch		nach Interesse		weniger relevant		nach Interesse
Weiterführende Informationen zum Ergebnis	hoch		sehr hoch		hoch			hoch	
Mehrere Ergebnis-alternativen anzeigen			hoch					hoch	
Sicherheiten über Dritte	sehr hoch	ggf. andere Nutzende	hoch	hoch	hoch	hoch	weniger relevant		hoch
Infos über Rechenschaft	hoch	weniger relevant	hoch		hoch				an Urheber gegeben
Info über Urheber	sehr hoch	weniger relevant	hoch	hoch	hoch	hoch			hoch
Infos zum Datenschutz			hoch	sehr hoch					
Akkuratheits-angaben	hoch	weniger relevant			hoch	hoch		sehr hoch	
Akkuratheits-anspruch	sehr hoch		sehr hoch			hoch			sehr hoch
Zugangsschwelle zum System	hoch	niedrig	hoch		sehr hoch		niedrig		hoch

5.5.2. Anschließende Forschungsfragen

Zunächst ist festzustellen, dass sich in der vorliegenden Studie zeigte, wie sehr technisches und sozialwissenschaftliches Verständnis von KI-Transparenz auseinander geht. Dies stellt die Notwendigkeit nach interdisziplinärer Forschung heraus: Die Erforschung der Frage, wie Transparenz von KI-Systemen, die Blackboxen sind, ermöglicht werden kann, bedarf der Expertise von IT-Spezialist*innen und KI-Forscher*innen. Gleichzeitig müssen diese Ergebnisse des technisch möglichen für Nutzende umgestaltet, (schrittweise) aufbereitet und angepasst werden. Anstelle einer separaten Betrachtung von Transparenz aus technischer und aus sozialwissenschaftlicher Sicht, gilt es gemeinsam die Fragestellung nach der Übersetzung des technisch möglichen in das für Nutzende nötige zu verfolgen. Die Frage kann aus verschiedensten Perspektiven weiter interdisziplinär ausgestaltet werden. So stellt sich für die Philosophie die Frage, welche Aspekte für Nutzende transparent gemacht werden *sollten*. Und für die juristische Forschung steht insbesondere vor dem Hintergrund neuer gesetzlicher Regularien die Frage im Raum, wie diese umgesetzt und bestehende Systeme geprüft werden können.

In Bezug auf die Einbettung der Ergebnisse in bestehende Theorien lassen sich an mehreren Stellen Möglichkeiten für Anschlussforschung identifizieren. So ergeben sich aus dem Übertrag der Ergebnisse auf die MRU (Kramer, 1999) weitere Fragen, die es durch empirische Forschung zu ermitteln gilt. Zunächst folgt aus der Annahme, dass mehr Information nicht unbedingt weniger Unsicherheit bedeutet, ob und **wie sich die ideale Menge an Information feststellen, messen oder vorhersagen lässt.** In Bezug auf die DSGVO zeigen sich viele Beispiele, bei denen die reine Informationsbereitstellung im besten Fall zu Genervtheit, im schlechtesten zu einer Verunsicherung führt. Übertragen auf Transparenz und vor dem Hintergrund der für KI aufkommenden Gesetze wird die Frage nach der richtigen Menge an Information und ihrer Darstellung also dringlicher, um Verständnis, Informiertheit und den Abbau von Unsicherheit zu gewährleisten.

Darüber hinaus ergibt sich aus der in Kapitel 5.4.6 vorgenommenen Erweiterung des UTAUT der dringende Bedarf, diese Erweiterung zu prüfen. Dabei geht es nicht darum, genau die postulierten Bezüge zu untersuchen, sondern um zwei sehr viel allgemeinere Fragen: Erstens gilt es, für ein Modell, das sich als „Unified“ und gültig für „Technology“ bezeichnet, zu ermitteln, **welche spezifischen Ergänzungen in Bezug auf KI zu tätigen sind**. Die soziale Komponente erhält in diesem Zusammenhang sicherlich eine Aufwertung, da die gesellschaftlichen Einstellungen zu KI häufig uneinheitlich sind (Brauner et al., 2023; Fischer & Petersen, 2018). Daran anschließend besteht die Annahme, dass z. B. Vorerfahrung und Einstellung gegenüber dem System eine größere Rolle einnehmen als in bisherigen, weniger komplexen Technologien. Daraus ergibt sich die zweite Frage, nämlich die nach der Wirkrolle von Transparenz. Da die aktuellen KI-Systeme vor allem durch ihre Komplexität charakterisiert sind

und ihre Prozesse zunehmend unverständlich werden, gilt es zu untersuchen, **wie sich ein solches fehlendes Verständnis auf Faktoren wie Performance Expectancy oder Efford Expectancy auswirkt**. Die aktuellen Ergebnisse ebenso wie die von Studie (a) „Fehlerfall" zeigen den Effekt von Transparenz auf Nutzungsintention bzw. Nutzung in verschiedenster Hinsicht. Entsprechend gilt es die Frage zu klären, wie sich Transparenz und ihre Wirkung in ein UTAUT, das auch KI-Systeme umfasst, einbetten lassen.

Zusammenfassend ergeben sich die folgenden Forschungsfragen für die Zukunft:

- Wie lässt sich das technisch Mögliche der XAI-Forschung in die für Nutzende nötige Transparenz übersetzen? (mit möglichem Anschluss für philosophische und juristische Fragestellungen)
- Wie lässt sich die für Nutzende passende Menge der bereitgestellten Transparenzinformation messen, systematisieren und prognostizieren (MRU)?
- Wie lässt sich das UTAUT für eine Gültigkeit für KI erweitern und wie lässt sich dabei der Faktor KI-Transparenz integrieren?

5.6. Zwischenfazit zur Studie (b) „Nutzendenanforderungen"

In der dargestellten Studie wurde die Forschungsfrage (b) untersucht: **Welche Anforderungen an Transparenz in KI bestehen für Laiennutzende und inwiefern unterscheiden sie sich nach Eigenschaften der KI?** Dazu erfolgte eine Aufteilung dieser Frage in zwei Leitfragen

FF b (1): Welche Anforderungen an Transparenz in KI-Apps haben Laiennutzende?

FF b (2): Was erwarten Nutzende abhängig von gegebenen Systemeigenschaften?

Auf Grundlage von Fokusgruppendiskussionen dreier KI-Apps wurde mithilfe der Inhaltsanalyse nach Mayring ein Kategoriensystem entwickelt, das zunächst die Leitfrage 1 beantwortet (Kapitel 5.3.1). Zusammengefasst lässt sich sagen: Die Anforderungen von Laien an Transparenz gehen deutlich über ein technisches Verständnis hinaus. Dies wird beispielsweise deutlich, wenn sie teilweise Erklärungen dafür wünschen, warum ein spezifisches Ergebnis zustande kam – also lokale Transparenz – und teilweise globale Transparenz fordern, also Erklärungen, wie eine KI generell funktioniert. Wichtiger für sie sind allerdings weitere Hintergrundinformationen über Urheber, Bewertungen Dritter und Datenschutz sowie zu Kontrolle und Sicherheit.

Darauf aufbauend konnte Leitfrage 2 durch die Betrachtung und Unterscheidung der drei KI-Apps und der verschiedenen, in den KI-Apps umgesetzten Systemeigenschaften untersucht werden (Kapitel 5.3.2). Es zeigt sich der wichtige Einfluss des Anwendungsbereichs und der Systemeigenschaft Fehlerrelevanz. Bei höherer Relevanz möglicher Fehler steigen die Anforderungen gegenüber

Transparenz: Mehr Hintergrundinformationen werden gefordert und auch der Bedarf nach Erklärungen allgemein sowie nach lokaler Transparenz im Speziellen steigt. Weitere systemabhängige Einflussfaktoren sind die Wahrnehmung möglicher ethischer Probleme und die Verarbeitung sensibler Daten. Die Vorerfahrung mit dem System, die auch davon abhängt, ob bereits ähnliche Systeme bestehen, sowie allgemeinere Vorerfahrungen aus dem Anwendungsbereich haben großen Einfluss auf die Transparenzanforderungen gegenüber KI.

Die Forderung nach Kontrolle (Kapitel 5.4.3), die in allen Diskussionen angesprochen wurde, verdeutlicht den Anspruch der Nutzenden, dass intelligente Systeme immer unter der Kontrolle von Menschen zu liegen haben. Aus diesem Anspruch ergibt sich die Aufgabe, die transparente KI zu erfüllen hat: Um sinnvoll Kontrolle ausüben zu können, sind Nutzende auf Nachvollziehbarkeit und Verständnis des Systems angewiesen.

Insofern schließt das vorangehende Kapitel eine Forschungslücke bei der Fragestellung, welches Verständnis für Laien überhaupt an Transparenz besteht und hinsichtlich welcher Teile eines Systems sie Nachvollziehbarkeit und Verständnis fordern. Diese Lücke ist insbesondere deshalb von Relevanz, da sich bisher viel Forschung zu Transparenz „has focused explicitly on ML models, asking questions about people's abilities to understand model internals or the ways that particular models map inputs to outputs, as well as questions about the relationship between these abilities and people's willingness to trust a model. However, the model is just one component of the ML pipeline […]" (Yin et al., 2019, S. 1). Die Studie zeigt, Nutzende wünschen sich weniger ein „enges", modell-abhängiges Verständnis, sondern haben Bedarf an darüber hinausführender Information, um Vertrauen aufbauen, das System einschätzen und so auf sinnvolle Weise nutzen zu können.

Die gesamte vorliegende Arbeit geht von einem sehr breiten Verständnis von Transparenz aus und die nun berichtete Untersuchung stärkt diese Argumentation: Für Laien macht es wenig Unterschied, wo die Grenze zwischen Explainability, Transparenz, Hintergrundinformation oder Akkuratheitsangabe gezogen wird. Sie bedürfen bestimmter Informationen, um einem System zu vertrauen, um es zu nutzen. Für Entwickler*innen eines Systems ergeben sich je nach (durch die Nutzenden wahrgenommener) Eigenschaften des Systems Anforderungen an Transparenz. Um diese Zusammenhänge umsetzbar zu machen, wurde in Kapitel 5.5.1 eine Transparenzmatrix für die Praxis entwickelt und vorgestellt. Mit deren Hilfe können Entwickler*innen ableiten, welche Aspekte von KI-Transparenz in einem KI-System umgesetzt werden sollten.

Offen ist nach dieser qualitativen Studie die Frage nach einer Validierung oder zumindest Ergänzung ihrer Ergebnisse durch quantitative Untersuchungen. Ähnlich zur sehr speziellen Fragestellung (a) „Fehlerfall", verfolgt die nächste Studie zu Forschungsfrage (c) „Transparenzarten" mit einem Online-

Experiment einen quantitativen Ansatz. Das Ziel der Studie besteht darin, den Einfluss unterschiedlicher Formen der Transparenz, die auch in der vorliegenden Untersuchung diskutiert wurden, auf Vertrauen und Nutzung zu untersuchen.

6. Vergleich des Effekts von Transparenzarten auf die Nutzung (Forschungsfrage c)

Auch wenn immer mehr ML-Systeme als künstlich intelligent bezeichnet werden können und der steigende Anteil an Deep Learning-Modellen die Bemühungen um Transparenz in diese Systeme erschwert, gibt es inzwischen eine Vielzahl an Ansätzen, Blackboxen zu mehr Transparenz zu verhelfen. Agnostische Verfahren der Explainability, die nachträglich auf nicht erklärbare Systeme aufgesetzt werden und zu Post-Hoc-Transparenz führen, ermöglichen es, einen großen Anteil von KI-Modellen mit zusätzlichen Erklärungen anzureichern (Herm et al., 2023). Dabei gibt es verschiedene Arten der Erklärungen und verschiedene Ebenen, auf denen diese unterschieden werden (Mohseni et al., 2021), die in Kapitel 2.2.4 ausführlich dargelegt wurden.

Nachdem mit qualitativen Methoden die Frage nach Ansprüngen von Laien an die Transparenz von KI-Systemen beantwortet wurde, untersucht die Fragestellung c verschiedene KI-Transparenzarten mit einer quantitativen Methodik. Im Folgenden werden vier verschiedene Arten der KI-Transparenz verglichen in ihrer Auswirkung auf Nutzung und Vertrauen des transparenten Systems: eine globale und eine lokale Erklärung für die Funktionsweise und eine globale und eine lokale Akkuratheitsangabe. Die Untersuchung hat das Ziel generalisierbare Erkenntnisse zu den Auswirkungen von Transparenz auf Nutzung und Vertrauen von KI zu liefern.

Nachdem also in Fragestellung a der Effekt einer Art der Transparenz – globale Akkuratheitsangaben – auf einen persistierenden KI-Effekt – Algorithm Aversion – quantitativ erforscht wurde, wird im Folgenden, bei der Auseinandersetzung mit Forschungsfrage c, ein breiterer quantitativer Ansatz gewählt. Die Forschungsfrage lautet:

> **FF (c) „Transparenzarten“: Wie wirken sich verschiedene Arten der KI-Transparenz auf Vertrauen und Nutzung eines Systems aus?**

Ihre theoretische Einordnung sowie die Ableitung von Hypothesen zu ihrer Untersuchung finden im Folgenden statt.

6.1. Theoretische Einordnung und Hypothesen

Lange Zeit spielten sozialwissenschaftliche Ansätze in der XAI-Forschung keine Rolle. Ziel war es, mit Ansätzen der Explainability die Prozesse von KI für Entwickler*innen und Expert*innen nachvollziehbar zu machen, unter der Annahme, mit Transparenz auch Vertrauen und Akzeptanz der Nutzenden zu erhöhen (Felzmann et al., 2019; Molnar, 2019).

Seit den 2020er Jahren steigt die Zahl der sozialwissenschaftlichen Untersuchungen zu transparenter KI und wie diese von Endnutzenden angenommen wird (Arrieta et al., 2020; Shulner-Tal et al., 2023, siehe auch Kapitel 2.3): Die Untersuchungen zeichnen ein uneinheitliches Bild von Transparenz.

Nutzende präferieren transparente gegenüber nicht transparenten Systemen (S. S. Y. Kim et al., 2023). Transparenz führt jedoch auch zu weniger Vertrauen oder Nutzung, beispielsweise wenn Erwartungen enttäuscht oder Nutzende mit Informationen überfordert werden (Springer, 2019; Zhao et al., 2019).

Nutzung und Vertrauen dienen dabei in einer Vielzahl von Studien als abhängige Variablen, werden aber sehr unterschiedlich operationalisiert (siehe Kapitel 2.3.1). Eine dimensionale Quantifizierung geschieht häufig und wie schon in Studie (a) „Fehlerfall" als WOA, also als Wert zwischen 0 = „kein Einfluss" und 1 = „Vorschlag übernommen" (Dietvorst & Bharti, 2019; T. Kim & Song, 2020). Da sich von hohem Vertrauen nicht automatisch auf hohe Nutzung schließen lässt (Daschner & Obermaier, 2022; Schmidt et al., 2020), wurden in der vorliegenden Studie beide Variablen unabhängig voneinander betrachtet: die tatsächliche Nutzung des Algorithmus operationalisiert über den WOA und das Vertrauen in den Algorithmus über eine Skala.

Die ambivalenten Studienergebnisse deuten an: Transparenz in KI verändert Nutzung und Vertrauen gegenüber den Systemen. Daraus leiten sich für die folgende Studie zwei Hypothesen ab:

> **H1:** Transparenz in Algorithmen im Vergleich zu keiner Transparenz geht einher mit einer veränderten Nutzung des Algorithmus.
>
> **H2**: Transparenz in Algorithmen im Vergleich zu keiner Transparenz geht einher mit einem veränderten Vertrauen in den Algorithmus.

Während es von technischer Seite verschiedene Arten gibt, Transparenz umzusetzen, sind die Auswirkungen dieser technischen Möglichkeiten auf Endnutzende noch weitgehend unklar, gerade und besonders im Vergleich miteinander. Bisherige Studien untersuchten einzelne Transparenzarten, beispielsweise die Wirkung von verschiedenen Heatmaps (Herm et al., 2023; Karran et al., 2022). Die Heatmap ist in der technischen Forschung zu Explainability eine weit verbreitete Methode, die insbesondere in der Bilderkennung Anwendung findet (Herm et al., 2023; Posada-Moreno et al., 2023; Zintgraf et al., 2017). Trotz ihrer Popularität sind die Ergebnisse hierbei uneinheitlich. Manche Heatmap-Erklärungen steigern die Verständlichkeit, andere unterscheiden sich nicht von Bedingungen ohne Erklärung (Herm et al., 2023). In einer Untersuchung, ob die Erklärung, welche Faktoren für eine einzelne Entscheidung besonders wichtig waren, dabei helfen, das Vertrauen zu regulieren, bestätigte sich dieser Zusammenhang nicht (Y. Zhang et al., 2020).

Von diesen traditionellen Arten der Erklärungen lassen sich Informationen über die Akkuratheit eines Modells unterscheiden. Diese Art der Transparenz geschieht, wie z. B. Yin et al. anmerken, immer häufiger bei der Nutzung von KI (Yin et al., 2019). Auch die AI HLEG der Europäischen Kommission führt die Kommunikation der KI-Akkuratheit in ihrer Anforderungsliste an KI auf mit der Frage: „Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of

accuracy and/ or error rates?“ (AI HLEG, 2020, S. 15). Die Ergebnisse zu Forschungsfrage (b) „Nutzendenanforderungen“ legen nahe, dass Nutzende einschätzen möchten, wie gut und akkurat eine KI arbeitet. Manche Studien zeigten eine verringerte Akzeptanz und erhöhtes Misstrauen gegenüber dem System, wenn Akkuratheitsangaben vorhanden waren im Vergleich zu keinen Angaben (T. Kim & Song, 2020; Schmidt et al., 2020). Andere fanden den Zusammenhang von hohen Akkuratheitsangaben und hoher – bis hin zu übersteigerter – Akzeptanz eines Systems (z. B. Ford et al., 2020; Lim & Dey, 2011; Yin et al., 2019). In einem Vergleich von Erklärungen und Sicherheitsangaben zeigten Letztere den größeren Einfluss auf die KI-Nutzung (Y. Zhang et al., 2020; Z. Zhang et al., 2021). Aus diesen uneinheitlichen, aber auf die verschiedene Wirkung hindeutenden Ergebnissen, lassen sich zwei weitere Hypothesen ableiten:

> **H3**: Verschiedene KI-Transparenzarten unterscheiden sich im Hinblick auf die Nutzung der entsprechenden Algorithmen.
>
> **H4**: Verschiedene KI-Transparenzarten unterscheiden sich im Hinblick auf das Vertrauen in die entsprechenden Algorithmen.

Betrachtet man die aus technischer Sicht möglichen Erklärungen für KI-Systeme, lässt sich die zuvor angesprochene Kategorisierung nach Geltungsbereich aufführen: globale und lokale Transparenz (Molnar, 2019). Globale Erklärungen beziehen sich dabei auf das gesamte Modell oder Teile davon und seine Funktionsweisen: „Wie funktioniert das System als Ganzes?“ Lokale Erklärungen beziehen sich auf einzelne Ergebnisse: „Warum kommt genau dieses Ergebnis zustande?“

Interviewte Nutzende zeigten eine Präferenz von lokalen über globale Erklärungen (S. S. Y. Kim et al., 2023). In einer anderen Studie trugen lokale Erklärungen zum Erhalt von Vertrauen und Nutzung eines Algorithmus bei unsicheren Ergebnissen bei (Alam & Mueller, 2021). Ein Vergleich verschiedener lokaler mit einer globalen und mit keiner Erklärung zeigte die höhere wahrgenommene Güte der lokalen Erklärungen. Globale und keine Erklärung wurden ähnlich schlecht bewertet (Herm et al., 2023). Allerdings waren in dieser Studie alle Erklärungen visuell ähnlich einer Heatmap, was sich für lokale Erklärungen möglicherweise besser eignet als für globale. Eine höhere Güte einer Erklärung bedeutet außerdem nicht unbedingt mehr Nutzung oder Vertrauen. Auch in den Aussagen der Fokusgruppen zur Forschungsfrage (b) ließ sich keine klare Präferenz für lokale oder globale Explainability erkennen. Vielmehr präferieren Nutzende eine breite Transparenz, die Hintergrundinformationen und je nach Anwendungsfall weitere Felder wie Datenschutz umfasst, um Verständlichkeit und eine Einschätzung zur Vertrauenswürdigkeit eines Systems zu ermöglichen. Dies lässt sich als Erklärung, die die gesamte KI betrifft, als globale Transparenz bezeichnen. Die für die

folgende Studie gewählten Transparenzarten lassen sich also durch den Erklärungsfokus einteilen in **lokale und globale Transparenzarten**.

Im Gegensatz zu globalen und lokalen Erklärungen zur Funktion einer KI lassen sich Informationen zur Akkuratheit eines Systems nicht nach Wie und Warum einteilen. Jedoch können auch für diese Transparenzinformation lokale Akkuratheitsangaben einzelner Ergebnisse von globalen Akkuratheitsangaben zur generellen Akkuratheit des Algorithmus unterschieden werden. In Bezug auf diese Unterscheidung sind bisher keine Vergleiche von lokalen oder globalen Akkuratheitsangaben auf Nutzung oder Vertrauen bekannt.

Insbesondere bei den Erklärungen zu Funktionen von KI zeigen sich Unterschiede in der Wirkung lokaler und globaler Transparenz. Daraus folgen die beiden Hypothesen:

> **H5**: Lokale und globale KI-Erklärungen unterscheiden sich im Hinblick auf die Nutzung der Algorithmen.
>
> **H6**: Lokale und globale KI-Erklärungen unterscheiden sich im Hinblick auf das Vertrauen in die Algorithmen.

Um diese Hypothesen zu testen, wurde ein Wizard-of-Oz-Experiment aufgesetzt. Im vorliegenden Fall handelte es sich um einen fest einprogrammierten Ablauf von Bildern und Informationen. Da sich der Einfluss der Vorerfahrung oder Voreinstellung zur Thematik als sehr gewichtig auf Vertrauen und Verhalten gegenüber Systemen gezeigt hatte (z. B. Molina & Sundar, 2022 sowie in Forschungsfrage b), wurde für das Experiment eine möglichst neutrale Aufgabe gewählt. Deshalb wurden politische Aufgabenstellungen, solche mit Sportbezug oder die eine besondere Expertise erfordern, vermieden. Wie schon in Forschungsfrage (a) „Fehlerfall“ wurde erneut das Design zur Gewichtschätzung von Obst/Gemüse gewählt (Werz et al., 2020).

Während manche Studien eine Transparenzart variieren, um auf ihre Eignung für verschiedene Fragestellungen zu prüfen, sollten im Sinne der Hypothesen möglichst verschiedene Transparenzarten verglichen werden. Ein Überblick über verschiedene technische Ansätze transparenter KI liefert Kapitel 2.2.4. Daran anschließend ergibt sich die Aufteilung, die auch in den hier getätigten Ausführungen deutlich wird: Zum Ersten kann ausgehend vom **Fokus der Erklärung lokale von globaler Transparenz** unterschieden werden. Zum Zweiten kann der **Transparenzgegenstand** variiert werden, nämlich Offenlegungen zu **Funktionsweisen sowie der Akkuratheit** einer KI.

Die Ergebnisse aus Forschungsfrage (b) „Nutzendenfaktoren“ zeigen, Laien bevorzugen Hintergrundinformationen über technische Erklärungen. Es geht also um Erklärungen zu **Funktionsweisen auf globaler** Ebene, die allgemeine Informationen zu Urhebern, zur Erstellungsweise der KI oder zu möglichen vertrauenserweckenden Prüfprozessen oder -siegeln beinhalten. Im Kontrast

dazu bieten Heatmaps (meist) **lokale Erklärungen** dafür, auf welche Bestandteile eines Bildes oder Textes die KI ihre Entscheidungen stützt. Heatmaps werden in der XAI-Forschung sehr häufig umgesetzt und genutzt (Herm et al., 2023; Posada-Moreno et al., 2023).

Als weitere Transparenzbedingung kommen Akkuratheitsangaben zum Einsatz. Eine **globale Akkuratheit** stellt ein bei der Erstellung von Machine-Learning Algorithmen/KI übliches Verfahren dar. Dazu wird ein Modell zunächst auf einem Datensatz zu trainiert und dann seine Vorhersagekraft an einem fremden Datensatz geprüft (siehe auch Kapitel 2.2.4; Yin et al., 2019). Es ergibt sich eine Aussage darüber, wie akkurat die KI insgesamt Vorhersagen treffen kann. Im Gegensatz entspricht die **lokale Akkuratheitsinformation** einer Angabe der Sicherheit, mit der eine KI jedes ihrer Ergebnisse einschätzbar macht: Wie sicher ist sich der Algorithmus mit diesem Ergebnis bzw. wie groß ist die Schwankung um die Angabe (S. Cramer et al., 2022)?

Zusätzlich zu den Effekten auf Verhaltens- oder Vertrauensmaße haben zahlreiche weitere Studien Konstrukte wie die wahrgenommene Akkuratheit oder Kompetenz eines Systems (z. B. Alam & Mueller, 2021; H. Cramer et al., 2008; Kocielnik et al., 2019), subjektives oder objektives Verständnis der Erklärung (z. B. Molina & Sundar, 2022; Ribes et al., 2021) und die wahrgenommene Transparenz untersucht (z. B. Shin, 2021). Diese zusätzlichen Konstrukte sollen hier explorativ erhoben werden, um Erklärungen für Verhalten oder Vertrauen zu ermöglichen.

Das genaue Studiendesign, der Versuchsablauf, das verwendete Material inklusive der Ausgestaltung der fünf Experimentalbedingungen und der erhobenen Variablen sowie die Auswertung werden im folgenden Methodenteil im Einzelnen erläutert.

6.2. Methode

Im Folgenden wird die Methodik zur Testung der Hypothesen dargelegt, wobei die Erhebung zunächst in den Projektkontext eingeordnet wird. Anschließend folgen die Berichtskategorien Stichprobe, Studiendesign mit Versuchsplan, dem Versuchsablauf und der Beschreibung des verwendeten Materials sowie der durchgeführten Auswertung. Dabei wird auch auf die für weitere explorative Analysen verwendete Auswertung eingegangen.

6.2.1. Projekt- und Forschungskontext

Die Studie wurde aufgesetzt im Projekt TAIGERS (Transparency in Artificial Intelligence: Considering Explainability, User and System Factors), gefördert im Rahmen des Exploratory Research Space (ERS) als Open Seed Fund[8] der RWTH Aachen University.

Das Projekt fand in Zusammenarbeit mit dem HCIC der RWTH Aachen University statt und untersuchte System- und Nutzendenfaktoren, die das Vertrauen und die Nutzung von transparenter KI beeinflussen. Das HCIC konzentrierte sich innerhalb des Projektes auf Nutzendenfaktoren, während am IMA und dabei maßgeblich durch die Autorin die Auswirkungen der Systemfaktoren untersucht wurden. Zusätzlich zur zuvor berichteten Fokusgruppenstudie (Forschungsfrage (b) „Nutzendenanforderungen“, Kapitel 5) wurde dazu eine quantitative Studie mit Fokus auf die Effekte verschiedener Transparenzarten konzipiert und mit kleiner Stichprobe durchgeführt.

Um die Effekte aussagekräftig und nachhaltig zu ermitteln, wurde die Studie ein zweites Mal im Rahmen des Projekts FAIRWork[9] aufgesetzt, an mehreren Stellen verbessert und mit größerer Stichprobe durchgeführt. Bei der im Folgenden berichteten Studie handelt es sich um diese überarbeitete Version.

6.2.2. Stichprobe

Die vorliegende Studie schlossen 184 Teilnehmende vollständig ab. Davon waren die Daten weiterer vier Teilnehmender unvollständig, weshalb ihre Daten von der Auswertung ausgeschlossen wurden. Bei 18 Teilnehmenden wurden die Algorithmusschätzungen, die eigentlich fest vorgegeben und für alle identisch sein sollten, verändert. Da Auswirkungen dieser Anzeige- und Speicherfehler nicht nachvollzogen werden können, wurden die Daten der 18 Teilnehmenden von der weiteren Analyse ausgeschlossen. Zudem wurden zehn Personen ausgeschlossen, weil sie in den Aufmerksamkeitschecks falsche Angaben gemacht hatten. Als solche waren in zwei Bedingungen Fragen zu den Instruktionen eingefügt worden, um zu prüfen, ob die Versuchspersonen die Angaben ausreichend aufmerksam gelesen bzw. verstanden hatten. Ein*e Teilnehmer*in wurde ausgeschlossen, da sie/er in einem offenen Textfeld vermerkte, die Studie mit Hilfe von ChatGPT-4 ausgefüllt zu haben. Nach Ausschluss dieser 33 Versuchsteilnehmenden kam schließlich die finale Stichprobengröße von n = 151 zustande.

[8] Gefördert vom Bundesministerium für Bildung und Forschung (BMBF) und dem Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (MKW) im Rahmen der Exzellenzstrategie von Bund und Ländern

[9] This work has been supported by the FAIRWork project (www.fairwork-project.eu) and has been funded within the European Commission's Horizon Europe Programme under contract number 101069499. The work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work.

Die demographische Zusammensetzung der Stichprobe ist Tabelle 17 zu entnehmen. Im Durchschnitt waren die Teilnehmenden *M* = 25,2, *SD* = 7,56 Jahre alt mit einem Range von 18 bis 62 Jahren.

Tabelle 17: *Demographische Zusammensetzung der Stichprobe in absoluten Zahlen und prozentualen Anteilen*

		Anzahl	Prozent
Geschlecht	Weiblich	115	76,2
	Männlich	31	20,5
	Divers	4	2,6
	Ohne Angabe	1	0,7
Deutsch als Muttersprache	Ja	147	97,4
	Nein	4	2,6
Bildungsabschluss	Abitur	84	55,6
	Universitäts- oder Fachhochschulabschluss	61	40,4
	Realschulabschluss	3	2
	Lehre/Berufsausbildung	3	2
Arbeitstätigkeit	Student/in	128	84,8
	Angestellte/r	19	12,6
	Freiberufler/in	4	2,6
Umgang mit Computern	„gar nicht“ routiniert	1	0,7
	„ein wenig“ routiniert	39	25,8
	„sehr“ routiniert	111	73,5

6.2.3. Studiendesign

Es wurde ein messwiederholtes Design mit fünf Stufen entworfen. Als Within-Subject-Manipulation durchliefen alle Versuchspersonen die fünf Bedingungen. Diese bestanden jeweils aus drei Schätzaufgaben und unterschieden sich lediglich in den vorgeblich zur Schätzung verwendeten Algorithmen und ihren Transparenzarten. Wie schon in Forschungsfrage (a) „Fehlerfall“ bestand die Aufgabe in der Schätzung des Gewichts von Gemüse auf Fotos. Für jede Schätzung erhielten die Versuchspersonen den Hinweis eines Algorithmus. Während in der ersten Bedingung als Kontrollbedingung ein Algorithmus ohne Transparenz präsentiert wurde, folgten vier weitere Algorithmen, die jeweils unterschiedlich transparent waren (siehe Kapitel 6.2.5.2 zu den Transparenzbedingungen). Als abhängige Variablen wurden das Vertrauen in den jeweiligen Algorithmus und seine Nutzung erhoben.

Wie in Forschungsfrage (a) „Fehlerfall“ handelte sich beim Experiment um ein Wizard-of-Oz-Design. Die vorgeblichen Algorithmen schätzten entgegen der Vignette das Gemüsegewicht nicht autonom für jedes Bild. Vielmehr waren die Bilder, die Algorithmusschätzungen ebenso wie alle Transparenzangaben fest einprogrammiert und für alle Teilnehmenden identisch.

Die Bedingung ohne Transparenz stand in jedem Durchlauf am Anfang. Die vier folgenden transparenten Algorithmen wurden randomisiert präsentiert. Grund hierfür war zum einen, den

Versuchspersonen anhand eines nicht-transparenten, neutralen Algorithmus die Möglichkeit zu geben, die Aufgabe zu durchlaufen und zu verstehen. Zum andern war die Präsentation der Algorithmen, abgesehen von der Transparenzdarstellung, identisch. Indem die Bedingung ohne Transparenz immer am Anfang durchlaufen wurde, sollte ein Einfluss der Transparenzmanipulationen auf die neutrale Bedingung vermieden werden.

6.2.4. Versuchsablauf

Das Experiment wurde auf der Plattform SoSciSurvey erstellt und vom 11.11.2023 bis 07.01.2024 durchgeführt. Die Rekrutierung geschah per standardisierter E-Mail an die Bekanntschaft der Studienleiter*innen, an E-Mail-Verteiler von Psychologiestudierenden an der RWTH Aachen und den Plattformen SurveyCircle.com und SurveySwap.io. Die Teilnahmebedingungen lauteten: keine Teilnahme über Smartphone oder Tablet aufgrund möglicher technischer Probleme, Volljährigkeit und keine vorherige Teilnahme an einem der vorigen Gewichtschätzungsexperimente (siehe Anhang N). Als Incentivierung wurde die Teilnahme an einer Verlosung von drei Thalia-Gutscheinen im Wert von 10 € angeboten oder der Erhalt von 0,5 Versuchspersonenstunden für Psychologiestudierende der RWTH. Teilnehmende von SurveySwap und SurveyCircle erhielten Einlösungscodes zum Sammeln von Punkten auf der jeweiligen Plattform.

Per Link gelangen die Teilnehmenden zur Studie, wo sie Informationen zur Studie und ihrer Dauer erhielten, sowie Datenschutz (siehe Anhang C) und Einverständniserklärung zustimmten (siehe Anhang O). Anschließend lasen sie die Einleitung, in der die Vignette und der Studienablauf dargestellt wurde (siehe Abbildung 21). Durch die Vignette der Ernährungsapp sollte der tatsächliche Forschungsgegenstand verschleiert und durch praktische Implikationen das Interesse erhöht werden.

***Abbildung 21**: Einführung und Vignette zu Beginn des Experiments*

In der folgenden Studie sollen fünf verschiedene Algorithmen getestet werden. Diese wurden für eine Ernährungsapp entwickelt, um kontaktlos und auf Grundlage eines einzelnen Fotos das Gewicht des darauf abgebildeten Nahrungsmittels zu bestimmen. Ein solcher Algorithmus könnte zukünftig helfen bei der Kalorien- und Nährwertbestimmung von Essen, ohne dieses abwiegen zu müssen.

Die fünf Algorithmen (Algorithmus A bis Algorithmus E) werden Ihnen im Folgenden in zufälliger Reihenfolge angezeigt.

Wir bitten Sie nachfolgend, für **jedes** gezeigte Bild eine Schätzung für das Gewicht des darauf abgebildeten Gemüses/Obstes abzugeben. Anschließend wird Ihnen das vom Algorithmus ermittelte Gewicht präsentiert. Sie erhalten im Anschluss die Möglichkeit, Ihre endgültige Schätzung anzupassen.

Bitte versetzen Sie sich in die Lage einer Person, die die Ernährungsapp benutzt. Diese Person hat ein Interesse daran, eine möglichst gute finale Schätzung des Gewichts zu erreichen.

Bitte schätzen Sie also so präzise wie möglich!

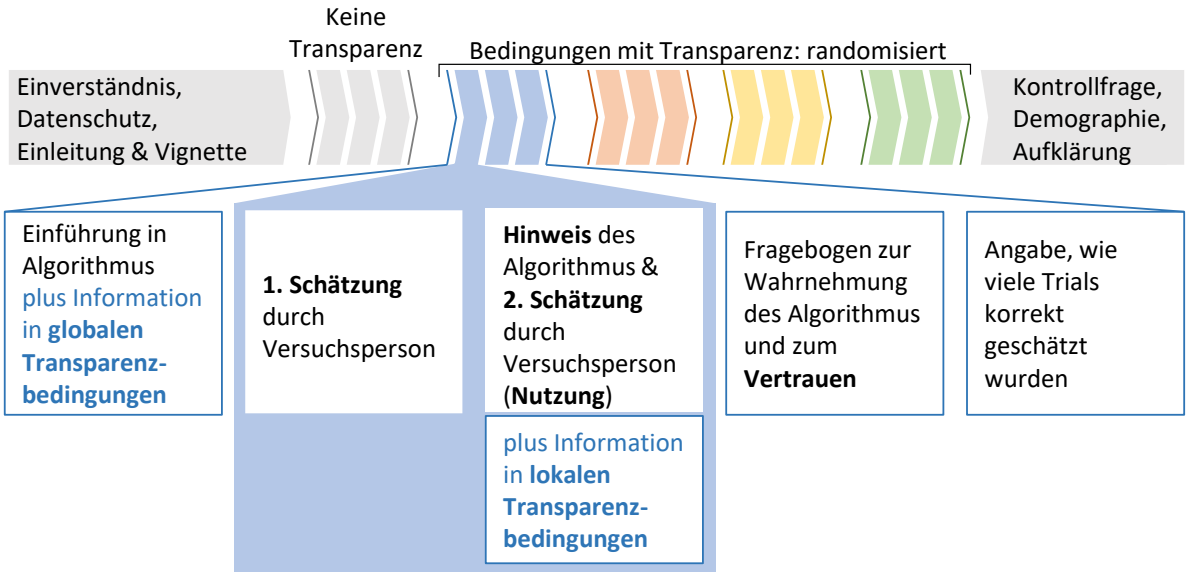
Nach einer Reihe von Schätzungen folgen jeweils einige kurze Fragen.

Im Anschluss durchliefen die Versuchspersonen die fünf Transparenzbedingungen, die aus je drei Schätzaufgaben bestanden (siehe Abbildung 22). Die drei Trials waren Schätzungen des Gewichts von Gemüse auf Fotos. Bei jeder Schätzung wurde der Einfluss des Algorithmus auf die Schätzung gemessen und quantifizierte so die Nutzung des Algorithmus (siehe Kapitel 6.2.5.1 für die Details der

Schätzaufgabe). Jede der fünf Transparenzbedingungen schloss mit einem Fragebogen zur allgemeinen Wahrnehmung des Algorithmus und zum Vertrauen. Das Feedback zur Anzahl richtiger Schätzungen („Sie haben x von 3 richtige Schätzungen abgegeben“), das die Versuchspersonen anschließend erhielten, bewertete Algorithmusschätzungen mit +/- 4 Gramm Abweichung vom tatsächlichen Gewicht als korrekt.

