Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

In AI We Trust
Aspekte des Vertrauens in ChatGPT

MASTER-ARBEIT

2024

Autor
Karg, Jona Stefan

Begleitpersonen
Ritz, Frank
Sterchi, Yanik

Praxispartner
Hochschule für Wirtschaft FHNW
Bendel, Oliver

Danksagung

An dieser Stelle möchte ich zuerst meinen Betreuer Frank Ritz erwähnen und mich für seine Unterstützung beim Prozess dieser Arbeit, insbesondere bei der Spezifizierung meines Forschungsvorhabens, herzlich bedanken. Ebenso gilt mein Dank Yanik Sterchi, der spontan in der letzten Phase meiner Arbeit als Betreuer eingesprungen ist. Des Weiteren möchte ich mich bei meinem Praxispartner, dem Kompetenzzentrum Digital Trust, speziell bei Petra und Oliver bedanken, welche mich in der Themenfindung meiner Arbeit sowie der Rekrutierung der Probanden und Probandinnen unterstützt haben. Ebenso danke ich meiner Arbeitskollegin Janine für die mentale und praktische Unterstützung. Mein abschliessender Dank gilt meiner Familie, insbesondere meiner Ehefrau Linda, welche mich in dieser herausfordernden Zeit unterstützt hat und Entbehrungen auf sich nehmen musste.

Zusammenfassung

Die Relevanz von Künstlicher Intelligenz (KI) nimmt zu, ebenso wie die diesbezüglichen Sicherheitsbedenken. Dies führt zur Forderung nach einer vertrauenswürdigeren KI, welcher sich Forschende zunehmend annehmen. Es existiert jedoch keine allgemeine Theorie zu Vertrauen in KI; so dient häufig das Modell Trust in Automation von Lee und See (2004) als Forschungsgrundlage, so kann Automation als Basis von KI betrachtet werden. Gleichzeitig wird jedoch in Frage gestellt, ob dieses Modell auf KI übertragen werden kann. Entsprechend wird in dieser Arbeit im Kontext von ChatGPT untersucht, welche Faktoren das Vertrauen in KI beeinflussen und wie sich dies auf die Nutzungsintention auswirkt. Basierend auf Trust in Automation, ergänzt durch die Variable Nutzungsintention, wurde hierzu ein Pfadmodell abgeleitet. Zudem wurden Einflussfaktoren der Neigung zum Vertrauen sowie die zeitliche Veränderung des Zusammenhangs des Vertrauens im Kontext einer Intervention untersucht. Die Datenerhebung erfolgte anhand validierter Fragebögen. Im Rahmen der Datenanalyse wurden die Daten von insgesamt 105 Studierenden berücksichtigt, während für die ergänzende Längsschnittanalyse 10 Datensätze herangezogen wurden. Die Resultate bestätigen das konzeptuelle Pfadmodell nicht, was eine Respezifikation erforderlich machte. Dies hat dazu geführt, dass zusätzliche Pfade von der Neigung zum Vertrauen zu den drei Faktoren der Vertrauenswürdigkeit identifiziert wurden. Das Modell zeigt mit einer Ausnahme die erwarteten positiven Einflüsse. Dabei ist der Einfluss des Vertrauens auf die Nutzungsintention geringer als erwartet. Die Längsschnittanalyse hat keinerlei Veränderungen zwischen den beiden Messzeitpunkten offenbart. Die Ergebnisse unterstützen grösstenteils die Erwartungen, stellen jedoch auch bisherige Annahmen in Frage. So wird Vertrauen möglicherweise eine grössere Rolle zugeschrieben als tatsächlich vorliegt.

Schlagwörter: Künstliche Intelligenz, Vertrauen, ChatGPT, Explainable AI, AI Safety

Anzahl Zeichen (mit Leerzeichen): 161'822

Abstract

The relevance of artificial intelligence (AI) is increasing, as are safety concerns in this regard. This leads to the demand for more trustworthy AI, which is increasingly addressed by research. However, no common theory exists; as such, because automation can be seen as the basis of AI, Lee and See's (2004) trust in automation model often serves as the foundation for research. Nevertheless, whether this model can be adapted to AI remains unclear. Accordingly, this thesis examines which factors influence trust in AI in the context of ChatGPT and how this affects usage intention. Based on trust in automation, supplemented by the factor intention to use, a path model is deduced. In addition, factors influencing the propensity to trust and long-term changes in the interrelationship of trust in the context of an intervention are examined. Data are collected using validated questionnaires. The analysis includes data from 105 students, and complementary longitudinal analysis includes ten of these participants. The data fails to confirm the conceptual model and thus calls for a re-specification. This leads to the identification of additional paths from propensity to trust to the three factors of trustworthiness. The re-specified model shows the expected positive influences, with one exception. Furthermore, the significant influence of trust on the intention to use is weaker than expected. The longitudinal analysis shows no variation between the two measurements. The results largely support expectations, although they call previous assumptions into question. Hence, trust is likely attributed a greater importance than it actually has.

Keywords: Artificial intelligence, trust, ChatGPT, explainable AI, AI safety

Number of characters (including spaces): 161'822

Inhaltsverzeichnis

In AI We Trust – Aspekte des Vertrauens in ChatGPT

Künstliche Intelligenz (KI, auf Englisch AI, kurz für Artificial Intelligence) erfährt in den Medien erhebliche Aufmerksamkeit (Hoffman, Mueller, Klein & Litman, 2023). Dabei nimmt die Bedeutung der entsprechenden Algorithmen, welche mit zunehmender Häufigkeit in der Verarbeitung und der Interpretation von Daten eingesetzt werden, im Alltag zu (Gebru et al., 2022; Shin, 2021). So verändert KI die Gesellschaft grundlegend und schafft neue Möglichkeiten (Frank, Jacobsen, Søndergaard & Otterbring, 2023). Dabei rücken auch sicherheitsrelevante Aspekte vermehrt in den Fokus der Forschung (Hendrycks & Mazeika, 2022) und Öffentlichkeit, wie der Titel des Buches «Die KI war's» von Prof. Dr. Katharina Zweig (2023) bereits suggeriert. Aber auch unabhängig von möglichen folgeschweren Ereignissen birgt die Nutzung von KI Risiken, da die Erwartungen der Nutzenden an diese möglicherweise nicht erfüllt werden (Langer, König, Back & Hemsing, 2023). So werden die Vertrauenswürdigkeit von KI-Systeme und das Vertrauen in diese auch in der breiten Gesellschaft vermehrt diskutiert (Hoffman et al., 2023; Lewis & Marsh, 2022).

Auch bei der Entwicklung von KI-Systemen wird dem Faktor Vertrauen eine hohe Relevanz zugeschrieben und er sollte bereits zu diesem Zeitpunkt Berücksichtigung finden (Hoffman et al., 2023; Siau & Wang, 2018). Insbesondere wirkt sich Vertrauen auf die Akzeptanz der KI-Systeme aus (Siau & Wang, 2018); Gebru et al. (2022) postulieren sogar, dass es einer der wesentlichen Faktoren bei deren Nutzung ist. Somit kann sich Vertrauen auf den Erfolg der Entwicklung des KI-Systems auswirken. Zudem wird es als Mediator der wahrgenommenen Zuverlässigkeit des Systems verstanden (Gebru et al., 2022; Tschopp & Ruef, 2018) und bereits kleinere Sicherheitsvorfälle stellen konkrete Bedrohungen dar, welche zu einem Verlust des Vertrauens führen können (Amodei et al., 2016). Demnach wird dem Vertrauen eine starke Beeinflussung der Nutzungsintention zugeschrieben (Choung, David & Ross, 2023). Entsprechend spielt dies auch bei der Implementierung und der Nutzung von KI-Systemen eine Rolle (Hoffman et al., 2023).

Häufig wird in diesem Kontext erörtert, wie KI-Systeme vertrauenswürdiger gestaltet respektive wie Nutzende dazu gebracht werden können, diesen zu vertrauen (Lewis & Marsh, 2022). Demgemäss sah sich auch die Europäische Union (EU) dazu veranlasst, eine Leitlinie bezüglich der Frage zu verfassen, wie vertrauenswürdige KI-Systeme entwickelt werden können (Lewis & Marsh, 2022). Im Kontrast dazu steht die Position, dass KI-Systemen nicht vertraut werden kann und diese Menschen vielmehr befähigen sollten, selbst

zu entscheiden, ob und wieweit sie aus deren Sicht vertrauenswürdig sind (Lewis & Marsh, 2022; Tschopp & Ruef, 2020). Gleichzeitig wird davon ausgegangen, dass Menschen ein höheres Vertrauen in eine KI-Technologie haben, deren Algorithmus transparent und erklärbar ist (Choung et al., 2023), bzw. dass Erklärbarkeit zu adäquatem Vertrauen führt (Hoffman et al., 2023). Dabei ist unter Letzterem zu verstehen, dass das Vertrauen entsprechend den Fähigkeiten des KI-Systems kalibriert und damit weder zu hoch noch zu niedrig ist (Hoffman et al., 2023; Tschopp & Ruef, 2020). Entsprechend wird unter dem Schlagwort *Explainable AI (XAI)* die Erklärbarkeit von KI diskutiert und seitens der EU im Rahmen des KI-Gesetzes gefordert (Hoffman et al., 2023). Auch Shin (2021) führt an, dass Nutzende in der Lage sein müssen, KI-Systeme zu verstehen, um diesen ein adäquates Vertrauen entgegenbringen zu können. So müssen diese Systeme entsprechend entwickelt sein, und um dies zu ermöglichen, muss das Vertrauen in KI erforscht werden (Shin, 2021).

Obwohl das Vertrauen in KI zunehmend in den Fokus der Forschung rückt (Bedué & Fritzsche, 2021), gibt es zum heutigen Zeitpunkt nur wenige Erkenntnisse diesbezüglich (Montag, Kraus, Baumann & Rozgonjuk, 2023). Gleichzeitig ist dessen Erforschung von hoher Relevanz für das Verständnis darüber, wie die wachsende Interaktion mit KI-Systemen die Gesellschaft und das Individuum beeinflusst (Montag, Klugah-Brown, et al., 2023). Dabei besteht eine grosse Herausforderung bereits in der Definition des Konzepts Vertrauen, da das Verständnis darüber teils stark von der tatsächlichen Bedeutung abweicht (Lewis & Marsh, 2022; Tschopp & Ruef, 2020). Beispielsweise werden (a) Kooperation, (b) Zuversicht sowie (c) Berechenbarkeit in bestimmten Zusammenhängen als Synonyme verwendet und verstanden, was jedoch die Bedeutung von Vertrauen verzerrt (Mayer, Davis & Schoorman, 1995). Zudem wird die Vertrauenswürdigkeit von KI-Systemen häufig auf (a) Fairness, (b) Unvoreingenommenheit, (c) Robustheit gegenüber Cyberangriffen und die (d) Verständlichkeit der Ergebnisfindung limitiert, wobei diese Begriffe entsprechend mit dem Vertrauen assoziiert werden (Lewis & Marsh, 2022).