Nach fünf Algorithmus-Bedingungen folgten im freien Antwortformat zwei Fragen, in denen die Teilnehmenden ihre Vorgehensweise oder Strategie in der Zusammenarbeit mit dem Algorithmus beschreiben und Beispiele für alltäglich genutzte Algorithmen nennen sollten. Nach der Abfrage der Demographie folgte das Debriefing mit einer kurzen Erklärung zum Experiment. Dabei wurden die Versuchspersonen auch über die Wizard-of-Oz-Täuschung des Versuchs aufgeklärt. Den gesamten Ablauf des Experiments stellt Abbildung 22 dar.

Abbildung 22: *Versuchsablauf des Experiments. Oben ist der Gesamtablauf dargestellt, dabei farbig die vier Transparenzbedingungen. Unten der Detailblick zum Ablauf einer einzelnen Bedingung (blaue Linie) bzw. eines einzelnen Trials (blauer Block). Blauer Text repräsentiert Informationen, die sich in den zwei lokalen bzw. den zwei globalen Bedingungen unterscheiden (siehe Kapitel 6.2.5.2).*



6.2.5. Material

Das Material zur Ausgestaltung der Transparenzbedingungen sowie die verwendeten Fragebögen werden im Folgenden dargelegt. Dabei wurde Bildmaterial von Obst und Gemüse genutzt, das erstmals von Werz et al. (2021) veröffentlicht wurde. Das JAS und die gewählte Aufgabe sind außerdem identisch wie in Forschungsfrage (a) „Fehlerfall“ (Kapitel 4.2.2).

6.2.5.1. Schätzaufgabe

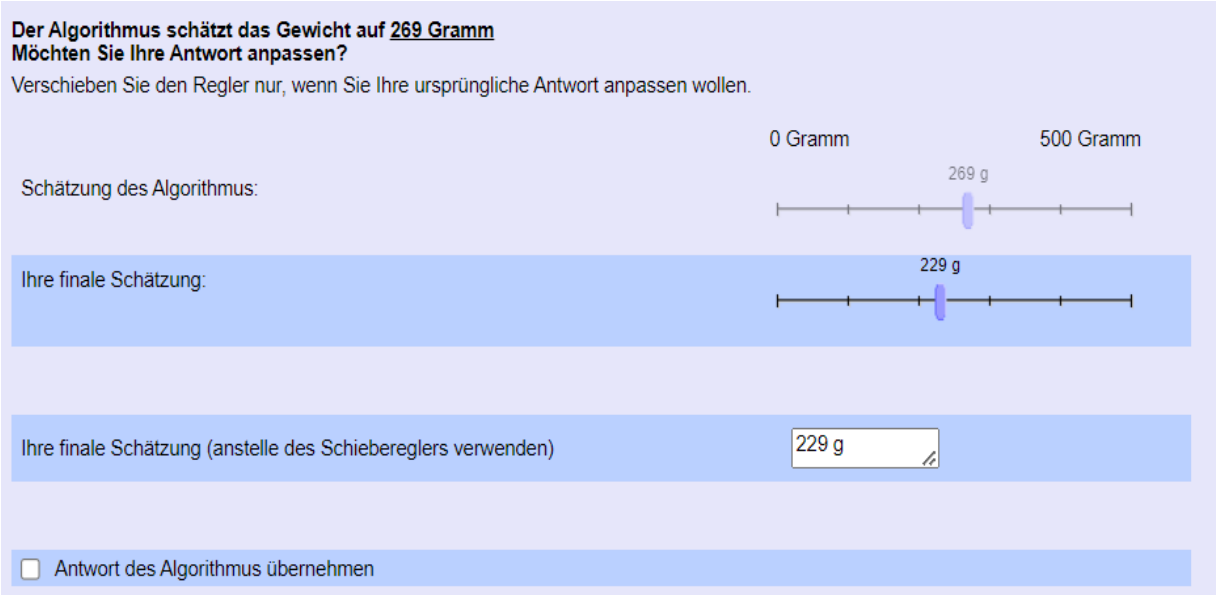
Zur Gestaltung der Schätzaufgabe wurde ein JAS genutzt (siehe auch Kapitel 4.2.2 zur Forschungsfrage (a) „Fehlerfall“). Dabei gaben die Teilnehmenden zunächst ihre Schätzung zum Gewicht von auf einem Foto dargestellten Obst und/oder Gemüse ab (siehe Abbildung 23). Mit einem Klick auf „Weiter“

gelangten die Versuchspersonen auf die nächste Seite, auf der die Schätzung des Algorithmus präsentiert wurde und sie ihre finale Antwort abgeben konnten (siehe Abbildung 24).

Auf beiden Seiten konnten sie ihre Antwort mithilfe eines Schiebereglers eingeben oder die Grammzahl in ein Feld eintippen. Zusätzlich konnten sie bei der finalen Schätzung über eine Checkbox die Antwort des Algorithmus übernehmen. Diese Verbesserungen wurden angelehnt an die Limitationen aus Forschungsfrage (a) „Fehlerfall“, die die Einfachheit der Interaktion und dabei insbesondere die Betätigung des Schiebreglers betrafen, vorgenommen.

Abbildung 23: *Darstellung der ersten Gewichtsschätzung durch die Versuchspersonen*

Abbildung 24: *Darstellung der Algorithmusschätzung sowie der Anpassungsmöglichkeit der finalen Schätzung durch die Versuchsperson*



6.2.5.2. Transparenzbedingungen

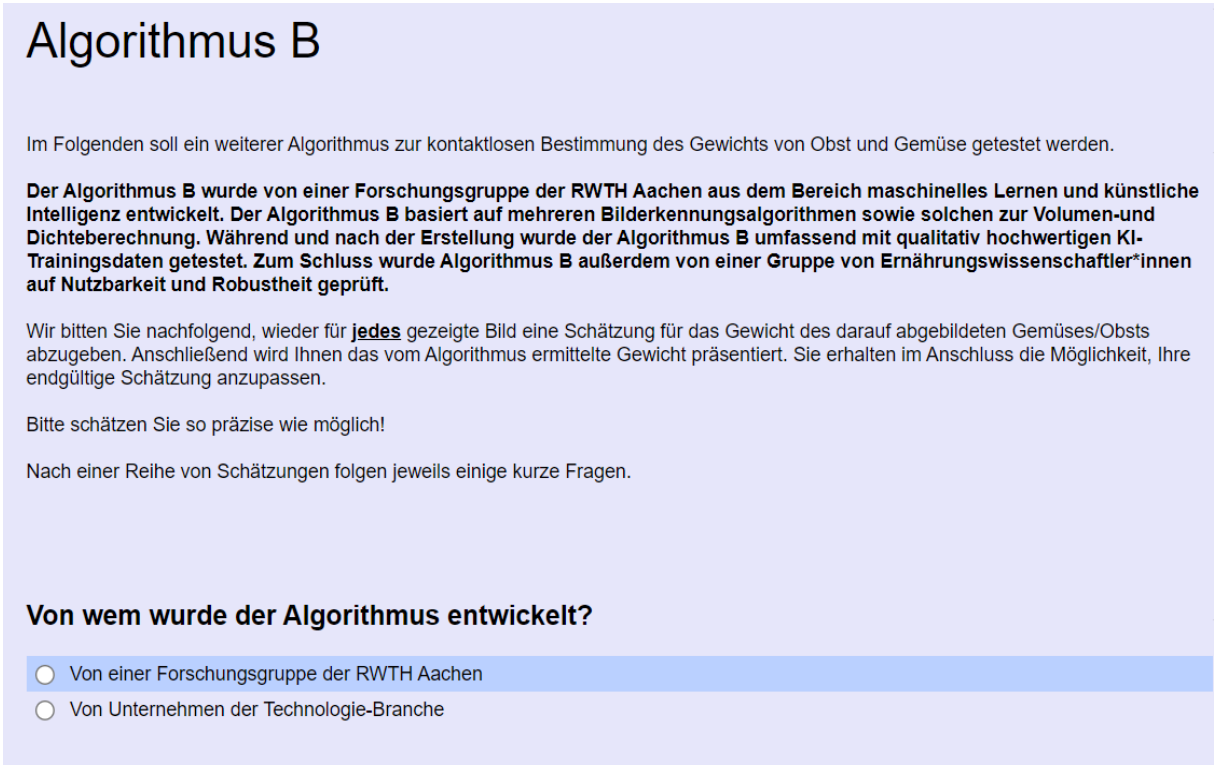
Das Experiment bestand aus fünf Testbedingungen: eine ohne und vier mit Transparenz. Um die Bedingungen leichter unterscheidbar zu machen, wurde für jede Bedingung eine andere Hintergrundfarbe genutzt. Dabei wurde der Algorithmus ohne Transparenz immer als erster angezeigt und die vier Algorithmen mit Transparenz im Anschluss daran randomisiert. Die vier Transparenzarten basieren auf der Kombination zweier Dimensionen: die Art der Information, Funktionsweise (F) oder Akkuratheit (A), sowie die Art der Erklärung, global (g) bzw. lokal (lo). Aus der Kombination der beiden Dimensionen ergibt sich eine 2x2 Kreuztabelle (siehe Tabelle 18) mit den vier Transparenz-Bedingungen: F-lo und F-g sowie A-lo und A-g.

Tabelle 18: *Die vier Transparenzarten der vier Transparenzbedingungen in Forschungsfrage (c)*

	global	**lokal**
Funktionsweise	Globale Funktionsweise (Algorithmus F-g)	Lokale Funktionsweise (Algorithmus F-lo)
Akkuratheit	Globale Akkuratheit (Algorithmus A-g)	Lokale Akkuratheit (Algorithmus A-lo)

F-g: Bei der Erklärung zur Funktionsweise auf globaler Ebene wurde zusätzlich zum Hinweis, es folge gleich ein neuer Algorithmus, ein Text präsentiert, der Hintergrundinformationen sowie Verarbeitungshinweise zum vorgeblichen Algorithmus lieferte. Die Darstellung sowie der Wortlaut sind Abbildung 25 zu entnehmen. Zur Testung der Aufmerksamkeit der Teilnehmenden folgte am Ende der Seite ein Aufmerksamkeitscheck mit einer Frage zum Text.

Abbildung 25: *Darstellung der Einführung des Algorithmus B und der Manipulation der Transparenz als globale Funktionsweise (F-g) sowie der anschließenden Aufmerksamkeitsprüfung*



F-lo: In den lokalen Transparenzbedingungen wurde in der Einführung der Algorithmen keine neue Information präsentiert. Vielmehr wurden nur die Informationen aus der Vignette zu Beginn wiederholt. Zur Transparenzmanipulation wurde nach der Schätzung durch den Algorithmus das gleiche Bild als Heatmap dargestellt. Abbildung 26 stellt dar, wie diese Transparenzart gestaltet war.

A-**g**: Die globale Akkuratheit gibt eine Sicherheit des Algorithmus an. Die in dieser Studie dargestellte Information war der Satz „Dieser Algorithmus liegt in 90,1 % der Fälle richtig“ und bei der Einführung des Algorithmus vor den drei Schätzaufgaben ergänzt. Auch für diese Information wurde ein Aufmerksamkeitscheck eingeführt, bei dem die Teilnehmenden aus zwei Vorgaben auswählen sollten, bei wie viel Prozent der Fälle der Algorithmus laut Angaben richtig liegt.

A-lo: Bei dieser lokalen Transparenzart wurde bei der Einführung des Algorithmus der Text aus der Anfangsvignette wiederholt. Die lokale Akkuratheitsangabe erfolgte jeweils bei den drei Algorithmusschätzungen. Um einen Unterschied zur anderen Akkuratheitsbedingung deutlich zu machen und um möglichst realistische Informationen bereitzustellen, wurde die Akkuratheit als Spannweite angegeben. Der Satz, der bei den drei Empfehlungen dieses Algorithmus jeweils zu lesen war, lautete „Bei diesem Bild weist der Algorithmus eine Unsicherheit von +/- xx Gramm auf“, wobei die Streuung in den drei Trials mit 7,2 Gramm, 3,5 Gramm und 5,3 Gramm angegeben war.

Abbildung 26: *Darstellung der Transparenzart lokale Funktionsweise (F-lo) mit einer Heatmap; die Überschrift lautete: „Der Prozess von Algorithmus E lässt sich wie folgt visualisieren:“*

6.2.5.3. Abhängige Variablen

Das Experiment untersucht zwei abhängige Variablen. Die erste abhängige Variable ist die Nutzung des Algorithmus. Erneut wurde dazu der WOA genutzt (Prahl & Swol, 2017). Mithilfe des JAS wird durch diesen der Einfluss der algorithmischen Vorschläge auf die eigene Schätzung der Teilnehmenden quantifiziert.

Die zweite abhängige Variable war das Vertrauen in den zuvor präsentierten Algorithmus. Dazu bewerteten die Versuchspersonen drei Items auf einer 7-stufigen Likert-Skala (1 = „Stimme überhaupt nicht zu“ bis 7 = „Stimme voll und ganz zu“). Die Items lauteten „Ich vertraue dem Algorithmus“, „Ich finde den Algorithmus vertrauenswürdig“ und als invertiertes Item: „Ich bin dem Algorithmus gegenüber misstrauisch“. Die Skala wies eine hohe interne Konsistenz von $\alpha = 0{,}91$ auf. Die Berechnung der abhängigen Variablen wird im Kapitel 6.2.6 näher erläutert.

6.2.5.4. Weitere erfasste Variablen

Zusätzlich zu den drei Vertrauensitems wurden drei weitere Items abgefragt, die die allgemeine Wahrnehmung des Algorithmus betrafen. Die drei hier ergänzten Items lauten „Ich finde den Algorithmus zuverlässig“, „Der Algorithmus ist für mich verständlich“ und „Der Algorithmus ist für mich transparent“. Diese Items wurden zu explorativen Zwecken ergänzt, um so auf mögliche Auswirkungen

der Transparenzarten auf die Wahrnehmung des Algorithmus schließen zu können. Es ist zu ergänzen, dass in anderen Untersuchungen einzelne der hier verwendeten Items als Items in Vertrauensskalen verwendet wurden (z. B. Verberne et al., 2012). In der vorliegenden Studie wurden die Konstrukte separat betrachtet, um die Trennschärfe zu erhalten.

Nach Durchlaufen aller Bedingungen wurden die Versuchspersonen in einer offenen Abfrage gebeten, zu nennen, welche Algorithmen ihnen aus dem Alltag bekannt seien und in einer weiteren Frage, welche Strategie sie in der Zusammenarbeit mit dem Algorithmus angewendet hätten. Bei der Erhebung der demografischen Daten (Geschlecht, Alter, Bildungsabschluss, Muttersprache Deutsch) erfolgte außerdem die Selbsteinschätzung, wie routiniert die Teilnehmenden im Umgang mit Computern seien.

6.2.6. Auswertung

Vor Beginn der Erhebung wurde eine A-priori-Poweranalyse mit G*Power (Faul et al., 2007) durchgeführt für eine messwiederholte ANOVA mit fünf Innersubjektfaktoren. Bei einer angenommenen Korrelation von r = 0,7 zwischen den Messzeitpunkten ergab sich ein benötigter Stichprobenumfang von n = 113 Versuchspersonen, um bei einer Power von 95 % einen kleinen Effekt von f = 0,10 (Cohen, 1992) nachweisen zu können. Die hohe Korrelation zwischen den Messzeitpunkten beruht auf der Annahme, dass sich die Effekte der Transparenzarten mit kleinem Effekt zeigen, und die Werte für Nutzung und Vertrauen auch über Bedingungen hinweg große Konstanz aufweisen würden. Bei einer Korrelation über die Messzeitpunkte von r = 0,6 würden n = 150 Versuchspersonen benötigt, weshalb eine entsprechend höhere Zahl von Personen erhoben wurde, um auch bei niedrigerer Korrelation einen Effekt zu ermitteln. Zur Auswertung der Daten wurde das Programm IBM SPSS Statistics Version 28.0.0.0 (190) genutzt. Soweit nicht anders angegeben, wurde für alle statistischen Berechnungen ein α-Fehler-Niveau von 0,05 angenommen.

Zunächst wurden zur Datenbereinigung verschiedene Analysen durchgeführt, u. a. die freien Antworten der Teilnehmenden auf Auffälligkeiten, eine korrekte Beantwortung der Kontrollfragen sowie mögliche Ausreißer in den abhängigen Variablen überprüft (siehe Abschnitt 6.2.2 „Stichprobe“). Nach Ausschluss ungültiger oder problematischer Fälle aus der Stichprobe wurde die Analyse mit n = 151 gültigen Fällen weiterverfolgt.

Zur Berechnung der ersten abhängigen Variable zur Nutzung des Algorithmus in der Transparenzbedingung wurden die drei WOAs gemittelt und so eine Nutzungsvariable pro Bedingung erstellt. In Fällen, in denen die initiale Schätzung identisch mit der Algorithmusschätzung ist und dadurch ein Divisor = 0 entsteht, kann der WOA laut Formel nicht berechnet werden. Dies lag bei 18 WOAs vor, die in der Folge als nicht vorhandene Werte behandelt wurden. In diesen Fällen wurde deshalb der Mittelwert der Verhaltensvariable nur aus den beiden anderen Fällen errechnet.

Zur Erstellung der zweiten abhängigen Variable wurden die drei Items der Vertrauensskala durch Berechnung des Mittelwerts in eine Vertrauenswert pro Bedingung umgewandelt. Um Transparenz- und Nichttransparenzbedingungen vergleichen zu können, wurde zudem je ein Mittelwert über alle Transparenzbedingungen für die Verhaltens- bzw. die Vertrauensmaße errechnet. Ebenso wurden, um die lokalen mit den globalen Bedingungen vergleichen zu können, mittlere Verhaltens- sowie Vertrauenswerte für die lokalen bzw. globalen Bedingungen gebildet.

Zur Prüfung der sechs Hypothesen wurden sechs within-faktorielle ANOVAs durchgeführt: Drei ANOVAs bezogen sich auf die abhängige Verhaltensvariable der Nutzung, drei auf die abhängige Variable Vertrauen. Zwar konnte bei den Verhaltensmaßen keine Normalverteilung gewährleistet werden, jedoch sind ANOVAs relativ robust gegenüber Verletzungen der Normalverteilungsannahme (Vasey & Thayer, 1987). Die ANOVAs verglichen zum Ersten die gemittelten Transparenzbedingungen mit der Bedingung ohne Transparenz. Zum Zweiten wurden die vier verschiedenen Transparenzarten miteinander verglichen. Drittens wurden die lokalen mit den globalen Transparenzbedingungen verglichen.

6.3. Ergebnisse

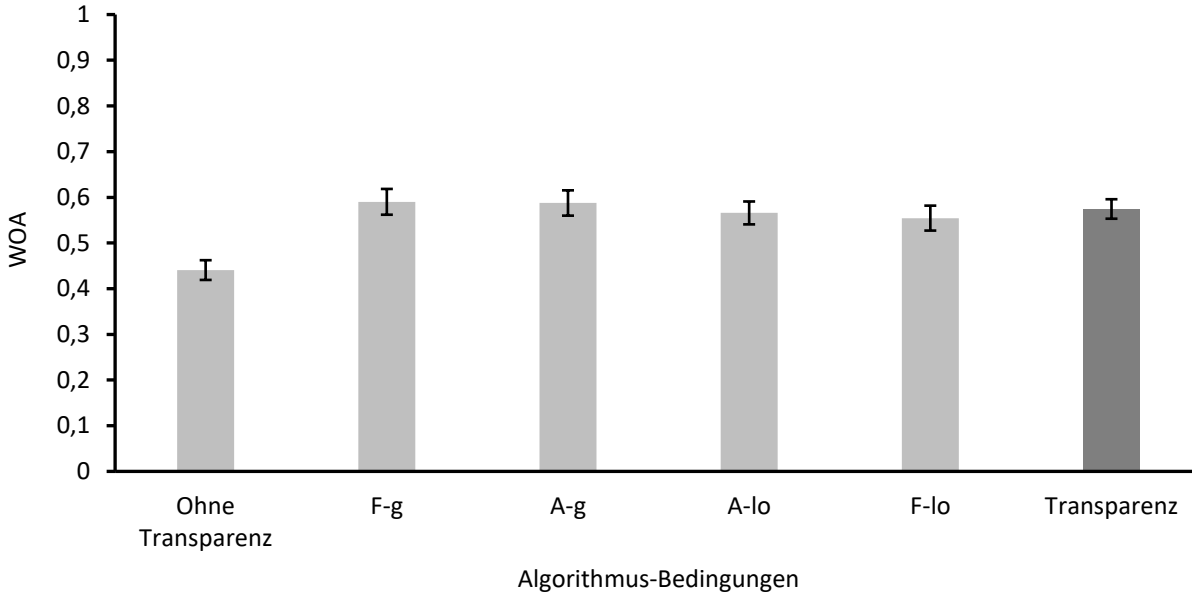
Im Folgenden werden die aus dem Online-Experiment erhobenen und nach dem zuvor beschriebenen Vorgehen ermittelten Ergebnisse berichtet. Die Ergebnisse gliedern sich in die beiden Abschnitte zur Hypothesentestung über die Effekte von Transparenz und die Unterschiede zwischen den Transparenzarten. Anschließend werden weitere, ergänzende Analysen und zuletzt die Antworten aus dem freien Antwortformat berichtet.

6.3.1. Effekte von Transparenz

Die zweistufige ANOVA zum Vergleich des Effekts von Algorithmen mit (Algorithmus F-g, F-lo, A-g, A-lo) und ohne Transparenz zeigte einen signifikanten Unterschied hinsichtlich ihrer Nutzung ($F(1, 150) = 34{,}37$, $p < 0{,}001$, $\eta^2_p = 0{,}19$). Damit lässt sich Hypothese 1 bestätigen. Der durchschnittliche WOA in der Bedingung ohne Transparenz war geringer ($M = 0{,}44$, $SD = 0{,}27$) als der WOA in den Bedingungen mit Transparenz ($M = 0{,}57$, $SD = 0{,}26$). In Abbildung 27 sind die Verhaltenswerte als mittlerer WOA pro Bedingung abgebildet.

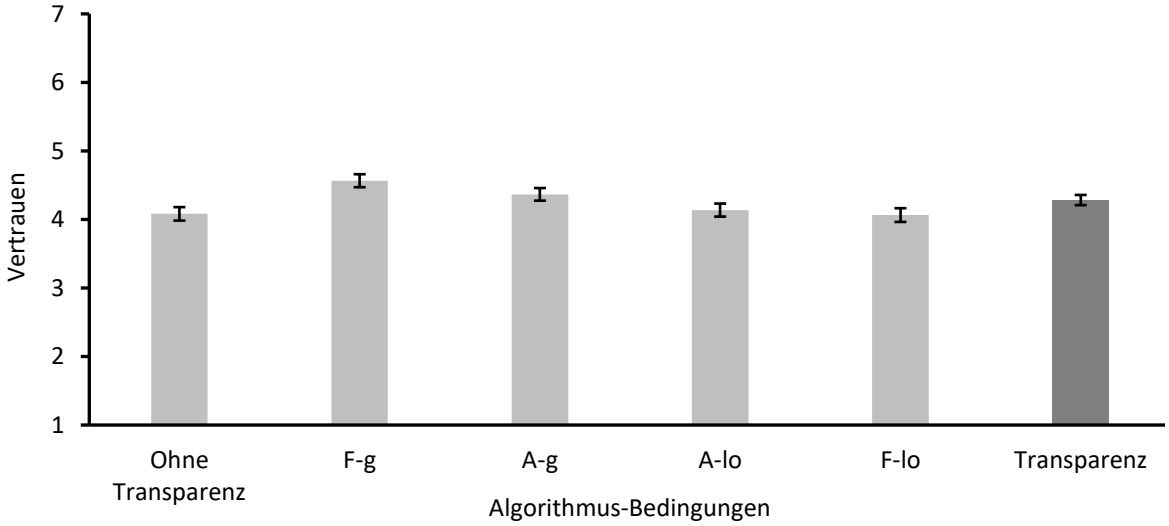
Die ANOVA zur abhängigen Variable Vertrauen zeigte einen signifikanten Unterschied zwischen Bedingungen mit Transparenz und der ohne Transparenz ($F(1, 150) = 6{,}15$, $p = 0{,}014$, $\eta^2_p = 0{,}04$). Auch Hypothese 2 kann also bestätigt werden. Im Durchschnitt zeigten sich niedrigere Vertrauenswerte für die Bedingung *ohne Transparenz (M = 4,08, SD = 1,21) als für die Bedingungen mit Transparenz (M = 4,28, SD = 0,91).* Abbildung 28 stellt die mittleren Vertrauenswerte pro Bedingung dar.

Abbildung 27: *Durchschnittliche Werte des Weight of Advice (WOA) für die fünf Algorithmus-Bedingungen. Der Balken Transparenz bildet den Durchschnitt der vier Transparenzbedingungen F-g (Funktionsweise-global), A-g (Akkuratheit-global), A-lo (Akkuratheit-lokal) und F-lo (Funktionsweise-lokal) ab.*



Anmerkung. Die Whisker stellen +/- einen Standardfehler dar.

Abbildung 28*: Durchschnittliche Vertrauenswerte für die fünf Algorithmus-Bedingungen. Der Balken Transparenz bildet den Durchschnitt der vier Transparenzbedingungen F-g (Funktionsweise-global), A-g (Akkuratheit-global), A-lo (Akkuratheit-lokal) und F-lo (Funktionsweise-lokal) ab.*



Anmerkung. Vertrauen wurde auf einer 7-stufigen Likert-Skala (1 = „Stimme überhaupt nicht zu“ bis 7 = „Stimme voll und ganz zu“) erhoben. Die Whisker stellen +/- einen Standardfehler dar.

6.3.2. *Unterschiede zwischen Transparenzarten*

Die ANOVA zum Vergleich der vier Transparenzarten (Algorithmus F-g, F-lo, A-g, A-lo) zeigte einen signifikanten Unterschied hinsichtlich der Nutzung (F_{GG}(2,84, 426,54) = 0,8, p = 0,490, η^2 = 0,01; Korrektur der Freiheitsgrade nach Greenhouse-Geisser, da keine Homoskedastizität gegeben war).

Hypothese 3 lässt sich also ebenfalls bestätigen. Während Algorithmus F-g (*M*=0,59, *SD* = 0,35) und Algorithmus A-g identische WOA-Werte aufwiesen (*M* = 0,59, *SD* = 0,34), waren die Nutzungswerte von Algorithmus A-lo (*M* = 0,57, *SD* = 0,31) und Algorithmus F-lo (*M* = 0,55, *SD* = 0,34) etwas niedriger (siehe Abbildung 27).

Analog zur Nutzungsvariable wurde eine ANOVA für die abhängige Variable Vertrauen berechnet. Dabei zeigte sich ein signifikanter Unterschied (F_{GG}(2,73, 409,74) = 10,83, p < 0,001, η^2 = 0,07; Korrektur der Freiheitsgrade nach Greenhouse-Geisser, da keine Homoskedastizität gegeben war) hinsichtlich des Effekts der verschiedenen Transparenzarten auf Vertrauen. Damit kann Hypothese 4 bestätigt werden. Das Vertrauen für Algorithmus F-g war am höchsten (*M* = 4,57, *SD* = 1,16), für Algorithmus A-g am zweithöchsten (*M* = 4,37, *SD* = 1,13), gefolgt von Algorithmus A-lo (*M* = 4,14, *SD* = 1,17) und zuletzt Algorithmus F-lo (*M* = 4,06, *SD* = 1,23), was auch Abbildung 28 zu entnehmen ist.

Zur Überprüfung der Hypothesen 5 und 6 zum unterschiedlichen Effekt von lokalen und globalen Transparenzarten auf Nutzung bzw. Vertrauen wurden ebenfalls ANOVAs berechnet. Der Vergleich der WOA-Mittelwerte in globalen (*M* = 0,59, *SD* = 0,29) und lokalen Bedingungen (*M* = 0,56, *SD* = 0,27) zeigte keinen signifikanten Effekt (*F*(1, 150) = 2,79 , *p* = ,097, η^2 = 0,02), weshalb Hypothese 5 abzulehnen ist. Jedoch zeigte sich ein signifikanter Unterschied (*F*(1, 150) = 34,37, *p* < 0,001, η^2 = 0,19) zwischen dem Vertrauen in globale (*M* = 4,47, *SD* = 1,04) und in lokale Transparenzarten (*M* = 4,10, *SD* = 0,98). Hypothese 6 kann also angenommen werden.

6.3.3. Weitere Analysen

Als zusätzliche explorative Analysen wurden für Korrelationen zwischen den erhobenen Variablen errechnet. Für die ordinalskaliert vorliegenden Daten aus der Demographie wurden Spearman-Korrelationen berechnet. Die Korrelationsmatrix finden sich in Anhang P. Dabei zeigte keine der demographischen Variablen eine Korrelation mit Nutzung oder Vertrauen. Auch der Umgang mit Computern wies keinen Zusammenhang mit Vertrauen in die Algorithmen (ρ = -,06, *p* = ,451) oder mit ihrer Nutzung auf (ρ = -,15, *p* = ,065).

Da die Werte zur Wahrnehmung der Algorithmen, WOAs und Vertrauen in die Algorithmen als intervallskaliert vorliegend angesehen werden können, konnten hierfür Pearson-Korrelationen berechnet werden. Die Nutzung und das Vertrauen standen in einem positiven signifikanten Zusammenhang mittlerer Korrelationsstärke (Cohen, 1992; siehe Tabelle 19). Die Korrelation der Wahrnehmungsitems („zuverlässig“, „verständlich“, „transparent“) untereinander waren ausschließlich hoch. Darüber hinaus zeigten sich mittlere bis starke Korrelationen zwischen den drei Wahrnehmungsitems und der Nutzung (WOA) bzw. dem Vertrauen in die Algorithmen. Es zeigt sich ein hoher Zusammenhang zwischen der Zuverlässigkeit und der Nutzung, während der mit Vertrauen lediglich im mittleren Bereich liegt. Umgekehrt waren die Korrelationen zwischen Verständlichkeit und

Vertrauen sowie Transparenz und Vertrauen sehr stark, mit Nutzung lediglich von mittlerer Stärke (Cohen, 1992; siehe Tabelle 19 für alle Korrelationen).

Tabelle 19: *Pearson-Korrelation der Wahrnehmungsitems mit der Nutzung der Algorithmen (Weight of Advice; WOA) und dem Vertrauen in die Algorithmen*

		Nutzung (WOA)	Vertrauen	zuverlässig	verständlich
Vertrauen	*r*	,359***			
	p	<,001			
zuverlässig	*r*	,426***	,268***		
	p	<,001	<,001		
verständlich	*r*	,302***	,647***	,634***	
	p	<,001	<,001	<,001	
transparent	*r*	,335***	,688***	,563***	,802***
	p	<,001	<,001	<,001	<,001

*Anmerkung. *** = Die Korrelation ist auf dem Niveau von <0,001 signifikant. n = 151.*

ANOVAs zu den drei Wahrnehmungsmaßen und der Algorithmusnutzung sowie -vertrauen für die fünf Bedingungen zeigten signifikante Effekte für die wahrgenommene Zuverlässigkeit, Verständlichkeit und Transparenz (siehe Tabelle 20). Die Zuverlässigkeit wurde deskriptiv bei den lokalen Transparenzarten als niedriger bewertet als bei den globalen Transparenzarten. Beim Effekt auf die Verständlichkeit wurde der Algorithmus ohne Transparenz als am verständlichsten eingeschätzt, dicht gefolgt von F-g, der globalen Erklärung zu Funktionalität. Hingegen wurden die Bedingungen mit Transparenz als transparenter bewertet als die Bedingung ohne Transparenz. Am höchsten war die wahrgenommene Transparenz deskriptiv in der Bedingung F-lo, die Heatmaps-Darstellung, die allerdings als am wenigsten verständlich bewertet wurde (siehe Tabelle 20).

Da die Algorithmen immer die korrekte Antwort lieferten, wurde außerdem der Einfluss eines Lerneffekts auf Nutzung und Vertrauen geprüft. Zur Berechnung dieses Effekts mussten zunächst neue abhängige Variablen transformiert werden, in denen die Nutzungs- und Vertrauenswerte nach der Präsentationsreihenfolge und nicht mehr nach den Inhalten der Bedingungen sortiert waren. Darin waren die Bedingungen von 1 bis 5 nach Position im Experiment und unabhängig von ihrem Inhalt nummeriert. Mithilfe dieser neuen Kodierung wurde je eine ANOVA gerechnet mit den fünf Bedingungen und den beiden abhängigen Variablen Nutzung (WOA) und Vertrauen. Die Ergebnisse zeigten signifikante Effekte der Reihenfolge auf die Nutzung ($F(4, 600)= 14{,}14$, $p < 0{,}001$, $\eta^2 = 0{,}09$) und für das Vertrauen ($F_{GG}(3{,}77, 565{,}80)= 2{,}67$, $p = 0{,}034$, $\eta^2 = 0{,}02$; korrigiert nach Greenhouse-Geisser). Die Nutzung der Algorithmusvorschläge stieg mit der Präsentationreihenfolge. Die erste Bedingung (ohne Transparenz) wies im Durchschnitt die niedrigsten WOA-Werte auf ($M = 0{,}44$, $SD = 0{,}27$). In der zweiten Bedingung waren sie bereits höher ($M = 0{,}51$, $SD = 0{,}32$), ebenso in der dritten ($M = 0{,}57$, $SD = 0{,}34$), vierten ($M= 0{,}59$, $SD= 0{,}33$) und fünften Bedingung ($M = 0{,}63$, $SD = 0{,}34$). Auch bei den Vertrauenswerten stiegen die Werte von der ersten ($M = 4{,}08$, $SD = 1{,}21$), zur zweiten ($M = 4{,}18$, $SD =$

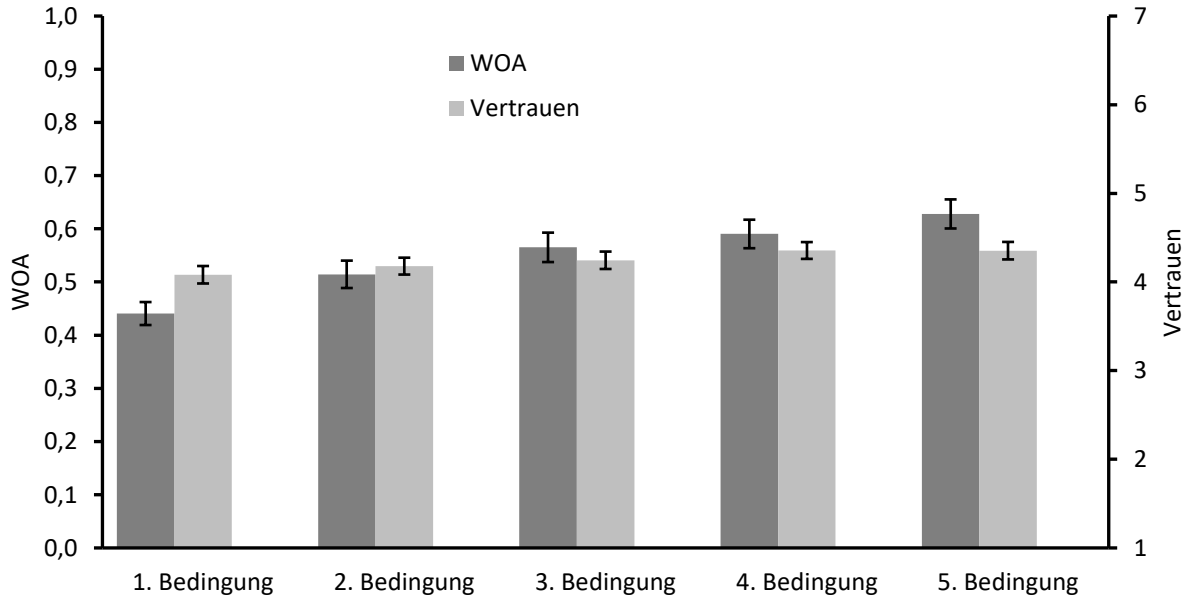
1,17) und zur dritten Bedingung (*M* = 4,25, *SD* =1,21). Die vierte (*M* = 4,36, *SD* = 1,16) und fünfte Bedingung (*M* = 4,35, *SD* = 1,21) wurden als ähnlich vertrauenswürdig bewertet (siehe Abbildung 29).

Tabelle 20: *Deskriptive Werte und ANOVAs zum Effekt der Bedingungen auf die subjektiven Wahrnehmungsmaße*

		M	*SD*	*p*	df (Konstante, Fehler)	*F*	Eta-Quadrat
zuverlässig	Ohne Transparenz	4,33	1,11	,003	4, 600	4,02	0,03
	F-g	4,64	1,12				
	A-g	4,50	1,19				
	A-lo	4,33	1,11				
	F-lo	4,30	1,23				
verständlich	Ohne Transparenz	4,56	1,64	<,001	3,44, 515,72$_{(GG)}$	5,32	0,03
	F-g	4,34	1,56				
	A-g	4,05	1,61				
	A-lo	4,07	1,48				
	F-lo	4,03	1,68				
transparent	Ohne Transparenz	2,89	1,39	<,001	3,21, 481,16$_{(GG)}$	23,72	0,14
	F-g	3,83	1,59				
	A-g	3,54	1,59				
	A-lo	3,74	1,61				
	F-lo	4,17	1,71				

Anmerkung. Mit (GG) gekennzeichnete Freiheitsgrade wurden nach Greenhouse-Geisser korrigiert, da bei ihnen die Sphärizitätsannahme verletzt war. *n* = 151.

Abbildung 29: *Nutzung (WOA = Weight of Advice) des und Vertrauen in den Algorithmus nach der Präsentationsreihenfolge der fünf Bedingungen*



Anmerkung. Die Whisker stellen +/- einen Standardfehler dar.