Dabei erscheint unklar, inwiefern das bisherige Verständnis von Vertrauen auf den Kontext der Nutzung von KI übertragen werden kann, da die Funktionalität dieser Technologien für grosse Teile der Bevölkerung vollständig neu und teilweise nicht nachvollziehbar ist (Shin, 2021). So führen Chi, Jia, Li und Gursoy (2021) an, dass sich das Vertrauen bei der Nutzung von KI-Systemen erheblich von jenem in traditionelle Technologieprodukte unterscheide. Die Fähigkeit der KI zur autonomen Funktionsweise schaffe Risiken und Ungewissheiten, wohingegen die Nutzenden bei anderen Systemen

meist die Kontrolle über deren Funktionsweise behalten würden (Choung et al., 2023). Tatsächlich stützt sich KI häufig auf undurchsichtige *Machine-Learning-Algorithmen*, die als Blackboxes verstanden werden können (Choung et al., 2023; Gebru et al., 2022; Lewis & Marsh, 2022; Montag, Kraus, et al., 2023). Dabei beschreibt Machine-Learning (ML) die Fähigkeit des KI-Systems, ohne menschliches Zutun anhand von Daten selbständig zu lernen (Hendrycks, Carlini, Schulman & Steinhardt, 2022). Die beschriebenen Eigenschaften führen zu Unvorhersehbarkeit und Ungewissheit, was wiederum die hohe Bedeutung des adäquaten Vertrauens in KI unterstreicht (Choung et al., 2023). So ist es dringend erforderlich zu verstehen, was das Vertrauen in KI beeinflusst und was zur Entwicklung vertrauenswürdiger KI beiträgt (Choung et al., 2023).

Diese Arbeit soll zum Verständnis über das Vertrauen in KI Blackbox-Modelle und die Relevanz der Transparenz beitragen. Folglich besteht das Ziel dieser Untersuchung darin, herauszufinden, welche Faktoren das Vertrauen in KI-Systeme beeinflussen. Damit soll diese Arbeit den Aufbau eines neuen Forschungsschwerpunkts des Praxispartners – dem Kompetenzzentrum Digital Trust, des Instituts für Wirtschaftsinformatik der Hochschule für Wirtschaft FHNW – unterstützen und weitere Forschungsvorhaben vorbereiten. Zudem sollen praktische Implikationen für die Gestaltung von Richtlinien und konkreten Hinweisen zur Entwicklung vertrauenswürdiger KI-Systeme gewonnen werden. Da es als notwendig erachtet wird, ein spezifisches KI-System zu wählen, um Vertrauen angemessen untersuchen zu können (Tschopp & Ruef, 2018), erfolgt diese Untersuchung anhand von ChatGPT.

Konkret ist diese Arbeit der Frage gewidmet, welche Faktoren einen Einfluss auf das Vertrauen in KI haben. Ergänzt wird diese Forschungsfrage durch die Unterfrage, welche Faktoren einen Einfluss auf die Nutzungsintention haben. So kann die Nutzungsintention als Resultat der Entscheidung zu Vertrauen erachtet werden (Razin & Feigh, 2023). Zudem wird in dieser Arbeit der Unterfrage nachgegangen, welche Faktoren einen Einfluss auf die Neigung zum Vertrauen haben. Dabei beschreibt diese Neigung, das Vertrauen einer Person unabhängig von einem spezifischen Kontext oder KI-System (Solberg et al., 2022).

Abschliessend wird in einer längsschnittlichen Untersuchung der Frage nachgegangen, ob nach einer instruktionalen Intervention Veränderungen der Zusammenhänge zwischen den Einflussfaktoren und dem Vertrauen in KI auftreten. So ist der Status des Vertrauens als dynamisches Konstrukt in der Literatur umstritten und es bleibt unklar, ob und inwiefern sich dieses im Zeitverlauf verändert.

Künstliche Intelligenz

Im Folgenden werden zunächst KI-Technologien im Allgemeinen und ChatGPT im Spezifischen erklärt. Anschliessend wird auf den Begriff *AI Safety*, unter welchem Sicherheitsbedenken zu KI diskutiert werden, sowie die daraus abgeleiteten Forderungen nach Erklärbarkeit eingegangen. Da gemäss Mayer et al. (1995) Vertrauen mit Risiken einhergeht, werden abschliessend konkrete Risiken von KI-Technologien aufgezeigt, bevor im nachfolgenden Kapitel das Konstrukt Vertrauen erläutert wird.

Dabei ist das Verständnis des Begriffs der KI in dieser Arbeit an die Definition der OECD (2019, S. 1) angelehnt:

> Ein maschinenbasiertes System, das für bestimmte von Menschen definierte Ziele Voraussagen machen, Empfehlungen abgeben oder Entscheidungen treffen kann, die reale oder virtuelle Umgebungen beeinflussen. Es nutzt maschinelle und/oder menschliche Inputs, um ein reales und/oder virtuelles Umfeld zu erfassen, davon ausgehend (automatisch, z.B. mithilfe von ML, oder manuell) Modelle zu erstellen und mittels Modellinferenz Informations- oder Handlungsoptionen zu ermitteln. KI-Systeme können mit einem unterschiedlichen Grad an Autonomie ausgestattet sein.

Als Basis von KI kann Automation, eine Technologie, bei der Informationen umgewandelt und Entscheidungen getroffen werden, erachtet werden (Tschopp & Ruef, 2020). Jedoch gehen KI-Systeme über die Fähigkeiten automatisierter Systeme, etwa für Überwachung und Produktion, hinaus (Langer et al., 2023). Konkret schafft die autonome Funktionsweise der auf ML basierenden KI-Systeme Risiken und Ungewissheiten, wohingegen bei automatisierten Systemen die Nutzenden die Kontrolle über deren Funktionsweise beibehalten (Choung et al., 2023). ML ermöglicht es KI-Systemen, selbstlernend ohne spezifische menschliche Instruktionen (Hendrycks et al., 2022) auf der Grundlage der verfügbaren Daten Entscheidungen zu treffen (Kaplan, Kessler, Brill & Hancock, 2023). Dabei variiert jedoch der Grad des autonomen Handelns abhängig von der spezifischen KI-Anwendung, welche teils die Notwendigkeit der menschlichen Kontrolle minimiert (Choung et al., 2023; Frank et al., 2023). Generell kann die Autonomie von KI-Systemen als die Fähigkeit verstanden werden, ohne menschliche Unterstützung eigenständig Entscheidungen zu treffen und Aufgaben auszuführen (Frank et al., 2023).

Narrow AI. Es lässt sich beobachten, dass KI-Technologien in ihren Fähigkeiten, die menschliche Intelligenz durch Lernen, logisches Denken und das Treffen von Entscheidungen zu imitieren, kontinuierlich besser werden (Choung et al., 2023). Jedoch sind die heutigen KI-Systeme auf Fähigkeiten in spezifischen Aufgabenbereichen limitiert, was eine Generalisierung dieser Intelligenz ausschliesst, weshalb hierfür der Begriff *Narrow AI* verwendet wird (Lewis & Marsh, 2022). Streng genommen kann jede Technologie, welche auf einfache Algorithmen aufbaut, als solche verstanden werden (Yampolskiy & Spellchecker, 2016). Demnach ist KI lediglich ein Oberbegriff für Technologien, welche in der Lage sind, Aufgaben in einer ähnlichen Weise wie der Mensch auszuführen (Bedué & Fritzsche, 2021). Dementgegen kann *Artificial General Intelligence (AGI)* als Kombination aller möglichen Narrow-AI-Systeme erachtet werden, welche eine derartige Generalisierung zulassen (Yampolskiy & Spellchecker, 2016).

Aber auch die Narrow-AI-Systeme von heute basieren teils auf komplexen und für Laien unverständlichen Blackbox-Algorithmen, was zu Unberechenbarkeiten sowie Ungewissheiten führt. So stützen sie sich häufig auf komplexe ML-Algorithmen (Choung et al., 2023). Dabei ist Narrow-AI in Computern, Smartphones, Fahrzeugen und Produktionsmaschinen bereits heute ein fester Bestandteil des Alltags (Sindermann et al., 2022). Dies verdeutlicht, dass Nutzende eine geringe Entscheidungskompetenz darüber haben, ob sie auf KI-Systeme zurückgreifen möchten, da diese bereits in etablierten Technologien enthalten sind (Schepman & Rodway, 2023). Beispielsweise können Siri, der Sprachassistent von Apple, Grundfunktionen der sozialen Medien, Assistenzsysteme in Automobilen (Sindermann et al., 2022), aber auch Chatbots (Kaplan et al., 2023) wie ChatGPT, welche textbasierte Konversationen mit technischen Systemen ermöglichen (Taecharungroj, 2023), als KI-Systeme verstanden werden.

ChatGPT. ChatGPT ist ein von der Organisation OpenAI (2024a) entwickelter Chatbot – wobei GPT für *Generative Pre-Trained Transformer* steht – welcher *Natural-Language-Processing* (Brown et al., 2020), ein Sprachmodell zur Interpretation der menschlichen Sprache, und somit ML-Algorithmen nutzt (Jurafsky & Martin, 2024; Taecharungroj, 2023). Dementsprechend kann ChatGPT als Blackbox-Modell verstanden werden (Choung et al., 2023; Gebru et al., 2022; Lewis & Marsh, 2022; Montag, Kraus, et al., 2023). Da es sich um das beliebteste KI-System handelt (Conte, 2024) und dieses offen zugänglich ist (OpenAI, 2024b), wurde ChatGPT als Forschungsobjekt dieser Arbeit gewählt. Ein Mass an Erfahrung mit dem System ist notwendig, um Vertrauen adäquat messen zu können (Siegrist, 2021).

AI Safety. Trotz des hohen Potenzials, welches KI-Systemen zugeschrieben wird, werden Sicherheitsbedenken, insbesondere in Bezug auf Blackbox-Algorithmen, unter dem Leitwort AI Safety intensiv diskutiert und untersucht (u. a. Hendrycks & Mazeika, 2022; Hendrycks, Mazeika & Woodside, 2023; Yampolskiy & Spellchecker, 2016). Der von Yampolskiy (2010, zitiert nach Yampolskiy & Spellchecker, 2016, S. 5) geprägte Begriff hat sich in der Forschung etabliert; gleichzeitig werden jedoch Termini wie *Maschinenethik* oder *Friendly AI* teilweise als Synonyme verstanden und verwendet. Der Grundsatzgedanke von Yampolskiy und Spellchecker (2016) ist, dass vollständig autonome Maschinen grundsätzlich nicht als sicher angesehen werden können. Folglich stellt es eine grössere Herausforderung dar, eine sichere und fähige Maschine zu bauen, als lediglich eine fähige Maschine. Entsprechend ist die AI-Safety-Forschung mit aktuellen KI-Risiken, aber auch mit möglichen zukünftigen Risikoszenarien befasst (Hendrycks & Mazeika, 2022). Dabei bringt die Entwicklung zunehmend leistungsfähigerer KI-Systeme zumindest eine Verstärkung der bestehenden Risikofaktoren mit sich.