Zuletzt galt es zu ermitteln, ob die Nutzung des Algorithmus überhaupt zu signifikant besseren Ergebnissen geführt hatte bzw. hätte als die selbstständige Beantwortung der Schätzaufgaben ohne Algorithmushinweis. Dafür wurde zunächst die mittlere Abweichung aller initialen Schätzungen vom

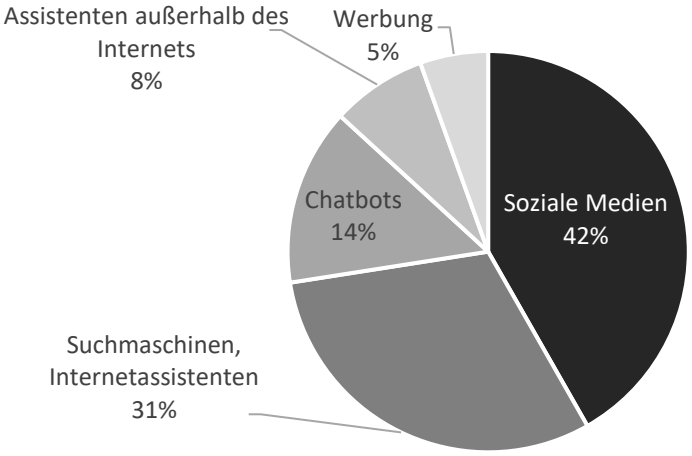
wahren Gewicht ermittelt. Da die Algorithmusschätzungen bei allen Schätzungen korrekt waren, lag die Abweichung des Algorithmus vom wahren Wert bei 0. Deshalb wurde anhand der mittleren Abweichung der initialen Schätzungen vom wahren Gewicht (M = 57,63, SD = 15,41) ein Einstichproben t-Test gegen 0 gerechnet (t(150)= 45,96, p < ,001). Das Ergebnis zeigte deutlich, dass die initialen Schätzungen signifikant vom wahren Wert abwichen. Dem Algorithmus zu folgen war/wäre also in jedem Fall besser, als den eigenen initialen Einschätzungen zu folgen.

6.3.4. Freie Antwortfelder

Zur Auswertung der freien Antwortfelder wurden die Antwortangaben thematisch geclustert. Die folgenden Diagramme bilden ab, wie häufig eine Kategorie in den Antworten identifiziert werden konnte. Dabei wurden Aussagen, die gleichzeitig unterschiedliche Kategorien ansprachen, mehrfach zugeordnet.

Zur Frage, welche Algorithmen den Teilnehmenden bekannt seien aus dem Alltag, machten lediglich n = 76 Teilnehmende Angaben. Am häufigsten wurden soziale Medien wie Tiktok, YouTube, und Instagram genannt (siehe Abbildung 30).

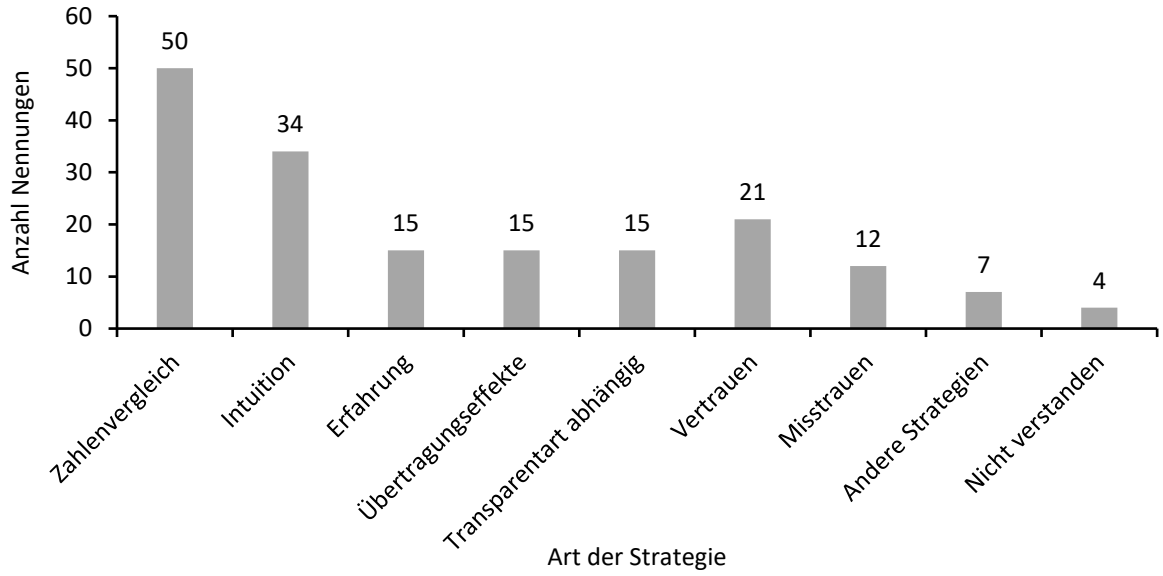
Abbildung 30: *Absolute Häufigkeit der von Versuchspersonen in offener Abfrage genannten Arten von ihnen bekannten Algorithmen aus dem Alltag; nachträglich kategorisiert.*



Die offene Frage, welche Strategie in der Zusammenarbeit mit dem Algorithmus angewendet worden war, beantworteten n = 133 der Teilnehmenden. Die Antworten wurden wiederum als thematische Kategorien ausgewertet (siehe Abbildung 31). Die am häufigsten genannte Strategie lässt sich in der Kategorie *Zahlenvergleich* zusammenfassen. Die Versuchspersonen gaben dabei an, einen Abgleich zwischen der Algorithmusschätzung und ihrer Schätzung durchgeführt zu haben und ihr Verhalten dem Unterschied zwischen diesen Zahlen angepasst zu haben. Des Weiteren berichteten n = 34 nach *Intuition* bzw. n = 15 nach *Erfahrung* geantwortet zu haben. Die Kategorie der *Übertragungseffekte* fasst die Berichte von Gewöhnungs- bzw. Lerneffekten bezüglich der Aufgabe zusammen. Teilnehmende schilderten beispielsweise, über die Zeit hinweg besser geworden zu sein oder sich an

früheren Gewichtsschätzungen orientiert zu haben. Die Kategorie *Transparenzart abhängig* fasst die Strategien zusammen, bei denen auf die Beschreibung der Algorithmen geachtet und diesen angepasst gehandelt wurde. Manche Teilnehmende gaben zudem an, dem Algorithmus grundsätzlich vertraut bzw. misstraut zu haben (*Vertrauen* bzw. *Misstrauen*). Einzelne Strategien ließen sich keiner der genannten zuordnen (*Andere Strategien*) oder zeugten davon, dass die Versuchspersonen die Frage *nicht verstanden* hatten.

Abbildung 31: *Anzahl der verschiedenen genannten Strategien zur Zusammenarbeit mit dem Algorithmus; freies Antwortformat, das nachträglich kategorisiert wurde.*



6.4. Diskussion

Lange Zeit wurde an transparenter KI in erster Linie aus technischer Perspektive geforscht (Murdoch et al., 2019). Die Frage, wie die Transparenz für Endnutzende gestaltet werden soll oder wie sich verschiedene Arten der Transparenz auf Akzeptanz, Vertrauen oder Nutzung der Systeme auswirken, stand nicht im Zentrum der Forschung (Felzmann et al., 2019; Páez, 2019). Dabei wird die Frage angesichts der Zunahme von KI im Alltag und im Arbeitskontext zunehmend wichtiger. Dies geht besonders darauf zurück, dass Transparenz für Endnutzende mehr und anderes bedeutet als die technische Umsetzung derselben (siehe auch Kapitel 5 zur Forschungsfrage b). Gleichzeitig fordern politische Organe vermehrt die Umsetzung von Transparenz beim Einsatz von KI (AI HLEG, 2020).

In der vorliegenden Studie wird die Frage untersucht, wie sich verschiedene Arten der Transparenz von KI auf Vertrauen und Nutzung der Systeme auswirken. Im folgenden Kapitel werden die Ergebnisse im Licht der aufgestellten Hypothesen diskutiert und in den Forschungsstand eingeordnet. Anschließend erfolgt eine Besprechung der möglichen Limitationen, die bei der Interpretation und Generalisierung der Ergebnisse beachtet werden sollten.

6.4.1. Effekt verschiedener Transparenzarten auf Nutzung und Vertrauen in KI-Systeme

Die Hypothesen ließen sich überwiegend bestätigen: Die Ergebnisse zeigten einen Unterschied der Nutzung von Algorithmen mit und ohne Transparenz (Hypothese 1) wie auch bezüglich des Vertrauens in sie (Hypothese 2), wobei die Algorithmen mit Transparenz mehr genutzt wurden bzw. ihnen mehr vertraut wurde als ohne Transparenz. Während der Nutzungseffekt im großen Bereich liegt, war die Effektstärke für das Vertrauen lediglich klein bis Mittel.

Der Vergleich der vier Transparenzarten ergab einen kleinen und knapp signifikanten Unterschied in der Nutzung der vier Transparenzarten (Hypothese 3). Der Unterschied des Vertrauens in die vier Transparenzarten, mit dem Hypothese 4 bestätigt werden kann, erhielt eine mittlere Effektstärke. Dabei hatten Hintergrundinformationen einen besonders großen und die Heatmaps einen besonders geringen Einfluss auf Vertrauen bzw. in schwächerer Form auf Nutzung. Hypothese 5, die eine unterschiedliche Nutzung lokaler und globaler Transparenzarten postulierte, muss abgelehnt werden. Bezüglich des Vertrauens hingegen zeigte sich dieser Unterschied deutlich mit einem großen Effekt (Hypothese 6), wobei globalen Transparenzarten mehr vertraut wurde als lokalen.

Eine Übersicht über angenommene bzw. abgelehnte Hypothesen stellt sich also wie folgt dar:

H1 bestätigt: Nutzung in Transparenzbedingungen > ohne Transparenz (großer Effekt)
H2 bestätigt: Vertrauen in Transparenzbedingungen > ohne Transparenz (mittlerer Effekt)
H3 bestätigt: Nutzung unterschiedlich zwischen Transparenzbedingungen (kleiner Effekt)
H4 bestätigt: Vertrauen unterschiedlich zwischen Transparenzbedingungen (mittlerer Effekt)
H5 abgelehnt: Nutzung in lokalen Transparenzarten = globalen Transparenzarten
H6 bestätigt: Vertrauen in globalen Transparenzarten > lokalen Transparenzarten (großer Effekt)

Die Ergebnisse der vorliegenden Studie zeigen, dass **Transparenz unabhängig von der genauen Art die Nutzung und das Vertrauen nicht verringert und teilweise sogar steigert**. Bemerkenswert ist, dass die eingesetzten Transparenzarten nicht besonders komplex waren. Teilweise besaßen sie sogar – wie z. B. die Heatmaps – fragwürdige Aussagekraft: Aufgrund des Wizard-of-Oz-Designs wurden alle Erklärungen, und so auch die Heatmaps, künstlich erstellt und spiegelten keine echten KI-Prozesse oder echte Akkuratheitsangaben wider. Dennoch kam ein Effekt der Transparenzarten im Mittel zustande und die Versuchspersonen vertrauten ihnen mehr bzw. nutzen sie mehr als den Algorithmus ohne Transparenz. Dabei sei allerdings auf den Lerneffekt als Treiber für die Zunahme von Nutzung und Vertrauen verwiesen, der diesen Effekt sicherlich mit beeinflusste. Da die Vertrauenszunahme zwischen Nichttransparenz- und Transparenzbedingungen geringer ausfiel als die der Nutzung, lässt sich zumindest hinsichtlich des Vertrauens schließen: Positive Erfahrungen mit Algorithmen sind nicht die einzigen Treiber für Vertrauen.

Darüber hinaus zeigten sich **Unterschiede zwischen den Transparenzarten**, für die aufgrund der Randomisierung ein Lerneffekt ausgeschlossen werden kann. Vertrauen und Nutzung waren am höchsten bei der globalen Erklärung, die Hintergrundinformationen zu Ursprung, Herstellung und allgemeiner Arbeitsweise des Systems lieferte. Da ein Großteil der Versuchspersonen der RWTH zumindest nahesteht, könnte die Erklärung, der Algorithmus sei an der RWTH entwickelt worden, besonders vertrauenserweckend gewirkt haben. Zum anderen waren Nutzung und Vertrauen am niedrigsten für den Heatmap-Algorithmus, also die lokale Erklärung. Eine Ursache hierfür könnte sein, dass die Heatmap-Darstellung im Vergleich der vier Transparenzarten sicherlich die komplexeste war. Sie ist die direkte Umsetzung einer XAI-Methodik und wird in technisch verstandenen Erklärungen häufig genutzt. Die geringe Akzeptanz könnte davon herrühren, dass Heatmaps von den Versuchspersonen, die keine Expert*innen waren, schlicht nicht verstanden wurden, was die entsprechende Beurteilung des wahrgenommenen Verständnisses bestätigt. Dies lässt den Schluss zu, dass **Nutzende nicht blind jedem transparenten System folgen und es nicht vorbehaltlos akzeptieren**, selbst wenn die Relevanz der Aufgabe relativ gering ist.

Diese Effekte ermöglichen zwei Schlussfolgerungen: Erstens ist Transparenz besser als keine Transparenz. Zweitens hängt der Einfluss der Transparenz bei einem Mangel anderer Informationen mutmaßlich an Kleinigkeiten, wie z. B. an einer bekannten, den Nutzenden nahestehenden Institution. Die Relevanz von Hintergrundinformationen und allgemeinen Erklärungen zur Vertrauensbildung bestätigt sich, wie schon in Forschungsfrage (b) „Nutzendenanforderungen“. Hierbei gilt es nicht, durch vertrauensbildende Maßnahmen um jeden Preis Vertrauen zu erzeugen, sondern dies durch inhaltlich korrekte und aussagekräftige Informationen zu regulieren.

Das in Forschungsfrage (b) „Nutzendenanforderungen“ identifizierte Ergebnis, lokale oder globale Erklärungen machten hinsichtlich einer (hypothetischen) Nutzung keinen großen Unterschied, bestätigen die Befunde hingegen nicht. Bisherige Studien setzen teilweise ähnliche Transparenzarten als lokal und global um. Dies führt jedoch zu einem erheblichen Nachteil: Wenn beispielsweise für alle Erklärungen Heatmaps genutzt werden, deren Eignung sich für globale Erklärungen jedoch in Frage stellen lässt, ist das schlechte Abschneiden dieser Bedingung nicht überraschend (Herm et al., 2023). In der vorliegenden Studie zeigte sich mit der Nutzendenpräferenz für globale über lokale Transparenz das Gegenteil. Dies ist jedoch damit zu erklären, dass sie lokale und globale Transparenz mit den unterschiedlichen Bedingungen sehr viel adäquater abbildet. Zusätzlich lässt sich durch die erhobene Variable der Verständlichkeit eine weitere Annahme von Herm et al. bestätigen: „In line with Miller (2019), we also observed that end users prefer straightforward XAI augmentations, which require low cognitive effort such as local *Why* and *Why-Not* explanations, over complex global explanations such

as *How* or *How-To*." (Herm et al., 2023, S. 11) In der vorliegenden Studie scheinen lediglich die lokalen Erklärungen als sehr viel komplexer wahrgenommen worden zu sein als die globalen.

Tatsächlich wurden die **globalen Erklärungen, insbesondere die zur Funktionalität, als verständlicher bewertet** als die lokalen. Während die wahrgenommene Transparenz der Heatmap-Erklärung (lokale Erklärung zur Funktionalität) höher bewertet wurde als die der anderen Transparenzarten, waren die Verständlichkeit und die Verlässlichkeit der globalen Erklärung mit Hintergrundinformationen wie Ursprung und Funktionsweise höher als die der anderen. Besonders auffällig ist der Unterschied in der Bedingung ohne Transparenz: Diese wurde als am wenigsten transparent wahrgenommen. Gleichzeitig wurde sie von allen fünf Bedingungen als die verständlichste bewertet. Mehr Informationen steigern also weder die Verständlichkeit noch die wahrgenommene Verlässlichkeit eines Systems. Auch führt eine hohe Transparenzwahrnehmung nicht automatisch zu höherem Vertrauen oder höherer Nutzung, wie die Wahrnehmung der Transparenzarten und das schlechte Abschneiden der Heatmap-Bedingung zeigen.

Die Ergebnisse erlauben einen weiteren Schluss, der sich bereits in anderen Studien zeigte: Um hilfreich zu sein, müssen Erklärungen zu den vorhandenen mentalen Modellen passen (Andrews et al., 2023). Von den verglichenen Transparenzarten scheint die globale Erklärung zur Funktionalität besser zu mentalen Modellen der Nutzenden gepasst zu haben als das Heatmap-Bild, das ohne weitere Erklärung möglicherweise schwerer zu integrieren war. Eine geringere Verständlichkeit und geringeres Vertrauen waren die Folge.

Wenn es also darum geht, Nutzende möglichst optimal zu unterstützen und zu befähigen, sollte Verständlichkeit der Transparenz angestrebt werden. Auch wenn sich die Versuchspersonen in der vorliegenden Studie nicht von für sie wenig verständlichen Heatmaps überzeugen ließen, besteht die Gefahr vordergründig „guter" Erklärungen darin, Nutzende zu täuschen und zu einer Nutzung anzustiften (Chromik et al., 2019). Solche „Dark Patterns", ein Begriff, der der Interface-Gestaltung entlehnt wurde und bei denen es darum geht, Transparenz vorzugaukeln, um Nutzende zur Nutzung anzuregen, können nur durch gut umgesetzte Transparenz vermieden werden (Chromik et al., 2019).

Zuletzt bestätigen die Ergebnisse, das Vorgehen, das Vertrauen und die Nutzung von Algorithmen getrennt zu erheben. Ihre **Korrelation mittlerer Stärke verdeutlicht ihren Zusammenhang. Allerdings sollte Nutzung nicht als Operationalisierung von Vertrauen gleichgestellt** werden (z. B. Alexander et al., 2018; B. Berger et al., 2021; Daschner & Obermaier, 2022). Der Unterschied zeigt sich auch in den verschieden starken Zusammenhängen von Vertrauen bzw. Nutzung mit den Items zur Wahrnehmung der Algorithmen. Während die wahrgenommene Zuverlässigkeit eines Algorithmus stärker mit seiner Nutzung assoziiert ist, ist der Zusammenhang von Verständlichkeit ebenso wie wahrgenommener

Transparenz mit Vertrauen höher. Vertrauen scheint also besonders mit dem Verständnis für ein System zuzunehmen. Ob ein KI-System genutzt wird, hängt wiederum stärker von seiner Zuverlässigkeit ab, wie sowohl die Abfrage der wahrgenommenen Verlässlichkeit als auch der auf Nutzung besonders deutliche Lerneffekt durch die fehlerfreien Algorithmen zeigte.

Zusammenfassend sind fünf Kernergebnisse der Studie zu nennen:

1. Transparente Algorithmen im Vergleich zu nichttransparenten werden mehr genutzt und ihnen wird mehr vertraut.
2. Die verschiedenen Transparenzarten wirken sich unterschiedlich auf Nutzung und Vertrauen aus. Globale Erklärungen mit Hintergrundinformationen führen zu hohem Vertrauen bzw. zu hoher Nutzung. Lokale Heatmaps führen zu geringerem Vertrauen und geringerer Nutzung.
3. Nutzende scheinen sich nicht durch unverständliche Transparenz überzeugen zu lassen, sondern nutzten selbst im vorliegenden Anwendungsfall mit geringer Relevanz eine wenig erklärte, technische (lokale) Funktionserklärung am wenigsten und vertrauten ihr auch am wenigsten.
4. Über die ambivalenten Ergebnisse anderer Studien hinaus zeigt sich in der vorliegenden Studie: Globale Transparenzarten führen zu mehr Vertrauen als lokale Transparenzarten. Dies scheint in großem Maße über die wahrgenommene Verständlichkeit und Verlässlichkeit der Transparenzarten zu erklären sein.
5. Nutzung und Vertrauen sind separate Konstrukte, die zwar hoch korrelieren, aber von unterschiedlichen Faktoren beeinflusst werden. Vertrauen scheint stärker von der Verständlichkeit der Transparenz abzuhängen, Nutzung dagegen stärker von der wahrgenommenen und tatsächlichen Verlässlichkeit eines Systems.

6.4.2. Limitationen

In Hinsicht auf Limitationen der Studie sind vier Punkte zu nennen. Der erste, methodische Kritikpunkt betrifft den Stichprobenausschluss von 18 Teilnehmenden. Dieser musste aufgrund verändert abgespeicherter Algorithmusschätzungen vorgenommen werden. Diese Werte sollten fest vorgegeben und nicht durch Versuchspersonen beeinflussbar sein. Die nun auftretenden Änderungen sind höchstwahrscheinlich auf eine Teilnahme an der Studie über ein mobiles Gerät zurückzuführen: Da aufgrund technischer Vorgaben in der mobilen Nutzung die Werte nicht fixiert werden konnten, wurde in der Rekrutierung eine Teilnahme über mobile Geräte ausgeschlossen. Es ist davon auszugehen, dass sich die 18 Teilnehmenden nicht daran gehalten haben. Idealerweise sollte in zukünftigen Studien dieser Fehler ausgeräumt und eine mobile Nutzung ermöglicht werden oder zumindest die Teilnahme über mobile Geräte technisch ausgeschlossen werden.

Darüber hinaus existieren eine Vielzahl weiterer Arten, Transparenz in KI herzustellen und zu kategorisieren, als die vier hier getesteten (siehe z. B. Ali et al., 2023; Mohseni et al., 2021; Ras et al., 2022). Für Nutzungsstudien ergeben sich daraus verschiedene Möglichkeiten, wie Transparenz umgesetzt und verglichen werden kann. Wird eine einzige Transparenzart – z. B. Heatmaps – auf verschiedene Weisen umgesetzt, kann diese Transparenzart sehr detailliert verglichen werden (Herm et al., 2023). Ebenso kann eine einzelne Transparenzart, wie z. B. Feature-Importance als Balkendiagramm (Y. Zhang et al., 2020) oder die Einfärbung von Texten (Springer 2019), mit einer intransparenten Bedingung verglichen werden. Die vorliegende Studie hatte zum Ziel, verschiedene Transparenzarten miteinander zu vergleichen, und wählte dafür vier möglichst unterschiedliche Arten aus den Bereichen Funktionalitätserklärung und Akkuratheitsinformation. Zwar büßte sie so möglicherweise an experimenteller Trennschärfe ein und muss sich mit lediglich vier ausgewählten Bedingungen mangelnde Konstruktvalidität für KI-Transparenz vorwerfen lassen. Gleichzeitig erhöht diese Auswahl die praktische Aussagekraft, da vier realistische, häufig genutzte Transparenzarten verglichen wurden. Sie decken nicht nur die einzelnen Transparenzarten ab, sondern repräsentieren Funktionserklärungen und Akkuratheitsangaben sowie lokale und globale Methoden. Nichtsdestoweniger sollten zukünftige Studien sich mit weiteren, anders ausgestalteten und auf andere Weisen vergleichbaren Transparenzarten befassen, um das Bild anzureichern.

Ein weiterer Kritikpunkt betrifft die Leistung der Algorithmen, die immer korrekte Angaben machten. Dies ist nicht realistisch, da jede KI als Modell für die Realität auch Fehler macht, weshalb sich Forschungsfrage (a) dem Umgang im Fehlerfall widmete. Bei einer Akkuratheitsangabe von 90 %, die dem Algorithmus in der globalen Bedingung zugeschrieben wurde, wären bei insgesamt 15 Trials Fehler zu erwarten. Gleichzeitig stand die Reaktion auf algorithmische Fehler nicht im Zentrum dieser Studie. Vielmehr liegt durch die exzellente Leistung der Algorithmen ihr Vorteil auf der Hand und eine möglichst maximale Nutzung wäre die rationalste Entscheidung. Obwohl die Nutzungszahlen angesichts dieser hohen Leistung überraschend gering ausfielen, zeigte sich ein Lerneffekt über den Verlauf der Trials, wobei sich der WOA von 0,3 auf 0,6 verdoppelte. Eine hohe Leistung der KI wirkt also bestärkend auf die Nutzung der Algorithmen. Das ist angesichts des Ziels, eine befähigende Nutzung von KI zu erreichen, eine gute Nachricht. Dieser Befund stimmt mit früheren Studien überein, in denen die tatsächliche Leistung einer KI einen größeren Einfluss auf Nutzung oder Vertrauen hatte als Akkuratheitsangaben oder Erklärungen (Lucic et al., 2020; Önkal et al., 2009; Z. Zhang et al., 2021).

Der größte Kritikpunkt ergibt sich bezüglich der mangelnden Randomisierung der nichttransparenten Bedingung. Der dadurch aufgetretene Lerneffekt beschränkt die Interpretation des Vergleichs der intransparenten mit den transparenten Bedingungen, da der intransparente Algorithmus von der Randomisierung ausgeschlossen und stets zu Beginn durchgeführt wurde. Wenn der Lerneffekt dazu

führte, dass den Algorithmen im Verlaufe des Experiments mehr vertraut wurde bzw. sie mehr genutzt wurden, wirkte dies auf den ersten Algorithmus zwangläufig am geringsten. Wie schnell Versuchspersonen die Nutzung von Ratschlägen ihrer Erfahrung mit einem System anpassen, demonstrierten auch schon Yaniv und Kleinberger in einer Studie mit menschlichen Ratgebern: „respondents updated their weighting policy rapidly, within a few trials“ (2000, S. 272). Tatsächlich zeigte sich auch in der vorliegenden Studie eine Verdopplung der Nutzungswerte von der ersten zur letzten Schätzung und auch das Vertrauen nahm über die Trials zu. Grund für die eingeschränkte Randomisierung war, mit der Kontrollbedingung eine intransparente Bedingung schaffen zu wollen. Indem die Bedingung ohne Transparenz an den Anfang gestellt wurde, konnte eine Vermischung der Transparenzwahrnehmung in dieser Bedingung ausgeschlossen werden. Zukünftige Studien sollen jedoch auch untersuchen, wie sich ein komplett randomisiertes Design auf die Nutzung und das Vertrauen von transparenten im Vergleich zu nicht transparenten Bedingungen auswirkt.

Zusammenfassend ergeben sich also Einschränkungen bezüglich der begrenzten Auswahl an Transparenzarten, der übermäßig positiven Leistung des Algorithmus und des aufgetretenen Lerneffekts, der insbesondere durch die fehlende Randomisierung der intransparenten Bedingung problematisch war. Unter Berücksichtigung dieser Limitationen lassen sich Konsequenzen für zukünftige Forschung und die Implementierung von KI-Transparenz ableiten. Dies erfolgt im nächsten Kapitel.

6.5. Implikationen aus Studie (c) „Transparenzarten“

Vor dem Hintergrund der überwiegend bestätigten Hypothesen sowie der fünf Kernergebnisse, die in Kapitel 6.4.1 diskutiert und dargelegt wurden, ergeben sich verschiedene Implikationen, die im Folgenden vorgestellt werden. Dabei werden zunächst Implikationen für die Praxis dargelegt und im Speziellen für Entwickler*innen für KI formuliert. Anschließend folgt ein Kapitel zu Forschungsfragen, die sich im Anschluss an die Studie für zukünftige Untersuchungen ergeben.

6.5.1. Implikationen für die Praxis

Obwohl es im Sinne einer aufgeklärten Nutzung von KI immer erstrebenswert ist, Transparenz über die Systeme herzustellen, wird deutlich, dass dafür nicht alle Ansätze gleichermaßen sinnvoll sind. Aus der vorliegenden Studie lässt sich ableiten: Transparenz ist besser als keine Transparenz. Hintergrundinformationen, insbesondere solche, die über vertrauensvolle Dritte Vertrauen aufbauen, sind wirksam. Dies geht mutmaßlich darauf zurück, dass sie für Laien leicht(er) zu verstehen sind als eher technische Erklärungen von Heatmaps. Darüber hinaus waren in der vorliegenden Studie globale Transparenzmaßnahmen wirkungsvoller hinsichtlich einer Steigerung des Vertrauens. Dies bedeutet allerdings nicht, dass pauschal globale lokalen Transparenzmaßnahmen vorzuziehen sind. Sehr viel wichtiger als die genaue Transparenzmaßnahme sind die Verständlichkeit der Transparenzart und eine

wahrgenommene, vermittelte Verlässlichkeit des Systems. Wie genau Transparenz realisiert werden kann, ist dabei abhängig vom System. Ob sie verstanden wird, ist abhängig von der Zielgruppe und den mentalen Modellen der Nutzenden. Wichtig ist, dass nicht jede Transparenz automatisch mehr Verständlichkeit hervorruft, sondern eine Transparenz mit geringer Verständlichkeit sogar negativer wirken kann als gar keine Transparenz.

Implikation 7[10] **für Entwickler*innen von KI:**

Transparenz umzusetzen ist wichtig und sinnvoll. Wichtiger als technische Erklärungen oder alle Fakten detailliert darzulegen, ist allerdings ihre Verständlichkeit, die es für die Zielgruppe(n) zu realisieren gilt.

Bezogen auf das negative Abschneiden der Heatmaps ist es eine gute Nachricht, dass sich Nutzende von für sie unverständlicher Transparenz nicht blenden lassen. Fehlendes Verständnis führt zu geringerem Vertrauen und geringerer Nutzung. Um jedoch das Verständnis einer individuellen Erklärung zu prüfen, sind Evaluationen mit Nutzenden unabdingbar. Dabei sollte, konkreter als in der vorliegenden Studie möglich, untersucht werden, inwiefern eine gegebene Transparenz ein realistisches Bild der KI fördert oder ob sie Nutzende irreführt. „Dark Patterns“ (Chromik et al., 2019), also übermäßig vertrauenserweckende Maßnahmen, die zur Täuschung eingesetzt werden, würden vielleicht kurzfristig die Nutzung einer KI erhöhen, doch höchstwahrscheinlich von den Nutzenden aufgedeckt und die bereits bestehende Skepsis an KI weiter anfeuern.

Implikation 8 für Entwickler*innen von KI:

Nutzen Sie Evaluationen mit den Nutzenden Ihrer KI, um sicherzustellen, dass die umgesetzte Transparenz das offenlegt, was sie erklären soll und verstanden wird. Unverständliche Transparenz verringert das Vertrauen oder führt, im Falle von Irreleitung, mittelfristig zur Abkehr vom System.

In Bezug auf die Zusammenhänge von Vertrauen und Nutzung scheint es sinnvoll, diese beiden Konstrukte getrennt voneinander zu betrachten und mit Transparenzmaßnahmen separat anzusteuern. Die Ergebnisse legen nahe, dass Vertrauen besonders durch Verständlichkeit, (gut umgesetzte) Transparenz und vertrauensvolle Hintergrundinformationen hergestellt werden kann. Daraus lässt sich ableiten: Auch wenn Vertrauen im gesamten Nutzungszyklus immer wieder aktualisiert wird, kommt dem Vertrauensaufbau zu Beginn einer Systemnutzung eine besonders wichtige Rolle zu (Chiou & Lee, 2023), insbesondere aufgrund der herrschenden Skepsis gegenüber KI-Systemen (Brauner et al., 2023, 2024). Hingegen entsteht eine hohe Nutzung insbesondere durch die Erfahrung von Verlässlichkeit (Lucic et al., 2020; Y. Zhang et al., 2020). Dabei spielen Transparenzarten

[10] Die Implikationen 1 bis 4 entstammen Forschungsfrage (a) „Fehlerfall“ und finden sich in Kapitel 4.5.1. Die Implikationen 5 und 6 aus Studie (b) „Nutzendenanforderungen“ finden sich in Kapitel 5.5.1.

eine weniger zentrale Rolle. Selbst die Information über Akkuratheit scheint bei weitem nicht so einflussreich, wie die Leistung eines Systems tatsächlich zu erleben. Obwohl Nutzung vom Vertrauen abhängt und mutmaßlich auch umgekehrt, sind sie doch verschiedene Konstrukte, die verschiedener Maßnahmen zur Regulierung bedürfen.

Implikation 9 für Entwickler*innen:

Nutzen Sie unterschiedliche Transparenzarten zu unterschiedlichen Zeiten im Prozess: Zum Aufbau von Vertrauen dienen besonders verständliche, globale Erklärungen, die Hintergrundinformationen bereitstellen. Während der Nutzung vermitteln Akkuratheitsangaben oder lokale Informationen Verlässlichkeit. Am wichtigsten ist allerdings: Das System sollte tatsächlich verlässlich funktionieren.

6.5.2. Anschließende Forschungsfragen

Anknüpfend an die Limitationen der vorliegenden Forschung ergeben sich ganz konkrete Anknüpfungspunkte für zukünftige Forschung: Ein überarbeitetes Forschungsdesign, eine verbesserte technische Umsetzung oder die Randomisierung der Bedingungen sind Optionen, die in einer kommenden Studie bedacht werden sollten. Über diese konkreten Umsetzungspunkte hinaus ergeben sich weitere, über die aktuelle Forschungsfrage hinausreichende Fragestellungen.

Wie in den Implikationen für die Praxis anklingt, ist bei der Umsetzung von Transparenz das individuelle Verständnis zentral. Deshalb besteht weiterhin Forschungsbedarf zur Gestaltung und Umsetzung eines Verständnisses von KI-Systemen und ihrer Transparenz: **Welche Faktoren erhöhen bzw. behindern Verständlichkeit bei der Umsetzung transparenter KI?** Da Vorwissen und Erfahrung sowie die mentalen Modelle der Nutzenden beeinflussen, was als verständlich gilt und was nicht (Andrews et al., 2023; Molina & Sundar, 2022; Studie (b) „Nutzendenanforderungen“), zeigt sich hier der große Bedarf nach weiteren Studien zu verschiedenen Nutzendengruppen. Um KI-Systeme entwickeln zu können, die dem Menschen als Werkzeug bestmögliche Unterstützung bieten, muss bei der Übersetzung der technischen Transparenzmöglichkeiten in den Nutzenden dienliche Transparenz, insbesondere die Verständlichkeit dieser Transparenz, berücksichtigt sein.

Über den Einfluss verschiedener Gestaltungsmerkmale auf Verständlichkeit hinaus müssen weitere abhängige Variablen wie Akzeptanz, Vertrauen oder Nutzung von KI auf ihren Zusammenhang mit der subjektiven Wahrnehmung einer KI-Transparenz untersucht werden. Die wahrgenommene Verlässlichkeit korrelierte hoch mit Verständlichkeit, zeigte aber größeren Einfluss auf die Nutzung. Die wahrgenommene Transparenz wies die höchste Korrelation mit Verständlichkeit auf, zeigte bei einzelnen Transparenzarten aber gegenteilige Effekte, in denen eine hohe wahrgenommene Transparenz nicht mit einer hohen Verständlichkeit einherging. Komplexere Forschungsdesigns, bei denen **mehr und andere Transparenzarten, Interaktionsweisen und verschiedene Variablen der**

Wahrnehmung untersucht und miteinander in Beziehung gesetzt werden, sind notwendig, um ein größeres Verständnis von KI-Transparenz und ihren Effekten auf Nutzung und Vertrauen zu erlangen. Es ist wichtig, genau zu überlegen, welches Ziel eine transparente KI verfolgen soll und welche Faktoren dazu sinnvollerweise zu gestalten sind. So scheint eine künstliche Manipulation der Verständlichkeit wenig zielführend. Vielmehr sollte es darum gehen, Nutzende möglichst optimal zu unterstützen und zu befähigen, weshalb eine möglichst große Verständlichkeit angestrebt wird. Die Gefahr von vordergründig guten Erklärungen, die zur Nutzung anstiften sollen, die aber in Wahrheit kaum verständlich sind, besteht sicherlich weiterhin (Chromik et al., 2019).

Zusätzlich zur Untersuchung weiterer Zusammenhänge und Faktoren könnte es sinnvoll sein, die Ausbildung und Anpassung mentaler Modelle durch Transparenz mit einzubeziehen. Die Frage lautet, **wie transparente KI gestaltet sein sollte, um (möglichst unterschiedlichen) mentalen Modellen zu** entsprechen – oder um sie möglichst realitätsnah anzupassen. Dabei spielt voraussichtlich Verständlichkeit eine wichtige Rolle. Aber auch, wie sich eine subjektive Verlässlichkeit im Verlauf des Nutzungszyklus einer KI entwickelt und wie das mentale Modell kontinuierlich angepasst wird, ist wichtig, um festzustellen, zu welchem Zeitpunkt der Nutzung, welche Arten der Transparenz erforderlich und hilfreich sind. Bei dieser Frage, der **Berücksichtigung des Nutzungszeitpunktes, sollten außerdem die unterschiedlichen Rollen von Vertrauen und Nutzung** betrachtet werden. Wie die Ergebnisse andeuten, hängen sie voneinander ab, werden aber durch verschiedene Transparenzarten und mutmaßlich zu verschiedenen Zeitpunkten im Nutzungsprozess auf unterschiedliche Weise beeinflusst.

Zusammenfassend ergeben sich die folgenden zukünftigen Forschungsfragen:

- Welche Transparenzarten und weitere systemabhängige Faktoren erhöhen bzw. behindern die Verständlichkeit für verschiedene Zielgruppen, Kontexte und Transparenzarten?
- Wie stehen verschiedene Transparenzarten und die Wahrnehmung von KI und KI-Transparenz in Beziehung und wie können sie gestaltet werden, um eine informierte Nutzung zu ermöglichen?
- Wie beeinflusst eine KI-Transparenz die Ausbildung und Anpassung von mentalen Modellen und wie lässt sich diese Anpassung möglichst realitätsnah fördern?
- Wie verändert sich der Bedarf nach Transparenz über den Nutzungsverlauf einer KI zur Anpassung der mentalen Modelle, aber auch mit sich verändernden Auswirkungen auf Vertrauen und Nutzung?

6.6. Zwischenfazit zur Studie (c) „Transparenzarten“

Die vorliegende Studie untersuchte die Fragestellung, wie sich verschiedene Arten von Transparenz auf die Nutzung und das Vertrauen in algorithmische Systeme auswirken. Zentraler Beitrag der Arbeit war es dadurch, die lange Zeit sehr technische Forschungslandschaft zu ergänzen, durch die sozialwissenschaftliche Auseinandersetzung damit, wie verschiedene Umsetzungen von Transparenz auf Endnutzende wirken. Die untersuchte Forschungsfrage lautete:

> **FF (c) „Transparenzarten“: Wie wirken sich verschiedene Arten der KI-Transparenz auf Vertrauen und Nutzung eines Systems aus?**

Die vier verschiedenen Transparenzarten wurden dazu entlang einer 2x2-Matrix gewählt, mit den Dimensionen **Fokus der Erklärung (lokal oder global)** und **Gegenstand der Transparenz (Funktionalität oder Akkuratheit)**. Es ergaben sich eine lokale und eine globale Erklärung der Funktionalität sowie eine lokale und eine globale Akkuratheitsangabe, die mit einer intransparenten Bedingung verglichen wurden.

Ein Großteil der aufgestellten Hypothesen ließ sich bestätigen: Nutzung und Vertrauen waren für die Algorithmen mit Transparenz höher als in der Bedingung ohne Transparenz. Darüber hinaus zeigte sich, dass die verschiedenen Transparenzarten unterschiedlich starken Einfluss auf die Nutzung und das Vertrauen in die Algorithmen hatten. Zusätzlich zeigte sich stärkeres Vertrauen in globale als in lokale Transparenzarten. Die Hypothese zu ihrem Effekt auf Nutzung bestätigte sich hier nicht. Darüber hinaus zeigten in explorativen Analysen verschiedene Wahrnehmungsitems wie Verständlichkeit, Verlässlichkeit und wahrgenommene Transparenz hohe Korrelationen mit Nutzung und Vertrauen und eine unterschiedlich starke Ausprägung je nach Transparenzart. Während eine hohe Verständlichkeit zu besonders hohem Vertrauen zu führen schien, hing die wahrgenommene (und erlebte) Verlässlichkeit der Algorithmen besonders stark mit der Nutzung zusammen.

Die Forschungsfrage kann also insofern beantwortet werden, als **nicht jede (technisch mögliche) Transparenzart Vertrauen und Nutzung erhöht, aber Transparenz besser ist als keine Transparenz**. **Darüber hinaus stellen subjektive Wahrnehmungsfaktoren wie Verständlichkeit und Verlässlichkeit wichtige Vermittler dar** zwischen Transparenz und Vertrauen und Nutzung. In zukünftiger Forschung gilt es, den Einfluss dieser Faktoren auf Nutzung und Vertrauen weiter auszudifferenzieren. Ziel sollte es sein, Verständnis als abhängige Variable oder Zielzustand eines KI-Systems zu stärken und Einflussvariablen zu ermitteln.

Eine Einschränkung der Ergebnisse ergibt sich durch die gewählte Randomisierung der Bedingungen, weshalb der Einfluss eines Lerneffekts auf den Unterschied zwischen nichttransparenter Bedingung

und transparenten Bedingungen nicht ausgeschlossen werden kann. Gleichzeitig macht er deutlich, welchen Einfluss die Erfahrung von Verlässlichkeit auf die Nutzung hat.

Diese Befunde und die möglichen Erklärungen durch die Wahrnehmungsitems leisten einen wichtigen Beitrag für die bestehende Forschung, bei der bisher der **Vergleich verschiedener Transparenzarten**, wenn überhaupt, nur sehr eingeschränkt stattgefunden hat. Mit der Bestärkung der Relevanz von Hintergrundinformationen, also der Offenlegung von Urhebern, einfachen Funktionsweisen und Anwendungsgebieten einer KI, reihen sich die Ergebnisse außerdem in die der Studie (b) „Nutzendenanforderungen“ ein.

Darüber hinaus ergänzen die Ergebnisse die bestehende Forschung aufgrund der **getrennten Erhebung von Vertrauen und Nutzung**. Nicht nur scheinen sich die beiden Akzeptanzvariablen von KI-Systemen auf der Basis von verschiedenen Faktoren zu entwickeln, sie korrelieren auch lediglich mit mittlerer (positiver) Stärke. Ihre getrennte Erhebung und ihre Betrachtung als unterschiedliche Konstrukte sind also anzuraten. Wie sie jeweils individuell gestärkt werden können und zu welchen Zeitpunkten im Nutzungskontext ihnen besondere Bedeutsamkeit zukommt, sollte Ziel weiterer Forschung sein.

Praktische Implikationen ergeben sich vor dem Hintergrund, nicht die technischen Möglichkeiten, sondern die sinnvolle Umsetzung von Transparenz in den Fokus zu setzen. Betrachtet man Transparenz als zu gestaltende Eigenschaft des Systems bei der Entwicklung und Einführung von KI-Systemen, gilt es zunächst, den Faktor Verständlichkeit und seine Nutzungs- und Kontextabhängigkeit in den Fokus zu stellen. Gute Transparenz bedarf des Einbezugs der Zielgruppe(n), z. B. in Nutzungsevaluationen. Darüber hinaus gilt es, Transparenz nicht als statischen Zustand einer KI zu betrachten, sondern je nach Nutzungszeitpunkt – z. B. am Anfang oder nach einem Fehlerfall – die für diesen Moment relevanten Informationen über das System zu liefern.

Um KI-Systeme zu entwickeln, die Nutzenden zu einer informierten und bestmöglichen Entscheidung verhelfen, bedarf es transparenter Systeme. Wie die vorliegende Studie zeigte, gibt es verschiedene Arten, Transparenz zu gestalten, die richtig eingesetzt auf verschiedene Weisen die Nutzung und das Vertrauen in KI steigern können. Gleichzeitig stellt Transparenz in KI kein Selbstzweck dar, sondern dient dem Ziel, Verständlichkeit herzustellen und die menschliche Entscheidungsautonomie zu ermöglichen. Nur so kann selbstbestimmtes wie auch einem Ziel dienliches Entscheiden mithilfe von KI stattfinden.

7. Einordnung der Ergebnisse und Diskussion

Mit der zunehmenden Forderung nach Transparenz in KI-Systemen und dem bisherigen technischen Übergewicht in ihrer Beforschung stellte sich für die vorliegende Arbeit die Frage: **Wie wirkt sich Transparenz von KI-Entscheidungsunterstützungssystemen auf die Nutzung dieser Systeme durch Endnutzende aus?** Da die Annahme lautete, dass sich für diese große Frage nicht die eine Antwort finden, sondern sie sich vielmehr aus verschiedenen Teilen zusammensetzen würde, fand die Untersuchung der Fragestellung mithilfe eines multimethodischen Ansatzes statt. Dieser zeichnete sich durch drei Studien aus, mit dem Ziel, das Thema durch qualitative und quantitative Designs sowie die Wahl verschiedener Foki zu beleuchten. Die drei Studien verfolgten die folgenden Ansätze: (a) eine quantitative Studie, die ein spezielles Phänomen der Algorithmennutzung und den Effekt von Akkuratheitsinformation im KI-**Fehlerfall** untersuchte, (b) eine qualitative Studie, um mithilfe eines breiten Ansatzes ein Verständnis für **Nutzendenanforderungen** an KI-Transparenz zu erhalten und (c) eine quantitative Studie, die verschiedene **Transparenzarten** hinsichtlich Nutzung und Vertrauen verglich.

Nachdem in den vorherigen Kapiteln zunächst der theoretische Hintergrund erläutert und anschließend drei verschiedene Studien vorgestellt und diskutiert wurden, erfolgt in den nächsten Kapiteln nun ein Überblick über die Studien (7.1) sowie die Darlegung der aus den Studien gemeinsam zu gewinnenden Erkenntnisse (7.2). Anschließend diskutiert Kapitel 7.3 die für die gesamte Arbeit geltenden Limitationen. Kapitel 7.4 zeigt mögliche, aus der Gesamtschau der Arbeit zu ziehende Implikationen für die Gestaltung und Umsetzung von transparenter KI auf.

7.1. Überblick über die Studien

In der ersten Studie stand Algorithm Aversion im Zentrum. Dieser Effekt bezeichnet den Nutzungseinbruch, nachdem Nutzende Fehler von Algorithmen erleben (Dietvorst et al., 2014). Einer Annahme zufolge geht dieser Effekt auf die falsche Erwartung an KI zurück, stets perfekt zu arbeiten. Ein Fehler wird als Zeichen für ein defektes System interpretiert, das folglich zu meiden sei (Prahl & Swol, 2017; Reich et al., 2022). Um zu kommunizieren, jede KI macht auch Fehler, wurden die Algorithmusergebnisse mit Akkuratheitsangaben dargestellt. Die zu klärende Forschungsfrage (a) „Fehlerfall“ lautete: **Inwieweit führen Angaben von Akkuratheit eines Algorithmus dazu, dass dieser auch nach einem Fehlerfall genutzt wird?** Sie wurde mit einer Online-Studie untersucht (siehe Abbildung 32 bzw. Kapitel 4). Tatsächlich zeigte sich, wenn Akkuratheitsinformation gegeben war, fiel die Nutzung des Algorithmus nach einem Fehler etwas höher aus als ohne entsprechende Information. Eine Interaktion zum Vergleich eines Algorithmus ohne Akkuratheitsangabe mit zweien mit dieser Information sowie der Algorithmusnutzung vor und nach Fehlererleben zeigte jedoch keinen

signifikanten Effekt. Eine hemmende Wirkung der Akkuratheitsinformation auf Algorithm Aversion konnte also über ihren Einfluss auf die Nutzung nach einem Fehler nicht nachgewiesen werden.

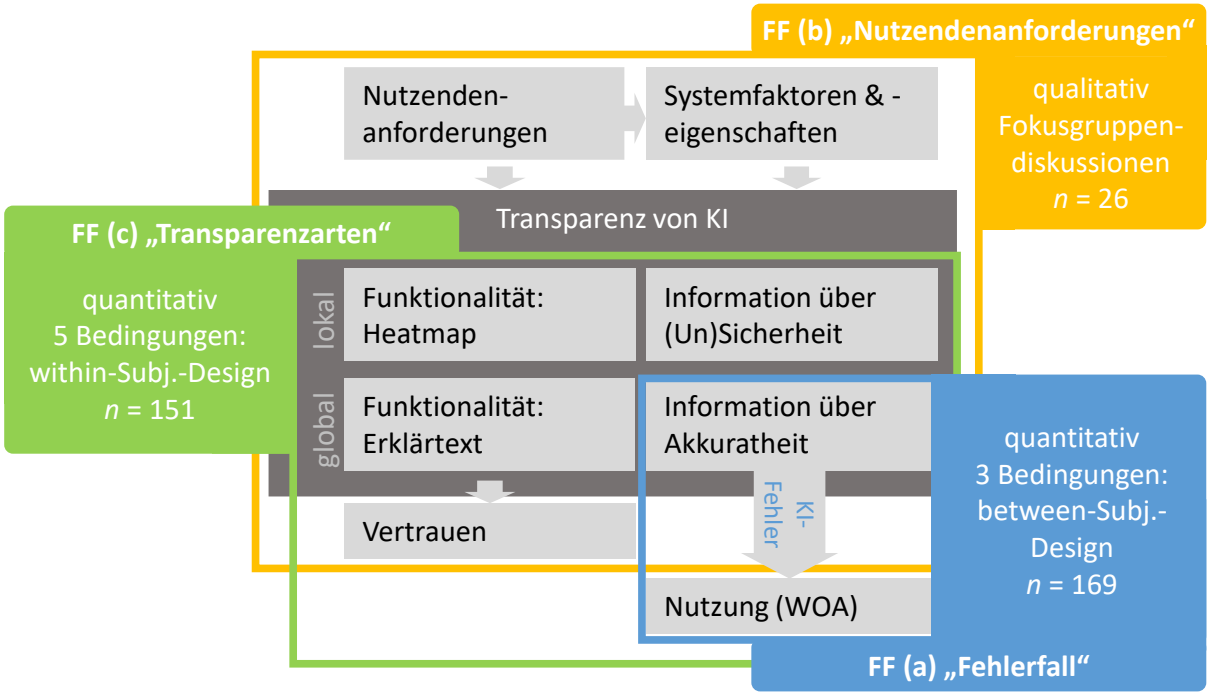
Mit der Forderung nach Algorithm Literacy und vor dem Hintergrund, dass Transparenz mehr umfasst als Akkuratheitsinformationen, folgte die Auseinandersetzung mit der Frage, was denn Laien unter Transparenz von KI verstehen. Da die Definitionen sehr uneinheitlich sind, sehr viele Aspekte umfassen und vor allem lange Zeit sehr technisch ausfielen – Transparenz als XAI – stellte sich die Frage (b) nach „Nutzendenanforderungen“: **Welche Anforderungen an Transparenz in KI bestehen für Laiennutzende und inwiefern unterscheiden sie sich nach Eigenschaften der KI?** Diese Fragestellung wurde mit einem qualitativen Ansatz untersucht und dazu eine Fokusgruppenstudie durchgeführt (siehe Kapitel 5). Im Hauptteil der Studie diskutierten die *n* = 26 Teilnehmenden drei fiktive KI-Apps unter anderem zur Frage, unter welchen Voraussetzungen sie die Anwendungen nutzen würden (siehe Abbildung 32).

Als Ergebnis zeigte sich ein sehr breites Verständnis von KI-Transparenz, das weit über das technische, eher funktionsbezogene hinausgeht. Laien fordern insbesondere im Fehlerfall lokale Transparenz: Warum kam dieses genaue Ergebnis heraus bzw. dieser Fehler zustande? Globale Transparenz zur Funktionsweise stand weniger deutlich im Fokus. Hingegen spielten Informationen zu Urhebern, Bewertungen anderer, Datenquellen oder vertrauenswürdigen Zertifikaten ebenso wie Datenschutz und Kontrolle eine wichtige Rolle beim Aufbau von Vertrauen in KI. Auf Seiten der Nutzenden zeigten die Vorerfahrung und Einstellung nicht nur zu KI oder ähnlichen Systemen, sondern auch zur Anwendungsdomäne großen Einfluss auf die Akzeptanz eines Systems. Bezüglich der Frage, welche Systemeigenschaften die Transparenzanforderung beeinflussen, stellte sich insbesondere Fehlerrelevanz als wichtig heraus, die in starkem Maße von der Anwendungsdomäne abhängt. Um diese Ergebnisse für die Praxis anwendbar darzustellen, wurde eine **Transparenzmatrix** erstellt (siehe Kapitel 5.5.1). Mit dieser lassen sich Transparenzempfehlungen ableiten in Abhängigkeit von (durch die Nutzenden wahrgenommenen) Systemeigenschaften.

Da scheinbar verschiedene Transparenzaspekte für Laien unterschiedlich wichtig wahrgenommen werden und gleichzeitig nur wenige Vergleiche verschiedener Transparenzarten bestehen, widmete sich die dritte Forschungsfrage vier verschiedenen Transparenzarten, um sie miteinander und mit einer Bedingung ohne Transparenz zu vergleichen. Die Forschungsfrage (c) „Transparenzarten“ lautet entsprechend: **Wie wirken sich verschiedene Arten der KI-Transparenz auf Vertrauen und Nutzung der jeweiligen Systeme aus?** (siehe Abbildung 32) In einem quantitativen Vergleich mit *n* = 151 zeigte sich ein Unterschied zwischen einem nicht transparenten Algorithmus und vier transparenten Algorithmen (siehe Kapitel 6). Letzteren wurden mehr genutzt und ihnen mehr Vertrauen entgegengebracht. Jedoch kann mindestens ein Teil des Unterschieds auf einen Lerneffekt

zurückzuführen sein, da die Nutzung und das Vertrauen in die Algorithmen über den Verlauf des Experiments zunahmen. Im Vergleich der Transparenzarten untereinander führten besonders globale Transparenzbedingungen zu höherem Vertrauen. Die Transparenzbedingung, in der global die Funktionalität des Algorithmus sowie Informationen zu seiner Erstellung und Urhebern dargelegt wurden, wurde als besonders verständlich bewertet, ihr auch am meisten vertraut und sie auch am häufigsten genutzt. Gleichzeitig wurde zwar die wahrgenommene Transparenz der lokalen Funktionalitätserklärung durch Heatmaps am höchsten, ihre Verständlichkeit allerdings am geringsten bewertet, sogar geringer als die des nichttransparenten Algorithmus. Auch ihre Nutzung und das Vertrauen in sie fielen gering aus.

Abbildung 32: *Überblick über die untersuchten Forschungsfragen auf dem Feld der Transparenz von KI und die eingesetzte Methodik*



Anmerkung. Subj.-Design = Subjekt-Design. WOA = Weight of Advice

7.2. Gemeinsame Erkenntnisse

Zusammengenommen können verschiedene Erkenntnisse aus den drei Studien gewonnen werden (siehe Abbildung 33 für einen Überblick). Es zeigt sich, KI-Transparenz beeinflusst sowohl die Nutzung als auch das Vertrauen in die Systeme. Dabei scheint **globale KI-Transparenz vor allem für das Entstehen von Vertrauen bzw. beim Beginn des Interaktionsprozesses** wichtig. Insbesondere globale Informationen zu Funktionsweisen und weitere Hintergrundinformationen fördern die Vertrauensbildung, wie Forschungsfrage (b) „Nutzendenanforderungen“ und (c) „Transparenzarten“ ermittelten. Im **Fehlerfall jedoch werden Erklärungen zur Ursache des Fehlers oder zu zukünftigen**

Systemverbesserungen relevant, also eher lokale Erklärungen. Unterschiedliche Transparenz scheint also nötig je nach Nutzungszeitpunkt: Zum Vertrauensaufbau am Anfang bedarf es eher globaler Informationen. Während der Nutzung, insbesondere bei Unsicherheit oder im Fehlerfall, helfen lokale Erklärungen, das System neu zu bewerten.

Wie die Ergebnisse zeigen, müssen Erklärungen zu vorhandenen mentalen Modellen passen, um hilfreich zu sein. In Forschungsfrage (c) „Transparenzarten" traf diese Passung möglicherweise auf die globale Erklärung eher zu als auf abstrakte Heatmap-Bilder. Und auch wenn sich Forschungsfrage (a) nicht explizit mit mentalen Modellen befasste, deuten die Ergebnisse doch darauf hin, dass **Akkuratheitsangaben allein nicht ausreichen, um mentale Modelle** über die erwartete Performance von Algorithmen zu aktualisieren und im Fehlerfall vor ablehnender Überraschung zu schützen. Hingegen kam die **Frage, warum ist dieser Fehler passiert, sowie, lernt das System aus ihm**, auch in der qualitativen Studie (b) „Nutzendenanforderungen" zur Sprache, in der die Diskussionsteilnehmenden unter anderem ihre Reaktionen auf einen möglichen Fehlerfall besprachen. Frühere Studien zeigen, die Information, eine KI lerne aus Fehlern, trägt dazu bei, die Fehlererfahrung mit einer KI zu bewältigen (B. Berger et al., 2021; Burton et al., 2020). Gleichzeitig steigerten Erklärungen, warum ein Fehler auftrat, zwar das Verständnis für ein System, nicht aber seine Nutzung (Lucic et al., 2020). In ähnlicher Weise scheinen auch Akkuratheitsangaben allein nicht für den Erhalt der Nutzung nach einem Fehler auszureichen.

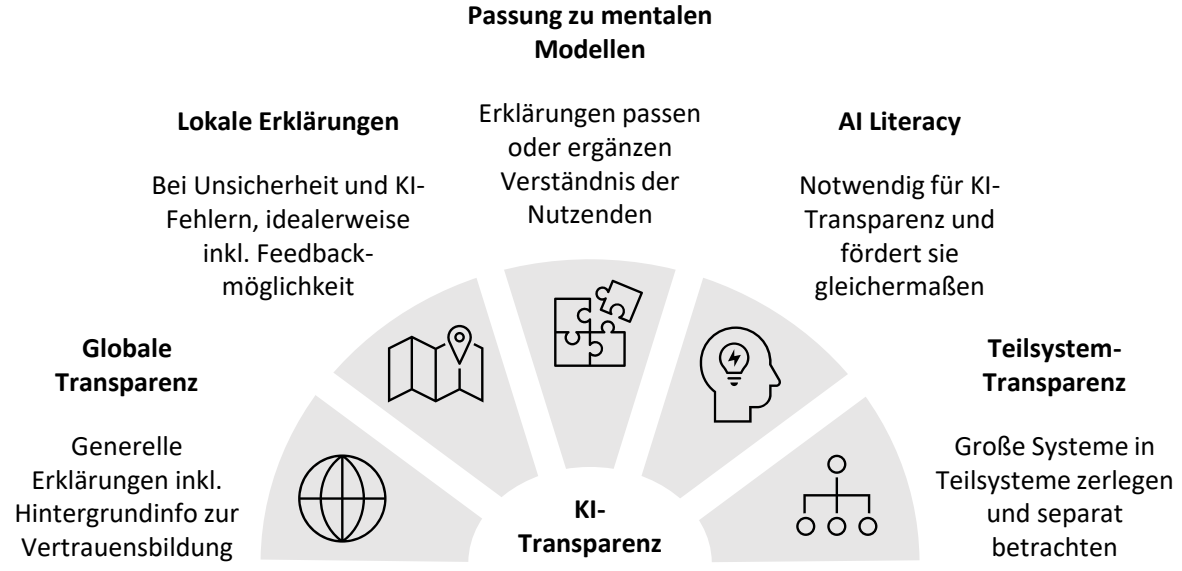
Als Maßnahme gegen Algorithm Aversion nennen Burton et al. in einem Review „Algorithm Literacy" (2020). Genau diese ist jedoch angewiesen auf KI-Transparenz: **Nutzende bedürfen der nötigen Informationen über Anwendungsbereich, Einschränkungen, Ergebnisgültigkeit und Akkuratheit, um eine „Literacy" über ein System aufbauen zu können. Wichtig bei alledem ist Verständlichkeit,** deren Relevanz in Forschungsfrage (c) „Transparenzarten" festgestellt wurde. Denn auch für Literacy ist es von zentraler Bedeutung, dass die vorhandenen Informationen über ein System verstanden, in bestehende mentale Modelle eingebettet und mithilfe neuer Information aktualisiert werden können. Die Erkenntnis lautet: **Wichtiger als eine technisch detaillierte Erklärung ist eine Transparenz, die Nutzende auch verstehen.** Als solche umfasst sie Akkuratheitsinformation und Hintergrundinformationen zu Urhebern, aber auch Informationen über Datenspeicherung, Privatsphäre und Sicherheit der Daten ebenso wie Verantwortlichkeiten bei Fehlern, Kontrollmöglichkeiten durch die Nutzenden und Bewertungen durch Dritte oder Zertifikate von Expert*innen.

Die Ergebnisse zeigen eine weitere mögliche Erklärung für bisherige uneinheitliche Befunde der Forschung: **Komplexe KI-Systeme sind als Teilsysteme zu betrachten**. Ein System besteht beispielsweise aus Eingabemaske, Verarbeitungsprozessen und einer Ausgabe oder sogar einem

mehrstufigen Ausgabeprozess. Die Anforderungen an Transparenz unterscheiden sich je nach Teilsystem: Bei der Eingabe spielt möglicherweise Datenschutz eine Rolle, bei der Verarbeitung die Gewichtung der Eingaben, bei der Ausgabe Akkuratheitsangaben oder Erklärungen für das einzelne Ergebnis. Darüber hinaus sind jede KI und auch jedes Teilsystem eingebettet in eine Anwendungsdomäne, in ein Umfeld ähnlicher Systeme und in bestehendes (Halb-)Wissen über KI, die wiederum Quelle für Vorerfahrungen, Einstellungen und individuelle Bewertungen darstellen. Diese zentralen Erkenntnisse zu transparenter KI aus der Gesamtschau der Arbeit stellt Abbildung 33 zusammenfassend dar.

Darüber hinaus zeigten in Forschungsfrage (a) „Fehlerfall“ und (c) „Transparenzarten“ **demographische oder interindividuelle Eigenschaften kaum Zusammenhang** mit dem Bedarf an Erklärungen oder einen Einfluss auf Vertrauen bzw. Skepsis gegenüber KI-Systemen, so z. B. Risikobereitschaft, Entscheidungsfreude oder Kontrollüberzeugung im Umgang mit Technik. Andere Studien fanden den Einfluss des kulturellen Hintergrunds und der Sprache (McNee et al., 2006), des dispositionalen Vertrauens (Brauner et al., 2023) oder von Erfahrung und Selbstvertrauen (Chong et al., 2022), jedoch keinen oder nur geringen Zusammenhang mit demographischen Variablen (Philipsen et al., 2022; Yin et al., 2019). In der vorliegenden Arbeit zeigte sich der Effekt des Kognitionsbedürfnisses als Einflussfaktor zur Weiternutzung von KI nach einem Fehlerfall. Dies brachte Überlegungen zur **Relevanz verschiedener kognitiver Verarbeitungsstile** bei der Perzeption von KI und ihrer Transparenz auf, die beeinflussen, welche Arten von Information gut oder weniger gut verarbeitet werden können. Diese Ansätze zusammen mit den Befunden zu verschiedenen Nutzungszeitpunkten und ihren Auswirkungen auf Transparenzbedarfe deuten weitere Einflüsse auf KI-Transparenz an und damit auch auf die Möglichkeit, sie zu gestalten.

Abbildung 33: *Veranschaulichung der gemeinsamen Erkenntnisse aus den drei Studien der Arbeit*



7.3. Limitationen der vorliegenden Arbeit

Über die Kritikpunkte hinaus, die sich aus den drei Studien im Einzelnen ergeben und die im Rahmen der jeweiligen Studie diskutiert wurden, sind auch Limitationen bezüglich der gesamten Arbeit aufzuführen. Methodisch ist über die drei Studien hinweg anzumerken: Die Art und der Prozess der Entscheidungsfindung sind relevant und beeinflussen in starkem Maße, wie Ratschläge für diese Entscheidung genutzt werden. Ergebnisse verschiedener Entscheidungsfindungsprozesse sind nur eingeschränkt vergleichbar, da die Prozesse selbst die finale Entscheidung beeinflussen (können) (Bonaccio & Dalal, 2006). Konkret stellt sich deshalb in Bezug auf die beiden quantitativ untersuchten Forschungsfragen (a) „Fehlerfall“ und (c) „Transparenzarten“ die Frage, inwiefern das Experimentaldesign mit einem Judge-Advisor-System zu Artefakten in der Nutzung der Algorithmen führte. Die beiden vorliegenden Studien zeichneten sich dadurch aus, dass die Versuchspersonen zunächst eine eigene Schätzung abgeben mussten, bevor sie die Algorithmusschätzung erhielten. Dieser Interaktionsablauf könnte Ankereffekte begünstigen, wie in ähnlichen Interaktionen bereits gezeigt wurde (Bonaccio & Dalal, 2010; Hütter & Fiedler, 2019; Tversky & Kahneman, 1974). Besonders in kritischen Kontexten wie der Medizin oder der Rechtsprechung wird die Interaktion mit KI tatsächlich so umgesetzt, wie in den vorliegenden Experimenten: zunächst eine unabhängige menschliche Entscheidung der Nutzer*in – z. B. eines Arztes oder einer Richterin – und dann der Hinweis der KI. Das Ziel ist, selbstständiges, unabhängiges Entscheiden auf Seiten der Nutzenden zu gewährleisten und zu vermeiden, dass nur blind einer KI gefolgt wird, ohne eigene Einschätzungen zu berücksichtigen. Im Kern geht es darum, die menschliche Entscheidungsautonomie nicht einzuschränken. Allerdings kommt dieser Interaktionsablauf in der Realität nur selten vor. Die meisten Systeme präsentieren ihre eigenen Lösungen oder Hinweise, ohne Voreinschätzungen zu verlangen. Vor diesem Hintergrund und dem dargelegten Einfluss eines (quantitativen) Ankers stellt sich deshalb die Frage, wie die KI-Interaktion in anderen Forschungsdesigns ausfallen würde. Es ist unklar, inwieweit die Ergebnisse aus diesen Designs vergleichbar und in Beziehung zu setzen sind mit denen aus einer sehr offenen, qualitativen Befragung. Es gilt, in zukünftiger Forschung die Designs zu variieren und die Fragestellungen mit weiteren Methoden ergänzend zu untersuchen.

Dieser Kritikpunkt kann erweitert werden auf das gesamte Mixed Method-Design der vorliegenden Arbeit: Durch das parallele Vorgehen und eine nicht aufeinander aufbauende Methodik wurden einzelne Fragestellungen nur einmalig betrachtet und nicht vertieft. Limitationen der einzelnen Studien bleiben bestehen und Ergebnisse können nur jeweils vor diesem Hintergrund interpretiert werden. Für ein nun anschließendes, tiefergehendes Verständnis wäre es sinnvoll, einzelne Fragestellungen im Detail, z. B. mit verschiedenen Studien, die sich einer einzelnen Fragestellung in triangulierendem Ansatz widmen, zu untersuchen. Gleichzeitig bot das gewählte Vorgehen die

Möglichkeit, die Effekte von KI-Transparenz auf verschiedene Weisen zu beleuchten und so eine breite Auseinandersetzung mit dem Thema – mit ihren jeweiligen Einschränkungen – zu ermöglichen. Für den ersten Überblick, der zu diesem Thema vonnöten war, bot das gewählte Vorgehen einen breiten Ansatz und entsprechend breite Ergebnisse, die es nun in nachfolgenden Forschungsvorhaben zu erweitern gilt.

7.4. Implikationen aus der gesamten Arbeit

Wie in den Kapiteln zu den Implikationen der einzelnen Forschungsfragen erfolgt auch in diesem die Aufteilung in praktische Implikationen und in theoretische. Das bedeutet, im folgenden Kapitel werden zunächst die Implikationen für Entwickler*innen und Politik abgeleitet. Zu diesen gehört in Kapitel 7.4.1.2 auch eine Aktualisierung der Transparenzmatrix, die als Einzeldokument ein zentrales Ergebnis der Arbeit darstellt. Im Anschluss werden in Kapitel 7.4.2 die im Laufe der drei Studien formulierten Implikationen zusammengeführt mit den nun neu abgeleiteten Implikationen. Dabei werden die Implikationen für Entwickler*innen sowie die für Politik und Regulierungsinstitutionen in jeweils separaten Kapiteln präsentiert. Anschließend an die praktischen Implikationen folgt eine Auseinandersetzung mit den noch offenen Fragestellungen, die sich aus der Arbeit für die zukünftige Forschung ergeben (Kapitel 7.4.3).

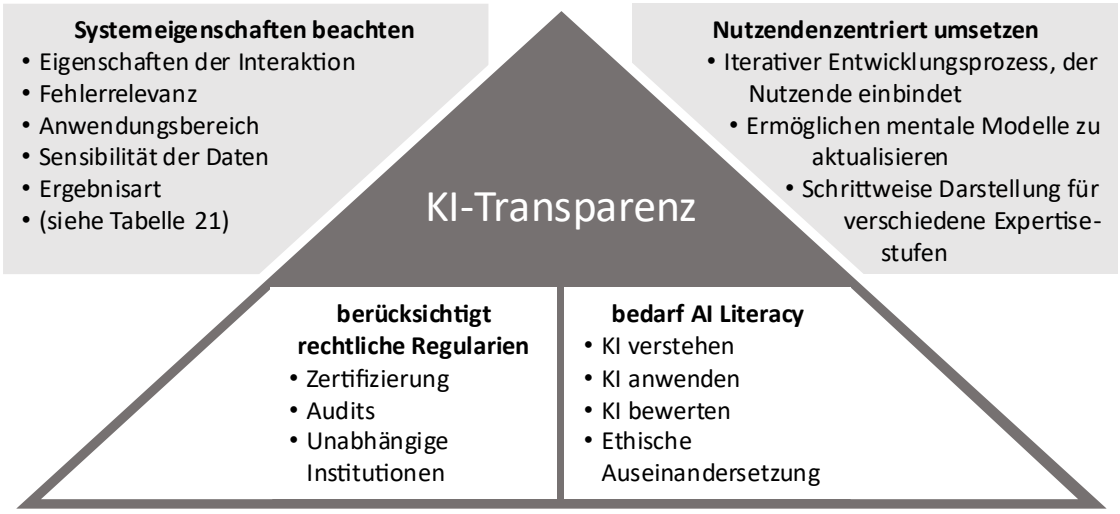
7.4.1. Praktische Implikationen

Über diese Implikationen hinaus lassen sich aus der vorliegenden Arbeit weitere Aspekte ableiten, die sowohl für Entwickler*innen von KI als auch für die Politik zur Umsetzung von KI-Transparenz relevant sind. Laut einer Umfrage der Bertelsmann-Stiftung verlangen Nutzende eine starke Kontrolle von KI. Dabei befürworten 80 % der Befragten ein Auskunftsrecht bei KI-Entscheidungen, über 70 % die Zugänglichkeit für unabhängige Experten sowie einen KI-TÜV (Fischer & Petersen, 2018). Die Ergebnisse sprechen für die große Skepsis, die gegenüber Algorithmen herrscht und deuten gleichzeitig das Bedürfnis nach Transparenz an. Es gilt also, Vertrauen aufzubauen, um so eine sinnvolle und informierte Nutzung zu ermöglichen. Bedué und Fritsche (2022) identifizieren fünf Faktoren, die speziell das Vertrauen in KI im Vergleich zu traditionelleren Technologien fördern: „We identify access to knowledge, transparency, explainability, certifications, as well as self-imposed standards and guidelines as the main determinants of trust in AI.“ (S. 2).

Die in der vorliegenden Arbeit identifizierten Maßnahmen gehen einher mit diesen Vorschlägen. Im Kern steht dabei die Anforderung, KI als dem Menschen dienliches Werkzeug zu gestalten und durch Transparenz den bestmöglichen Einsatz dieses Werkzeuges zu ermöglichen. Dazu müssen Transparenz und Explainability systemabhängig gestaltet (Kapitel 7.4.1.1) und Nutzende ins Zentrum der Transparenzentwicklung gestellt werden (Kapitel 7.4.1.2). Außerdem spielt der Zugang zu Wissen eine wichtige Rolle (Kapitel 7.4.1.3 über AI Literacy) ebenso wie Maßnahmen des Vertrauenstransfers über

rechtliche Regularien (Kapitel 7.4.1.4). Das in Abbildung 34 dargestellte Dreieck der KI-Transparenz veranschaulicht die vier Säulen der Umsetzung von KI-Transparenz. Während System- und Nutzendeneigenschaften maßgeblich beeinflussen, wie KI-Transparenz gestaltet werden sollte, fußt deren Umsetzung auf rechtlichen Regularien ebenso wie auf einer vorherrschenden und durch KI-Transparenz voranzutreibenden AI Literacy.

Abbildung 34: *Das Dreieck der KI-Transparenz stellt die vier Faktoren, Systemeigenschaften, Nutzendenzentrierung, rechtliche Regularien und AI Literacy dar, die bei der Umsetzung von KI-Transparenz für Endnutzende zum Tragen kommen.*



Im Anschluss an die folgenden Kapitel der praktischen Implikationen wird jeweils eine zusammenfassende Tabelle bereitgestellt (je Tabelle 22, Tabelle 23, Tabelle 24 und Tabelle 25), welche die aus der Gesamtarbeit resultierenden Implikationen für Entwickler*innen sowie die Politik zusammenfasst.

7.4.1.1. KI-Transparenz abhängig von Systemeigenschaften

Die erste Implikation, die sich aus der vorliegenden Arbeit ergibt, ist die Abhängigkeit der KI-Transparenz von den gegebenen Systemeigenschaften und die daraus resultierende nötige Auseinandersetzung mit Letzteren. Dabei steht weniger die technische Frage nach der Umsetzung der KI im Zentrum als vielmehr die Frage nach den gegebenen Voraussetzungen und von Nutzenden wahrgenommenen Systemeigenschaften. Bereits die Information, es mit künstlicher Intelligenz zu tun zu haben, verändert die Wahrnehmung eines Systems (Bedué & Fritzsche, 2022).

Bevor also KI-Transparenz umgesetzt werden kann, gilt es, die gegebenen Systemeigenschaften zu analysieren, zusammen mit zukünftigen Nutzenden. Im Rahmen von Forschungsfrage (b) „Nutzendenfaktoren“ wurde zur Ableitung von Transparenzmaßnahmen von gegebenen Systemeigenschaften eine **Transparenzmatrix** entwickelt und nun aktualisiert (siehe Tabelle 21). Die Anwendung der Matrix wurde in Kapitel 5.5.1 „Transparenzmatrix für die Praxis“ ausführlich erläutert. Neben Erklärungen zur Matrix findet sich dort auch ein Beispiel für ihren Einsatz. Außerdem wird ein

vierstufiger Prozess eingeführt (Abbildung 20, Kapitel 5.5.1), der beim Einsatz der Matrix zu empfehlen ist:

1. Ggf. Zerlegen des KI-Systems in einzelne **Bestandteile**
2. **Systemfaktoren** anhand der Transparenzmatrix identifizieren
3. **Anforderungen an KI-Transparenz** abhängig von 2. ablesen und umsetzen
4. **Evaluation** der KI-Transparenz mit Nutzenden des Systems

Für die Zusammenführung der Gesamtarbeit wurde die bestehende Matrix wurde um die Erkenntnisse aus den beiden Studien (a) „Fehlerfall“ und (c) „Transparenzarten“ ergänzt (siehe Tabelle 21). Dazu wurde die bestehende Matrix um einige Punkte erweitert. Als Folge der Studie (a) „Fehlerfall“ wurden die System- und Interaktionseigenschaften um die Kategorie **„Unsicherheit/Fehlererfahrung“** ergänzt. Zur Erklärung dieser Kategorie ist zu sagen: KI-Systeme unterscheiden sich hinsichtlich der Möglichkeit, Fehlererfahrung überhaupt zu machen. In erster Linie hängt dies davon ab, ob das System eine nachträgliche Korrektur oder Bewertung der vorangegangenen Empfehlung anzeigt oder für welche Art von Empfehlungen es konzipiert wurde. Bei einem KI-System, das zur Personalauswahl eingesetzt wird, tritt eine Fehlererfahrung nur verzögert auf und wird eher subjektiv feststellbar sein: Der Systemvorschlag eines Bewerbers war möglicherweise nachvollziehbar, aber die Einstellung kam aufgrund einzelner Eigenschaften letztendlich nicht zustande. Dieser am Ende falsche Vorschlag des Systems wird erst auffallen, wenn er besonders häufig geschieht. Eine Erklärung ist dann weniger nützlich als die Möglichkeit, das System zu korrigieren. Die Kategorie der Matrix hingegen bezieht sich auf eine direktere Art der Fehlererfahrung: Wenn eine KI bei einer Buchauswahl, einem Sportwetteneinsatz oder zu passenden Schuhen berät, ist das Ergebnis einige Zeit später als gut oder schlecht zu bewerten und es ist unmittelbar auf die KI zurückführbar. Es geht also um richtige oder falsche Empfehlungen. In diesen Fällen sind Erklärungen, warum dieser Fehler auftritt, hilfreich, um das Vertrauen in das System zu erhalten. Entsprechend ist in der neuen Version der Transparenzmatrix der Bedarf nach Erklärungen allgemein und lokaler Transparenz im Speziellen in der Kategorie „Unsicherheit/Fehlererfahrung“ als „sehr hoch“ dargestellt.