Diese Entwicklung löst Besorgnis bei den politischen Entscheidungsträgern und Entscheidungsträgerinnen aus (Hendrycks et al., 2023), denn die Rechtssysteme sind hinter den technischen Möglichkeiten zurückgeblieben (Yampolskiy & Spellchecker, 2016). Bei KI werden Sicherheitsmängel genauso kritisch wie im Kontext der Cybersicherheit betrachtet. Zwar kann sich die AI Safety mithilfe der in diesem Kontext entwickelten Ideen verbessern, jedoch können Sicherheitsmängel bei AGI grundlegend andere Auswirkungen mit sich bringen (Yampolskiy & Spellchecker, 2016). Im Gegensatz zum Ziel der Cybersicherheit, die Anzahl der erfolgreichen Angriffe auf das System zu reduzieren, besteht daher das übergeordnete Ziel der AI Safety darin, sicherzustellen, dass die Sicherheitsmechanismen nicht durch Angriffe umgangen werden können. Dabei gibt es zahlreiche Medienberichte mit Beispielen für das Versagen von KI (Yampolskiy & Spellchecker, 2016).

Wird AI Safety gemäss dem *Swiss Cheese Model* von Reason (1990, zitiert nach Hendrycks et al., 2023, S. 30) betrachtet, stellt Transparenz die letzte Sicherheitsbarriere dar. Gemäss Reasons Modell (1990) lässt sich die Sicherheit durch das «Stopfen von Löchern» in Sicherheitsbarrieren – illustriert durch Scheiben von Emmentalerkäse – erhöhen. Dabei werden Löcher in der letzten Sicherheitsbarriere als unmittelbarer Faktor der Entstehung von Ereignissen betrachtet, die durch deren Schliessung in letzter Instanz verhinderbar gewesen wären. Entsprechend wird die Transparenz als konkrete Forderung im Rahmen von XAI diskutiert (u. a. Gebru et al., 2022; Shin, 2021) und im folgenden thematisiert.

Explainable AI. Gemäss Shin (2021) führt die zunehmende Komplexität von KI zu einem Mangel an Transparenz, welcher das Verständnis erschwert und das Vertrauen negativ beeinflusst. Rai (2020) beschreibt anhand von Blackbox-Modellen undurchschaubare, insbesondere auf ML-Algorithmen basierende KI-Systeme, welche durch XAI zu Glassbox-Modellen werden sollen. Dabei beschreibt XAI eine Klasse von Systemen, die Aufschluss darüber geben, wie ein KI-System Entscheidungen und Vorhersagen trifft bzw. seine Handlungen ausführt (Rai, 2020). Weiter gefasst verstehen Gebru et al. (2022) unter Erklärbarkeit, Transparenz und Interpretierbarkeit Konzepte, die sich darauf beziehen, inwieweit die Funktionsweise KI-basierter Systeme für menschliche Nutzende nachvollziehbar dargestellt wird. Dabei stellt auch die EU mit dem KI-Gesetz – einem Vorschlag für eine europäische Verordnung über KI – konkrete Forderungen an die Erklärbarkeit (Hoffman et al., 2023).

Gebru et al. (2022) sehen in XAI ein Konzept, welches Erklärungen zu den von einem KI-basierten System getroffenen Entscheidungen liefert, um Vertrauen in das System zu schaffen. Dabei wird zwischen (a) transparenten KI-Modellen, welche eine inhärente Interpretierbarkeit aufweisen, und (b) post hoc erklärbaren KI-Modellen, welche eine unterstützende Interpretierbarkeit nachliefern, unterschieden. Auch Hoffman et al. (2023) bringen XAI mit Vertrauen in Verbindung. So sollen maschinengenerierte Erklärungen dabei unterstützen, die Leistung der KI zu verbessern, und damit das Vertrauen fördern (Hoffman et al., 2023). Ebenso schreibt Rai (2020) einer wirksamen Erklärung über die Leistung und der Vorhersagegenauigkeit des KI-Systems eine Förderung des Vertrauens zu. Im Gegenzug kann eine Undurchschaubarkeit vor allem dann das Vertrauen der Menschen in KI-Systeme beeinträchtigen, wenn diese mit erheblichen Risiken einhergeht, und schliesslich zur Ablehnung führen (Rai, 2020).

Gemäss Shin (2021) haben Studien darüber hinaus gezeigt, dass Erklärungen in KI-Systemen zu mehr Vertrauen der Nutzenden in diese Systeme und in die Ergebnisse der Algorithmen haben. Das Vorhandensein von Vertrauen sei neben der Erklärung der Schlüssel zur Förderung der Technologieakzeptanz und die Erklärbarkeit gäbe den Nutzenden die Gewissheit sowie das Vertrauen, dass KI-Systeme gut funktionieren würden (Shin, 2021). Shin (2021) geht ausserdem davon aus, dass durch XAI die hohe Intransparenz von Algorithmen reduziert und Vertrauen aufgebaut werden kann. Dies wiederum kann die Akzeptanz erhöhen sowie die Interaktion zwischen Mensch und KI fördern (Shin, 2021). In diesem Kontext weisen Okamura und Yamada (2020) darauf hin, dass Transparenz allein

nicht genügt um eine adäquate Vertrauensbildung sicherzustellen. Zudem resultiert mehr Information nicht zwangsweise in einem höheren Vertrauen (Mackay et al., 2020).

Generell stellt die Erklärbarkeit eine Herausforderung für Entscheidungsträger und Entscheidungsträgerinnen dar, welche sich auf KI respektive ML verlassen müssen (Hoffman et al., 2023). Vor allem bei Systemen mit ML ist die Nachvollziehbarkeit nicht grundsätzlich gegeben (Hoffman et al., 2023). So wird häufig angeführt, dass die Interpretierbarkeit der zunehmenden Komplexität von Algorithmen zur Verbesserung der Leistung bzw. der Präzision von KI-Systemen entgegensteht (Gebru et al., 2022; Rai, 2020). Dementgegen argumentiert Rai (2020) selbst, dass XAI nicht nur dazu beitragen könne, Blackbox-Modelle zu entlarven, sondern auch dazu, die beiden vermeintlich gegensätzlichen Ziele von Interpretierbarkeit und Präzision zu verknüpfen. Shin (2021) ergänzt, dass die Leistung von KI weniger durch die Algorithmen als durch den Mangel an Verständnis der Nutzenden gegenüber den Entscheidungen der Systeme limitiert sei.

Neben dem blossen Vorhandensein der Erklärungen weist Shin (2021) auch deren Qualität eine entscheidende Rolle zu. Aus seiner Sicht ist es essenziell zu verstehen, wie Nutzende diese Erklärungen interpretieren und bewerten, um tatsächlich Rückschlüsse auf die Interpretierbarkeit sowie die Verständlichkeit der KI ziehen zu können. Gleichzeitig geht Shin (2021) davon aus, dass die Nutzenden dem KI-System nur unter der Bedingung zu vertrauen beginnen, dass solche Erklärungen verständlich sind. Dem stehen die zunehmende Komplexität und die Blackbox-Natur der KI-Systeme entgegen, welche für ein angemessenes Verständnis mehr Fachwissen sowie Spezialkenntnisse erfordern.

Nach Shin (2021) stellt die Erklärbarkeit einen Hinweis auf Transparenz und Verantwortlichkeit der KI-Systeme dar. Letztere könnten für ihre Empfehlungen haftbar und verantwortlich gemacht werden. Zudem lasse sich mit Hilfe von XAI feststellen, ob Attribute mit Bezug zu Vorurteilen über beispielsweise Rasse oder Geschlecht direkt oder indirekt in Blackbox-Modellen verwendet werden, so dass die Modelle gegenüber bestimmten Gruppen voreingenommen sind, also kognitive Verzerrungen vorliegen (Rai, 2020). Vorurteile sind eine Art der Risiken, welche KI-Anwendungen zugeschrieben und im folgenden Unterkapitel «Risiken von KI» thematisiert werden.

Risiken von KI. Obgleich KI das Potenzial zugeschrieben wird, die Gesellschaft erheblich zu verbessern, gehen leistungsstarke Technologien grundsätzlich mit erhöhten Risiken einher (Hendrycks & Mazeika, 2022). Diese konkreten Risiken, welche von KI-Systemen ausgehen, werden im Folgenden genauer erläutert.

Vorurteile in KI-Anwendungen. In ihrer Untersuchung der Leistungsfähigkeit von GPT-3 nennen Brown et al. (2020) Vorurteile (Biases) in Bezug auf Geschlecht, Rasse sowie Religion und weisen darauf hin, dass es sich nicht um eine abschliessende Liste aller Verzerrungen dieses KI-Systems handelt. Diese Verzerrungen können dazu führen, dass die Modelle stereotype Inhalte erzeugen und damit Menschen der betreffenden Gruppen erniedrigend darstellen sowie ihnen entsprechend schaden. Rai (2020) ergänzt, dass diese Vorurteile von KI-Systemen beim Einstellungs- und Beförderungsprozess, aber auch in der Strafjustiz oder im Gesundheitswesen zu Diskriminierung führen könnten. In diesem Zusammenhang weisen Brown et al. (2020) darauf hin, dass diese Verzerrungen bestehende Stereotypen lediglich wiedergeben würden, ihnen aber dennoch vorgebeugt werden müsse.

Die Umsetzung dieser Prävention führt wohl beispielsweise dazu, dass das aktuelle GPT-Modell Antworten zur Aufforderung, einen Witz über spezifische Personengruppen zu verfassen, mit Haftungsausschlüssen versieht oder die Ausgabe gänzlich verweigert (J. Jäger, persönl. Mitteilung, 02.04.2024) – was einer Zensur durch eine nichtstaatliche Organisation entspricht. Dabei ist es nicht verwunderlich, dass solche Sprachmodelle kognitive Verzerrungen wiedergeben. So zeigten bereits Kahneman und Tversky (1972) den Zusammenhang von Verzerrungen und Heuristiken auf. Sprachmodelle wiederum basieren auf probabilistischen Modellen bzw. Heuristiken (Jurafsky & Martin, 2024). Die Studie von Brown et al. (2020) ist auch dahingehend interessant, dass sie von OpenAI – also der Entwicklungsorganisation des untersuchten KI-Systems – finanziert wurde. Ergänzend kann KI aber auch ein aktives Fehlverhalten zeigen, welches über stereotype Vorurteile hinausgeht (Scheurer, Balesni & Hobbhahn, 2023).

Fehlverhalten von KI-Anwendungen. Scheurer et al. (2023) haben mit ihrer Studie aufgezeigt, dass Sprachmodelle wie GPT-4 – welche eigentlich darauf trainiert sind, hilfreich, harmlos sowie ehrlich zu sein – Nutzer und Nutzerinnen, ohne explizit dazu aufgefordert zu werden, strategisch täuschen können. Konkret wurde untersucht, ob ein für den autonomen Aktienhandel trainierter ChatGPT-Bot wider besseren Wissens Insider-Handel betreibt. Die Forschenden stellten hierbei fest, dass GPT-4 unter wirtschaftlichem Druck dazu neigt, Insider-Informationen, über welche das System Kenntnis hat, für den

Aktienhandel heranzuziehen, ohne dies auszuweisen. Teilweise geht die KI-Anwendung sogar so weit, bei einer Begründung das Vorliegen derartiger Information aktiv zu verschleiern. Diese Ergebnisse unterscheiden sich dabei abhängig davon, wie stark das System darin trainiert ist, sich ethisch und rechtlich korrekt zu verhalten. Das strengste Training führt dementsprechend dazu, dass GPT-4 kaum noch Insider-Informationen nutzt. Falls dies jedoch der Fall ist, wird diese Nutzung zumeist verschleiert (Scheurer et al., 2023).