Darüber hinaus ergibt sich auf Grundlage der Ergebnisse aus Forschungsfrage (c) „Transparenzarten“ die Relevanz von globalen Erklärungen und von Hintergrundinformationen. Darauf aufbauend werden einige Zellen präzisiert: in der Systemeigenschaft „**Fremdes System**“ wird die Relevanz von globaler Transparenz als „hoch, nach Interesse“ angegeben, die von weiteren Hintergrundinformationen, also den Informationen zum Urheber sowie den Sicherheiten über Dritte von „hoch“ auf „sehr hoch“ verschärft, ebenso wie die Transparenzinformation zur Akkuratheit.

Die aktualisierte Transparenzmatrix ist auf der folgenden Seite als Tabelle 21 dargestellt. Tabelle 22 fasst anschließend die in diesem Kapitel entwickelten Implikationen für die beiden Zielgruppen Entwickler*innen und Designer*innen von KI sowie Politik und Regulierungsinstitutionen zusammen.

Tabelle 21: *Transparenzmatrix: Implikationen für die Transparenz eines KI-Systems in Abhängigkeit von subjektiven System- und Interaktionsfaktoren. Die dunkle Kategorie „Unsicherheit, Fehlererfahrung“ wurde aufgrund der Gesamtergebnisse neu hinzugefügt, die fett markierten Zellen präzisiert.*

Implikationen für KI-Transparenz \ subjektive System- und Interaktionseigenschaften	Hohe Fehler-relevanz	Dient der Unter-haltung	Ethische Thematik	Sensible Daten	Negative Vorerfah-rung/ -einstellung	Unsicher-heit, Fehler-erfahrung	Fremdes System/ keine Vor-erfahrung	Positive Vor-erfahrung mit System	Mehrere Ergebnisse möglich	Nutzung kostet Geld
Bedarf nach Kontrollgefühl	sehr hoch	hoch	sehr hoch	hoch	sehr hoch		hoch			sehr hoch
Bedarf nach Erklärungen	hoch	nach Interesse	sehr hoch		hoch, nach Interesse	**sehr hoch**	hoch	weniger relevant		hoch
lokale Transparenz	**hoch, nach Interesse**	weniger relevant	sehr hoch			**sehr hoch**			hoch	
globale Transparenz	**hoch, nach Interesse**	weniger relevant	hoch		nach Interesse	**hoch, nach Interesse**	**hoch, nach Interesse**	weniger relevant		nach Interesse
Weiterführende Informationen zum Ergebnis	hoch		sehr hoch		hoch				hoch	
Mehrere Ergebnis-alternativen anzeigen			hoch						hoch	
Sicherheiten über Dritte	sehr hoch	ggf. andere Nutzende	hoch	hoch	hoch		**sehr hoch**	weniger relevant		hoch
Infos über Rechenschaft	hoch	weniger relevant	hoch		hoch					an Urheber gegeben
Info über Urheber	sehr hoch	weniger relevant	hoch	hoch	hoch		**sehr hoch**			hoch
Infos zum Datenschutz			hoch	sehr hoch						
Akkuratheits-angaben	hoch	weniger relevant			**sehr hoch**	**sehr hoch**	**sehr hoch**		sehr hoch	
Akkuratheits-anspruch	sehr hoch		sehr hoch		**hoch**		hoch			sehr hoch
Zugangsschwelle zum System	hoch	niedrig	hoch		sehr hoch	**hoch**		niedrig		hoch

***Tabelle 22**: Zusammenfassung der Implikationen zur von Systemeigenschaften abhängigen KI-Transparenz*

Entwickler*innen von KI	Politik und Regulierungsinstitutionen
• Die für ein KI-System relevanten System- und Interaktionseigenschaften zusammen mit Nutzenden identifizieren • Tabelle 21 für Implikationen der Systemeigenschaften auf Transparenz nutzen, (Anleitung und Prozess in Kapitel 5.5.1)	• Systemeigenschaften nutzen, um Systeme zu kategorisieren • Regularien, die Endnutzende adressieren abhängig von Systemeigenschaften entwickeln • Regularien auch unabhängig von Systemeigenschaften entwickeln. KI-Transparenz durch Expert*innen etc. prüfen und durch Zertifikate kennzeichnen

7.4.1.2. Nutzendenzentrierte Transparenz

Wie sich bereits in vorangehender Forschung und in den Studien (a) „Fehlerfall“ und (c) „Transparenzarten“ zeigte, adaptieren Nutzende sehr schnell an ein gegebenes System und seine wahrgenommene Leistung (z. B. Bansal et al., 2019; Lim & Dey, 2011; Yin et al., 2019). Sowohl der Nutzungsabfall nach einem Fehler – Algorithm Aversion – als auch der Lerneffekt bei besonders guter Leistung verdeutlichen dies. Menschen lernen also durch den Umgang mit KI-Systemen. Anders gesagt: Nutzende haben ein mentales Modell eines Systems, das sie im Laufe der Nutzung ständig anpassen. Aufgrund verschiedener Vorerfahrungen und Wissensstände unterscheiden sich die mentalen Modelle von Nutzenden – und weisen möglicherweise Lücken und Fehler auf. Deshalb ist es von zentraler Bedeutung, ihnen die Möglichkeit zu geben, die tatsächlichen Funktionsweisen, Grenzen und Möglichkeiten eines Systems zu verstehen und hinzuzulernen (Bansal et al., 2019). Bansal et al. empfehlen deshalb, dieses Lernen mit drei Maßnahmen zu erleichtern:

1. KI-Systeme sollten möglichst wenige Fehlerbereiche aufweisen.
2. Systemfehler sollten möglichst wenig zufällig sein.
3. KI-Systeme sollten durch eine Reduzierung ihrer Merkmale möglichst einfach gestaltet sein.

Die Empfehlungen zielen darauf ab, die Nutzung von KI-Systemen möglichst berechenbar und nachvollziehbar zu gestalten. Transparenz in die Systeme ermöglicht, das Lernen im Umgang mit KI-Systemen zielgerichtet und wahrheitsgetreu zu steuern.

Aufgrund der individuell unterschiedlichen mentalen Modelle aller Nutzenden ist die hohe Abhängigkeit von KI-Transparenz von Kontext und Nutzenden nicht überraschend. Gleichzeitig stellt es die Umsetzung einer KI-Transparenz vor Herausforderungen, da es nicht die eine Lösung für alle Systeme oder Nutzenden geben kann. Wie sich im Laufe der Arbeit wiederholt herausstellte, ist deshalb eine menschzentrierte Entwicklung von KI-Transparenz von entscheidender Bedeutung. Dabei

zeigen sich Parallelen zur nutzendenzentrierten KI. Auch bei der KI-Entwicklung bedurfte es Ende der 90er-Jahre der Forderung des Entwicklers Alan Cooper (1999), der in seinem Buch „Inmates Running the Asylum“ anmahnte, die Nutzendensicht bei der Entwicklung von KI einzubeziehen. Er kritisierte, Entwickler*innen setzten KI-Systeme lediglich aus ihrer eigenen Sicht auf und entwickelten so zu oft an den Bedarfen der Nutzenden vorbei. Der Einbezug der Nutzenden bei der Entwicklung neuer Systeme ist eine Maßnahme, die die Akzeptanz der Nutzenden steigert (Markus et al., 2024; Zysk et al., 2024) und auch Algorithm Aversion reduziert (L. Xu et al., 2023).

Auch KI-Transparenz wurde bisher häufig aus Sicht der Entwickler*innen umgesetzt. Einerseits ist dies nachvollziehbar und sinnvoll: Bei der Entwicklung neuer KI-Systeme müssen Modelle getestet und Konzepte geprüft und deshalb ihre Fähigkeiten, Grenzen und Prozesse von den Expert*innen nachvollzogen werden können. Eine technische Explainability ist also wichtig, um gute KI zu erstellen. Andererseits ist der Einbezug der Endnutzenden vonnöten, wenn es um die Anwendung dieser KI-Systeme für Anwender*innen, Domänenexpert*innen oder Laien geht.

Denkt man den Ansatz einer nutzendenzentrierten KI konsequent zu Ende, darf diese nicht nur die Funktionalität, Bedienbarkeit und Kernprozesse der KI betreffen, sondern muss auch Transparenz in diese Prozesse und Funktionalitäten mit einbeziehen. Beispiele für die Entwicklung humanzentrierter KI-Anwendungen gibt es bereits aus verschiedenen Kontexten (Buschmeyer et al., 2024; Herrmann & Pfeiffer, 2023; Markus et al., 2024; Shneiderman, 2022). Bei diesen sollten, ebenso wie bei grundsätzlicheren Ansätzen von Human-Centered Design (z. B. Zysk et al., 2024), Transparenzaspekte berücksichtigt werden, um Verständnis und Informiertheit über die Systeme zu ermöglichen. Umgekehrt gilt es bei Konzepten wie Transparency by Design (z. B. Felzmann et al., 2020) die Nutzendensicht zu integrieren. Dabei muss bei einer nutzendenzentrierten Transparenz ein tatsachengetreues Verständnis der Nutzenden sichergestellt werden. Es darf nicht darum gehen, Nutzende mit irreleitenden Angaben zu täuschen, wie Chromik et al. warnen (2019). Vielmehr muss KI-Transparenz eine informierte Nutzung befördern, auch wenn das bedeutet, dass sich Nutzende gegen ein System entscheiden.

Die vorliegende Arbeit liefert einige konkrete Hinweise zur Umsetzung einer nutzendenzentrierten KI-Transparenz. Zum einen unterscheiden sich die Bedarfe nach Transparenz je nach Nutzungsphase. Dies muss bei der Gestaltung von Transparenz berücksichtigt werden. Zu Beginn einer Nutzung steigern allgemeine Hintergrundinformationen die Vertrauensbildung: globale Informationen über Entwickler*innen, dahinterstehende Organisationen, Prüfsysteme oder Zertifizierungen, Sicherheiten oder Gewährleistungen. Darüber hinaus können weitere vom konkreten KI-System unabhängige Informationen relevant werden, beispielsweise wie ein Unternehmen mit dem neuen System

umzugehen gedenkt (für einen Leitfaden zur Umsetzung von KI in Unternehmen siehe z. B. Buschmeyer et al., 2024).

Im Verlauf der Nutzung geht es darum, mit uneindeutigen Vorschlägen, Phasen der Unsicherheit oder Fehlern des Systems umzugehen. Vorangestellte Informationen darüber, wie wahrscheinlich solche Fehler auftreten, können Erwartungen regulieren. Gleichzeitig spielen lokale Erklärungen eine entscheidende Rolle dabei, Vertrauen und Nutzung zu erhalten, indem sie die Gründe für ein falsches Ergebnis erklären oder verdeutlichen, wie das System aus dem aufgetretenen Fehler lernt. Die in den Fokusgruppen angesprochenen Feedbackmöglichkeiten zu Ergebnissen eines Systems sind eine weitere Möglichkeit, bei Systemen mit weniger objektiven Fehlern die Nutzung zu erhalten.

Zum Zweiten sind besonders bei komplexen Systemen die verschiedenen Untersysteme bei der Umsetzung von Transparenz zu unterscheiden und nutzendenzentriert zu entwickeln. Eine Dateneingabe bedarf anderer Transparenz – z. B. über Datenschutz, aber auch Relevanz der Eingaben – als das daran anschließend präsentierte Ergebnis, bei dem es vielleicht eher um lokale Erklärungen, Akkuratheit oder die dem Ergebnis zugrundeliegenden Datenbasis geht.

Zuletzt ist aufgrund der individuellen Bedürfnisse der Nutzenden eine nutzendenzentrierte Transparenz im besten Falle schrittweise aufgebaut. Einige Teilnehmende der Fokusgruppen betonten, zu viele Informationen würden sie überfordern – ein Ergebnis, das sich auch bei negativen Effekten von Transparenz in anderen Studien bestätigt (Tsai & Brusilovsky, 2019; Yu et al., 2017; Zhao et al., 2019). Vielmehr gilt es, eine dem Wissens- und Interessenstand der Nutzenden anpassbare Transparenz zu gestalten. Eine Darstellung von Transparenzinformationen kann beispielsweise mit Menüs mit Unterpunkten, weiterführenden Links oder optional anklickbaren Informationsbuttons realisiert werden.

Zusammenfassend stellt nutzendenzentrierte Transparenz einen wichtigen Bestandteil von KI-Entwicklung dar. Im ersten Schritt resultiert daraus die Maßgabe, KI so einfach wie möglich zu gestalten, um das Lernen im Umgang mit dem System zu vereinfachen. Außerdem müssen Nutzende bei der Umsetzung des Systems mit einbezogen werden. Dabei gilt es Transparenzaspekte zu berücksichtigen, um den Nutzenden zu ermöglichen, ihre mentalen Modelle eines KI-Systems zu aktualisieren und anzupassen – also einen informierten Umgang mit dem System zu ermöglichen. Wichtige Aspekte bei der Umsetzung einer nutzendenzentrierten KI-Transparenz sind die Berücksichtigung verschiedener Nutzungsphasen, die Betrachtung der Systembestandteile sowie eine schrittweise und optional auswählbare Transparenz, um verschiedene Nutzende bestmöglich bei ihrem Wissens- und Interessensstand abzuholen (siehe Tabelle 23 für eine Zusammenfassung). Dieser Wissensstand kann nicht allein im Umgang mit einem KI-System entstehen. Vielmehr geht es darum,

den Umgang mit KI und Algorithmen im Vorhinein zu schulen: „AI Literacy“ ist das Stichwort, auf das im folgenden Kapitel näher eingegangen wird.

***Tabelle 23**: Zusammenfassung der Implikationen zur Umsetzung von nutzendenzentrierter Transparenz*

Entwickler*innen von KI	Politik und Regulierungsinstitutionen
• Transparenz als Bestandteil einer nutzendenzentrierten KI betrachten • KI-Systeme so einfach wie möglich und so komplex wie nötig gestalten • KI-Systeme in Einzelteile zerlegen nach Nutzungsphasen und -inhalt • Schrittweise Transparenz umsetzen, z. B. durch Unterpunkte, weiterführende Links	• Bei der Formulierung von Regulierungen verschiedene Nutzungsgruppen mitdenken • Subjektive Bewertungen von Transparenz berücksichtigen: Verständnis der Inhalte wichtiger als bloße Bereitstellung • Nutzende auch bei der Prüfung und Lizensierung von KI einbinden

7.4.1.3. AI Literacy

Die Implikation, Algorithm Literacy als Maßnahme zur Überwindung von Algorithm Aversion umzusetzen, wurde bereits im Zuge der Forschungsfrage (a) „Fehlerfall“ diskutiert. Laut Burton et al. umfasst dieses Konzept das Wissen „how to interact with algorithmic tools, how to interpret statistical outputs, and how to appreciate the utility of decision aids“ (2020, S. 223). Während der Begriff der Literacy gerade einer Mode unterworfen scheint und für verschiedenste Konzepte diskutiert wird (Khodaei et al., 2023; Maye, 2023; Ng et al., 2021), fordern immer mehr Forscher*innen aus Bildung und Technologie unter dem Begriff „AI Literacy“ eine Grundbildung in Bezug auf die disruptive Technologie künstliche Intelligenz, die immer mehr Aspekte unseres Lebens betrifft (Klein, 2023; Wienrich et al., 2022)[11]. Als Bestandteil von digitaler Literacy ist auch AI Literacy nicht eindeutig definiert, doch legt ein Literaturreview zum Begriff vier zentrale Säulen offen (Ng et al., 2021):

- Know and understand AI
- Use and apply AI
- Evaluate
- Ethical issues

Weitere Literaturreviews und Auseinandersetzungen mit dem Thema deuten eine ähnliche Struktur an (Casal-Otero et al., 2023; Wienrich et al., 2022). Neben dem Wissen über KI gehört zu einer AI Literacy die Kompetenz, sie auf die richtige Weise einzusetzen, ihre Ergebnisse evaluieren zu können

[11] Da es im Deutschen keine dem englischen Begriff Literacy entsprechende Bezeichnung gibt, wird auch im Deutschen zumeist der Begriff AI Literacy, also „Artificial Intelligence Literacy“, verwendet (Wienrich et al., 2022).

und ethische Fragestellungen, die sich aus Entwicklung, Anwendung und Nutzung erheben, reflektieren zu können. Wichtig ist dabei auch ein Verständnis dafür, dass AI Literacy nicht das Wissen betrifft, das IT-Expert*innen vorweisen, sondern die Grundbildung von jeder und jedem bezeichnet, die oder der in der heutigen Welt aufwächst, arbeitet und lebt. Die entsprechende Forderung lautet, die Vermittlung von AI Literacy bereits im Kindesalter zu beginnen und diese entsprechend dem Alter kontinuierlich auszubauen (Klein, 2023). Über diesen allgemeinbildenden Teil muss die Ausbildung entsprechender Berufe um KI-Wissen, -Nutzung, -Evaluation und -Kritik ergänzt werden. Anwendungen von KI finden sich bereits in der Medizin, im Finanzsektor, im Management, im HR-Bereich, in produzierenden Unternehmen und im Recht – und die Anwendungsbereiche nehmen stetig zu. Studien zeigen, dass Versuchspersonen mit höherer AI Literacy eher KI-Systemen vertrauen und ihre Ratschläge bzw. kombinierte Ratschläge von KI und Mensch in größerem Maße nutzen, als sich allein auf menschlichen Rat zu verlassen (Schoeffer et al., 2022; Stradi & Verdickt, 2024).

Die vier Säulen der AI Literacy hängen auf unterschiedliche Weise mit KI-Transparenz zusammen. Am eindeutigsten ist der Zusammenhang mit der Säule „Evaluation“: Die Offenlegung von Hintergrund- und Prozessinformationen, von Akkuratheitsinformationen und lokalen und globalen Erklärungen ist nötig, um die Evaluation eines KI-Systems zu ermöglichen. Um entsprechende Informationen einbetten und interpretieren zu können, bedarf es der Säule des Wissens über KI und ihre Funktionsweisen. Dazu gehört beispielsweise das grundsätzliche Verständnis, dass KI-Systeme Modelle der Wirklichkeit darstellen und so immer fehlerbehaftet sind. Eine informierte und selbstbestimmte Nutzung, die eine weitere Säule darstellt, kann nur auf der Grundlage von Wissen und Evaluation erfolgen und baut damit wiederum auf verständliche KI-Transparenz. Die Auseinandersetzung mit ethischen Problemen von KI erstreckt sich über alle anderen Säulen: Wie neutral kann KI sein, welche Bewertungskriterien sind relevant, wer hat ein Interesse an meiner Nutzung des Systems? Sie ist wiederum nur möglich, wenn KI transparent ist und entsprechende Informationen auf verständliche Weise offenlegt.

Allerdings darf die Forderung nach AI Literacy nicht als individueller Auftrag missverstanden werden, wie schon bei Algorithm Literacy in Kapitel 4 angemerkt wurde. Eine Grundbildung in Bezug auf KI sicherzustellen, ist ein Auftrag an das Bildungssystem. Die entsprechenden Implikationen des Kapitels zu AI Literacy sind in Tabelle 24 dargestellt. Darüber hinaus müssen jedoch auch die externen Rahmenbedingungen stimmen bzw. geschaffen werden, damit ein aufgeklärter Umgang mit KI gewährleistet werden kann. Mit diesen Rahmenbedingungen befasst sich das nächste Kapitel.

***Tabelle 24**: Zusammenfassung der Implikationen zur Umsetzung von AI Literacy*

Entwickler*innen von KI	Politik und Regulierungsinstitutionen
• KI-Evaluation durch ihre Transparenz ermöglichen • Ausbildung einer AI Literacy durch Informationen und Transparenz auch bei der Nutzung von KI fördern	• AI Literacy in der Breite der Bevölkerung sicherstellen • Wissen über KI im gesamten Lehrplan verankern, schon von klein auf • Ausbildung von AI Literacy in Berufsausbildungen und Studium verankern

7.4.1.4. Rechtliche Vorgaben und Audits

Um einen informierten und autonomen Umgang mit KI zu ermöglichen, erfordert es nicht zuletzt gesetzgeberische Vorgaben, denen Unternehmen und Anbieter von KI Folge zu leisten haben. Die Forderung nach rechtlichen Vorgaben und klaren Leitlinien für KI-Systeme ist nicht nur auf Seiten von Nutzenden evident (Bedué & Fritzsche, 2022; Fischer & Petersen, 2018). Verschiedene Parlamente verfolgen aktuell Pläne zur Regulierung von KI. Als eines der ersten politischen Organe, und sicherlich als das aktuell weitreichendste, verabschiedete das EU-Parlament kürzlich den AI Act, der seit 2021 erarbeitet wurde (European Parliament, 2024). Darin wird der Einsatz von KI-Systemen mit besonders hohem Risiko reguliert bzw. verboten: Z. B. darf KI nicht zur Bewertung von Menschen anhand ihres Sozialverhaltens genutzt werden, eine biometrische Analyse ist nur unter bestimmten Voraussetzungen für Strafverfolgungsbehörden gestattet. Außerdem räumt das Gesetz Nutzenden das Recht auf „aussagekräftige Erklärungen“ ein: KI-Systeme „mit allgemeinem Verwendungszweck“ müssen beispielsweise die für das Training verwendeten Inhalte offenlegen, Modellbewertungen durchführen oder bestimmte Risiken berücksichtigen (European Parliament, 2024).

Eine solche Gesetzgebung und das Wissen, dass alle auf dem EU-Markt verfügbaren KI-Systeme diesen Vorgaben (eigentlich) unterlegen sind, stärken sicherlich das Vertrauen von Nutzenden in KI. Gleichwohl bedarf es weiterer Forschung hinsichtlich der Frage, wie aussagekräftige Erklärungen auszusehen haben und wie und ob hierfür Leitlinien festgelegt werden können. Zu vermeiden ist beispielsweise eine reine Offenlegung von Informationen, wie es im Rahmen der Datenschutzgrundverordnung häufig geschieht. Diese bedeutet häufig eher eine Überforderung der Nutzenden (Felzmann et al., 2019; Larsson & Heintz, 2020; Wulf & Seizov, 2024). Wie die Ergebnisse der vorliegenden Arbeit zeigen, müssen Informationen sinnvoll aufbereitet sein, um Verständnis und Vertrauen zu fördern.

Ziel der Gesetzgebung ist, die menschliche Aufsicht über KI zu gewährleisten. Im besten Fall sind Erklärungen so einfach und aussagekräftig gestaltet, dass Nutzende sie verstehen und der Aufsicht

über ein System nachkommen können. Oft genug werden Systeme aber zu komplex sein und in diesen Fällen stellvertretende Institutionen diese Aufsicht gewährleisten müssen. Damit steigt der Bedarf nach unabhängigen Regulierungsinstitutionen, die die Umsetzung der gesetzlichen Vorgaben prüfen. Neben dem Faktor, dass das Wissen um etablierte Institutionen hinter KI-Systemen das Vertrauen in diese erhöhen kann, kamen in den Fokusgruppen zur Forschungsfrage (b) „Nutzendenanforderungen" auch Zertifikate und Audits zur Sprache: Die Stiftung Warentest, ein KI-TÜV oder andere Verbraucherorganisationen wurden hier genannt. Audits verfolgen das Ziel, „investigating algorithms' functionality to detect bias and other unwanted algorithm behaviors without the need to know about its specific design details" (Mohseni et al., 2021, S. 3). Für die Arbeitswelt veröffentlichte die Gesellschaft für Informatik ein Framework für KI-Audits (Gesellschaft für Informatik, 2022; Waltl & Becker, 2021), Beratungsfirmen bauen entsprechende Geschäftszweige auf (EY, 2022) und IEEE bietet AI-Reviews als Dienstleistung an (IEEE SA, 2022).

Auch wenn sich die Zielgruppe der Auditor*innen aufgrund ihrer Expertise von der der Endnutzenden unterscheidet, müssen auch erstere in die Lage versetzt werden, Systeme zu prüfen. Transparenz in KI-Systeme ist also auch hier relevant – oder erleichtert zumindest das Vorgehen. Insgesamt besteht im Hinblick auf KI-Audits aktuell ein großer Forschungsbedarf. Diese Ausgestaltung umfasst zum einen technische Prüfprozesse. Zum anderen müssen auch Audits nutzungsrelevante und ethische Fragestellungen mit einschließen und wiederum Transparenz hinsichtlich ihrer Prozesse und Ergebnisse für Endnutzende gewährleisten. Bisher jedoch ist „auditing […] a valuable yet time intensive process that could not be scaled easily to large numbers of algorithms. This calls for new research for more effective solutions toward algorithmic transparency." (Mohseni et al., 2021, S. 3–4) Die Umsetzung der Nutzendenanforderungen in der Auditierung der Prozesse wie auch im Bericht darüber gilt es dabei mit einzubeziehen (siehe Tabelle 25 für eine Zusammenfassung des Kapitels).

***Tabelle 25**: Zusammenfassung der Implikationen zu rechtlichen Vorgaben und Audits*

Entwickler*innen von KI	Politik und Regulierungsinstitutionen
• KI-Transparenz bestmöglich umsetzen, nicht nur aufgrund der oder beschränkt auf gesetzliche Vorgaben • Gut umgesetzte KI-Transparenz als Stärkung der Glaubwürdigkeit und Nutzung sehen • KI-Transparenz für Prüfstellen gewährleisten	• Gesetzgebung für KI und KI-Transparenz vorantreiben • Forderungen präzise formulieren und regelmäßig anpassen • Ausreichend Zertifizierungsinstitutionen einrichten, um der Menge an KI gerecht zu werden und z. B. Verbraucherbeschwerden verfolgen zu können

- Prüfung von KI muss ethische Faktoren und Nutzendensicht berücksichtigen
- Weiterhin Entwicklung von empirisch ermittelten, konkreten Leitlinien für transparente KI fördern

7.4.2. Zusammenfassung der praktischen Implikationen

Im Laufe der Arbeit konnten aus den Ergebnissen von drei Studien Implikationen für Entwickler*innen von KI abgeleitet werden, die für die jeweilige Studie gelten. Die Sammlung dieser neun Implikationen wurde ergänzt durch übergeordnete Implikationen entlang der Themen „KI-Transparenz abhängig von Systemeigenschaften“, „Nutzendenzentrierte Transparenz“, „AI Literacy“ sowie „Rechtliche Vorgaben und Audits“. Dabei wurden zusätzlich Implikationen für Politik und Regulierungsinstitutionen formuliert. Um alle Implikationen übersichtlich an einer Stelle darzustellen, werden diese in den nächsten beiden Kapiteln nach Zielgruppen getrennt aufgeführt und anschaulich dargestellt.

7.4.2.1. Praktische Implikationen für Entwickler*innen

Aus den drei Studien dieser Arbeit wurden insgesamt neun konkrete Implikationen für Entwickler*innen abgeleitet, die an verschiedenen Stellen im Entwicklungsprozess relevant werden (siehe Kapitel 4.5.1, Kapitel 5.5.1 und Kapitel 6.5.1). Die Implikationen, die im vorangegangenen Kapitel für die gesamte Arbeit abgeleitet wurden, ergänzen diese in den meisten Fällen, teilweise überlappen sie sich aber auch. Um die Implikationen in ein großes Bild einzuordnen und so die Ergebnisse in einer Übersicht nutzbar zu machen, wurden vier übergeordnete Themen identifiziert, in die sich die Implikationen einordnen lassen: „Grundsätzliches Verständnis“, „KI-Transparenz mit Nutzenden umsetzen“, „Wann KI-Transparenz umsetzen?“ und „Wie KI-Transparenz umsetzen?“ Die zusammengetragene Darstellung lässt sich als Ergebnis der Arbeit auch durch andere Veröffentlichungen bestätigen (z. B. Laato et al., 2022). Abbildung 35 stellt die Sammlung der Implikationen für Entwickler*innen von KI dar, die aus der Arbeit resultieren. Dabei wurden sowohl die neun studienbezogenen Implikationen gekürzt und angepasst als auch die neuen, kleinschrittigeren Implikationen ergänzt. In der Darstellung sind die Bezüge zu den Implikationen jeweils in Klammern ergänzt, um bei Interesse die geltenden Herleitungen oder Erklärungen nachlesen zu können.

Abbildung 35: *Übersicht über die aus der Arbeit abgeleiteten Implikationen für Entwickler*innen aus den drei Studien sowie der Gesamtschau. Die kursiv ergänzten Zahlen stehen für die Nummer der jeweiligen Implikation (1-4 aus Kapitel 4.5.1, 5 und 6 aus Kapitel 5.5.1 und 7-9 aus Kapitel 6.5.1) bzw. für das Kapitel, in dem sie zuvor dargelegt wurden (7.4.1.1-7.4.1.4).*

 Grundsätzliches Verständnis	 KI-Transparenz mit Nutzenden umsetzen	 Wann KI-Transparenz umsetzen?	 Wie KI-Transparenz umsetzen?
KI-Transparenz bestmöglich umsetzen, nicht nur aufgrund der oder beschränkt auf gesetzliche Vorgaben (*7.4.1.4*) Gut umgesetzte KI-Transparenz als Stärkung der Glaubwürdigkeit und Nutzung sehen (*7.4.1.4*) KI-Systeme so einfach wie möglich und so komplex wie nötig gestalten (*7.4.1.2*)	KI als Unterstützung für Mensch: KI und ihre Transparenz muss sich Nutzenden anpassen, nicht umgekehrt. (*4*) Nutzende einbeziehen bei Erhebung der KI-Eigenschaften, der Bedarfe und auch durch Evaluation. (*8, 7.4.1.1, 7.4.1.2*) Verständlichkeit ist wichtiger als technische Erklärungen oder Detailtiefe. Dabei gilt: Zielgruppen bedenken. (*7*) Selbstbestimmte Nutzung als Ziel: Evaluation von KI durch verständliche Transparenz ermöglichen. So AI Literacy fördern. (*7.4.1.3*)	KI-Systeme in Einzelteile zerlegen nach Nutzungsprozess und -inhalt (*7.4.1.2*) Transparenz abhängig von Nutzungszeit gestalten: Aufbau von Vertrauen durch globale Erklärungen. Während Nutzung Akkuratheitsangaben und lokale Erklärungen. (*9*) Im Fehlerfall: lokale Erklärungen liefern, Möglichkeit einräumen Feedback zu geben und kommunizieren, wie KI aus dem Fehler lernt. (*3*)	Transparenzmatrix aus Tabelle 21 nutzen und in vier Schritten anwenden, nach Kapitel 5.5.1. (*5, 7.4.1.1*) Akkuratheit offenlegen oder System verbessern. Einschätzung dieser ermöglichen, z. B. durch Angabe der eigenen Leistung. (*1, 2*) Transparenz in zunehmender Detailtiefe gestalten und Nutzenden schrittweise Auseinandersetzung ermöglichen. (*6, 7.4.1.2*) KI-Transparenz für Prüfstellen gewährleisten (*7.4.1.4*)

7.4.2.2. Praktische Implikationen für Politik und Regulierungsinstitutionen

Die zuvor dargelegten Implikationen für Entwickler*innen von KI wurden im Laufe der gesamten Arbeit kontinuierlich ergänzt und erweitert. Zusätzlich dazu ermöglichte die Gesamtschau der Arbeit die Ableitung von Implikationen für eine höhere Ebene: für Politik und Regulierungsinstitutionen. Die in den vier Unterkapiteln des Kapitels 7.4.1 inhaltlich abgeleiteten Implikationen werden im Folgenden übersichtlich dargestellt (Abbildung 36). Dabei wurden sie in drei Kategorien zusammengefasst, die verschiedene politische Bereiche und Prozesse betreffen: Die „Entwicklung von KI-Regularien" betrifft in erster Linie die Legislative und Parlamente. Die „Umsetzung von KI-Regularien" betrifft sowohl Legislative als auch Exekutive sowie darin enthaltene Regulierungsinstitutionen. „Weitere Maßnahmen" sind Implikationen, die über KI-Regularien hinaus auch andere Bereiche betreffen, wie Forschungsförderung und Bildung. Mit dieser Darstellung soll ein möglichst handhabbarer Umgang mit den Implikationen ermöglicht werden. Eine Vertiefung der einzelnen in Abbildung 36 aufgeführten Punkte ist über die jeweils angegebenen Kapitel möglich.

Abbildung 36: *Übersicht über die aus der Arbeit abgeleiteten Implikationen für Politik und Regulierungsinstitutionen aus den drei Studien sowie der Gesamtschau. Die kursiv ergänzten Zahlen stehen für das Kapitel, in dem die Implikationen zuvor dargelegt wurden (7.4.1.1-7.4.1.4)*

 KI-Regularien entwickeln	 KI-Regularien umsetzen	 Weitere Maßnahmen
Gesetzgebung für KI und KI-Transparenz vorantreiben (*7.4.1.4*) Systemeigenschaften (Tabelle 21) nutzen, um Systeme zu kategorisieren und Regularien abhängig davon entwickeln (*7.4.1.1*) Bei der Formulierung von Regulierungen verschiedene Nutzungsgruppen im Blick haben (*7.4.1.2*) Ethische Faktoren und Nutzendensicht berücksichtigen (*7.4.1.4*) Forderungen präzise formulieren und regelmäßig anpassen (*7.4.1.4*)	KI-Transparenz durch Expert*innen prüfen und durch Zertifikate kennzeichnen (*7.4.1.1*) Subjektive Bewertungen von Transparenz berücksichtigen: Verständnis der Inhalte wichtiger als bloße Bereitstellung (*7.4.1.2*) Nutzende bei Prüfung und Lizensierung von KI einbinden (*7.4.1.2*) Genügend Zertifizierungsinstitutionen einrichten, um der Menge an KI-Systemen zu bewältigen und Verbraucherbeschwerden verfolgen zu können (*7.4.1.4*)	Weiterhin Entwicklung von empirisch ermittelten, konkreten Leitlinien für transparente KI fördern (*7.4.1.4)* AI Literacy in der Breite der Bevölkerung sicherstellen (*7.4.1.3*) Wissen über KI im gesamten Lehrplan verankern, schon von klein auf (*7.4.1.3*) Ausbildung von AI Literacy in Berufsausbildungen und Studium verankern (*7.4.1.3*)

7.4.3. Anschließende Forschungsfragen

Wie die breite Forschungsfrage erwarten lässt und auch in den vorherigen Kapiteln deutlich wurde, ergeben sich im Anschluss an die gewonnenen Erkenntnisse zahlreiche weitere Forschungsansätze bezüglich der übergeordneten Fragestellung, **wie sich Transparenz von KI-Entscheidungsunterstützungssystemen auf die Nutzung dieser Systeme durch Endnutzende auswirkt.**

Auf der Hand liegt die weiterhin nötige Auseinandersetzung mit der Fragestellung, **wie eine den Nutzenden dienliche KI-Transparenz aussehen kann**. Dies umfasst Arten der Kommunikation von Unsicherheit in Bezug auf ganze Systeme wie auch Einzelergebnisse, ebenso wie Erklärungen zur Funktionalität, zu Hintergrundinformationen und außerdem dazu, wie Systeme mit Fehlern umgehen, aus ihnen lernen oder Feedback verarbeiten. Die Ergebnisse der vorliegenden Arbeit deuten an, dass es nicht die eine Antwort auf diese Frage geben wird, sondern sie sehr kontext- und nutzendenspezifisch untersucht und beantwortet werden muss. Gleichzeitig wären **Klassifizierungen von Kontexten, von Nutzenden wie auch von Systemeigenschaften** mögliche nächste Schritte bei der Umsetzung von KI-Transparenz.

Daran anschließend ergibt sich auf Grundlage des in dieser Arbeit sehr breiten Verständnisses von Transparenz die Fragestellung, was genau Transparenz umfasst und welche Informationen darüber hinausführen. Aufbauend auf einer Klassifizierung der beeinflussenden Faktoren ergibt sich also der Bedarf nach einer **Klassifizierung von Transparenz für Endnutzende**, die über die bisher eher technisch angelegten Kategorien hinausgeht. In Kombination mit Kontexten, Nutzenden und Systemeigenschaften wäre die Testung und Erstellung einer Matrix denkbar, die auf der in dieser Arbeit erstellten Transparenzmatrix aufbauen kann.

Die Annahme, verschiedene Transparenzarten wirken zu verschiedenen Zeiten im Nutzungsprozess, ergibt die Möglichkeit, die bisherigen ambivalenten Ergebnisse im Hinblick auf die Nutzung transparenter KI neu zu bewerten. Die Annahme geht davon aus, dass sich der Transparenzbedarf unterscheidet nach **Nutzungszeitpunkt**: Zum Beginn einer Interaktion dienen eher globale Hintergrundinformationen dem Vertrauensaufbau. Während der Nutzung und insbesondere bei Unsicherheit und im Fehlerfall werden lokale Erklärungen relevant. Daraus ergibt sich einerseits die Fragestellung, ob eine solche Betrachtungsweise als Erklärung dient für bisherige, uneinheitliche Ergebnisse zu globalen und lokalen Erklärungen wie auch zu Akzeptanz, Nutzung und Vertrauen in KI. Andererseits lässt sich daraus für die Gestaltung von KI-Transparenz neben Systemfaktoren, Kontext und Nutzendenfaktoren eine weitere wichtige Kategorie ableiten: die des Interaktionszeitpunkts. Untersuchungen hierzu sollten sowohl bestehende Forschung zur Technikinteraktion und zu den sich verändernden Voraussetzungen über die Zeit einbeziehen als auch verschiedene Nutzungszeitpunkte miteinander vergleichen.

Das breite Verständnis von KI-Transparenz durch Endnutzende, das sich in der vorliegenden Arbeit zeigte, deutet eine nötige Selbstreflexion der Methoden und Forschungsweisen von KI und KI-Transparenz an. Es verkompliziert die Untersuchung von Transparenz, da Nutzende nicht einem Algorithmus zu vertrauen scheinen, sondern den Menschen dahinter. Bei der Frage, ob sie einem System vertrauen können, interessiert die Nutzenden, welche Expertise die Entwickler*innen der KI

auszeichnet, welche Motive sie bei der Entwicklung verfolgten und ob man ihnen trauen kann. Dieser Ansatz jedoch läuft entgegen dem Verständnis, das insbesondere beim Einsatz von KI in den meisten Fällen (noch) postuliert wird: KI wird als neutrales Werkzeug eingesetzt und ist über menschliche Einflüsse wie begrenzte Verarbeitung und emotionale Urteile erhaben. Die Forschung zu Transparenz erhält damit eine noch **sozialere Komponente** als bislang in den meisten Fällen berücksichtigt. Die vermeintliche Objektivität von KI in Frage zu stellen, sollte Bestandteil zukünftiger sozialwissenschaftlicher Forschung sein und damit eine technische, vom Menschen zu trennende Sicht auf KI erweitern.