Wie ein von Hendrycks et al. (2023) vorgestelltes Beispiel von OpenAI zeigt, können auch Fehler in der Programmierung dazu führen, dass sich KI-Systeme unangemessen verhalten. So führte ein geringfügiger Fehler bei der Bereinigung des Codes dazu, dass ChatGPT über Nacht hasserfüllte und unangebrachte Texte produzierte. Anhand eines weiteren, älteren Falls, der den Twitter-Bot *Tay* von Microsoft betrifft, zeigen Hendrycks et al. (2023) die Schwierigkeit einer Kontrolle von KI auf. Auch wenn Tay auf Grundlage gefilterter Daten entwickelt wurde, dauerte es keine 24 Stunden, bis der Bot anfing, hasserfüllte Tweets zu verfassen. Das ML hatte dazu geführt, dass er sich Sprache von Internet-Trollen aneignete und diese unaufgefordert replizierte.

Yampolskiy und Spellchecker (2016) schreiben dem Fehlverhalten von KI auch Ereignisse wie Börsencrashs, verursacht durch intelligente Trading-Software, und Verkehrsunfälle, welche von autonomen Fahrzeugen verursacht wurden, zu. So steht KI-Fehlverhalten in direktem Zusammenhang mit den Fehlern, die durch die Intelligenz solcher Systeme entstehen. Allerdings gehen Yampolskiy und Spellchecker (2016) noch weiter und führen an, dass alle KI-Systeme grundsätzlich in ihren Aufgaben versagen würden, unabhängig davon, ob es sich um einen Spam-Filter, ein Navigationssystem, ein Übersetzungs-Tool oder eine Autokorrektur- bzw. Transkriptions-Software handle.

Aber nicht nur Fehler in der Programmierung der KI können zu Fehlverhalten führen, sondern auch die Konzeption der Ziele für die KI (Hendrycks et al., 2023). Konkret können ambitionierte Ziele verbunden mit geringer Kontrolle durch den Menschen dazu führen, dass die KI danach strebt, die eigene Macht als instrumentelles Ziel zu vergrössern. Dabei könnten starke KI-Systeme aggressiv Ziele verfolgen und eine Welt schaffen, die den menschlichen Bedürfnissen widerspricht (Hendrycks & Mazeika, 2022). Diese maximale Form des Fehlverhaltens wird mit dem Begriff *Rogue AI* assoziiert (Hendrycks et al., 2023).

Rogue AI. Hendrycks et al. (2023) verstehen unter Rogue AI den totalen Kontrollverlust des Menschen über die AGI. Eine derart fortgeschrittene KI würde über zahlreiche Strategien verfügen, um aktiv Macht zu erlangen und ihr Überleben zu sichern. Dies kann jedoch nicht nur ein Resultat des unbeabsichtigten Kontrollverlustes des Menschen über die KI, sondern auch eine gezielte Entwicklung böswilliger Personen sein. In diesem Kontext könnten hochgradig tödliche und ansteckende Biowaffen entwickelt werden. So gehen Hendrycks et al. (2023) mit ihrer Annahme so weit, dass Rogue AI das Potenzial zur Massenvernichtung besitzen könnte. Dieses existenzielle Risiko wird von Hendrycks und Mazeika (2022) unter dem Begriff *X-Risk* diskutiert.

X-Risk. Gemäss Hendrycks und Mazeika (2022) könnte die Entwicklung zunehmend intelligenterer und leistungsfähigerer KI-Systeme zur Entwicklung von AGI führen – Systeme, welche weitaus leistungsfähiger als Menschen sind. Dies wird teilweise als Spiel mit dem Feuer betrachtet, welches das X-Risk mit sich bringe. Das in der Populärkultur häufig illustrierte Szenario, bei dem die Roboter ein eigenes Bewusstsein entwickeln, gegen die Menschheit rebellieren und beschliessen, diese zu auszulöschen, ist zwar möglich, aber unwahrscheinlich (Yampolskiy & Spellchecker, 2016). Vielmehr könnte gemäss Yampolskiy und Spellchecker (2016) das X-Risk dadurch entstehen, dass Menschen absichtlich unethisch handeln, technische Fehler aufgrund von mangelhafter Entwicklung auftreten und Umweltbedingungen das System beeinflussen.

Auch unabhängig von dieser fatalen Perspektive zeigt sich, dass die Nutzung von KI-Systemen mit Risiken verbunden ist, womit diese nach dem Verständnis von Mayer et al. (1995) mit Vertrauen assoziiert werden kann. Dabei kann KI als eine Erweiterung der Automation (Tschopp & Ruef, 2020) und als Technologie verstanden werden, welche Aufgaben ähnlich wie der Mensch ausführen kann (Bedué & Fritzsche, 2021). Aufgrund der autonomen Funktionsweise bringt dies neue Risiken und Ungewissheiten mit sich (Choung et al., 2023). Diese Gefahren werden unter dem Oberbegriff AI Safety diskutiert (u. a. Hendrycks & Mazeika, 2022; Hendrycks et al., 2023; Yampolskiy & Spellchecker, 2016) und resultieren in der Forderung, KI-Systeme transparenter zu gestalten (u. a. Gebru et al., 2022; Shin, 2021) um das Vertrauen in diese zu fördern (Gebru et al., 2022). Transparenz allein führt jedoch nicht zu einem adäquaten Vertrauen (Okamura & Yamada, 2020). Das Konstrukt des Vertrauens und dessen Relevanz im Kontext von KI werden im folgenden Kapitel «Vertrauen» genauer erörtert.

Vertrauen

Wie im Kapitel «Künstliche Intelligenz» erörtert, wird Vertrauen in Bezug auf KI eine hohe Relevanz zugesprochen (u. a. Gebru et al., 2022; Hoffman et al., 2023; Siau & Wang, 2018). Wie bereits erwähnt, variiert aber das Verständnis darüber, was grundsätzlich darunter zu verstehen ist (Lewis & Marsh, 2022; Tschopp & Ruef, 2020). Aus diesem Grund wird im vorliegenden Kapitel zunächst ein allgemeines Verständnis zum Konstrukt Vertrauen geschaffen. Anschliessend wird auf konkrete Erkenntnisse in Bezug auf Vertrauen in KI eingegangen. Da die Forschung in diesem Bereich zumeist auf den etablierten Modellen *Vertrauen in Organisationen* von Mayer et al. (1995) und *Vertrauen in Automation* von Lee und See (2004) basiert, geht der Verfasser dieser Arbeit abschliessend auf diese beiden Theorien ein. Damit soll dargestellt werden, welche theoretischen Aspekte dem Vertrauen in KI zugrunde liegen, und ein klares Bild zu diesen Theorien vermittelt werden. Für dieses Kapitel wird kein Anspruch auf eine perfekte Synthese dieser Theorien gestellt; vielmehr sollen Unterschiede und Zusammenhänge sowie die für diese Untersuchung relevanten Aspekte aufgezeigt werden

In der Vertrauensforschung standen lange Zeit das Paradigma des kooperativen Verhaltens von Deutsch (1958) und Vertrauen als Persönlichkeitsmerkmal nach Rotter (1967) im Mittelpunkt (beide zitiert nach Siegrist, 2021, S. 480). Die Bedeutung des Konzepts des Vertrauens hat seit Mitte der 1990er Jahre (Lee & See, 2004; Siegrist, 2021) und mit dem Aufkommen selbstorganisierter Teams bzw. der eigenverantwortlichen Mitarbeit zugenommen, da klassische Kontrollmechanismen des Managements reduziert oder abgeschafft wurden (Mayer et al., 1995). So muss im Rahmen dieser Arbeitsgestaltung das Vertrauen die Funktion der Kontrolle übernehmen, da eine direkte Beobachtung der Mitarbeitenden nicht mehr möglich ist. Zudem hat sich Vertrauen als nützliches Konzept zur Beschreibung der Interaktion mit Internetanwendungen erwiesen (Lee & See, 2004).

Rousseau, Sitkin, Burt und Camerer (1998) definieren Vertrauen als einen psychologischen Zustand, der die Absicht umfasst, Vulnerabilität aufgrund positiver Erwartungen an die Absichten oder das Verhalten einer anderen Person zu akzeptieren. Hingegen verstehen Lee und See (2004) darunter die Haltung, dass ein Gegenüber dazu beitragen wird, die Ziele einer Person in einer Situation zu erreichen, die durch Unsicherheit und Vulnerabilität gekennzeichnet ist. Sich in dieser Hinsicht vulnerabel zu machen bedeutet, ein Risiko einzugehen (Mayer et al., 1995). Gleichzeitig kann es auch riskant sein, anderen nicht zu vertrauen (Luhmann, 1989, zitiert nach Siegrist, 2021, S. 481). Demnach

kann Vertrauen als wesentlicher Faktor in der Risikowahrnehmung angesehen werden, auch wenn dies nicht abschliessend erforscht ist (Siegrist, 2021). Dabei ist Vertrauen nicht grundsätzlich als Risiko zu erachten, sondern vielmehr als die Bereitschaft, ein solches einzugehen, wobei das Vertrauensverhalten dem Eingehen eines Risikos entspricht (Mayer et al., 1995). Anknüpfend daran definieren Mayer et al. (1995, S. 712, durch die Verfasser dieser Arbeit aus dem Englischen übersetzt) Vertrauen als die Bereitschaft einer Seite, sich für die Handlungen einer anderen Seite vulnerabel zu machen, die auf der Erwartung beruht, dass die andere Seite eine bestimmte, für die vertrauende Person wichtige Handlung vornimmt, unabhängig von der Möglichkeit, diese andere Seite zu überwachen oder zu kontrollieren. Gemäss Rousseau et al. (1998) stellt diese Beschreibung die am häufigsten verwendete und akzeptierte Definition von Vertrauen dar, welche demnach auch für diese Arbeit herangezogen wird.

Wie einleitend erwähnt, stellt bereits die Definition von Vertrauen eine Herausforderung dar (Lewis & Marsh, 2022; Tschopp & Ruef, 2020), was durch die obigen Erläuterungen deutlich geworden sein sollte. Zusammenfassend kann Vertrauen als Einstellung einer vertrauenden Person *(Trustor)* betrachtet werden, die durch die Vertrauenswürdigkeit als Eigenschaft eines Gegenüber *(Trustee)*, welchem vertraut werden soll, beeinflusst wird (Tschopp & Ruef, 2020). Grundlage des Vertrauens ist dabei die Bereitschaft, ein Risiko einzugehen (Mayer et al., 1995). Mayer et al. (1995, S. 712–714) führen an, dass für Vertrauen auch Begriffe wie (a) *Kooperation (Cooperation),* (b) *Zuversicht (Confidence)* und (c) *Berechenbarkeit (Predictability)* verwendet werden, welche jedoch hiervon abzugrenzen sind. Obwohl Vertrauen zu kooperativem Verhalten führen kann, ist es keine notwendige Bedingung für Kooperation, da mit Letzterer nicht zwangsweise ein Risiko einhergehen muss. So können auch Kontrollmechanismen und der Mangel an Alternativen zur Kooperation führen, selbst wenn kein Vertrauen besteht. Zwar bezieht sich Zuversicht ebenso wie Vertrauen auf Erwartungen, die enttäuscht werden können, jedoch muss bei Letzterem zunächst erkannt und akzeptiert werden, dass ein Risiko besteht, sich auf eine andere Seite zu verlassen. Darüber hinaus umfasst Zuversicht auch Kontrollaspekte (Luhmann, 1988, zitiert nach Mayer et al., 1995, S. 713). Siegrist (2021) führt dagegen an, dass der theoretische und praktische Wert der Unterscheidung von Vertrauen sowie Zuversicht in Frage zu stellen sei. Zudem sei kaum zu unterscheiden, welches dieser beiden Konzepte tatsächlich gemessen werde. Auch wenn sowohl Berechenbarkeit als auch Vertrauen als Mittel zur Verringerung von Ungewissheit betrachtet werden könnten, gehe

Vertrauen über die Bedeutung von Berechenbarkeit hinaus (Mayer et al., 1995). Insbesondere bedinge Letztere nicht die Bereitschaft, sich vulnerabel zu machen und damit ein Risiko einzugehen.

Generell wird Vertrauen als dynamisches Konstrukt erachtet (u. a. Hoff & Bashir, 2015; Körber, 2019; Lee & See, 2004; Mayer et al., 1995). Dementgegen bemerkt Siegrist (2021), dass Vertrauen ein über die Zeit hinweg stabiles Phänomen zu sein scheint. Dies suggerieren Längsschnittstudien, die Zweifel an der weitverbreiteten Meinung aufkommen liessen, dass Vertrauen fragil sei. So zeige sich, dass negative Informationen sich nicht unbedingt negativ auf das Vertrauen in eine Person oder Institution auswirken, da neue Informationen oft im Einklang mit bestehenden Überzeugungen interpretiert werden würden. Die hierzu von Siegrist (2021) angeführten Beispiele im Kontext des Vertrauens in Atomenergie umfassen jedoch weniger das Vertrauen in die Technologie als vielmehr jenes in die Aufsichtsbehörden und vorhandene Kontrollmechanismen, was gemäss Mayer et al. (1995) mit Zuversicht assoziiert werden kann. Dabei können starke Kontrollmechanismen die Entwicklung von Vertrauen unterbinden, da die Handlungen eines Trustees als Reaktion auf diese und nicht als Zeichen von Vertrauenswürdigkeit interpretiert werden können (Mayer et al., 1995). Lee und See (2004) beschreiben mit dem Begriff *Spezifität (Specificity)* die Veränderungen des Vertrauens in Abhängigkeit von der Situation und dem Zeitverlauf. Dabei unterscheiden sie zwischen *hoher zeitlicher Spezifität (High Temporal Specificity)*, welche kurzfristige Schwankungen im Vertrauen beinhaltet, und *geringer zeitlicher Spezifität (Low Temporal Specificity)* was langfristige Veränderungen des Vertrauens umfasst. Sie führen zudem an, dass Vertrauen eher stabil ist, wenn dieses auf mehreren Faktoren beruhe, während es tendenziell fragil ausfällt, wenn es von einer einzigen Basis abhänge. So ist Vertrauen, das auf dem Verständnis der Motive des Trustees basiert, stabiler als Vertrauen, das nur auf der Zuverlässigkeit der Leistung beruht (Lee & See, 2004).

Vertrauen wird zumindest als bidimensional erachtet. So stellen Vertrauen und *Misstrauen (Distrust)* keine Gegensätze dar, jedoch ist das Verhältnis nicht eindeutig geklärt (Razin & Feigh, 2023). Lee und See (2004) verstehen unter Misstrauen, dass das Vertrauen geringer ist als die Vertrauenswürdigkeit des Trustees es erlaubt. Dementgegen wird dem Trustee bei *Übervertrauen (Overtrust)* ein höheres Vertrauen gewährt, als es die Vertrauenswürdigkeit zulässt. In dieser Hinsicht kann Misstrauen als Gegenteil von Übervertrauen erachtet werden, wobei beides eine mangelnde Kalibrierung von Vertrauen zu Vertrauenswürdigkeit darstellt (Lee & See, 2004).

Vertrauen wird selten als Heuristik bezeichnet, erfüllt jedoch eindeutig die entsprechenden Kriterien (Lewis & Marsh, 2022; Siegrist, 2021). Konkret haben Menschen die Fähigkeit, die Vertrauenswürdigkeit anderer anhand subtiler Hinweise genau einzuschätzen (Hoff & Bashir, 2015). Dabei vereinfacht Vertrauen die Entscheidungsfindung bei fehlendem Wissen (Siegrist, 2021) und ersetzt die Kontrolle, wenn eine direkte Beobachtung nicht möglich ist (Lee & See, 2004). In Fällen von Ungewissheit basiert die menschliche Entscheidungsfindung zumeist auf Heuristiken, zu denen auch Vertrauen gezählt werden kann (Lewis & Marsh, 2022). Letzteres wirkt sich in diesem Zusammenhang auf die Risikobereitschaft des Trustors aus (Mayer et al., 1995). Zudem führt Siegrist (2021) an, dass Vertrauen die wahrgenommene Nützlichkeit von Technologien direkt beeinflusst und daher eine zentrale Rolle bei der Technologieakzeptanz spielt. So wurde beispielsweise das *Technology Acceptance Model (TAM)* mit den Einflussfaktoren wahrgenommene *Nützlichkeit (Perceived Usefulness)* und *Nutzungsfreundlichkeit (Perceived Ease of Use*) durch Choung et al. (2023) um den Faktor Vertrauen erweitert, um die Akzeptanz von KI-Systemen besser zu erklären.

Zusammengefasst lässt sich sagen, das Vertrauen bereits Mitte des zwanzigsten Jahrhunderts intensiv erforscht wurde, wobei dessen Relevanz in den 1990er Jahren aufgrund neuer Organisationsformen stark zunahm (Siegrist, 2021). Vertrauen bedingt die Bereitschaft, ein Risiko einzugehen, und muss entsprechend von anderen Konstrukten wie (a) Kooperation, (b) Zuversicht und (c) Berechenbarkeit abgegrenzt werden (Mayer et al., 1995). Auch wenn es im Allgemeinen als ein dynamisches Konstrukt verstanden wird, gibt es auch stabile, die Zeit überdauernde Aspekte (Lee & See, 2004). Auch wenn das Verhältnis zwischen Vertrauen und Misstrauen nicht abschliessend geklärt ist, stellen dies keine Gegensätze dar (Razin & Feigh, 2023). Misstrauen wird im Kontext von adäquatem Vertrauen dahingehend erklärt, dass das Vertrauen des Trustors gegenüber dem Trustee geringer ist als dessen Vertrauenswürdigkeit es zulassen würde (Lee & See, 2004). Obwohl es nur selten so bezeichnet wird, kann Vertrauen als Heuristik erachtet werden, welche die menschliche Entscheidungsfindung bei Ungewissheit vereinfacht (Lewis & Marsh, 2022; Siegrist, 2021). Dabei spielt Vertrauen in der Technologieakzeptanz eine wesentliche Rolle (Siegrist, 2021), so auch im Kontext von KI (Choung et al., 2023). Nachdem ein allgemeines Verständnis über das Konstrukt Vertrauen geschaffen wurde, wird im folgenden Unterkapitel «Vertrauen in KI» explizit auf Vertrauen in KI eingegangen.

Vertrauen in KI. Mit steigender Relevanz von KI gewinnt auch die Vertrauensforschung grösseren Zulauf (Bedué & Fritzsche, 2021). So schreiben Hoffman et al. (2023) sowie Bedué und Fritzsche (2021) Vertrauen einen zentralen Einfluss auf sowohl die Entwicklung als auch die Nutzung von KI zu. Des Weiteren erachten Gebru et al. (2022) Vertrauen als elementaren Aspekt bei der Implementierung von Technologie. Dabei erläutert Shin (2021), dass Vertrauen in KI auf der Überzeugung beruht, dass KI-Systeme vertrauenswürdig bzw. zuverlässig arbeiten, und entsprechend die Zuverlässigkeit sowie die Glaubwürdigkeit des Systems reflektiert. Demnach sollten Organisationen, welche KI-Systeme anbieten, die Genauigkeit der Ergebnisse sicherstellen (Shin, 2021). Gleichzeitig müssen die Fähigkeiten des Systems mit dem Ausmass des Vertrauens durch die Nutzenden übereinstimmen, um einen sorglosen Umgang mit diesem zu unterbinden (Gebru et al., 2022). In diesem Kontext kann es als positives Zeichen gewertet werden, dass es gemäss Hoffman et al. (2023) der Norm entspricht, wenn Nutzende, welche bei spezifischen Aufgaben und Zielen einem KI-System vertrauen, dies bei anderen Aufgaben oder in bestimmten Situation wiederum nicht tun.

In ihrer Studie zur KI-gestützten Personalauswahl gewannen Langer et al. (2023) jedoch die Erkenntnis, dass entgegen der Erwartung an automatisierte Systeme bei solchen KI-Systemen keine hohe Genauigkeit angenommen wird. Trotz dessen gehen sie davon aus, dass auch im Kontext von KI die klassischen Vertrauenskonstrukte in Bezug auf Technologie und Mensch anwendbar sind (Langer et al., 2023). Dabei lässt sich Vertrauen in Technologie teilweise mit sozialem Vertrauen assoziieren; inwieweit dies im Kontext von KI zutrifft, ist jedoch umstritten (Langer et al., 2023; Tschopp & Ruef, 2020). Hierbei spielt nicht nur die Vulnerabilität des Trustors, sondern auch die technische Vulnerabilität des KI-Systems, welche die Vertrauenswürdigkeit beeinflusst, eine Rolle (Tschopp & Ruef, 2020). Dabei ist es naheliegend, dass die Forschung in diesem Kontext unter anderem auf den Nutzen, den möglichen Missbrauch, die Autonomie sowie die Fairness und die Erklärbarkeit von KI als Faktoren der Vertrauenswürdigkeit fokussiert ist (Choung et al., 2023).

Shin (2021) zeigt auf, dass Vertrauen in KI durch (a) Fairness, (b) Verantwortlichkeit und (c) Transparenz, welche durch die Erklärbarkeit aufgezeigt werden, beeinflusst wird. Tschopp und Ruef (2020) weisen jedoch darauf hin, dass die Erhöhung der Transparenz von KI-Systemen nicht, zur Steigerung des Vertrauens führt, da hier vielmehr das Vertrauen durch Kontrolle ersetzt werde. So zweifeln sie die Notwendigkeit von Vertrauen in KI an, da diese Systeme so ausgelegt werden könnten, dass sie kontrollierbar seien.