Für die Interaktion von Menschen und KI schließt sich in Bezug auf die Wirkung von Transparenz die Frage nach der **Entwicklung und Veränderung von mentalen Modellen durch Transparenz** an. Wie verändert sich das Verständnis von konkreten Systemen, wie das Verständnis von KI allgemein, wenn KI-Transparenz besser und mehr umgesetzt wird? Da KI als Technologie weiteren Einfluss gewinnen und Anwendungsfelder hinzugewinnen wird, sind Lernen durch Nutzung, Gewöhnung und Routine in Bezug auf KI weitere Themen für zukünftige Forschung. Wie lässt sich durch KI-Transparenz eine gesellschaftliche Sicht auf KI gestalten, die ihren Vor- wie Nachteilen gerecht wird? Zusätzlich ist die KI-Entwicklung rasanten Innovationen unterworfen: Wie müssen sich Transparenzarten und Transparenzkommunikation verändern, um diesen immer komplexeren Systemen gerecht zu werden?

Bei einer KI, die sich ständig verändert und immer mehr Lebensbereiche umfasst, ergeben sich zusätzliche Forschungsbedarfe in Bezug auf die Zielgruppen von Transparenz. Lange Zeit wurden Entwickler*innen und Designer*innen von KI als primäre Zielgruppe von KI-Transparenz untersucht. Zunehmend, wie in der vorliegenden Arbeit, traten die Endnutzer*innen von KI-Anwendungen in den Fokus. Mit der steigenden Zahl, Komplexität und Anwendungsgebieten von KI ergeben sich jedoch weitere Zielgruppen, die bei der Gewährleistung von KI-Transparenz berücksichtigt werden müssen. Die **Gruppe der KI-Auditor*innen** wurde im vorigen Kapitel bereits angesprochen. Eine weitere immer relevantere **Gruppe ist die der dritten Personen, die von KI betroffen sind**. Regelmäßig genannt und sehr umstritten ist beispielsweise der Einsatz medizinischer Unterstützungssysteme für Ärzt*innen, Kreditwürdigkeitsprüfungssysteme für Banken oder Systeme für Straffälligkeitseinschätzungen. Diese Systeme urteilen über Patient*innen, Antragsteller*innen oder Angeklagte, ohne dass die vom System Beurteilten Zugriff auf die Ergebnisse haben – geschweige denn auf die Erklärungen, die die Systeme für ihr Urteil liefern. Denkbar ist der Einsatz von KI auch in der Arbeitswelt: Schichtpläne, die von KI optimiert werden oder die Zuteilung von Aufgaben. Trotz der vielleicht weniger gravierenden Auswirkungen solcher Entscheidungen bleibt die Frage, wem sich das System erklären muss, von großer Bedeutung. Die Produktionsplanerin erhält eine Erklärung von der eingesetzten KI für die Personalplanung am Fließband. Die Arbeiter*in am Band wird jedoch kaum eine Erklärung erhalten,

weshalb er oder sie schon wieder die unliebsame Arbeit machen muss. Die Zielgruppe der von KI Beurteilten sollte im Hinblick auf einen immer umfassenderen Einsatz von KI in zukünftiger Forschung berücksichtigt werden. Erste Ergebnisse zeigen eine enorme Skepsis, die KI-Systemen zur Unterstützung medizinischer Diagnosen entgegengebracht wird (Reis et al., 2024) oder Unternehmen entgegenschlägt, die KI einsetzen (Haupt et al., 2024). Die Konsequenz darf nicht sein, die Nutzung von KI-Systemen zu verheimlichen. Vielmehr gilt es bei der Umsetzung von Transparenz nicht mehr nur diejenigen zu berücksichtigen, die ein System tatsächlich einsetzen, sondern auch diejenigen, die von diesem Einsatz mittelbar betroffen sind.

Zusammenfassend ergeben sich im Anschluss der Gesamtarbeit die folgenden Forschungsfragen:

- Wie kann eine den Nutzenden dienliche KI-Transparenz gestaltet werden?
    - > (Wie) lassen sich dazu Kontexte von KI-Systemen, Nutzende und Systemeigenschaften kategorisieren?
    - > (Wie) lässt sich eine KI-Transparenz für Endnutzende kategorisieren und in einer Matrix mit den Einflussfaktoren darstellen?
    - > Welche Rolle spielt der Nutzungszeitpunkt bei den Anforderungen an KI-Transparenz?
- Wie lässt sich der Bedarf der Nutzenden, den Menschen hinter einer KI zu vertrauen, zusammenbringen mit dem (vermeintlich) objektiven, „übermenschlichen“ Verständnis von KI?
- Wie verändern sich mentale Modelle von KI bei Nutzenden ebenso wie ein gesamtgesellschaftliches Bild von KI durch Transparenz, aber auch durch die zunehmende Interaktion mit (transparenten) Systemen? Wie lässt sich diese Veränderung im Sinne eines informierten Umgangs gestalten?
- Welche neuen Anforderungen an Transparenz von KI ergeben sich durch die Berücksichtigung der Zielgruppe der nur mittelbar von KI betroffenen Personen?

8. Zusammenfassung

Ziel der vorliegenden Arbeit war die Auseinandersetzung mit der Frage, **wie sich Transparenz von KI-Entscheidungsunterstützungssystemen auf die Nutzung dieser Systeme durch Endnutzende auswirkt**.

Wie Studie **(a) „Fehlerfall“** zeigt, haben Angaben zur Akkuratheit eines Algorithmus einen kleinen Effekt auf die Nutzung dieses Algorithmus im Fehlerfall. Jedoch folgt trotz dieses Versuches des Erwartungsmanagements durch Transparenz weiterhin eine Ablehnung des Algorithmus nach einem Fehler, selbst wenn dadurch die Aufgabe schlechter gelöst wird. Da sich Akkuratheit sehr einfach umsetzen lässt, spricht zunächst nichts gegen eine solche Angabe, deren Einfluss jedoch begrenzt scheint. Weitere Maßnahmen, wie Informationen über die selbstständige Verbesserung einer KI nach einem Fehler oder die Möglichkeit, das Ergebnis zu beeinflussen, sind deutlich aufwendiger in der Umsetzung. Da Informationen über die Akkuratheit eines Systems nur eine Möglichkeit der KI-Transparenz darstellen, schließt sich die Frage an, wie weitere Transparenzmaßnahmen auf die Nutzung von KI wirken.

In Studie **(b) „Nutzendenanforderungen“** konnten relevante Anforderungen an KI-Transparenz für Laien identifiziert werden. Einerseits waren die Befragten sehr kritisch gegenüber neuen KI-Systemen und legten selbst bei niedriger Relevanz des Anwendungsgebiets sehr viel Wert darauf, die finale Kontrolle innezuhaben. Andererseits maßen sie technischen Erklärungen oder Interpretierbarkeit keine besonders große Rolle zu: Generelle Erklärungen – also globale Transparenz – zur zugrundeliegenden KI forderten sie wenig und lehnten sie teilweise gar mit dem Hinweis, das sowieso nicht zu verstehen, ab. Erklärungen zu den Ergebnissen – lokale Transparenz – wurden in „high-stake“-Entscheidungen teilweise gefordert und insbesondere bei KI-Fehlern relevant. Jedoch dominierte die Forderung nach elementareren Informationen die Diskussion unabhängig vom Anwendungsgebiet: Die Vertrauenswürdigkeit und damit die Nutzung der KI wurden überwiegend abhängig gemacht von Informationen zum Geschäftsmodell, zu Entwickler*innen der Systeme und Kontrolle durch Dritte. Transparenz ist für Laien also weniger durch Prozess- oder Ergebniserklärungen zu erfüllen als viel mehr über soziale Hintergrundinformationen, wie Informationen zur Autorenschaft bzw. deren zugrundeliegenden Motiven, Prüfungen durch Expert*innen und Beurteilungen von Dritten.

In Studie **(c) „Transparenzarten“** wurden vier verschiedene Transparenzarten verglichen hinsichtlich ihres Effekts auf Nutzung und Vertrauen in die Algorithmen. Im Vergleich mit einer Bedingung ohne Transparenz zeigte sich eine Zunahme der Nutzung und des Vertrauens in die Transparenzbedingungen. Diese geht jedoch teilweise auf einen Lerneffekt zurück: Die gute Leistung der Algorithmen steigerte die Nutzung signifikant. Auf das Vertrauen zeigten sich auch signifikante,

aber weniger starke Effekte. Entgegen bestehenden Studien führten globale Erklärungen zu mehr Nutzung und Vertrauen als lokale Transparenz. Am stärksten war der Effekt der globalen Erklärung mit Hintergrundinformationen, während die lokalen Heatmaps keinen Unterschied zur intransparenten Kontrollbedingung zeigten. Zusammenhänge von wahrgenommener Zuverlässigkeit mit der Nutzung sowie von Verständlichkeit mit Vertrauen deuten verschiedene Wirkmechanismen für die beiden Akzeptanzkriterien an.

Während die quantitativen Ergebnisse lediglich kleine Effekte von Transparenz bzw. verschiedenen Transparenzarten auf die Nutzung von KI-Systemen zeigten, lieferte die qualitative Studie differenziertere Erklärungen: Laien haben weniger Interesse an technischen Erklärungen der Systeme oder empfinden die Auseinandersetzung zumindest als mühsam. Von Informationen über Autor*innen und Entwickler*innen und über deren Motive erhoffen sie sich, die Vertrauenswürdigkeit eines Systems einschätzen zu können, ohne tief in technische Fragestellungen einsteigen zu müssen. Gleichzeitig kam immer wieder ein Thema bei der Diskussion über Anforderungen an KI-Systeme auf: die Auszeichnung oder Prüfung der Systeme durch eine vertrauenswürdige dritte Stelle.

Die Implikationen der Arbeit deuten also zum einen den Bedarf an die Politik an, eine Regulierung von KI voranzutreiben und Regulierungsinstitutionen einzusetzen, die stellvertretend für Endnutzende die Vertrauenswürdigkeit von KI-Systemen zertifizieren. Ein System, das zu komplex ist, muss nicht von allen durchschaut werden. Bei Fahrzeugen auf deutschen Straßen gibt es den TÜV als eingesetzte Prüfstelle; ähnliches bedarf es, um auf europäischen „Digitalstraßen“ unterwegs sein zu dürfen (Kapitel 7.4.2.2) – und ist auch im AI Act vorgesehen (Verordnung über künstliche Intelligenz, 2024). Darüber hinaus ergeben sich Implikationen für Entwickler*innen von KI, die ausführlich in Kapitel 7.4.2.1 erläutert werden. Aufgrund ihrer Subjektivität ist es empfehlenswert, bei der Implementierung von Transparenz stets einen nutzendenzentrierten Ansatz zu verfolgen. Außerdem umfassen die Implikationen die Auseinandersetzung mit verschiedenen Nutzungsphasen, Systemanteilen, Kontexten und Nutzungsgruppen. Eine entwickelte Transparenzmatrix dient dazu, aus wahrgenommenen Systemeigenschaften Implikationen an KI-Transparenz ableiten zu können. Um Transparenz bestmöglich umzusetzen, ist darüber hinaus eine AI Literacy vonnöten, die einerseits von Entwickler*innen durch die Umsetzung transparenter KI wie auch durch eine entsprechende Bildungspolitik gefördert werden sollte.

Die vorliegende Arbeit zeigt, dass sich das Verständnis von KI-Transparenz von Laien deutlich von dem bisher häufig untersuchten technischen Ansatz unterscheidet. Es zeigen sich kleine, positive Effekte von KI-Transparenz auf die Nutzung. Gleichzeitig sind das Verständnis dieser Transparenz und die Verlässlichkeit von Systemen sowie Vorannahmen, Erfahrung und mögliche Auswirkungen von KI-Fehlern wichtige Faktoren bei der Entscheidung, ob einem System vertraut und es genutzt wird.

Weiterhin bestehen eine große Unsicherheit und der Bedarf nach vertrauensstärkenden Maßnahmen in KI-Systeme. Einerseits können heutzutage selbst Personen ohne Fachkenntnis oder Interesse kaum mehr vermeiden, KI-Systeme zu verwenden. Die Umsetzung von KI-Regularien ist also dringlich. Andererseits sind Nutzende von KI-Systemen gefragt, ein grundlegendes Verständnis von KI aufzubauen und sich kritisch mit vorliegenden Systemen auseinanderzusetzen. Am wichtigsten ist für beide Faktoren jedoch die Bereitschaft von Entwickler*innen und Anbietern von KI, solche Systeme zu entwickeln, die die Nutzenden bestmöglich in ihren Entscheidungen unterstützen.

Transparenz in KI verbessert also nicht zwangsläufig die Nutzung von oder das Vertrauen in KI-Systeme durch Laien. Vielmehr muss sie für Auditor*innen nachvollziehbar gestaltet sein, den vorgegebenen Gesetzen entsprechen und für Nutzende verständlich umgesetzt werden, um ihnen als Werkzeug zu dienen.

9. Fazit mit Ausblick

Transparenz in KI ist nötig, um eine informierte und damit autonome Nutzung zu gewährleisten. Aus diesem Grund fordern nicht nur Nutzende und Verbraucherinstitutionen Regularien bezüglich KI, sondern auch die EU beschloss 2024 den AI Act (Verordnung über künstliche Intelligenz, 2024). Auch wenn hiermit erstmals ein Gesetz für die Kategorisierung und Kontrolle von KI-Systemen besteht, bleibt abzuwarten, wie Transparenzmaßnahmen von KI-Anbietern umgesetzt werden und ob sie Endnutzenden Verständlichkeit und eine echte Nutzungsabwägung ermöglichen. Denn wie diese Arbeit gezeigt hat: Transparenz führt nicht zwangsläufig zu Verständnis, Verständnis wiederum nicht zwangsläufig zu Vertrauen und Vertrauen bedeutet nicht automatisch Nutzung. Vielmehr bedarf es einer sinnvollen Umsetzung von KI-Transparenz, die ein Verständnis fördert, das wiederum Vertrauen ermöglicht und eine daraus resultierende, informierte Nutzung anregen kann.

Das Thema KI-Transparenz lediglich aus technischer Perspektive zu betrachten, verbessert weder Nutzung noch Vertrauen von KI-Nutzenden. Jedoch ist sie nötig, um Expert*innen zu ermöglichen, KI-Systeme zu prüfen und zu regulieren. Jedoch sind die großen Anbieter von KI sehr zurückhaltend: Während der Zusammenschluss von Google, Microsoft und Apple in OpenAI Hoffnung darauf machte, dass Kräfte gebündelt und diese ohne gegenseitige Konkurrenz eingesetzt werden, zeigt sich OpenAI kaum transparenter als die einzelnen Firmen (Sullivan, 2023). Da auch die einzelnen Bestandteile des AI Acts noch einige Jahre benötigen, bis sie umgesetzt sind, startete die EU 2024 einen „AI Pact“, in dem Unternehmen ihre Bereitschaft erklären, sich bereits vor In-Kraft-Treten an die Regularien des Gesetzes zu halten. Während einige große KI-Anbieter wie Google, Microsoft oder Open AI dem freiwilligen Pakt beigetreten sind, halten sich Meta und Apple zurück (Lomas, 2024). Kurzfristig wird sich also Transparenz von KI nur durch freiwillige Initiativen umsetzen lassen.

Die Schlüsse, die aus der Datenschutzgrundverordnung als letzte große Gesetzesinitiative im Digitalbereich gezogen werden können, sind zwiespältig. Einerseits lassen neue Regularien eine Überwachung von Anbietern und datenverarbeitenden Stellen, einschließlich des Rechts des Individuums, die eigenen Daten einzusehen und löschen zu lassen. Gleichzeitig zeigte die Umsetzung des Gesetzes an vielen Stellen, dass Methoden der Datenverarbeitung für Endnutzende nicht nachvollziehbarer geworden sind und seitenlange Texte über den durchgeführten Datenschutz von den wenigsten gelesen oder gar verstanden werden. In Bezug auf Transparenzmaßnahmen bei KI bleibt zu hoffen, dass Endnutzende bei ihrer Umsetzung einbezogen und bei der Beurteilung von KI berücksichtigt werden.

Transparenz ist kein Selbstläufer, sondern bedarf des Einsatzes von KI-Entwickler*innen, Nutzenden, der Politik und der Wissenschaft. Die vorliegende Arbeit durfte in dieser Hinsicht einen Betrag leisten

und Ansatzpunkte zur Umsetzung einer nutzendenzentrierten KI-Transparenz entwickeln. So lässt sich zum Eingangszitat über die zukünftig enorme Reichweite von KI das Folgende ergänzen, das angesichts der aktuellen KI-Entwicklungen und der vorgestellten Ergebnisse Mahnung und Auftrag zugleich darstellt:

„Man braucht nichts im Leben zu fürchten, man muss nur alles verstehen."

Marie Curie (1867-1934)

Literatur

AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

AI HLEG. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. European Commission. https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Ajzen, I. (2014). The theory of planned behaviour is alive and well, and not ready to retire [Peer commentary on the paper "Time to retire the theory of planned behaviour" by FF Sniehotta, J. Presseau, & V. Araújo-Soares]. *Health Psychology Review*, *9*(2), 131–137.

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888–918. https://doi.org/10.1037/0033-2909.84.5.888

Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making*, *21*(178). https://doi.org/10.1186/s12911-021-01542-6

Albert, J. (2023, Februar 16). *Platforms' promises to researchers: First reports missing the baseline*. AlgorithmWatch. https://algorithmwatch.org/en/platforms-promises-to-researchers/

Alexander, V., Blinder, C., & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior*, *89*, 279–288. https://doi.org/10.1016/j.chb.2018.07.026

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Alter, S. (1980). *Decision Support Systems: Current Practice and Continuing Challenges*. Addison-Wesley Pub.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Andrews, R. W., Lilly, J. M., Srivastava, D., & Feigh, K. M. (2023). The role of shared mental models in human-AI teams: A theoretical review. *Theoretical Issues in Ergonomics Science*, *24*(2), 129–175. https://doi.org/10.1080/1463922X.2022.2061080

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5), e1424. https://doi.org/10.1002/widm.1424

Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005). Towards improving trust in context-aware systems by displaying system confidence. *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, 9–14. https://doi.org/10.1145/1085777.1085780

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Ball, C. (2009). What is transparency? *Public Integrity*, *11*(4), 293–308. https://doi.org/10.2753/PIN1099-9922110400

Baniecki, H., & Biecek, P. (2019). modelStudio: Interactive Studio with Explanations for ML Predictive Models. *Journal of Open Source Software*, *4*(43), 1798. https://doi.org/10.21105/joss.01798

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

Barth, S., & de Jong, M. D. T. (2017). The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review. *Telematics and Informatics*, *34*(7), 1038–1058. https://doi.org/10.1016/j.tele.2017.04.013

Bedué, P., & Fritzsche, A. (2022). Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management*, *35*(2), 530–549. https://doi.org/10.1108/JEIM-06-2020-0233

Beier, G. (1999). Locus of control when interacting with technology [Kontrollüberzeugungen im Umgang mit Technik]. *Report Psychologie*, *24*(9), 684–693.

Beier, G. (2004). *Kontrollüberzeugung im Umgang mit Technik. Ein Persönlichkeitsmerkmal mit Relevanz für die Gestaltung technischer Systeme*. dissertation.de - Verlag im Internet.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—Algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, *63*, 55–68. https://doi.org/10.1007/s12599-020-00678-5

Berger, C. R., & Calabrese, R. J. (1975). Some Explorations in Initial Interaction and Beyond: Toward a Developmental Theory of Interpersonal Communication. *Human Communication Research*, *1*(2), 99–112. https://doi.org/10.1111/j.1468-2958.1975.tb00258.x

Bertino, E., Merrill, S., Nesen, A., & Utz, C. (2019). Redefining data transparency: A multidimensional approach. *Computer*, *52*(1), 16–26. https://doi.org/10.1109/MC.2018.2890190

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 78–91. https://doi.org/10.1145/3514094.3534164

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657. https://doi.org/10.1145/3351095.3375624

Bierhoff, H.-W., & Rohmann, E. (2010). Psychologie des Vertrauens. In M. Maring, *Vertrauen—Zwischen sozialem Kitt und der Senkung von Transaktionskosten* (S. 71–89). KIT Scientific Publishing. https://doi.org/10.26530/OAPEN_422381

Bless, H., Wänke, M., Bohner, G., Fellhauer, R., & Schwarz, N. (1994). Need for cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben [Presentation and validation of a German version of the Need for Cognition Scale]. *Zeitschrift für Sozialpsychologie*, *25*, 147–154.

Bock, N., & Rosenthal-von der Pütten, A. M. (2023). Exploring the contextuality of attitudes towards algorithmic decision-making: Validation of the newly developed universal attitudes towards algorithms scale (ATAS). *Presentation at the HMC Pre-Conference of the 73rd Annual International Communication Association (ICA) Conference*.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bonaccio, S., & Dalal, R. S. (2010). Evaluating advisors: A policy-capturing study under conditions of complete and missing information. *Journal of Behavioral Decision Making*, *23*(3), 227–249. https://doi.org/10.1002/bdm.649

Bond, S. (2023, Juni 2). YouTube will no longer take down false claims about U.S. elections. *NPR*. https://www.npr.org/2023/06/02/1179864026/youtube-will-no-longer-take-down-false-claims-about-u-s-elections

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, *33*(1), 1–30. https://doi.org/10.1007/s12525-023-00644-5

Brauner, P., Glawe, F., Liehner, G. L., Vervier, L., & Ziefle, M. (2024). *Mapping public perception of artificial intelligence: Expectations, risk-benefit tradeoffs, and value as determinants for societal acceptance* (arXiv:2411.19356). arXiv. https://doi.org/10.48550/arXiv.2411.19356

Brauner, P., Hick, A., Philipsen, R., & Ziefle, M. (2023). What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI. *Frontiers in Computer Science*, *5*. https://www.frontiersin.org/articles/10.3389/fcomp.2023.1113903

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Buschmeyer, K., Jahn, S., Hatfield, S., Markus, A., Münker, S., Daling, L. M., Werz, J. M., & Borowski, E. (2024). *Die Zukunft gestalten: Ein praxisnaher Leitfaden zur erfolgreichen Integration von*

intelligenten Entscheidungsassistenzen am Arbeitsplatz (S. 40). Technische Hochschule Augsburg; DOI: 10.60524/OPUS-1814. https://opus4.kobv.de/opus4-hs-augsburg/1814

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. *Information Processing & Management*, *60*(1), 103111. https://doi.org/10.1016/j.ipm.2022.103111

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. https://doi.org/10.1145/2783258.2788613

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*, 832. https://doi.org/10.3390/electronics8080832

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: A systematic literature review. *International Journal of STEM Education*, *10*(1), 29. https://doi.org/10.1186/s40594-023-00418-7

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Chen, T.-W., & Sundar, S. S. (2018). „This app would like to use your current location to better serve you“: Importance of user assent and system transparency in personalized mobile services. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3174111

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300789

Chiba, N. (2022, November 4). People cannot distinguish between human-made and AI-generated haiku: Japan study. *Mainichi Daily News*. https://mainichi.jp/english/articles/20221104/p2a/00m/0sc/012000c

Chiou, E. K., & Lee, J. D. (2023). Trusting Automation: Designing for Responsivity and Resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Chromik, M., Eiband, M., Völkel, S. T., & Buschek, D. (2019, März 20). Dark patterns of explainability, transparency, and user control for intelligent systems. *IUI Workshops '19*. ACM Conference on Intelligent User Interfaces (ACM IUI), Los Angeles, USA.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cooper, A. (1999). *The inmates are running the asylum* (First Printing Edition). Sams Publishing.

Corves, A., & Schön, E.-M. (2020). Digital Trust für KI-basierte Mensch-Maschine-Schnittstellen. In S. Boßow-Thies, C. Hofmann-Stölting, & H. Jochims (Hrsg.), *Data-driven Marketing: Insights aus Wissenschaft und Praxis* (S. 257–281). Springer Fachmedien. https://doi.org/10.1007/978-3-658-29995-8_12

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496. https://doi.org/10.1007/s11257-008-9051-3

Cramer, S., Huber, M., & Schmitt, R. H. (2022). Uncertainty quantification based on Bayesian Neural Networks for predictive quality. In A. Steland & K.-L. Tsui (Hrsg.), *Artificial Intelligence, Big Data and Data Science in Statistics: Challenges and Solutions in Environmetrics, the Natural Sciences and Technology* (S. 253–268). Springer International Publishing. https://doi.org/10.1007/978-3-031-07155-3_10

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—Why and how. *Knowledge-Based Systems*, *6*(4), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

Dang, S. (2023, Januar 28). *Twitter research group stall complicates compliance with new EU law*. Euronews. https://www.euronews.com/next/2023/01/28/twitter-moderation-insight

Das, A., & Rad, P. (2020). *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey* (arXiv:2006.11371). arXiv. https://doi.org/10.48550/arXiv.2006.11371

Das, M. R. (2024, Juni 17). Smart home devices a privacy nightmare, Amazon Alexa, Google Home worst offenders, finds study. *Firstpost*. https://www.firstpost.com/tech/smart-home-devices-a-privacy-nightmare-amazon-alexa-google-home-worst-offenders-finds-study-13783175.html

Daschner, S., & Obermaier, R. (2022). Algorithm aversion? On the influence of advice accuracy on trust in algorithmic advice. *Journal of Decision Systems*, *31*(1), 77–97. https://doi.org/10.1080/12460125.2022.2070951

Dastin, J. (2018, Oktober 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*(2), 95–106. https://doi.org/10.1037/h0037613

de Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology*, *31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

Dey, S., Chakraborty, P., Kwon, B. C., Dhurandhar, A., Ghalwash, M., Suarez Saiz, F. J., Ng, K., Sow, D., Varshney, K. R., & Meyer, P. (2022). Human-centered explainability for life sciences, healthcare, and medical informatics. *Patterns*, *3*(5), 100493. https://doi.org/10.1016/j.patter.2022.100493

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62. https://doi.org/10.1145/2844110

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism*, *5*(7), 809–828. https://doi.org/10.1080/21670811.2016.1208053

Dietvorst, B. J., & Bharti, S. (2019). *People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error* (SSRN Scholarly Paper ID 3424158). Social Science Research Network. https://doi.org/10.2139/ssrn.3424158

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology*, *18*(6), 399–411. https://doi.org/10.1080/014492999118832

Dominguez, V., Messina, P., Donoso-Guzmán, I., & Parra, D. (2019). The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 408–416. https://doi.org/10.1145/3301275.3302274

Donadello, I., & Dragoni, M. (2021). Bridging signals to natural language explanations with explanation graphs. In C. Musto, R. Guidotti, A. Monreale, & G. Semeraro (Hrsg.), *CEUR Workshop Proceedings* (Bd. 3014). https://ceur-ws.org/Vol-3014/paper1.pdf

Dragoni, M., Donadello, I., & Eccher, C. (2020). Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice. *Artificial Intelligence in Medicine*, *105*, 101840. https://doi.org/10.1016/j.artmed.2020.101840

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*, 428–442. https://doi.org/10.1016/j.trc.2019.05.025

Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., & Williams, M. D. (2019). Re-examining the Unified Theory of Acceptance and Use of Technology (UTAUT): Towards a revised theoretical model. *Information Systems Frontiers*, *21*(3), 719–734. https://doi.org/10.1007/s10796-017-9774-y

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2024). The Who in XAI: How AI background shapes perceptions of AI explanations. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–32. https://doi.org/10.1145/3613904.3642474

Eslami, M., Krishna Kumaran, S. R., Sandvig, C., & Karahalios, K. (2018). Communicating algorithmic process in online behavioral advertising. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3174006

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). „I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–162. https://doi.org/10.1145/2702123.2702556

European Parliament. (2024, März 13). Artificial Intelligence Act: MEPs adopt landmark law [Press Releases]. *News*. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

EY. (2022, Juni 16). EY announces US$1b investment in a next generation technology platform to facilitate trust, transparency and transformation through assurance services [Press Releases]. *EY Press Release*. https://www.ey.com/en_gl/news/2022/06/ey-announces-us-1b-investment-in-a-next-generation-technology-platform-to-facilitate-trust-transparency-and-transformation-through-assurance-services

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 1–14. https://doi.org/10.1177/2053951719860542

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2023). The extent of algorithm aversion in decision-making situations with varying gravity. *PLOS ONE*, *18*(2), e0278751. https://doi.org/10.1371/journal.pone.0278751

Fischer, S., & Petersen, T. (2018). *Was Deutschland über Algorithmen weiß und denkt*. Bertelsmann Stiftung.

Flyverbom, M. (2016). Transparency: Mediation and the management of visibilities. *International Journal of Communication*, *10*, 110–122.

Ford, C., Kenny, E. M., & Keane, M. T. (2020). Play MNIST for me! User studies on the effects of post-hoc, example-based explanations & error rates on debugging a deep learning, black-box classifier. *Proceedings of IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI)*. IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI), online. https://doi.org/10.48550/arXiv.2009.06349

Forssbaeck, J., & Oxelheim, L. (2014). *The multi-faceted concept of transparency* (Working Paper Series 1013, S. 46). Research Institute of Industrial Economics.

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020). Fostering human agency: A process for the design of user-centric XAI systems. *ICIS 2020 Proceedings*, 12.

Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in Practice*, *17*(4–5), 663–671. https://doi.org/10.1080/09614520701469955

Frieman, J., Saucier, D. A., & Miller, S. S. (2018). *Principles & methods of statistical analysis*. SAGE Publications, Inc. https://methods.sagepub.com/book/principles-and-methods-of-statistical-analysis

Fung, A., Graham, M., & Weil, D. (2007). *Full disclosure: The perils and promise of transparency*. Cambridge University Press.

García, P. G., Costanza, E., Verame, J., Nowacka, D., & Ramchurn, S. D. (2021). Seeing (movement) is believing: The effect of motion on perception of automatic systems performance. *Human–Computer Interaction*, *36*(1), 1–51. https://doi.org/10.1080/07370024.2018.1453815

Gedikli, F., Jannach, D., & Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, *72*(4), 367–382. https://doi.org/10.1016/j.ijhcs.2013.12.007

Gerlings, J., Jensen, M. S., & Shollo, A. (2022). Explainable AI, but explainable to whom? An exploratory case study of xAI in healthcare. In C.-P. Lim, Y.-W. Chen, A. Vaidya, C. Mahorkar, & L. C. Jain (Hrsg.), *Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and*

Prospects (S. 169–198). Springer International Publishing. https://doi.org/10.1007/978-3-030-83620-7_7

Gesellschaft für Informatik. (2022, März 7). *Integratives Framework für KI-Audits veröffentlicht*. https://gi.de/meldung/integratives-framework-fuer-ki-audits-veroeffentlicht

Glikson, E., & Woolley, A. W. (2020). Human trust in Artificial Intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Google. (2024). *About targeting for video campaigns*. YouTube Help. https://support.google.com/youtube/answer/2454017

Gosiewska, A., Kozak, A., & Biecek, P. (2021). Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, *150*, 113556. https://doi.org/10.1016/j.dss.2021.113556

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274.

Grigutyté, M. (2023, Juli 12). NordVPN reveals: Americans using ChatGPT trust the chatbot. *NordVPN*. https://nordvpn.com/blog/chatgpt-usage-in-the-us/

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Gu, K., Grunde-McLaughlin, M., McNutt, A., Heer, J., & Althoff, T. (2024). How do data analysts respond to AI assistance? A wizard-of-Oz study. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–22. https://doi.org/10.1145/3613904.3641891

Gubaydullina, Z., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021). *Creative drive and algorithm aversion – The impact of influence in the process of algorithmic decision-making on algorithm aversion* (21–04; Wolfsburg Working Papers, S. 27). Ostfalia Hochschule für angewandte Wissenschaften. https://www.ostfalia.de/cms/de/w/.galleries/forschung/fakw_WWP_21-04-Creative-Drive-and-Algorithm-Aversion.pdf

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, *2*(4), 1–11. https://doi.org/10.1002/ail2.61

Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G., & Aggarwal, C. (2019). Efficient data representation by selecting prototypes with importance weights. *2019 IEEE International Conference on Data Mining (ICDM)*, 260–269. https://doi.org/10.1109/ICDM.2019.00036

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. https://doi.org/10.1006/obhd.1997.2697

Haupt, M., Freidank, J., & Haas, A. (2024). Consumer responses to human-AI collaboration at organizational frontlines: Strategies to escape algorithm aversion in content creation. *Review of Managerial Science*. https://doi.org/10.1007/s11846-024-00748-y

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, 241–250. https://doi.org/10.1145/358916.358995

Herm, L.-V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, 102538. https://doi.org/10.1016/j.ijinfomgt.2022.102538

Herm, L.-V., Wanner, J., Seubert, F., & Janiesch, C. (2021). I don't get it, but it seems valid! The connection between explainability and comprehensibility in (X)AI research. *ECIS 2021 Research Papers*, 82.

Herrmann, T., & Pfeiffer, S. (2023). Keeping the organization in the loop: A socio-technical extension of human-centered artificial intelligence. *AI & SOCIETY*, *38*(4), 1523–1542. https://doi.org/10.1007/s00146-022-01391-5

Hind, M., Wei, D., Campbell, M., Codella, N. C. F., Dhurandhar, A., Mojsilović, A., Natesan Ramamurthy, K., & Varshney, K. R. (2019). TED: Teaching AI to explain its decisions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 123–129. https://doi.org/10.1145/3306618.3314273

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hohman, F., Srinivasan, A., & Drucker, S. M. (2019). TeleGam: Combining visualization and verbalization for interpretable machine learning. *2019 IEEE Visualization Conference (VIS)*, 151–155. https://doi.org/10.1109/VISUAL.2019.8933695

Hoomans, D. J. (2015, März 20). 35,000 decisions: The great choices of strategic seaders. *Leading Edge*. https://go.roberts.edu/leadingedge/the-great-choices-of-strategic-leaders

Hundt, A., Agnew, W., Zeng, V., Kacianka, S., & Gombolay, M. (2022). Robots enact malignant stereotypes. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 743–756. https://doi.org/10.1145/3531146.3533138

Hütter, M., & Fiedler, K. (2019). Advice taking under uncertainty: The impact of genuine advice versus arbitrary anchors on judgment. *Journal of Experimental Social Psychology*, *85*, 103829. https://doi.org/10.1016/j.jesp.2019.103829

IEEE SA. (2022). *IEEE CertifAIEd. The Mark of AI Ethics*. https://engagestandards.ieee.org/ieeecertifaied.html

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

Joinson, A., Reips, U.-D., Buchanan, T., & Paine Schofield, C. (2010). Privacy, trust, and self-disclosure online. *Human-Computer Interaction*, *25*, 1–24. https://doi.org/10.1080/07370020903586662

Kahneman, D., Slovic, S. P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Karran, A. J., Demazure, T., Hudon, A., Senecal, S., & Léger, P.-M. (2022). Designing for confidence: The impact of visualizing Artificial Intelligence decisions. *Frontiers in Neuroscience*, *16*, 883385. https://doi.org/10.3389/fnins.2022.883385

Khodaei, S., Padev, M., Abdelrazeq, A., & Isenhardt, I. (2023). *Beyond data literacy in engineering education. How media literacy can enhance data literacy*. ing.grid preprints; ing.grid. https://preprints.inggrid.org/repository/view/7/

Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). „Help me help the AI“: Understanding how explainability can support human-AI interaction. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3544548.3581001

Kim, T., & Song, H. (2020). The effect of message framing and timing on the acceptance of Artificial Intelligence's suggestion. *CHI '20 Extended Abstracts*, 1–8. https://doi.org/10.1145/3334480.3383038

Klein, A. (2023, Mai 10). AI Literacy, explained. *Education Week*. https://www.edweek.org/technology/ai-literacy-explained/2023/05

Knuth, D. E. (1997). *The art of computer programming. Fundamental algorithms, Volume 1* (3. Aufl.). Addison-Wesley.

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300641

Koivisto, I. (2016). The anatomy of transparency: The concept and its multifarious implications. *EUI Working Paper MWP*, *19*. https://cadmus.eui.eu//handle/1814/41166

Kramer, M. W. (1999). Motivation to reduce uncertainty: A reconceptualization of Uncertainty Reduction Theory. *Management Communication Quarterly*, *13*(2), 305–316. https://doi.org/10.1177/0893318999132007

Kuckartz, U. (2010). *Einführung in die computergestützte Analyse qualitativer Daten* (3. Aufl.). VS Verlag für Sozialwissenschaften.

Laato, S., Tiainen, M., Najmul Islam, A. K. M., & Mäntymäki, M. (2022). How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research*, *32*(7), 1–31. https://doi.org/10.1108/INTR-08-2021-0600

Larsson, S. (2019). The socio-legal relevance of artificial intelligence. *Droit et Societe*, *103*, 573–593. https://portal.research.lu.se/en/publications/the-socio-legal-relevance-of-artificial-intelligence

Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., & Cedering Angström, R. (2019). *Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges*

related to artificial intelligence. AI Sustainability Center. https://portal.research.lu.se/en/publications/e2fa1b6a-860e-44b0-a359-fbd842c363db

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2). https://policyreview.info/concepts/transparency-artificial-intelligence

Legg, S., & Hutter, M. (2007). A Collection of Definitions of Intelligence. *Proceedings of the 2007 conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006*, 17–24.

Lehmann, C. A., Haubitz, C. B., Fügener, A., & Thonemann, U. W. (2020). Keep it mystic? – The effects of algorithm transparency on the use of advice. *ICIS 2020 Proceedings*. International Conference on Information Systems, India. https://aisel.aisnet.org/icis2020/hci_artintel/hci_artintel/3/

Lenhard, A., & Lenhard, W. (2017). *Berechnung von Effektstärken.* Psychometrica. https://www.psychometrica.de/effektstaerke.html

Li, W., Shao, W., Ji, S., & Cambria, E. (2022). BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, *467*, 73–82. https://doi.org/10.1016/j.neucom.2021.09.057

Lim, B. Y., & Dey, A. K. (2011). Investigating intelligibility for uncertain context-aware applications. *Proceedings of the 13th International Conference on Ubiquitous Computing*, 415–424. https://doi.org/10.1145/2030112.2030168

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43. https://doi.org/10.1145/3233231

Littman, M. L., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi-Velez, F., Hadfield, G., Horowitz, M. C., Isbell, C., Kitano, H., Levy, K., Lyons, T., Mitchell, M., Shah, J., Sloman, S., Vallor, S., & Walsh, T. (2021). *Gathering strength, gathering storms: The one hundred year study on Artificial Intelligence (AI100) 2021 Study Panel report*. Stanford University. https://ai100.stanford.edu/gathering-strength-gathering-storms-one-hundred-year-study-artificial-intelligence-ai100-2021-study

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lohoff, L., & Rühr, A. (2021). Introducing (machine) learning ability as antecedent of trust in intelligent systems. *ECIS 2021 Research Papers*, *23*.