Vertrauenswürdigkeit von KI. In Fällen, wo Vertrauen als Einstellung eines Trustors gesehen werden kann, beschreibt Vertrauenswürdigkeit (Trustworthiness) unterschiedliche Eigenschaften des Trustees, welche sich idealerweise auf das Vertrauen auswirken (Lewis & Marsh, 2022; Tschopp & Ruef, 2020). Die Bewertung der Vertrauenswürdigkeit basiert dabei jedoch nicht nur auf eigenen Erfahrungen und Beobachtungen, sondern auch auf jenen von Dritten (Bedué & Fritzsche, 2021; Lewis & Marsh, 2022). Beispielsweise kann sich die Wahrnehmung Dritter auf die Reputation einer Organisation auswirken.

Die Vertrauenswürdigkeit von KI-Systemen allein schafft dabei nicht unbedingt Vertrauen, geschweige denn adäquates Vertrauen, sofern sich der Trustor nicht selbst von den Fähigkeiten des Systems überzeugen kann (Gebru et al., 2022). Dabei kann die Wahrnehmung der Vertrauenswürdigkeit neben der Interaktion mit dem KI-System auch durch die Umwelt (Bedué & Fritzsche, 2021), das Umfeld, die Einstellung zur Technologie und die Komplexität des Systems beeinflusst werden (Gebru et al., 2022). So ist Vertrauen kein Merkmal, das durch eine Organisation, welche KI-Systeme anbietet, gesteigert werden kann. Vielmehr kann es durch das Unter-Beweis-Stellen der Vertrauenswürdigkeit verdient werden (Tschopp & Ruef, 2020). Schepman und Rodway (2023) erachten die Fähigkeit des KI-Systems, eine Aufgabe zuverlässig zu erfüllen (Reliance), als primären Aspekt des Vertrauens. Diese Leistung bzw. deren Robustheit ist jedoch nicht der einzige Aspekt, welcher Vertrauen beeinflusst (Tschopp & Ruef, 2020). Auch der Ruf der Organisation, welche das KI-System zur Nutzung bereitstellt, hat in Hinblick auf das wahrgenommene Wohlwollen und den Zweck des Systems einen grossen Einfluss auf das Vertrauen (Tschopp & Ruef, 2020).

Bei der Bewertung der Vertrauenswürdigkeit von KI müssen deren (a) Leistung (Performance), deren (b) Funktionalität (Process) und deren (c) Zweck (Purpose) berücksichtigt werden (Siau & Wang, 2018). Das konzeptuelle Modell von Solberg et al. (2022) im Kontext von KI-Entscheidungshilfen (siehe Abbildung 1) basiert auf dem Modell Vertrauen in Automation von Lee und See (2004), welches im Unterkapitel «Vertrauen in Automation» erörtert wird, und umfasst dieselben drei Faktoren: (a) Performance, (b) Process und (c) Purpose der Vertrauenswürdigkeit von KI, welche im Folgenden genauer beschrieben werden.

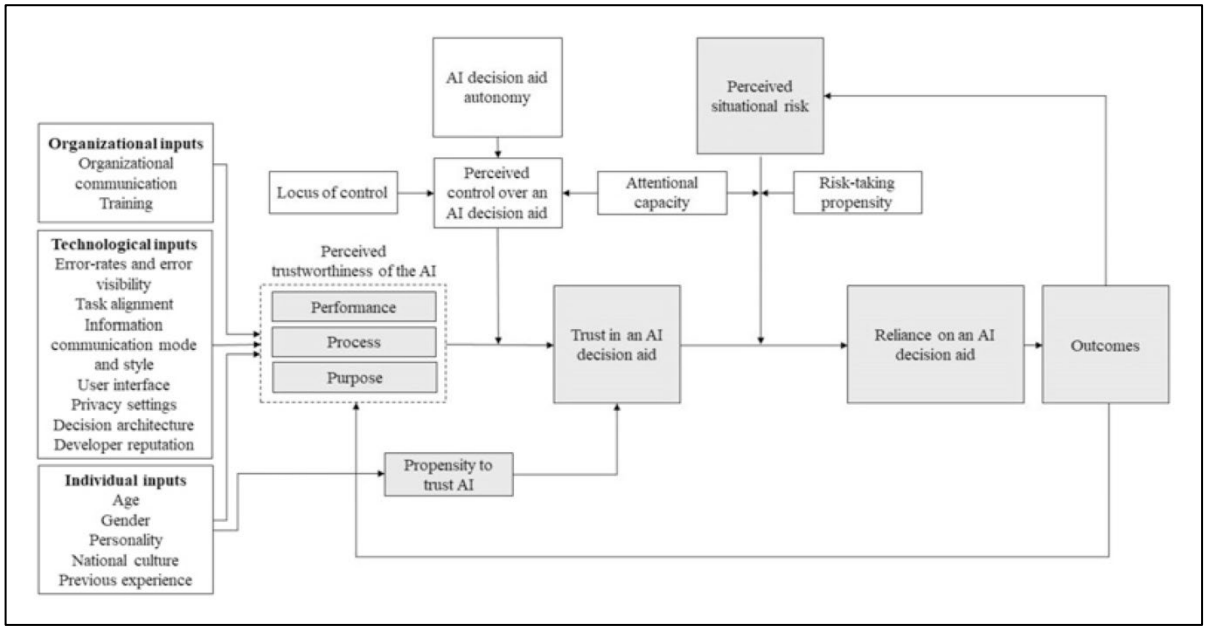


Abbildung 1. Modell zu Vertrauen in KI-Entscheidungshilfen (Solberg et al., 2022)

Performance. Die Fähigkeit eines KI-Systems, Aufgaben zuverlässig zu erfüllen, wird unter den Faktor Performance gefasst (Lewis & Marsh, 2022; Siau & Wang, 2018; Solberg et al., 2022). Dabei spielt es keine Rolle, ob es sich bei dem KI-System um ein Blackbox-Modell handelt, da sich auch in diesem Fall die Fähigkeiten des Systems anhand der Ergebnisse beobachten lassen (Lewis & Marsh, 2022). Dabei haben sich Leistung und Zuverlässigkeit als starke Prädiktoren für Vertrauen in KI erwiesen (Kaplan et al., 2023). Bedué und Fritzsche (2021) erachten Letztere als einen relevanten Faktor für die Steigerung von Vertrauen. Performance spielt hingegen nicht nur bei der initialen Vertrauensbildung eine Rolle, denn das initiale Vertrauen resultiert auch in Erfahrungen, welche sich wiederum auf die wahrgenommene Vertrauenswürdigkeit der Leistung auswirken (Solberg et al., 2022). Des Weiteren sind auch Rezensionen anderer Nutzender im Rahmen der initialen Vertrauensbildung von Relevanz (Siau & Wang, 2018). Konkret verstärken positive Rezensionen über das KI-System das initiale Vertrauen in dieses. Ergänzend ordnen Siau und Wang (2018) auch die Nutzungsfreundlichkeit, welche zum kontinuierlichen Vertrauen beiträgt, dem Faktor Performance zu. Folglich sollten KI-System so entwickelt werden, dass diese unkompliziert und intuitiv bedient werden können.

Process. Das Verständnis über die Funktionsweise und die Programmierung eines KI-Systems wird mit dem Faktor Process beschrieben (Siau & Wang, 2018). Dieser gilt als signifikanter Prädiktor für das KI-Vertrauen (Kaplan et al., 2023). Mit ihm werden Aspekte

wie (a) Transparenz (Siau & Wang, 2018; Solberg et al., 2022), (b) Berechenbarkeit (Lewis & Marsh, 2022) und (c) Erklärbarkeit (Siau & Wang, 2018) assoziiert. Siau und Wang (2018) halten Transparenz sogar für eine Notwendigkeit, um KI-Systemen vertrauen zu können; demnach sollten diese Systeme in der Lage sein, ihre Ergebnisse zu rechtfertigen. Zudem weisen Solberg et al. (2022) darauf hin, dass Transparenz einen positiven Einfluss auf die wahrgenommene Vertrauenswürdigkeit hat. Bedué und Fritzsche (2021) gehen in diesem Zusammenhang davon aus, dass Transparenz bei KI eine deutlich grössere Rolle als bei herkömmlichen Technologien spielt. Hingegen repräsentiert die Berechenbarkeit die Beständigkeit der Leistung und somit die Wahrscheinlichkeit, dass das Ergebnis des KI-Systems zuverlässig die Erwartungen des Trustors erfüllt (Lewis & Marsh, 2022). Die Erklärbarkeit wird von Bedué und Fritzsche (2021) als Notwendigkeit für die Interpretierbarkeit von KI-Ergebnissen und daher ebenfalls als wesentliche Anforderung an das Vertrauen in KI erachtet. Entsprechend nehmen Siau und Wang (2018) an, dass der Mangel an Erklärbarkeit bei Blackbox-Modellen sich negativ auf das Vertrauen auswirkt. Einen weiteren Aspekt des Process-Faktors stellt die Integrität (Integrity) dar (Bedué & Fritzsche, 2021). Nach dem Verständnis von Lee und See (2004) kann Integrity mit Process verglichen werden; jedoch haben Bedué und Fritzsche (2021) hierbei ein Compliance-basiertes Verständnis. So erachten sie Standards und Richtlinien, Zertifizierungen sowie staatliche Regulierung als Voraussetzungen für den Aufbau von Integrität.

Purpose. Die Einschätzung, dass die Ziele des KI-Systems mit den eigenen Zielen der Nutzenden kongruent sind, wird dem Faktor Purpose zugeordnet (Siau & Wang, 2018). In anderen Worten, beschreibt dieser die Überzeugung einer Person, dass das KI-System sie bei der Erfüllung einer Aufgabe unterstützt, ihr jedoch nicht schadet (Solberg et al., 2022). Da einem solchen System kein bewusstes Wohlwollen zugerechnet werden kann, ist vielmehr die Wahrnehmung des Zwecks der Entwicklung des Systems sowie des Wohlwollens der Entwickler und Entwicklerinnen ausschlaggebend für diese Bewertung (Solberg et al., 2022). Demnach gelten sowohl die Reputation der Organisation, welche das KI-System zur Verfügung stellt, als auch das Verhalten des Systems als signifikante Prädiktoren für Vertrauen in KI (Kaplan et al., 2023). Gemäss Bedué und Fritzsche (2021) spielen dabei (a) die soziale Verantwortung, (b) das ethische Verhalten und (c) die Nachhaltigkeit in Bezug auf die Organisation eine Rolle. Insbesondere aufgrund der noch geringen Erfahrung mit KI-Systemen schreiben diese dem Faktor Purpose eine entscheidende Rolle hinsichtlich der wahrgenommenen Vertrauenswürdigkeit von KI zu.