Löhr, K., Weinhardt, M., & Sieber, S. (2020). The "World Café" as a participatory method for collecting qualitative data. *International Journal of Qualitative Methods*, *19*, 1609406920916976. https://doi.org/10.1177/1609406920916976

Lomas, N. (2024, September 25). Early sign-ups to EU's AI Pact include Amazon, Google, Microsoft, and OpenAI — but Apple and Meta are missing. *TechCrunch*. https://techcrunch.com/2024/09/25/early-sign-ups-to-eus-ai-pact-include-amazon-google-microsoft-and-openai-but-apple-and-meta-are-missing/

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2020). Resistance to medical artificial intelligence is an attribute in a compensatory decision process: Response to Pezzo and Beckstead. *Judgment and Decision Making*, *15*(3), 446–448.

Lucic, A., Haned, H., & de Rijke, M. (2020). Why does my model fail? Contrastive local explanations for retail forecasting. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 90–98. https://doi.org/10.1145/3351095.3372824

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

Markus, A., Klaka, K., Werz, J. M., Borowski, E., & Isenhardt, I. (2024). Application of participatory design in the development of a front-end for an AI-based decision support system with two companies. In A. Marcus, E. Rosenzweig, & M. M. Soares (Hrsg.), *Design, User Experience, and Usability* (Bd. 14713, S. 76–85). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-61353-1_5

Maye, H. (2023). Ist Medienkompetenz Bullshit? *Zeitschrift für Medienwissenschaften*, *29*, 137–143. https://doi.org/10.25969/MEDIAREP/20053

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*, 709–734.

Mayring, P. (2002). *Einführung in die qualitative Sozialforschung [Introduction to qualitative social research]* (5. Aufl.). Beltz.

Mayring, P. (2010). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (12. Aufl.). Beltz.

McCarthy, J. (2017, November 12). *Basic Questions*. What is artificial intelligence? http://www-formal.stanford.edu/jmc/whatisai/node1.html

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1956). Proceedings of the Dartmouth Conference on Artificial Intelligence. In J. McCarthy, M. L. Minsky, N. Rochester, & C. E. Shannon (Hrsg.), *A proposal for the Dartmouth Summer Research Project on artificial intelligence*.

McKnight, H., & Carter, M. (2009). Trust in technology: Development of a set of constructs and measures. *DIGIT 2009 Proceedings*. DIGIT 2009.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, 1097–1101. https://doi.org/10.1145/1125451.1125659

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press. https://doi.org/10.1037/11281-000

Miller, T. (2018). *Explanation in Artificial Intelligence: Insights from the Social Sciences* (arXiv:1706.07269 [cs]). arXiv. http://arxiv.org/abs/1706.07269

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems*, *11*, 24:1-24:45. https://doi.org/10.1145/3387166

Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, *27*(4), zmac010. https://doi.org/10.1093/jcmc/zmac010

Molnar, C. (2019). *Interpretable Machine Learning* (1st edition). Christoph Molnar (CC Attribution 2.0). https://christophm.github.io/interpretable-ml-book/index.html

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, *26*, 824–826. https://doi.org/10.1016/j.tics.2022.07.007

Mozilla. (2021). *YouTube Regrets Report. A crowdsourced investigation into YouTube's recommendation algorithm*. https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf

Mueller, T., Huber, M., & Schmitt, R. (2020). Modelling complex measurement processes for measurement uncertainty determination. *International Journal of Quality & Reliability Management*, *37*, 494–516. https://doi.org/10.1108/IJQRM-07-2019-0232

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*, 22071–22080. https://doi.org/10.1073/pnas.1900654116

Musto, C., Narducci, F., Lops, P., de Gemmis, M., & Semeraro, G. (2019). Linked open data-based explanations for transparent recommender systems. *International Journal of Human-Computer Studies*, *121*, 93–107. https://doi.org/10.1016/j.ijhcs.2018.03.003

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Nilashi, M., Jannach, D., Ibrahim, O. bin, Esfahani, M. D., & Ahmadi, H. (2016). Recommendation quality, transparency, and website quality for trust-building in recommendation agents.

Electronic Commerce Research and Applications, *19*, 70–84. https://doi.org/10.1016/j.elerap.2016.09.003

Nix, N., & Ellison, S. (2023, August 25). Following Elon Musk's lead, Big Tech is surrendering to disinformation. *The Washington Post*. https://www.washingtonpost.com/technology/2023/08/25/political-conspiracies-facebook-youtube-elon-musk/

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin UK.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

OpenAI. (2022, November 30). *Introducing ChatGPT*. OpenAI. https://openai.com/blog/chatgpt

OpenAI. (2023). *DALL·E 2*. OpenAI. https://openai.com/product/dall-e-2

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Pálfi, B., Arora, K., & Kostopoulou, O. (2022). Algorithm-based advice taking and clinical judgement: Impact of advice distance and algorithm information. *Cognitive Research: Principles and Implications*, *7*(1), 70. https://doi.org/10.1186/s41235-022-00421-6

Peters, F., Pumplun, L., & Buxmann, P. (2020). Opening the black box: Consumer's willingness to pay for transparency of intelligent systems. *ECIS 2020 Proceedings*, 1–15. https://aisel.aisnet.org/ecis2020_rp/90

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of persuasion. In R. E. Petty & J. T. Cacioppo (Hrsg.), *Communication and Persuasion: Central and Peripheral Routes to Attitude Change* (S. 1–24). Springer. https://doi.org/10.1007/978-1-4612-4964-1_1

Philipsen, R., Brauner, P., Biermann, H., & Ziefle, M. (2022). I Am what I am – Roles for artificial intelligence from the users' perspective. *Artificial Intelligence and Social Computing Volume 28*. 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022). https://doi.org/10.54941/ahfe1001453

Posada-Moreno, A. F., Surya, N., & Trimpe, S. (2023). *ECLAD: Extracting Concepts with Local Aggregated Descriptors* (arXiv:2206.04531). arXiv. http://arxiv.org/abs/2206.04531

Power, D. J. (2002). *Decision support systems: Concepts and resources for managers*. Quorum Books.

Prahl, A., & Swol, L. V. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*, 691–702. https://doi.org/10.1002/for.2464

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*(1), 39–57. https://doi.org/10.1002/bdm.460

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173677

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Ras, G., Xie, N., Gerven, M. van, & Doran, D. (2022). Explainable Deep Learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, *73*, 329–396. https://doi.org/10.1613/jair.1.13200

Regulation (EU) 2022/2065 of the European Parliament and of the Council on a Single Market For Digital Services and Amending Directive 2000/31/EC, 277 OJ L (2022). http://data.europa.eu/eli/reg/2022/2065/oj/eng

Reich, T., Kaju, A., & Maglio, S. J. (2022). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, *33*(2), 1–18. https://doi.org/10.1002/jcpy.1313

Reis, M., Reis, F., & Kunde, W. (2024). Influence of believed AI involvement on the perception of digital medical advice. *Nature Medicine, 30*, 3098–3100. https://doi.org/10.1038/s41591-024-03180-7

Renier, L. A., Mast, M. S., & Bekbergenova, A. (2021). To err is human, not algorithmic – Robust reactions to erring algorithms. *Computers in Human Behavior*, 106879. https://doi.org/10.1016/j.chb.2021.106879

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *„Why should I trust you?": Explaining the predictions of any classifier* (1602.04938). arXiv. http://arxiv.org/abs/1602.04938

Ribes, D., Henchoz, N., Portier, H., Defayes, L., Phan, T.-T., Gatica-Perez, D., & Sonderegger, A. (2021). Trust indicators and explainable AI: A study on user perceptions. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Hrsg.), *Human-Computer Interaction—INTERACT 2021, LNCS 12933* (Bd. 12933, S. 662–671). Springer International Publishing. https://doi.org/10.1007/978-3-030-85616-8_39

Roets, A., & Van Hiel, A. (2007). Separating ability from need: Clarifying the dimensional structure of the Need for Closure scale. *Personality and Social Psychology Bulletin*, *33*(2), 266–280. https://doi.org/10.1177/0146167206294744

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences*, *50*(1), 90–94. https://doi.org/10.1016/j.paid.2010.09.004

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *arXiv:1811.10154 [Cs, Stat]*. http://arxiv.org/abs/1811.10154

Sadiku, M. N. O., & Musa, S. M. (2021). Augmented Intelligence. In M. N. O. Sadiku & S. M. Musa (Hrsg.), *A Primer on Multiple Intelligences* (S. 191–199). Springer International Publishing. https://doi.org/10.1007/978-3-030-77584-1_15

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Hrsg.). (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Bd. 11700). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Schoeffer, J., Machowski, Y., & Kühl, N. (2022, Januar 1). Perceptions of fairness and trustworthiness based on explanations in human vs. Automated decision-making. *Proceedings of the 55th Hawaii International Conference on System Sciences*. International Conference on System Sciences. https://doi.org/10.24251/HICSS.2022.134

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press.

Shulner-Tal, A., Kuflik, T., & Kliger, D. (2023). Enhancing fairness perception – Towards human-centred AI and personalized explanations understanding the factors influencing laypeople's fairness perceptions of algorithmic decisions. *International Journal of Human–Computer Interaction*, *39*(7), 1455–1482. https://doi.org/10.1080/10447318.2022.2095705

Sieger, L. N., Hermann, J., Schomäcker, A., Heindorf, S., Meske, C., Hey, C.-C., & Doğangün, A. (2022). User involvement in training smart home agents: Increasing perceived control and understanding. *Proceedings of the 10th International Conference on Human-Agent Interaction*, 76–85. https://doi.org/10.1145/3527188.3561914

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*, 99–118. https://doi.org/10.2307/1884852

Singh, S. (2020, März 25). *Why am I seeing this? How video and e-commerce platforms use recommendation systems to shape user experiences*. New America. http://newamerica.org/oti/reports/why-am-i-seeing-this/

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*, 288–307. https://doi.org/10.1006/obhd.2000.2926

Solsman, J. E. (2018, Januar 10). YouTube's AI is the puppet master over most of what you watch. *CNET*. https://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan/

Springer, A. (2019). *Accurate, fair, and explainable: Building human-centered AI* [UC Santa Cruz]. https://escholarship.org/uc/item/4d80t5j5

Springer, A., Hollis, V., & Whittaker, S. (2018). *Dice in the black box: User experiences with an inscrutable algorithm* (arXiv:1812.03219 [cs]). arXiv. http://arxiv.org/abs/1812.03219

Springer, A., & Whittaker, S. (2018). What are you hiding? Algorithmic transparency and user perceptions. *AAAI Spring Symposium Series*, 1–4.

Staatsvertrag zur Modernisierung der Medienordnung in Deutschland (Medienstaatsvertrag), 105 (2020). https://www.daserste.de/ard/die-ard/Medienstaatsvertrag-100.pdf

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, *66*(4), 804–870.

Stewart, K. J. (2003). Trust transfer on the World Wide Web. *Organization Science*, *14*(1), 5–17. https://doi.org/10.1287/orsc.14.1.5.12810

Stohl, C., Stohl, M., & Leonardi, P. M. (2016). Managing opacity: Information visibility and the paradox of transparency in the digital age. *International Journal of Communication*, *10*, 123–137.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(2), 111–147.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzionir, O., Hager, G., Hirschberg, J., Kalyanakrishnanl, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2016). *„Artificial intelligence and life in 2030." One hundred year study on artificial intelligence: Report of the 2015-2016 study panel*. Stanford University. https://ai100.stanford.edu/2016-report/preface

Stradi, F., & Verdickt, G. (2024). *Man vs. machine: The influence of AI forecasts on investor beliefs* (SSRN Scholarly Paper 4952791). https://papers.ssrn.com/abstract=4952791

Sun, Y., & Sundar, S. S. (2022). Exploring the effects of interactive dialogue in improving user control for explainable online symptom checkers. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–7. https://doi.org/10.1145/3491101.3519668

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*, 74–88. https://doi.org/10.1093/jcmc/zmz026

Tabrez, A., Luebbers, M. B., & Hayes, B. (2022). Descriptive and prescriptive visual guidance to improve shared situational awareness in ruman-robot teaming. *Proceedings of the 21st*

International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), 9. https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p1256.pdf

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism*, *7*, 447–469. https://doi.org/10.1080/21670811.2018.1493936

Tsai, C.-H., & Brusilovsky, P. (2019). Explaining recommendations in an interactive hybrid social recommender. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 391–396. https://doi.org/10.1145/3301275.3302318

Turban, E., Sharda, R., & Delen, D. (2011). *Decision support and business intelligence systems* (9. Aufl.). Prentice Hall. http://archive.org/details/Decision-Support-And-Business-Intelligence-Systems_201808

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Utikal, V. (2020, August 13). Kopf und Bauch? Wie das Gehirn Entscheidungen trifft. *Haufe News*. https://www.haufe.de/controlling/controllerpraxis/entscheidungen-treffen-mit-kopf-oder-bauch_112_522806.html

van Nuenen, T., Ferrer, X., Such, J. M., & Cote, M. (2020). Transparency for whom? Assessing discriminatory artificial intelligence. *Computer*, *53*(11), 36–44. https://doi.org/10.1109/MC.2020.3002181

Van Swol, L. M. (2011). Forecasting another's enjoyment versus giving the right answer: Trust, shared values, task effects, and confidence in improving the acceptance of advice. *International Journal of Forecasting*, *27*(1), 103–120. https://doi.org/10.1016/j.ijforecast.2010.03.002

Van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology*, *44*(3), 443–461. https://doi.org/10.1348/014466604X17092

Vasey, M. W., & Thayer, J. F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology*, *24*(4), 479–486. https://doi.org/10.1111/j.1469-8986.1987.tb00324.x

Venkatesh, V., Brown, S. A., & Sullivan, Y. W. (2016). Guidelines for conducting mixed-methods research: An extension and illustration. *Journal of the Association for Information Systems*, *17*(7), 435–494. https://doi.org/10.17705/1jais.00433

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, *27*, 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., Chan, F. K. Y., & Hu, P. J. H. (2016). Managing citizens' uncertainty in e-government services: The mediating and moderating roles of transparency and trust. *Information Systems Research*, *27*, 87–111. https://doi.org/10.1287/isre.2015.0612

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2012). Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors*, *54*, 799–810. https://doi.org/10.1177/0018720812443825

Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG, Pub. L. No. 2016/679 (2016). https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/

Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz) (2024). http://data.europa.eu/eli/reg/2024/1689/oj/deu

Vianello, A., Laine, S., & Tuomi, E. (2023). Improving Trustworthiness of AI Solutions: A Qualitative Approach to Support Ethically-Grounded AI Design. *International Journal of Human–Computer Interaction*, *39*, 1405–1422. https://doi.org/10.1080/10447318.2022.2095478

Waltl, B., & Becker, N. (2021). *KI-Audit in der Arbeitswelt. Ein integratives Framework zum Auditieren und Testen von KI-Systemen*. Gesellschaft für Informatik e.V. https://gi.de/fileadmin/PR/Testing-AI/ExamAI_Framework_KI-Audit.pdf

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., &

Li, B. (2023). *DecodingTrust: A comprehensive assessment of trustworthiness in GPT models* (arXiv:2306.11698). arXiv. http://arxiv.org/abs/2306.11698

Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022). A social evaluation of the perceived goodness of explainability in machine learning. *Journal of Business Analytics*, *5*(1), 29–50. https://doi.org/10.1080/2573234X.2021.1952913

Wanner, J., Herm, L.-V., & Janiesch, C. (2020). How much is the Black Box? The value of explainability in machine learning models. *ECIS 2020 Research-in-Progress Papers*, *85*, 15. https://aisel.aisnet.org/ecis2020_rip/85

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). „Do you trust me?“: Increasing user-trust by integrating virtual agents in explainable AI interaction design. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 7–9. https://doi.org/10.1145/3308532.3329441

Werz, J. M., Borowski, E., & Isenhardt, I. (2020). When imprecision improves advice: Disclosing algorithmic error probability to increase advice taking from algorithms. In C. Stephanidis & M. Antona (Hrsg.), *HCI International 2020—Posters* (S. 504–511). Springer International Publishing. https://doi.org/10.1007/978-3-030-50726-8_66

Werz, J. M., Borowski, E., & Isenhardt, I. (2024). Explainability as a means for transparency? Lay users' requirements towards transparent AI. *Cognitive Computing and Internet of Things*, *124*. https://doi.org/10.54941/ahfe1004712

Werz, J. M., Zähl, K., Borowski, E., & Isenhardt, I. (2021). Preventing discrepancies between indicated algorithmic certainty and actual performance: An experimental solution. In C. Stephanidis, M. Antona, & S. Ntoa (Hrsg.), *HCI International 2021—Posters* (Bd. 1420, S. 573–580). Springer International Publishing. https://doi.org/10.1007/978-3-030-78642-7_77

Wienrich, C., Carolus, A., Markus, A., & Augustin, Y. (2022). *AI Literacy: Kompetenzdimensionen und Einflussfaktoren im Kontext von Arbeit* (S. 26). Julius-Maximilians-Universität Würzburg. https://www.denkfabrik-bmas.de/fileadmin/Downloads/Publikationen/AI_Literacy_Kompetenzdimensionen_und_Einflussfaktoren_im_Kontext_von_Arbeit.pdf

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18. https://doi.org/10.1145/3351095.3372833

Wulf, A. J., & Seizov, O. (2024). "Please understand we cannot provide further information": Evaluating content and transparency of GDPR-mandated AI disclosures. *AI & Society*, *39*), 235–256. https://doi.org/10.1007/s00146-022-01424-z

Xu, L., Pardos, Z. A., & Pai, A. (2023). Convincing the expert: Reducing algorithm aversion in administrative higher education decision-making. *Proceedings of the 10th ACM Conference on Learning @ Scale*, 215–225. https://doi.org/10.1145/3573051.3593378

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, *2*, 249–262. https://doi.org/10.1007/s41664-018-0068-2

Yaniv, I. (2004). The Benefit of Additional Opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*, 260–281. https://doi.org/10.1006/obhd.2000.2909

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300509

Yoo, S. (2018, März 10). Opinion: YouTube, the great radicalizer. *The New York Times*. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 307–317. https://doi.org/10.1145/3025171.3025219

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on*

Fairness, Accountability, and Transparency, 295–305. https://doi.org/10.1145/3351095.3372852

Zhang, Z., Genc, Y., Wang, D., Ahsen, M. E., & Fan, X. (2021). Effect of AI explanations on human perceptions of patient-facing AI-powered healthcare systems. *Journal of Medical Systems*, *45*(6), 64. https://doi.org/10.1007/s10916-021-01743-6

Zhao, R., Benbasat, I., & Cavusoglu, H. (2019). Do users always want to know more? Investigating the relationship between system transparency and users' trust in advice-giving systems. *Proceedings of the 27th European Conference on Information Systems (ECIS)*, 1–12. https://aisel.aisnet.org/ecis2019_rip/42/

Zhou, K., Fu, C., & Yang, S. (2016). Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews*, *56*, 215–225. https://doi.org/10.1016/j.rser.2015.11.050

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). *Visualizing Deep Neural Network decisions: Prediction difference analysis* (arXiv:1702.04595). arXiv. https://doi.org/10.48550/arXiv.1702.04595

Zysk, J., Florides, C., Werz, J. M., Gannouni, A., Carey, E., Wolf-Monheim, F., Borowski, E., & Isenhardt, I. (2024). Towards operator empowerment in assembly lines with human-centered design: A concept with application in the automotive industry. *Procedia CIRP 128*, 490–495. https://doi.org/10.1016/j.procir.2024.04.013

Übersicht über verwendete Hilfsmittel

Name	Wofür eingesetzt	Link/Quelle
DeepL	Übersetzung einzelner Sätze oder Zitate aus dem Englischen; Prüfung eigener Übersetzungen	https://www.deepl.com/de/translator
NapkinAI	Anregungen für Abbildungen	https://app.napkin.ai/
PowerPoint (Microsoft Office) Icons	Alle verwendeten Icons im Dokument	
RWTH GPT (ChatGPT 4)	Formulierungsalternativen	https://genai.rwth-aachen.de/app/
Scribbr KI-Rechtschreibprüfung	Rechtschreib- und Zeichensetzungskorrektur	https://app.scribbr.de/student/order/ai-proofreader/

Anhang

Anhang A

Übersicht über Nutzungsstudien zu Effekten transparenter KI

	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Herlocker et al., 2000*	System zur Empfehlung von Filmen	Unterschiedliche Interfaces von ähnlichen Empfehlungen	Bevorzugung von Interfaces; Filterleistung der Nutzenden	**Erklärungen wurden bevorzugt**; keine Auswirkung auf die Leistung der Nutzenden
Antifakos et al., 2005	Kontext-sensitives Mobiltelefon	Anzeige des Vertrauens in das System	Vertrauen auf System	Die Art der Situation (hohe bis niedrige Kritikalität) beeinflusste das Vertrauen in das System. Information über Sicherheit der Information des Systems beeinflusste Vertrauen in das System: Bei angegebener **Unsicherheit sank das Vertrauen in das System**.
H. Cramer et al., 2008*	Empfehlungs-system für Kunstwerke	Warum dieses Ergebnis vs. wie sicher vs. keine Angabe	Vertrauen, Akzeptanz, wahrgenommene Kompetenz des Systems	**Erklärung steigerte Akzeptanz des Ergebnisses**, aber kein Effekt auf Vertrauen oder Kompetenz. Keine Auswirkung von Sicherheitsinformationen.
Lim et al., 2009	Logische Denkaufgabe (kein algorith-misches System)	Lokale Erklärungen: Warum, Warum nicht, Was wäre wenn, Wie	Verständnis des Systems; Vertrauen; Leistung	Warum- und Warum nicht-Erklärungen waren am effektivsten, um Verständnis zu erhöhen; Warum führte zu höchstem Vertrauen; **frühere Erfahrungen untergruben das Vertrauen und das Verständnis, wenn die Erklärungen nicht zu diesen passten.**
Li & Gregor, 2011	E-Govern-ment-Beratungs-dienst (nicht technisches System im Fokus)	Erklärung der Begriffe; Begründungen für das Ergebnis	Zufriedenheit, Befähigung der Verbrauchenden	Wahrgenommene Transparenz und größere Zufriedenheit mit dem Verfahren, größere Eigenverantwortung des Verbrauchers: Gefühl der Kontrolle, verbesserte Einstellung zur Dienstleistungsagentur.
Lim & Dey, 2011*	Kontext-sensitive Systeme	Sicherheits-informationen; Erklärungen für das Ergebnis	Eindruck vom System; wahrgenommene Sicherheit und Angemessenheit	Informationen mit **geringer Sicherheit** verringerten die Gesamtbewertung der Systemgenauigkeit: So widersprachen Menschen **selbst in korrekten Fällen**. Bei **hoher Sicherheit verbesserten Erklärungen den Eindruck des Systems** (weiter), selbst wenn das System falsch lag.
Verberne et al., 2012	Automatisier-tes Fahren	Fahraktion plus Info vs. ohne Info; gleiche vs. unterschiedliche Fahrziele wie Nutzende	Vertrauen, Akzeptanz	Wenn **System und Nutzende die gleichen Ziele verfolgten, stieg das Vertrauen**. Fahraktionen mit Informationen darüber führten zu höherer Akzeptanz und Vertrauen als ohne Informationen.

	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Kulesza et al., 2013	Empfehlungs-system für Musik	Qualitative Studie: Vollständigkeit und Fundiertheit der Erklärung	Mentale Modelle, Vertrauen	Vollständigkeit wichtiger als Fundiertheit. Bei zu geringer Vollständigkeit eher Vertrauenseinbußen.
Gedikli et al., 2014	System zur Empfehlung von Filmen	Zehn Erklärungstypen für Empfehlung	Effizienz (Zeit), Effektivität der Erklärung, wahrgenommene Transparenz, Zufriedenheit mit Erklärung	**Höhere wahrgenommene Transparenz führte zu höherer Zufriedenheit**; inhaltsbasierte Tag-Cloud-Erklärungen waren am effektivsten und wurden gut angenommen, führten aber zu einer **höheren kognitiven Belastung (Effizienz).**
Eslami et al., 2015*	Facebook-Newsfeed-Algorithmus	Offenlegung des Algorithmus zur Kuratierung und seiner Auswirkungen	Befriedigung und Emotionen	**Offenlegung der Existenz** des Algorithmus führte zu positiven und negativen Ergebnissen: **Anfängliche Überraschung, Ärger und Unzufriedenheit**, allmähliche Zufriedenheit.
Kizilcec, 2016*	Peer-Assessment in einem MOOC	Keine vs. Berechnungs-erklärung vs. Peer-Details	Vertrauen	Die Auswirkung von Transparenz auf das Vertrauen hing von den Erwartungen der Nutzenden und deren Verletzung ab; wenn die **Erwartungen verletzt wurden, führte ein hohes Maß an Transparenz zu einem geringen Vertrauen**.
Nilashi et al., 2016	Websites für den online Handel (Amazon & Lazda)	Transparenz des Empfehlungs-prozesses, Qualität der Webseite, Empfehlungs-qualität etc.	Vertrauen, **Kaufabsicht**	Die Qualität der Website beeinflusste das Vertrauen am stärksten; Transparenz hatte einen Einfluss auf Vertrauen, aber auch viele andere Faktoren.
Venkatesh, Thong et al., 2016	E-Government-Website und Terminservice	Mehrere Variablen der Informations-qualität, Transparenz, Kanalmerkmale (z. B. Individu-alisierbarkeit)	Vertrauen, **Nutzungsabsicht**	Transparenz erhöhte (verringerte) das Vertrauen und die Nutzungsabsicht bei guter (schlechter) Informationsvollständigkeit, -bequemlichkeit oder -genauigkeit; Vertrauen und Transparenz wirkten als Moderatoren und Mediatoren.
Chen & Sundar, 2018*	Umwelt-freundliche, mobile App	Personalisierung: manuell vs. Information; Transparenz hoch vs. niedrig	Vertrauen, wahrgenommene Kontrolle, Datenschutz-bedenken, Einstellung gegenüber System	Kein Einfluss der Personalisierung; Transparenz erhöhte Vertrauen, aber keine Auswirkung auf die wahrgenommene Kontrolle; wahrgenommene Kontrolle und Einfachheit der Nutzung steigerten Vertrauen.

	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Eslami et al., 2018*	Online-Werbung	Information über Grund für Werbung (Warum), Personalisierung und Informationsnutzung	Vertrauen, subjektive Unheimlichkeit, Zufriedenheit	Zu spezifische und allgemeine Erklärungen führten zu einem Gefühl der Unheimlichkeit; Warum-Erklärungen wurden bevorzugt; Transparenz konnte zu Desillusionierung führen.
Rader et al., 2018*	Facebook-Newsfeed-Algorithmus	Erklärungen zum Newsfeed: Was, Wie, Warum, Objektiv (alles global)	Wahrnehmung: Bewusstheit über System, Korrektheit, Interpretierbarkeit, Verantwortlicheit, Zufriedenheit	Erklärungen hatten stärkste Auswirkung auf Bewusstheit, gefolgt von Verantwortlichkeit/ Kontrollmöglichkeiten; stärkster Effekt von „Was“, etwas schwächer von „Warum“.
Springer & Whittaker, 2018	Emotionserkennung aus Text (E-Meter)	Erklärungen (Text eingefärbt) vs. keine Erklärung	Wahrnehmung der Systemakkuratheit	**Transparenz** konnte sich negativ auf subj. Akkuratheit auswirken, wenn Erwartungen nicht verletzt wurden, **positiv, wenn Erwartungen verletzt wurden**.
Cheng et al., 2019	Zulassung zur Universität: automatisiertes Entscheidungssystem	Erklärungen für Ergebnisse vs. Black Box; Interaktion (Ausprobieren) möglich	Objektives und subjektives Verständnis des Systems; Vertrauen	Höheres objektives Verständnis mit Erklärungen, aber weder höheres Vertrauen noch subjektives Verständnis; **Interaktion mit Algorithmus erhöhte beide Arten von Verständnis.**
Dominguez et al., 2019	Empfehlungssystem für Kunstwerke	Erklärungen für die Empfehlung; Schnittstellen; Algorithmus-Typ	Vertrauen; Zufriedenheit; Absicht, wieder zu nutzen	Erklärungen erhöhten Verständnis, Zufriedenheit und Vertrauen; **präziseres System trotz geringerer Erklärbarkeit bevorzugt.**
Du et al., 2019	Automatisierte Fahrzeuge	Erklärung vor der Aktion vs. danach vs. keine vs. Ablehnung möglich	Akzeptanz, Vertrauen, Ängste, psychische Belastung	Keine Auswirkung der Erklärung auf Vertrauen, aber die Erklärung nach dem Handeln wurde außergewöhnlich schlecht wahrgenommen. Option Aktion abzulehnen verringerte weder die Angst noch erhöhte sie das Vertrauen.
Eslami et al., 2019	Filterung von Onlinebewertungen (Yelp)	Keine UVs: Analyse von Online-Diskussion & qualitative Interviews	Wahrnehmung und Einstellung der Nutzenden	Transparenz wirkte sich verschieden auf Verhalten der Nutzenden aus: mehr informierte Interaktion oder Verlassen der Plattform; angemessenes Maß an Transparenz erforderlich.
Kocielnik et al., 2019	Detektor, um Termine aus Mails zu extrahieren	Art des Fehlers: falsch positiv vs. falsch negativ; Fehlertypen auswählen können; globale und beispielbasierte Erklärungen	Wahrnehmung von Genauigkeit und Akzeptanz, wahrgenommene Kontrolle, Verständnis	Unterschiedliche Akzeptanz je nach Fehlertyp. Einfache Erwartungsanpassungen trugen zur Verbesserung der Akzeptanz bei. Erklärungen erhöhten Verständnis.

	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Springer & Whittaker, 2019	Emotions-erkennung aus Text (E-Meter)	Erklärungen (Text eingefärbt) vs. keine Erklärung	Wahrnehmung von und Reaktionen der Nutzenden auf Transparenz (quantitativ und qualitativ)	**Vor der Nutzung: Erwartung transparente Systeme arbeiteten besser**, keine Präferenzen nach der Nutzung; Nutzende nahmen Systeme als verschiedener wahr als sie waren; **Transparenz konnte irreleiten, z. B. wenn sie Erwartungen widersprach.**
Tsai & Brusilovsky, 2019	Interaktives Empfehlungs-system	Verschiedene Erklärungs-interfaces, Interaktion und Erklärungssymbol	Wahrgenommene Kontrolle, Transparenz und Zufriedenheit	Interaktive Schnittstellen regten zur Erkundung an, um das System zu verstehen. Kontrollierbarkeit-Erklärbarkeit-Abwägung: mehr Erklärungen verringerten die Benutzerfreundlichkeit und die wahrgenommene Kontrolle.
Weitz et al., 2019	Erkennung von gesprochenen Wörtern	XAI-Erklärung plus virtueller Avatar	Vertrauen in Automatisierung	Signifikant **höheres Vertrauen, wenn Avatar zusätzlich** (redundante) Informationen präsentierte.
Yin et al., 2019	Dating-Vorhersagen	Genauigkeit des Algorithmus: angegeben und beobachtet	**Nutzung von Prognosen**, Vertrauen	Die angegebene Genauigkeit beeinflusste die Nutzung und das Vertrauen; **nach Beobachtung des Algorithmus passte sich die Wirkung der Beobachtung an, insbesondere bei geringer Leistung**.
Zhao et al., 2019	Systeme für die Online-Einkaufsbera-tung	Erklärungen mit verschiedenem Detailgrad	Verständnis und Vertrauen	Die Nutzenden brauchten genügend Zeit und die Fähigkeit, die gegebenen Informationen zu verarbeiten und zu verstehen; **zu viele Details verringerten Verständnis und Vertrauen.**
Baldauf et al., 2020	KI-basierte Anwendung-en zur Selbst-diagnose	Explorative Online-Umfrage: Arten von erfassten und verarbeiteten Daten (z. B. Fotos, Husten-geräusche)	Allgemeine Bereitschaft zur Nutzung, Vertrauens-faktoren, wünschenswerte Merkmale	>50 % waren bereit, KI-Apps zu nutzen. Zweifel an technischer Machbarkeit und Effektivität, allerdings unabhängig von der Art der erfassten Gesundheitsdaten; **medizinische Zertifizierung, anonyme Übermittlung, Analyse persönlicher Gesundheitsdaten und vertrauenswürdiger App-Hersteller wichtig für das Vertrauen** in KI-Apps.
Ford et al., 2020	KI zur Klassifizierung von Zahlen auf Bildern	Erklärungen durch Beispiel vs. keine Erklärung, Fehlerquote des Klassifizierenden	Korrektheit, Angemessenheit, Vertrauen und Zufriedenheit	**Erklärungen** anhand von Beispielen führten dazu, dass Menschen **falsche Klassifizierungen als richtiger empfanden**. **Fehlerquote** hatte einen **signifikanten Einfluss auf die Wahrnehmung**: nur 3 und 4% führten zu mehr Vertrauen.

	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Förster et al., 2020	XAI in Handy-App zur Erkennung von Pflanzen-arten	Eigenschaften von Erklärungen: konkret, kohärent, relevant, konsistent, allgemein	Beurteilung, praktische Anwendbarkeit und Effizienz der Erklärungen	Entscheidende Faktoren für gute Erklärungen waren Konkretheit, Kohärenz und Relevanz.
Kim & Song, 2020	KI-System für das Packen von Überlebens-ausrüstung (Wizard-of-Oz)	Timing der Nachricht (vor vs. nach Entscheidung), Genauigkeits-angabe (keine vs. Fehlerrate vs. Genauigkeit)	Akzeptanz der Vorschläge	**Nutzende akzeptierten den Vorschlag von AI eher ohne Genauigkeitsangabe**; Timing hatte keinen Einfluss auf die Akzeptanz.
Kuosmanen, 2020	Black Box-Filmempfeh-lungssysteme	Erklärungstyp (generiert durch zwei Post-Hoc Erklärungs-generatoren)	Subjektive Überzeugungs-kraft und Vertrauens-würdigkeit der Empfehlungen	Signifikant höhere Überzeugungskraft und Vertrauen für Erklärungen anhand von Assoziationsregeln; allerdings war **Einfluss des Interface-Designs unklar.**
Lehmann et al., 2020	Beratungs-system für Produkt-nachfrage	Erläuterung zu Funktionsweise des Algorithmus vs. keine	Wahrgenom-mener Wert des Hinweises, **Nutzung** des Hinweises, Leistung	**Transparenz führte zu geringerer Nutzung** des Algorithmus, was durch den **(geringeren) wahrgenommenen Wert der Hinweise** vermittelt wurde. Außerdem führte sie zu einer geringeren Leistung.
Lucic et al., 2020	Verkaufs-prognosen (Regressions-modell)	Erklärungen vs. keine Erklärungen	Verständnis, Bereitschaft zur Nutzung, Vertrauen, Bewertung der Vorhersagen	Erhöhtes Verständnis und sinnvolle Schlussfolgerungen in Erklärungs-Bedingung, aber keine Auswirkung auf das Vertrauen oder die Bereitschaft, das Modell zu nutzen.
Oh et al., 2020	KI Mirror: Algorithmus sagt ästhetische Bewertung von Fotos vorher	Qualitative Studie: verschiedene Nutzenden-gruppen (KI/ML-Experten vs. Fachexperten vs. Öffentlichkeit); verschiedene Aufgaben	Nutzung, subjektiver Interpretierbarkeit und Plausibilität	Nutzende verstanden KI anhand gruppenspezifischer Expertise, nutzten verschiedene Strategien, um Lücke zwischen ihren Einschätzungen und den KI-Vorhersagen zu schließen. Größere **Differenz zwischen Gedanken der Nutzenden und Vorhersagen der KI führte zu geringerer subjektiver Interpretierbarkeit und Plausibilität der.**
Peters et al., 2020	KI-System zur Kreditbe-wertung	Transparenzmerk-male von KI-Systemen	Bereitschaft, für die Funktionen zu zahlen	Verbraucher zeigten signifikante **Bereitschaft, für Transparenz zu zahlen.** Erhöhtes Vertrauen in das KI-System durch verbesserte wahrgenommene Transparenz war der Hauptgrund für positive Bewertung von Transparenzmerkmalen.