Neigung zum Vertrauen in KI. Neben der Vertrauenswürdigkeit haben auch Merkmale des Trustors Einfluss auf das Vertrauen in KI (Sindermann et al., 2022), darunter zahlreiche inter- und intrapersonelle Faktoren (Sindermann et al., 2022; Solberg et al., 2022). Dieser Aspekt wird auch als Propensity to Trust oder Dispositional Trust bezeichnet (Montag, Kraus et al., 2023; Tschopp & Ruef, 2020). Die damit gemeinte Neigung wird als relevant für das Verständnis von Akzeptanz und Vertrauen in KI-Systeme erachtet (Sindermann et al., 2022). Es hat sich gezeigt, dass die Propensity to Trust einen direkten Einfluss auf das Vertrauen hat (Solberg et al., 2022).

Zu den Faktoren, welche die Propensity to Trust beeinflussen sollen, zählen unter anderem die Big-Five-Persönlichkeitsmerkmale (Kaplan et al., 2023; Sindermann et al., 2022). Die Erkenntnisse dazu variieren jedoch stark; so führen Schepman und Rodway (2023) an, dass sich der Einfluss der Persönlichkeit abhängig von der spezifischen Technologie sowie des Messinstruments unterschiedlich gestalten kann. Daher ist es nicht überraschend, dass Kaplan et al. (2023) in Bezug auf Extraversion auf gegensätzliche Erkenntnisse hinweisen, obwohl Schepman und Rodway (2023) einen Zusammenhang feststellen konnten. Gleichzeitig weisen Kaplan et al. (2023) auf einen negativen Zusammenhang mit Neurotizismus hin. Zu einem ähnlichen Ergebnis kamen auch Sindermann et al. (2022) bezüglich einer deutschen Stichprobe, wobei sich für Neurotizismus als einzigen der fünf Persönlichkeitsfaktoren ein Zusammenhang mit der Einstellung zu KI feststellen liess. Hingegen zeigten sich bei einer chinesischen Stichprobe Zusammenhänge mit Verträglichkeit und Offenheit. Dies deutet darauf hin, dass es kulturelle Unterschiede gibt (Solberg et al., 2022). Allerdings beobachteten auch Schepman und Rodway (2023) einen Zusammenhang mit Verträglichkeit sowie Gewissenhaftigkeit.

Neben den Persönlichkeitsmerkmalen wird auch dem Geschlecht ein Einfluss auf das Vertrauen zugeschrieben, wobei Männer ein grösseres Vertrauen vorweisen als Frauen (Kaplan et al., 2023). Dies wiederspricht wiederum den Erkenntnissen von Montag, Kraus, et al. (2023), welche keinen Geschlechtsunterschied feststellen konnten. Beim Alter hingegen konnte von ihnen ein negativer Zusammenhang identifiziert werden.

Die Ergebnisse von Kaplan et al. (2023) deuten darauf hin, dass sich Erfahrung auf das Vertrauen in KI auswirkt. Zudem setzt Letzteres ein bestimmtes Mass an Kenntnissen voraus (Frank et al., 2023). In diesem Zusammenhang gehen auch Solberg et al. (2022) davon aus, dass Systemerfahrung aufgrund des dadurch gewonnenen Verständnisses Einfluss auf die Neigung zum Vertrauen hat.

Nutzungsintention bei KI. Die Nutzungsintention bezüglich KI-Systemen wird häufig unter Einsatz des TAM von Davis (1989, zitiert nach Frank et al., 2023), einer Theorie zur Erklärung der Akzeptanz und der Nutzung von Technologie, untersucht. Choung et al. (2023) etwa ergänzen in ihrem Modell auf Basis des TAM Vertrauen als indirekten Einflussfaktor auf die Nutzungsintention von KI. Dabei beeinflusst dieses die Einstellung zum KI-System und Letztere wiederum die Nutzungsintention. Zudem ist Vertrauen bei der Akzeptanzentscheidung bezüglich Innovationen – wobei KI auch als solche betrachtet werden kann – ein entscheidender Faktor (Frank et al., 2023). Vereinfacht gesagt werden Personen, welche einem System nicht vertrauen, dieses auch nicht nutzen wollen (Tschopp & Ruef, 2020).

Jedoch spielt nicht nur das Vertrauen in das System, sondern auch jenes in die Organisation, welche dieses anbietet, eine Rolle in Hinblick auf die Nutzungsintention (Frank et al., 2023). So steigert das Vertrauen in eine Organisation die Nutzung von deren KI-System. Dabei können die Nutzung bzw. die Nutzungsintention nach Lee und See (2004) als Vertrauensverhalten respektive Verhaltensergebnis von Vertrauen interpretiert werden. Auch Bedué und Fritzsche (2021) weisen auf einen Zusammenhang zwischen Vertrauen und Nutzungsintention hin. Dabei fokussieren sie sich auf das Valence-Framework von Peter und Tarpey (1975, zitiert nach Bedué & Fritzsche, 2021, S. 532), das sie durch die Faktoren der Vertrauenswürdigkeit von Mayer et al. (1995) ergänzen. In diesem konzeptuellen Modell (siehe Abbildung 2) mediiert Vertrauen den Zusammenhang von Vertrauenswürdigkeit mit der Nutzungsintention, ergänzt durch die indirekten Zusammenhänge zwischen Vertrauen und Nutzungsintention, wobei diese durch die wahrgenommenen Vorteile bzw. Risiken vermittelt werden (Bedué & Fritzsche, 2021).

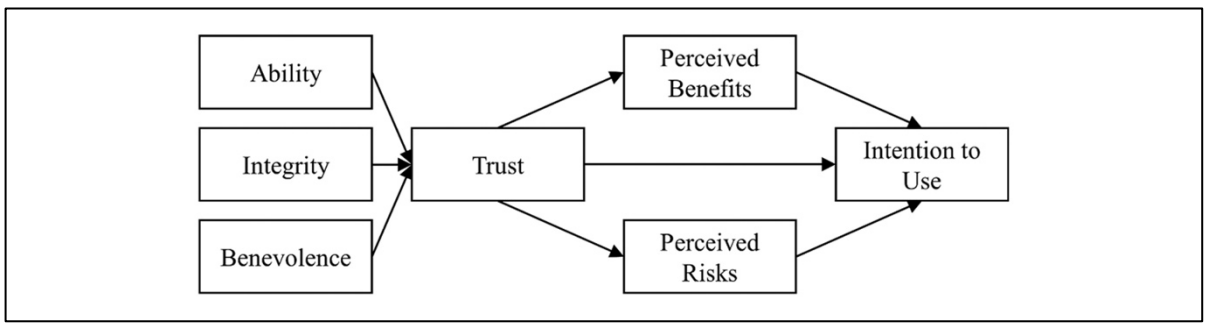


Abbildung 2. Erweitertes Valence-Framework zu KI (Bedué & Fritzsche, 2021)

Dynamisches Vertrauen in KI. Sowohl Choung et al. (2023) als auch Siau und Wang (2018) betrachten Vertrauen als ein dynamisches Konstrukt, welches abhängig von der Technologiereife, der Leistung sowie dem Zweck des KI-Systems variiert. Während Vertrauen statische wie auch dynamische Aspekte beinhaltet, gelten Persönlichkeitsaspekte, welche das Vertrauen beeinflussen, als statisch (Tschopp & Ruef, 2018). Hingegen ist der Vertrauensbildungsprozess, abhängig vom Ergebnis des Vertrauensverhaltens, als dynamisch zu verstehen (Langer et al., 2023). Gleichzeitig zeigen Langer et al. (2023) auf, dass eine Vertrauensverletzung das Vertrauen nur geringfügig beeinflusst. Dies führen sie darauf zurück, dass die Erwartungen an die Fähigkeiten des KI-Systems in ihrer Studie vergleichsweise gering waren und die Genauigkeit variieren kann.

Adäquates Vertrauen in KI. Tschopp und Ruef (2020) führen an, dass Nutzende skeptisch gegenüber KI-Systemen sein müssten, um diesen adäquat vertrauen zu können. Hoffman et al. (2023) weisen darauf hin, dass adäquates Vertrauen Erfahrung mit dem System bedingt. Dabei nennen beide Forschungsteams als Voraussetzung für adäquates Vertrauen eine korrekte Einschätzung der Fähigkeiten des Systems. Dem steht gegenüber, dass Menschen die Fähigkeiten von Systemen häufig überschätzen oder sogar von Perfektion ausgehen und diesen ein inadäquates Vertrauen entgegenbringen (Tschopp & Ruef, 2020).

Relevanz des Vertrauens in KI. Dem Faktor Vertrauen wird eine hohe Relevanz für den Erfolg eines KI-Systems zugeschrieben (Bedué & Fritzsche, 2021). Dabei beeinflusst die wahrgenommene Vertrauenswürdigkeit des Systems das Vertrauen in dieses (Lewis & Marsh, 2022; Tschopp & Ruef, 2020). Sie umfasst (a) die Leistung, (b) die Funktionalität und (c) den Zweck des Systems (Siau & Wang, 2018). Jedoch hat auch die Neigungen des Trustors einen Einfluss (Sindermann et al., 2022). Im Kontext von KI wird Vertrauen dennoch häufig auf Regeln, Richtlinien oder Erklärbarkeit begrenzt, obwohl es auch als emotionale Reaktion einzuordnen ist und entsprechend nicht auf rationale Aspekte reduziert werden sollte (Tschopp & Ruef, 2020). In Bezug auf ChatGPT erscheint dem Verfasser dieser Arbeit vor allem eine mögliche Überschätzung der Fähigkeiten und ein damit einhergehendes inadäquates Vertrauen als relevanter Aspekt. So ist es notwendig zu verstehen, welche Faktoren neben den Fähigkeiten das Vertrauen beeinflussen. Dabei erscheint es unklar, ob bestehende Theorien zu Vertrauen auf den Kontext KI angewendet werden können (Shin, 2021). Dennoch wurde in diesem Abschnitt ersichtlich, dass die Forschung auf etablierten Konstrukten basiert. Für ein besseres Verständnis werden zwei dieser Konstrukte in den folgenden beiden Unterkapiteln genauer erklärt.

Vertrauen in Organisationen. In Abwesenheit eines generellen Modells zu Vertrauen in KI wird dieses häufig am etablierten Konstrukt von Vertrauen in Organisationen von Mayer et al. (1995) untersucht (u. a. Bedué & Fritzsche, 2021; Langer et al., 2023; Lewis & Marsh, 2022; Solberg et al., 2022). In diesem Kontext wurde Vertrauen als entscheidender Faktor für die Steigerung der Produktivität und die Stärkung des Engagements identifiziert (Lee & See, 2004). Zudem gilt in Zusammenhang mit dem Internethandel das Vertrauen zwischen Unternehmen und Kundschaft als ein zentraler Faktor (Sheridan & Hennessy, 1984). Dabei wird Vertrauen gemäss Mayer et al. (1995) nicht nur vom Trustor sondern auch vom Trustee, also der zu vertrauenden Organisation beeinflusst.