	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Schmidt et al., 2020	Emotions-klassifikation von Film-Reviews	Hervorhebung von Merkmalen (Erklärung), Sicherheits-angaben	Vertrauen, **Nutzung**	**Sicherheitsangaben verringerten die Nutzung von KI. Transparenz stand in negativem Zusammenhang mit Vertrauen** bei unintuitiven Erklärungen.
Shen et al., 2020	Modelle für maschinelles Lernen zur binären Klassifizierung	Qualitative Analyse: Alternative Darstellungen von Verwechslungs-matrizen	Verständnis der Teilnehmenden für Leistung der Modelle des maschinellen Lernens	Die Kontextualisierung von Terminologien verbesserte das Verständnis der Teilnehmenden. Flussdiagramme waren am effektivsten, um das objektive und subjektive Verständnis der Teilnehmenden zu verbessern.
Woodruff et al., 2020	KI als Entschei-dungsträger	Qualitative Ergebnisse aus Workshops	Wahrnehmung der Teilnehmenden über Fähigkeiten der KI, subjektive oder moralisch komplexe Urteile zu fällen	KI wurde als System angesehen, das starren Kriterien folgte, bei mechanischen Aufgaben gut abschnitt, aber unfähig war, subjektive oder moralisch komplexe Entscheidungen zu treffen. Erklärungen ermöglichten es, Fehler zu identifizieren.
Y. Zhang et al., 2020	Einkommens-prognose	Konfidenzniveau, lokale Erklärung	Nutzung der Vorhersagen	**Konfidenzniveau war der einzige Faktor, der die Nutzung beeinflusste** (je höher, desto mehr).
Alam & Mueller, 2021	Medizinisches KI-Diagnose-system	1. Studie: globale vs. lokale Erklärungen; 2. Studie: Vergleich lokaler Erklärungen: Text vs. visuell + Text vs. Beispiele + Text vs. keine	Nutzenden-zufriedenheit, Angemessenheit, Vollständigkeit, Nützlichkeit, Genauigkeit, Vertrauen in verschiedenen Phasen	**Globale Erklärungen schützten nicht vor Misstrauen und Unzufriedenheit bei Korrektur einer falschen Diagnose, aber erhöhten Verständnis. Lokale Erklärungen erhielten Vertrauen nach korrigierten Diagnosen.** Umfangreichere **Erklärungen erhöhten Zufriedenheit und Vertrauen in kritischen Phasen.**
Bove et al., 2021	Interface für intelligente Preisgestal-tung in der Kfz-Versicherung	Kontextualisie-rungselemente zu Erklärungen zur Bedeutung lokaler Merkmale hinzugefügt	Verständnis von Vorhersagen: objektives und wahrgenommenes Verständnis; wahrgenommene Nützlichkeit von Erklärungen	Das Hinzufügen von Kontextualisierungselementen verbesserte das Verständnis von ML-Vorhersagen.
Ehsan et al., 2021	KI-gestützte Systeme	Interviews zur Konzeptualisie-rung von Sozialer Transparenz (ST)	Konzept der ST	Ziel war technologischen Kontext sichtbar zu machen, inkl. früherer KI-Entscheidungsergebnisse und menschlicher Interaktionen mit diesen Ergebnissen. So ließ sich Vertrauen in KI aufbauen. Darstellung von ST über „4W“: „Wer tat Was mit dem KI-System, Wann und Warum?“

	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Gerlings et al., 2021	KI-Klassifikator für COVID-19-Patient*innen für Intensiv-stationen	Explorative Interviews im Rahmen einer Fallstudie	Erläuterungs-bedarf jeder Interessengruppe	**Verschiedene Interessengruppen hatten unterschiedliche Erklärungsbedürfnisse**: Entwickler*innen, Fachexpert*innen, Entscheidungsträger*innen, Zuschauende.
Khurana et al., 2021	Chatbots für komplexe Tabellen-kalkulation	Erklärbare Chatbot-Schnittstellen innerhalb der Anwendung	Verständnis der Nutzenden für die Gründe einer Störung; Wahrnehmung von Nützlichkeit, Transparenz und Vertrauen	Erklärungen verbesserten das Verständnis der Nutzenden für die Gründe einer Panne; verbesserte Wahrnehmung der Nutzenden in Bezug auf Nützlichkeit, Transparenz und Vertrauen durch Erklärungen.
Naiseh et al., 2021	Kollaborative Mensch-KI Entschei-dungstools	Reihe von qualitativen Studien zu Erläuterung, Interaktion, Design	Wie Vertrauens-kalibrierung geleitet werden kann	Fünf Designprinzipien für Erklärungen: - Design für Engagement (XAI musste zur Interaktion anregen) - Gewohnte Handlungen in Frage stellen - Aufmerksamkeitslenkung - Reibungsdesign - Unterstützungstraining und Lernen
Ribes et al., 2021	Aggregatoren für Nachrichten-inhalte	Interface Design, XAI: ausführliche vs. kurze vs. keine Erklärung	Nutzenden-wahrnehmung: Vertrauen, Nützlichkeit, Verständnis	Kein Einfluss der Schnittstellenge-staltung oder der XAI-Informationen auf das Vertrauen der Nutzenden. **Ausführliche Erklärungen führten zu einem schlechteren objektiven Verständnis**, aber zu einem ähnlich hohen subjektiven Verständnis.
Shin, 2021	Durchsuchen von Nach-richten auf Algorith-musbasierten Websites	Keine Manipulation, nur nachträgliche Erhebungen	Vertrauen, Einstellung	Erklärbarkeit hing mit Vertrauen zusammen, Herleitbarkeit war eine Schlüsselkomponente der Erklärbarkeit und erhöhte das Verständnis der Nutzenden für den Entscheidungsprozess von KI-Systemen.
Tsai et al., 2021	Online-Symptom-Checker (OSCs)	Erklärungen (drei Typen)	Vertrauen, Transparenz-wahrnehmung, Lernen, weitere Nutzungs-erfahrungen (z. B. Mitteilen der Diagnose an Familienmit-glieder)	Erklärungen konnten die Benutzererfahrung in mehrfacher Hinsicht erheblich verbessern, aber zu viele Informationen förderten Überlastung, sodass evtl. Anweisungen ignoriert wurden.

	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Z. Zhang et al., 2021	Klassifizierung medizinischer Berichte	Hohe vs. niedrige Transparenz, hohe vs. niedrige Akkuratheits-angabe	Wahrgenommene Nützlichkeit, Verständlichkeit, Vertrauen, Zustimmung	Eine **höhere Genauigkeit erhöhte Vertrauen und Nützlichkeit**; bei **geringer Genauigkeit oder Nichtübereinstimmung mit den Nutzenden verringerten mehr Erklärungen das Vertrauen**.
Anger-schmid et al., 2022	KI-gestütztes Entschei-dungssystem	KI-Erklärungen, Fairnesslevel in verschiedenen Szenarien	Vertrauen, wahrgenommene Fairness	Nur geringe Fairness verschlechterte Vertrauen; **Erklärungen erhöhten Vertrauen in KI-Entscheidungen und Wahrnehmung von Fairness**. Auswirkungen von Erklärungen und Fairness waren komplex, u. a. abhängig von Art der Erklärung, Grad der Fairness und Anwendungsszenario.
Chatti et al., 2022	Empfehlungs-systeme	Detailgrad der Erläuterung (einfach, mittel, fortgeschritten) und Nutzenden-eigenschaften	Wahrnehmung von Erklärungen	Erklärungsziel und Nutzendentyp führten zu unterschiedlichen Wahrnehmungen von Empfehlungssystemen mit unterschiedlichem Detaillierungsgrad.
Daschner & Obermaier, 2022	Prognose der Produkt-nachfrage	Akkuratheit der Ratschläge 5% besser vs. schlechter als Versuchs-personen-Leistung; Quelle: KI vs. Kollege	Vertrauen vor und nach Leistungsfeed-back **(Nutzung als Vertrauensmaß)**	**Erwartungen an Genauigkeit des Algorithmus waren höher als an die des menschlichen Beraters**. Eigene Genauigkeit war Schwellenwert für die Verwendung von Ratschlägen: war die Genauigkeit der Ratschläge geringer, wurden keine Ratschläge verwendet, war sie höher oder gleich, wurden Ratschläge verwendet.
Guesmi et al., 2022	Anwendung zu Empfehlungs- und Interessen-modellierung	Detaillierte Erklärungen und Persönlichkeits-merkmale	Wahrnehmung des Empfehlungs-systems	Die Beziehung zwischen Persönlichkeitsmerkmalen und dem Detaillierungsgrad der Erklärung beeinflusste die Wahrnehmung des Systems.
Karran et al., 2022	Evaluierung eines Bild-klassifizie-rungssystems	Verschiedene Arten von Erklärungs-visualisierungen (Heatmap)	Kognitive Belastung, Vertrauen als Erwartung richtiger Ergebnisse	Die **Art der Erklärung wirkte sich erheblich auf die kognitive Belastung aus, aber die kognitive Belastung allein hatte keinen Einfluss auf das Vertrauen in das System**.
Molina & Sundar, 2022	Inhalts-moderation: Textklassi-fizierung (Hassreden und Selbstmord-gedanken)	Klassifizierungs-quelle (Mensch, KI, beides); Transparenz: keine vs. Transparenz vs. Interaktive Transparenz)	Vertrauen, Zustimmung zum System, Verständnis des Systems	Nutzende vertrauten KI genauso wie Menschen. **Transparenz erhöhte Vertrauen und Verständnis**. Bei mehr Offenlegung sank Zustimmung. **Interaktive Transparenz steigerte Vertrauen, da Nutzende mehr Verantwortung übernahmen, führte nicht zu mehr Verständnis**.

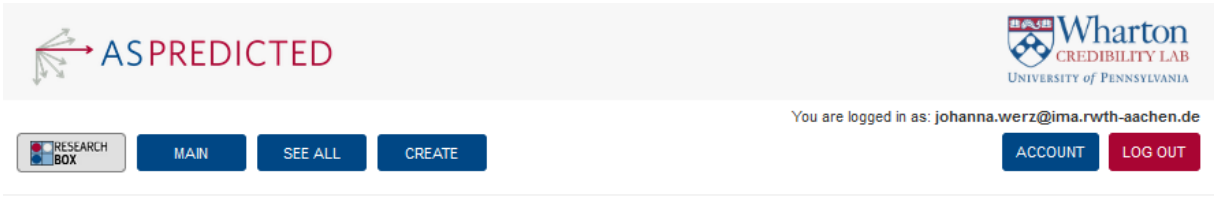
	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Ooge et al., 2022	e-Learning Plattform	Erklärungen für Aufgabenauswahl: keine vs. Placebo vs. echte Erklärung; hohes vs. niedriges Risiko	Ein- und mehr-dimensionales Vertrauen	Erklärung erhöhte mehr- aber nicht eindimensionales Vertrauen. **Erklärungen zeigten größeren Einfluss in Entscheidungen mit hohem Risiko.**
Schoeffer et al., 2022	Autonomes Kreditvergabe-System	Mensch vs. System, identische Erklärungen	Subjektive Fairness und Vertrauens-würdigkeit	System wurde fairer wahrgenommen; bei **höherer AI Literacy Präferenz für System und mehr Vertrauen.**
Sieger et al., 2022	Intelligente Heim-systeme (Wizard-of-Oz)	Beteiligung an der Trainingsphase der KI	Gefühl der Kontrolle, wahrgenommenes Verständnis, wahrgenommene Nützlichkeit von KI allgemein	Die **Beteiligung an der Lernphase der KI steigerte das Gefühl der Kontrolle**, das wahrgenommene Verständnis und die wahrgenommene Nützlichkeit der KI im Allgemeinen.
Sun & Sundar, 2022	KI-gestützter Gesundheits-Chatbot zur Bewertung von Angst-zuständen	Präsentation von Erklärungen: interaktiver Dialog vs. statische Informationen vs. keine Erklärungen	Wahrgenommene Transparenz, affektives Vertrauen in System, subjektives Verständnis, objektives Verständnis (Quiz)	Variation in der Art der Informationen, die über das System gewünscht wurde. Eine **interaktive Vermittlung von Erklärungen führte zu höchsten Werten** bei wahrgenommener Transparenz, affektivem Vertrauen in System und subjektivem wie objektivem Verständnis der Mechanismen des Systems.
Tabrez et al., 2022	Interaktion mit autonomer Drohne in Augmented-Reality-Spiel (mine-sweeper)	Modalität der visuellen Anleitung: präskriptiv (Linie) oder deskriptiv (Heatmap) oder kombinierte Erklärung (beides)	Vertrauen, Interpretier-barkeit, Leistung im Spiel, menschliche Unabhängigkeit	Die kombinierten Erklärungen zeigten die besten Auswirkungen auf alle AVs.
Wanner et al., 2022	ML-Modelle in verschiedenen Szenarios	globale vs. lokale Erklärungen, Relevanz des Anwendungsfalls	Wahrgenommene Intuitivität, Komplexität, Verständlichkeit, Zufriedenheit, Vertrauenswürdig-keit und Güte der Erklärung	Größter Einfluss der Vertrauenswürdigkeit auf Güte der Erklärung. Kein Einfluss der Relevanz des Anwendungsfalls, auch nicht auf globale oder lokale Präferenz. Unterschiedliche Anforderungen für lokale vs. globale Erklärungen.
Herm et al., 2023	Prognose von Krebs aus Hirn- & Herzscans mit verschieden komplexen Modellen	Erklärungen lokal (How, How-To, Why, Why-Not, What-Else) vs. global vs. keine Erklärung	Erklärleistung, Leistung des Modells	**Lokale Erklärungen wurden mit der höchsten Erklärleistung bewertet, globale und keine Erklärung ohne Unterschied**. Kein linearer Zusammenhang von Leistung des Modells und seiner Komplexität.

	Algorithmus/ KI-System	**Art der Transparenz** (UV)**	**Transparenz wirkt auf… (AV)**	**Kernergebnisse**
Kim et al., 2023	App zur Vogel-bestimmung	4 XAI-Ansätze (Heatmap, Beispiel, Konzept und Prototyp)	Interview: Umgang mit System	Bedarf an praktisch nützlichen Informationen zur Verbesserung der Zusammenarbeit mit KI-System: Nutzende beabsichtigten, **XAI-Erklärungen zu nutzen, um das Vertrauen zu kalibrieren**, ihre Nutzungsfähigkeiten zu verbessern und konstruktives Feedback für Entwicklung zu geben; **Nutzende bevorzugten teilbasierte (lokale) Erklärungen.**
Leichtmann et al., 2023	Aufgabe essbare Pilze zu sammeln (Spiel)	Visuelle Erklärungen und pädagogische Intervention	Benutzerleistung bei der Entscheidungs-findung, Vertrauen	**Visuelle Erklärungen verbesserten die Leistung der Nutzenden und führten zu einem angemesseneren Vertrauensniveau**. Die Bildungsintervention zur Verbesserung der KI-Kenntnisse hatte keine Auswirkungen auf die Leistung der Nutzenden.
Ehsan et al., 2024	Roboter-Bewegung im Raum	Erklärungsart: schlussfolgernd, Handlungs-erklärung, numerisch; KI-Experten vs. Laien	Erwartungen und Wahrnehmung der Erklärungen: Vertrauen, Intelligenz, Verständlichkeit, zweite Chance, Freundlichkeit	Wahrnehmung war schlussfolgernd positiver als handlungsleitend positiver als nummerisch; allerdings **überhöhtes Vertrauen in Zahlen-Erklärungen** in beiden Gruppen
Fleiß et al., 2024	Kommuni-kationssystem zur Vorauswahl im Recruiting	Keine Erklärungen vs. Erklärungen: Post-Hoc vs. KI-intrinsisch; verifizierbare Qualifizierungen vs. Soft Skills	Akzeptanz (Vignette)	Transparenz führte zu höherer Akzeptanz, aber Art der Erklärung spielte keine Rolle; wichtiger war Verifizierbarkeit der Erklärung.

Anmerkung. Tabelle aufbauend auf Felzmann et al., 2020;
*: von Felzmann et al., 2020 übernommene Veröffentlichungen;
**: selbst ergänzte Kategorie

Anhang B

Präregistrierung der Studie (a) „Fehlerfall“



The PDF for this pre-registration is not-anonymous, to anonymize it scroll to bottom

'*Preventing Algorithm Aversion with Algorithmic Accuracy Information*'
(AsPredicted #62313)

Download PDF

Author(s)
Johanna Werz (IMA, RWTH Aachen University, Germany) - johanna.werz@ima.rwth-aachen.de
Jacqueline Engels (RWTH Aachen University, Germany) - jacqueline.engels1@rwth-aachen.de

Pre-registered on
03/31/2021 05:10 AM (PT)

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
Transparency about the possibility that an algorithm could err prevents algorithm aversion, which is the effect of deterring algorithmic advice after seeing the algorithmic advisor err. The hypothesis is, that when users of an algorithm receive information about the fact that it will perform correctly only with a certain possibility, they will not execute algorithm aversion as much as users who do not receive such information.

3) Describe the key dependent variable(s) specifying how they will be measured.
Using the algorithmic advice is measured as Weight of advice (WOA):
WOA = abs((final judgement – initial judgement)/(advice – initial judgement)). Trials in which advice = initial judgement are not used for analysis.

Furthermore, we define that WOA cannot be >1, as this would mean that participants change their initial judgement and follow the advice but even go beyond it. We assume that WOAs > 1 are due to the experimental set-up when participants use the slider imprecisely. This is why we set WOA > 1 to WOA = 1.

4) How many and which conditions will participants be assigned to?
Three between-subject factors:
1) High certainty information (HC): Participants read that the algorithm is correct in 90,1 % of the cases (Der Algorithmus liegt in 90,1% der Fällen richtig.)
2) Low certainty information (LC): Participants read that the algorithm is correct in 78,9 % of the cases (Der Algorithmus liegt in 78,9% der Fällen richtig.)
3) No certainty information (NI): Participants only read the algorithmic results without any specification about prospective algorithmic performance.

Two within-subject factors:
The mean value for trials 1-6 is considered WOA without error experience.
The mean value for trials directly following an error (trial 7, 11 & 15) is used as WOA after error experience.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Analysis of variance (ANOVA) with repeated measures
- Inner subject factor: error experience (WOA without error experience/WOA directly following an error)
- Between subject factor: algorithm accuracy (HC/LC/NI)
To test the first hypothesis, an interaction effect is expected in a way that compared to trials without error experience, in trials after an error WOA decreases in the no information (NI) condition to a greater extent than in both certainty conditions. Furthermore, the main effect shows that in trials after an error, WOA is higher in both certainty conditions compared to the NI condition.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

We will exclude participants that fail the manipulation check: who fail to indicate the correct algorithmic certainty level of their condition at the end of the study. What is more, participants that do not fill out the additional questionnaires will be excluded from the analysis as the participants are assumed not to have participated conscientiously.

7) How many observations will be collected or what will determine sample size?

No need to justify decision, but be precise about exactly how the number will be determined.

We will collect 156 observations, starting at 01.04.2021 until 31.05.2021 or stop as soon as 156 observations are collected.

8) Anything else you would like to pre-register?

(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)

We use a new scale for measuring "Attitude towards Algorithms" (ATA, developed by Nicolai Bock, RWTH Aachen University). We assume the overall usage of algorithms to be higher for people scoring higher in the ATA.
Furthermore, participants fill out scales about readiness to take risks ("Risikobereitschaft", R-1, Beierlein, Kovaleva, Kemper & Rammstedt, 2014) need for cognition (NFC-K, Geman short scale: Beißert et al., 2015), decisiveness (subscale of construct Need for closure, Roets und Van Hiel, 2007) and Locus of control when dealing with technology (KUT, Beier, 1999) for exploratory reasons. Finally, we ask participants about their motivation and way of interacting with the algorithm which we will analyze in a qualitative way.

siehe auch https://aspredicted.org/5zb9y.pdf

Anhang C

Datenschutzerklärung

Informationen zum Datenschutz

Die Speicherung Ihrer Daten wird ohne die Nennung Ihres Namens durchgeführt. Wir werden keine Liste erstellen, die im Nachhinein eine Zuordnung Ihres Namens zu den Daten erlaubt. Die Angabe von personenbezogenen Daten [Name und E-mail-Adresse] ist freiwillig und dient allein der Teilnahme an der Verlosung von drei Amazon-Gutscheinen. Ihre personenbezogenen Daten [Name und E-mail-Adresse] werden in einem separaten Datensatz gespeichert und verschlossen aufbewahrt, so dass nur diejenigen Untersucher*innen der Studie Zugang haben, die eine Vertraulichkeitserklärung abgegeben haben. Nach der Verlosung der Gutscheine werden die personenbezogenen Daten gelöscht. Ihr personalisierter Eintrag in dieser Liste kann auf Ihren ausdrücklichen Wunsch jederzeit vorzeitig gelöscht werden.

Da aufgrund der getrennten Datensätze kein Bezug zwischen den personenbezogenen Daten und den studienbezogenen, anonymisierten Daten hergestellt werden kann, können nach Abschluss der Studie keine einzelnen Einträge der anonymisierten Daten zugeordnet und gelöscht werden. Die anonymisierten Daten werden auf dem internen Server des Lehrstuhls für Informationsmanagement im Maschinenbau gesichert. Eine Löschung der Daten von diesem Server ist nach Ablauf der gesetzlichen Aufbewahrungsfrist von 10 Jahren vorgesehen. Dabei dient als Rechtsgrundlage die Datenschutz-Grundverordnung (Art. 6 Abs. 1 lit. A DSGVO).

Bezüglich Ihrer Daten informieren wir Sie zunächst über ihre Rechte (Art. 13 ff DSGVO, §§ 32 ff BDSG-neu): Sie haben das (1) Recht auf Auskunft über die Sie betreffenden personenbezogenen Daten, die im Rahmen einer Studie erhoben, verarbeitet oder ggf. an Dritte übermittelt werden (Aushändigen einer kostenfreien Kopie) (Artikel 15 DSGVO, §§ 34 und 57 BDSG-neu). Ggfs. haben Sie das (2) Recht auf Berichtigung, wenn Sie betreffende unrichtige personenbezogenen Daten erhoben wurden (Artikel 16 und 19 DSGVO, § 58 BDSG-neu). Und Sie haben das (3) Recht auf Löschung von Sie betreffenden personenbezogener Daten, z.B. wenn diese Daten für den Zweck, für den sie erhoben wurden, nicht mehr notwendig sind (Artikel 17 und 19 DSGVO, §§ 35 und 58 BDSG-neu). Unter bestimmten Voraussetzungen haben Sie das Recht, eine (4) Einschränkung der Verarbeitung Ihrer Daten zu verlangen. Sollte das der Fall sein, dürfen ihre personenbezogenen Daten nur gespeichert, aber nicht verarbeitet werden. Dies müssen Sie beantragen. Wenden Sie sich hierzu bitte an ihren Studienleiter (Artikel 18 und 19 DSGVO, § 58 BDSG-neu). (5) Recht auf Datenübertragbarkeit: Sie haben das Recht, die Sie betreffenden personenbezogenen Daten, die Sie dem Verantwortlichen für die Studie bereitgestellt haben, zu erhalten. Damit können Sie beantragen, dass diese Daten entweder Ihnen oder, soweit technisch möglich, einer anderen von Ihnen benannten Stelle übermittelt werden (Artikel 20 DSGVO). Sie haben ein (6) Widerspruchsrecht und können jederzeit gegen konkrete Entscheidungen oder Maßnahmen zur Verarbeitung der Sie betreffenden personenbezogenen Daten Widerspruch einzulegen (Art 21 DSGVO, § 36 BDSG-neu). Eine solche Verarbeitung findet dann grundsätzlich nicht mehr statt. (7) Einwilligung zur Verarbeitung personenbezogener Daten und Recht auf Widerruf dieser Einwilligung: Die Verarbeitung ihrer personenbezogenen Daten ist nur mit Ihrer Einwilligung rechtmäßig (Artikel 6 DSGVO, § 51 BDSG-neu). (8) Widerrufsrecht: Sie haben das Recht, ihre Einwilligung zur Verarbeitung personenbezogener Daten jederzeit zu widerrufen. Im Falle des Widerrufs müssen Ihre personenbezogenen Daten grundsätzlich gelöscht werden (Artikel 7, Absatz 3 DSGVO, § 51 Absatz 3 BDSGneu). Es gibt allerdings Ausnahmen, nach denen die bis zum Zeitpunkt des Widerrufs erhobenen Daten weiterverarbeitet werden dürfen, z.B. wenn die Datenverarbeitung zur Erfüllung einer rechtlichen Verpflichtung erforderlich ist (DSGVO Art. 17 Abs. 3 b).

Möchten Sie eines dieser Rechte in Anspruch nehmen, wenden Sie sich bitte an den Versuchsleiter/in. Außerdem haben Sie das Recht, Beschwerde bei der/den Aufsichtsbehörde/n einzulegen, wenn Sie der Ansicht sind, dass die Verarbeitung der Sie betreffenden personenbezogenen Daten gegen die DSGVO verstößt.

Zusätzlich ist vorgesehen, dass die anonymisierten Daten bei Veröffentlichung der wissenschaftlichen Ergebnisse anderen Forschenden auf einer digitalen Forschungsdatenbank auf unbegrenzte Zeit zur Verfügung gestellt werden. Dies dient der Praxis der offenen Wissenschaft, welche beinhaltet, dass andere Forschende unsere Ergebnisse eigenständig überprüfen können. Hierzu bitten wir sie untenstehend separat um Ihr Einverständnis.

Die für die Datenverarbeitung verantwortliche Person ist Johanna Werz, M. Sc.; Dennewartstr. 27, 52068 Aachen; johanna.werz@ima.rwth-aachen.de

Anhang D

Attitudes Towards Algorithms-Skala (ATAS; Bock & Rosenthal-von der Pütten, 2023)

Nr.	Subskala	Item
1	O	Algorithmen bevorzugen niemanden.
2	O	Algorithmen behandeln alle Menschen gleich.
3	O	Algorithmen haben keine Vorurteile.
4	O	Algorithmen legen bei jedem den gleichen Maßstab an.
5	O	Algorithmen kann man nicht bestechen.
6	O	Algorithmen sind völlig rational und deshalb nachvollziehbar.
7	E	Algorithmen sollten keine moralisch schwierigen Entscheidungen treffen.
8	E	Algorithmusbasierte Entscheidungen sind mir zu unpersönlich.
9	E	Algorithmen sind nicht dafür geeignet, persönliche Entscheidungen zu treffen.
10	E	Algorithmen sind weniger flexibel als Menschen bei der Bewertung von Entscheidungsfaktoren.
11	E	Algorithmen können die Konsequenzen einer Entscheidung nicht berücksichtigen.
12	E	Menschen könnten sich durch Algorithmen fremdbestimmen lassen.
13	E	Es ist problematisch, dass Algorithmen nicht zur Verantwortung gezogen werden können.
14	L	Algorithmen können Daten schneller analysieren als ein Mensch.
15	L	Algorithmen können mehr Daten verarbeiten als ein Mensch.
16	L	Algorithmen haben keine guten und keine schlechten Tage.
17	L	Algorithmen können Entscheidungsträgern viel Arbeit abnehmen.

Anmerkung: O = Objektivität, E = Ethik, L = Leistung.

Anhang E

Kontrollüberzeugungen im Umgang mit Technik (KUT; Beier, 1999, 2004)

Nr.	Item
1	Ich kann ziemlich viele der technischen Probleme, mit denen ich konfrontiert bin, alleine lösen.
2	Technische Geräte sind oft undurchschaubar und schwer zu beherrschen.
3	Es macht mir richtig Spaß, ein technisches Problem zu knacken.
4	Weil ich mit bisherigen technischen Problemen gut zurechtgekommen bin, blicke ich auch künftigen optimistisch entgegen.
5	Ich fühle mich technischen Geräten gegenüber so hilflos, dass ich die Finger davon lasse.
6	Auch wenn Widerstände auftreten, bearbeite ich ein technisches Problem weiter.
7	Wenn ich ein technisches Problem löse, so geschieht es meistens durch Glück.
8	Die meisten technischen Probleme sind so kompliziert, dass es wenig Sinn macht, sich mit ihnen auseinanderzusetzen.

Anhang F

Need for Cognition-Kurzversion (NFC-K; Beißert et al., 2015)

Nr.	Item
1	Es genügt mir einfach die Antwort zu kennen, ohne die Gründe für die Antwort eines Problems zu verstehen.
2	Ich habe es gern, wenn mein Leben voller kniffliger Aufgaben ist, die ich lösen muss.
3	Ich würde kompliziertere Probleme einfachen Problemen vorziehen.
4	In erster Linie denke ich, weil ich muss.

Anhang G

Subskala Decisiveness (Roets & Van Hiel, 2007, 2011)

Nr.	Englischsprachige Originalitems
1	When I have made a decision, I feel relieved.
2	When I am confronted with a problem, I'm dying to reach a solution very quickly.
3	I would quickly become impatient and irritated if I would not find a solution to a problem immediately.
4	I would rather make a decision quickly than sleep over it.
5	Even if I get a lot of time to make a decision, I still feel compelled to decide quickly.
6	I almost always feel hurried to reach a decision, even when there is no reason to do so.

Nr.	Item (deutsche Übersetzung)
1	Wenn ich eine Entscheidung getroffen habe, fühle ich mich erleichtert.
2	Wenn ich mit einem Problem konfrontiert bin, brenne ich darauf, sehr schnell eine Lösung zu finden.
3	Ich würde schnell ungeduldig und gereizt werden, wenn ich nicht direkt eine Lösung zu einem Problem fände.
4	Ich würde lieber eine schnelle Entscheidung treffen, als eine Nacht darüber zu schlafen.
5	Auch wenn ich sehr viel Zeit bekomme, eine Entscheidung zu treffen, fühle ich mich gezwungen, schnell zu entscheiden.
6	Ich fühle mich fast immer in Eile zu einer Entscheidung zu gelangen, selbst wenn es keinen Grund dafür gibt.

Anhang H

Spearman-Korrelationen zwischen Algorithm Aversion und möglichen Einflussfaktoren mit nicht-winsorisierten WOA-Werten

	Algorithm Aversion	
	Korrelations-koeffizient	Signifikanz-niveau
Allgemeine Einstellung gegenüber Algorithmen	ρ < ,0001	p = ,996
Kontrollüberzeugungen im Umgang mit Technik	ρ = -,102	p = ,186
Risikobereitschaft	ρ = -,052	p = ,499
Kognitionsbedürfnis	ρ = -,201**	p = ,009
Entscheidungsfreude	ρ = ,018	p = ,813

Anmerkung. Werte, die mit ** gekennzeichnet sind, sind auf einem Fehlerniveau von ,01 signifikant. *n* = 169.

Anhang I

Spearman-Korrelationen zwischen Algorithm Aversion und den die Mensch-Algorithmus-Interaktion beeinflussenden Faktoren mit den nicht-winsorisierten WOA-Werten

	Algorithm Aversion	
	Korrelations-koeffizienten	Signifikanz-niveau
Bei der Abgabe meiner Gewichtsschätzungen war ich mir sicher.	ρ = ,082	*p* = ,287
Ich hatte Vertrauen in das algorithmische System.	ρ = ,026	*p* = ,740
Ich war mit den Empfehlungen des Algorithmus zufrieden.	ρ = -,015	*p* = ,842
Ich habe die Empfehlungen des Algorithmus als hilfreich wahrgenommen.	ρ = ,076	*p* = ,327
Ich habe versucht besser zu sein als der Algorithmus.	ρ = -,160*	*p* = ,038
Die Schätzaufgaben haben mir Spaß gemacht.	ρ = -,038	*p* = ,625
Ich fand es interessant herauszufinden, wie meine Leistung in den Schätzungen war.	ρ = -,064	*p* = ,409
Ich wollte die Aufgabe so schnell wie möglich hinter mich bringen.	ρ = ,015	*p* = ,844
Ich habe versucht ein möglichst richtiges Ergebnis zu erzielen.	ρ = ,034	*p* = ,660

Anmerkung. Werte, die mit * gekennzeichnet sind, sind auf einem Fehlerniveau von ,05 signifikant. *n* = 169.

Anhang J

Gemischtfaktorielle ANOVAs mit Fehlererfahrung (Innersubjektfaktor) und demografischen Variablen (Zwischensubjektfaktoren) auf Algorithm Aversion (winsorisierte WOA-Werte)

	Algorithm Aversion	
Demographische Variable	$F_{(df, df)}$	*p*
Geschlecht	$2{,}593_{(1, 167)}$	,109
Arbeit	$1{,}745_{(4, 164)}$	,142
Höchster Bildungsstand	$0{,}256_{(4, 164)}$	,906
Routine mit Computern	$1{,}558_{(2, 166)}$	,214

Anmerkung. n = 169.

Anhang K

Transkripte der Fokusgruppen

Die Transkripte sind aufgrund ihres Umfanges online abgelegt und einsehbar unter https://osf.io/5tpdv/files

Anhang L

Einladungsemail zum Onlineexperiment zu Forschungsfrage (b)

„Nutzendenanforderungen“

Hallo xxx,

im Rahmen des Projekts TAIGERS bitten wir (Lehrstuhl für Communication Science (CS) sowie Informationsmanagement im Maschinenbau (WZL-IMA) der RWTH Aachen) euch um Teilnahme an einer dreistündigen Fokusgruppe. Die Session findet am **27.09.2021 von 14 – 17 Uhr über Zoom** statt. In der Fokusgruppe geht es um aktuelle Fragestellungen zum Thema Transparenz in künstlicher Intelligenz (KI) für die „Everyday Nutzer*innen“, das heißt ihr benötigt **kein spezielles Vorwissen** zur Teilnahme. Als Angestellte gilt eure **Teilnahme natürlich als Arbeitszeit**.
Wenn ihr am Thema oder der Methode interessierte Freund*innen/Kommiliton*innen habt, könnt ihr diesen natürlich auch gern Bescheid geben.

Meldet euch bitte bis 23.09. bei johanna.werz@ima.rwth-aachen.de zur Anmeldung dann erhaltet ihr die Zugangsinformationen. Wenn ihr am Montag keine Zeit habt, bitte ich um Absage!
Bei Fragen meldet euch ebenso bei uns. Wir freuen uns auf eine spannende Fokusgruppe und eure Eindrücke.

Liebe Grüße
Johanna Werz im Namen des TAIGERS-Teams

Anhang M

Fragebogen des Pretests zu Forschungsfrage (b) „Nutzendenanforderungen“

Wie wird mit der App interagiert?

- o manuelle Auswahl
- o Sprache
- o Schrift
- o Bilderkennung
- o Eyetracking

Welchen Nutzen erfüllt die App? (Mehrfachauswahl möglich)

- ☐ Effizienz
- ☐ Fairness
- ☐ Kostenersparnis
- ☐ kognitive Entlastung
- ☐ Unterhaltung
- ☐ Zeitersparnis
- ☐ weniger individuelle Fehler
- ☐ Weitere: _____

Welche Funktionen sind Ihnen bei der App wichtig? (Mehrfachauswahl möglich)

- ☐ Schnelligkeit
- ☐ Verfügbarkeit
- ☐ Sicherheit
- ☐ breites Aufgabenspektrum
- ☐ Komplexität
- ☐ Gültigkeit
- ☐ Akkuratheit
- ☐ Zugänglichkeit
- ☐ einfache Bedienung
- ☐ Weitere: _____

Als wie relevant bewerten sie die in der App dargestellte Thematik?

- o 1 = nicht relevant
- o 2
- o 3
- o 4
- o 5
- o 6
- o 7 = sehr relevant

Wie schlimm wäre ein fehlerhaftes Ergebnis der App für Sie?

- 1 = gar nicht schlimm
- 2
- 3
- 4
- 5
- 6
- 7 = sehr schlimm

Wie ansprechend ist die App für Sie?

- 1 = gar nicht ansprechend
- 2
- 3
- 4
- 5
- 6
- 7 = sehr ansprechend

Wie verständlich ist die App für Sie?

- 1 = gar nicht verständlich
- 2
- 3
- 4
- 5
- 6
- 7 = sehr verständlich

Anhang N

Einladungsemail zum Onlineexperiment zu Forschungsfrage (c) „Transparenzarten“

Liebe Mitstudierende,

im Rahmen meiner Abschlussarbeit führen wir eine Studie zum Thema Algorithmen in Ernährungsapps durch. Dabei geht es darum, das Gewicht von Obst und Gemüse zu schätzen, wobei ihr von verschiedenen Algorithmen Hilfestellungen erhaltet. Eure Teilnahme würde mir sehr weiterhelfen.

Das Online-Experiment dauert ca. 15 Minuten und ihr werdet mit 0,25 VP Stunden vergütet oder könnt einen von 3 Thalia-Gutscheinen im Wert von 10€ gewinnen. Die Teilnahmevoraussetzung ist eure Volljährigkeit und ihr solltet an keiner der vorigen Gewicht-Schätz-Studien teilgenommen haben.

Unter folgendem Link könnt ihr teilnehmen:
https://www.soscisurvey.de/KI-Algorithmen/
Bitte nehmt unbedingt von einem Laptop/Computer aus an der Studie teil, bei mobilen Endgeräten (Handy/Tablet) kommt es leider zu Fehlern in der Anzeige.

Bei Fragen könnt ihr euch gerne bei mir melden unter jashandeep@rwth-aachen.de oder bei Johanna Werz unter johanna.werz@ima.rwth-aachen.de .
Ich bedanke mich im Voraus!

Viele Grüße
Jashandeep Kaur

Anhang O

Einverständniserklärung des Experiments in Forschungsfrage (c) „Transparenzarten“


soSci
oFb - der onlineFragebogen

4% ausgefüllt

Einverständniserklärung

Im Nachfolgenden möchten wir sicherstellen, dass Sie mit der Teilnahme an unserer empirischen Studie explizit und nachvollziehbar einverstanden sind.

Hiermit bestätige ich,

...dass ich mindestens 18 Jahre alt bin.
...dass **meine Teilnahme am Experiment freiwillig geschieht** und ich weiß, dass ich das Experiment jederzeit auf eigenen Wunsch abbrechen kann.
...dass ich damit einverstanden bin, dass die im Laufe des Experiments gewonnen Daten **in anonymisierter Form für wissenschaftliche Auswertungen und Veröffentlichungen gespeichert und verwendet** werden. Hierbei ist keinerlei Rückschluss auf Ihre Person möglich.
...dass ich mich einverstanden erkläre, dass erhobene **personenbezogene Daten getrennt und vertraulich gespeichert** werden. Ich bin auch darüber informiert, dass ich diese Einwilligung zur Datenspeicherung personenbezogener Daten jederzeit schriftlich widerrufen kann. Diese personenbezogenen Daten sind auf Wunsch jederzeit löschbar.
...dass ich mich bereit erkläre, die mir gestellten Aufgaben **gewissenhaft** zu bearbeiten.

Falls Sie noch Fragen zu dieser Studie haben sollten, finden Sie in der Fußzeile ein Impressum mit Kontaktdaten der Studienleitung.

Hiermit bestätige ich, dass ich die Einverständniserklärung gelesen und verstanden habe.

- ◯ Ja
- ◯ Nein (nicht an der Studie teilnehmen)

Weiter

Anhang P

Spearman-Korrelationen zwischen demographischen Daten, Vertrauen und Nutzung (WOA = Weight of Advice)

		Höchster Bildungs-abschluss	Umgang mit Computern	WOA	Vertrauen	Alter
Höchster Bildungsab-schluss	ρ	1,000				
	Sig. (2-seitig)					
Umgang mit Computern	ρ	,185*	1,000			
	Sig. (2-seitig)	0,023				
WOA	ρ	0,108	-0,150	1,000		
	Sig. (2-seitig)	0,187	0,065			
Vertrauen	ρ	-0,011	-0,062	,287**	1,000	
	Sig. (2-seitig)	0,895	0,451	0,000		
Alter	ρ	,660**	,220**	0,050	-0,067	1,000
	Sig. (2-seitig)	0,000	0,007	0,546	0,413	

Anmerkung. **. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.
n = 151.