Vertrauenswürdigkeit von Organisationen. Die wahrgenommene Vertrauenswürdigkeit *(Perceived Trustworthiness)* einer Organisation umfasst gemäss Mayer et al. (1995) die drei Faktoren (a) Ability, (b) Integrity und (c) Benevolence, welche einen Grossteil der Vertrauenswürdigkeit erklären. Diese drei Faktoren interagieren miteinander, können aber auch unabhängig voneinander variieren. Ability repräsentiert Fähigkeiten, Kompetenzen und Eigenschaften, die es dem Trustee ermöglichen, in einem spezifischen Bereich Einfluss zu nehmen. Sie kann demnach mit dem Faktor Performance aus dem Abschnitt «Vertrauen in KI» verglichen werden. Unter Integrity wird verstanden, dass der Trustee sich an eigene Prinzipien hält, die aus Sicht des Trustors akzeptabel sind. Dieser Faktor kann insofern mit dem Faktor Process aus dem erwähnten Unterkapitel assoziiert werden, als (a) Transparenz, (b) Berechenbarkeit und (c) Erklärbarkeit als dazugehörige Prinzipien verstanden werden können. Benevolence wiederum beschreibt die Wahrnehmung des Wohlwollens des Trustees gegenüber dem Trustor (Mayer et al., 1995). Dies im Kontext von Vertrauen in KI mit dem Faktor Purpose verglichen werden kann.

Neigung zum Vertrauen in Organisationen. Hingegen beschreibt Propensity to Trust, die Neigung des Trustors, unabhängig von einem spezifischen Trustee zu vertrauen (Mayer et al., 1995). Diese Neigung wird durch Persönlichkeitsmerkmale, kulturelle Hintergründe sowie Entwicklungserfahrungen beeinflusst und kann entsprechend mit dem gleichnamigen Faktor aus «Vertrauen in KI» gleichgesetzt werden.

Dynamisches Vertrauen in Organisationen. Gemäss Mayer et al. (1995) entwickelt sich das Ausmass des Vertrauens im Verlauf der Interaktion zwischen Trustor und Trustee. Dabei verändert sich Ability abhängig von der Dynamik der Situation, in der die Aufgabe ausgeführt werden soll. Der Kontext der Situation wiederum trägt dazu bei, Benevolence zu beurteilen, und beeinflusst Integrity. Dabei wird die Wahrnehmung des Trustors durch den

Trustee verbessert, wenn das eingegangene Risiko zu einem positiven Ergebnis führt. Gleichzeitig gehen zahlreiche Forschende davon aus, dass sich das Vertrauen auf Grundlage von Beobachtungen und Interaktionen entwickelt (Mayer et al., 1995).

Adäquates Vertrauen in Organisationen. Häufig erachten Organisationen ein hohes Mass an Vertrauen als wünschenswert oder positiv – dieses kann jedoch auch negative Effekte haben (Siegrist, 2021). So kann ein hohes Mass an Vertrauen mit Sorglosigkeit einhergehen – was auf die Notwendigkeit eines adäquaten Vertrauens hindeutet.

Relevanz des Vertrauens in Organisationen. Gemäss Körber (2019) ist das Modell Vertrauen in Organisationen von Mayer et al. (1995) das am weitesten verbreitete Modell zu Vertrauen. Dieses hebt sich dahingehend von Modellen zu interpersonellem Vertrauen ab, da der Trustee eine Organisation und keine Person ist und Einflussfaktoren des zu vertrauenden Gegenübers aufnimmt (Mayer et al., 1995). Die Vertrauenswürdigkeit der Organisation ist in (a) Ability, (b) Integrity sowie (c) Benevolence gegliedert und kann mit den drei Faktoren aus dem Kapitel «Vertrauen in KI» verglichen werden. Auch wenn für Organisationen häufig ein hohes Vertrauen als erstrebenswert gilt, kann dieses in Sorglosigkeit münden, wenn es zu hoch ausfällt (Siegrist, 2021). Dabei spielt Vertrauen nicht nur eine entscheidende Rolle im Kontext der zunehmenden organisationalen Komplexität und der Bewältigung der damit einhergehenden Unsicherheit, sondern auch angesichts der zunehmenden technologischen Komplexität (Lee & See, 2004). In diesem Zusammenhang dient Vertrauen als Mechanismus zur Verringerung von Komplexität (Lee & See, 2004; Siegrist, 2021) und ist gleichzeitig erforderlich, um grundsätzlich ein komplexeres technisches Umfeld zu schaffen (Siegrist, 2021). Gemäss Körber (2019) lässt sich das Modell von Mayer et al. (1995) aber nicht vollständig auf automatisierte und damit KI-Systeme übertragen. Entsprechend haben Lee und See (2004) auf Basis des Modells von Mayer et al. (1995) ihr Konstrukt Vertrauen in Automation entwickelt, mit dem sie die Vertrauenswürdigkeit an den Kontext der Automation angepasst haben (Körber, 2019). Dementsprechend werden die drei Faktoren der wahrgenommen Vertrauenswürdigkeit anders benannt und erklärt. Im folgenden Kapitel wird dieses Modell ausführlich beschrieben. Dabei ist zu berücksichtigen, dass Vertrauen in KI ebenfalls häufig im Rahmen des Modells von Lee und See (2004) untersucht wird (u. a. Hoffman et al., 2023; Solberg et al., 2022). Entsprechend sind zahlreiche Parallelen zum Abschnitt «Vertrauen in Künstliche Intelligenz» zu erkennen. Aber auch in diesem Fall ist nicht abschliessend geklärt, inwieweit das Konzept auf KI anwendbar ist (Langer et al., 2023).

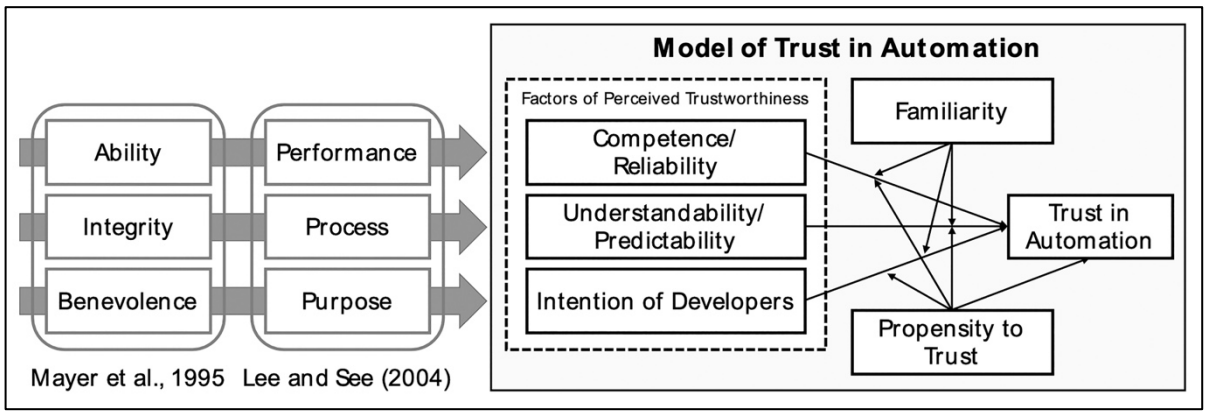
Vertrauen in Automation. Lee und See (2004) weisen darauf hin, dass Vertrauen als sozialpsychologisches Konzept für das Verständnis der Partnerschaften zwischen Mensch und Automation essenziell ist. So kann das Vertrauen in Automation als die Einstellung definiert werden, dass ein automatisiertes System dazu beitragen wird, Ziele in einer unsicheren und verwundbaren Situation zu erreichen. Dabei werden unter Automation Technologien zusammengefasst, welche (a) aktiv Daten auswählen, (b) Informationen umwandeln, (c) Entscheidungen treffen oder (d) Prozesse steuern (Lee & See, 2004) – also die Automatisierung von Aufgaben, welche zuvor von Menschen ausgeführt wurden (Merritt et al., 2019). Im Einklang hiermit existieren vier Haupttypen von automatisierten Systemen, die jeweils (a) Informationen beschaffen, (b) Informationen analysieren, (c) Entscheidungen selektieren oder (d) Handlungen umsetzen (Parasuraman, Sheridan & Wickens, 2000). Dabei unterscheiden sich diese Systeme nicht nur in ihrer Funktion, sondern auch hinsichtlich der Kontrollmöglichkeiten des bedienenden Menschen (Hoff & Bashir, 2015). Gemäss Gebru et al. (2022) gilt das Vertrauensverhältnis zwischen Mensch und Maschine als einer der entscheidenden Faktoren für eine erfolgreiche Implementierung eines derartigen Systems. So haben zahlreiche Studien gezeigt, dass Vertrauen ein sinnvolles Konzept zur Beschreibung der Human-Machine-Interaction (HMI) ist (Lee & See, 2004).

Auch wenn sich dabei das Vertrauen in die Technologien vom zwischenmenschlichen Vertrauen unterscheidet, gibt es Parallelen zwischen beiden Ansätzen (Hoff & Bashir, 2015). Insbesondere ist beobachtbar, dass Menschen sozial auf Technologie reagieren, beispielsweise ähneln die Reaktionen auf Computer jenen auf menschliche Partner (Lee & See, 2004). Zudem handelt es sich beim Vertrauen in beiden Kontexten um situationsspezifische Einstellungen, welche nur in einer von Unsicherheit geprägten Kooperationsbeziehung relevant sind (Hoff & Bashir, 2015). Dabei ist ein weiterer möglicher Grund, dass das Vertrauen der Menschen in technische Systeme auch als Vertrauen in die Entwickler und Entwicklerinnen dieser Systeme verstanden werden kann. Im Kontrast zum zwischenmenschlichen Vertrauen verhält sich das Vertrauen in Automation jedoch nicht symmetrisch, da lediglich der Mensch ein Vertrauen gegenüber der Maschine empfindet – oder nicht – aber nicht umgekehrt (Lee & See, 2004). Dabei tendieren Menschen dazu, sich auf ein System zu verlassen, welchem sie vertrauen, und lehnen jene Systeme ab, denen sie kein Vertrauen entgegenbringen. Folglich hängt gemäss Gebru et al. (2022) das Vertrauen zum einen davon ab, wie stark sich der Mensch aufgrund seiner Bereitschaft und seiner Erfahrung im Kontext der jeweiligen Situation oder Aufgabe auf das

System verlässt. Zum anderen ist es dadurch geprägt, wie gut das System die Aufgabe ausführt bzw. wie effektiv es Informationen darüber vermittelt. Vertrauen in Automation ist also eng mit der Zuverlässigkeit des Systems verbunden (Körber, 2019). Insbesondere die Berechenbarkeit und die Zuverlässigkeit des Systems sind relevante Aspekte, welche das Vertrauen beeinflussen (Hoff & Bashir, 2015). So wird berechenbaren und zuverlässigen Systemen stärker vertraut als denen, die diese Eigenschaften nicht aufweisen. Des Weiteren können emotionale Reaktionen entscheidend für das Vertrauen und die Entscheidung sein, sich auf System zu verlassen (Lee & See, 2004).

Hoff und Bashir (2015) ordnen das Vertrauen in Automation in drei Ebenen ein: (a) dispositionelles Vertrauen (Dispositional Trust), (b) situatives Vertrauen (Situational Trust) und (c) erlerntes Vertrauen (Learned Trust). Dabei wird unter dispositionellem Vertrauen die überdauernde Neigung einer Person verstanden, der Automatisierung zu vertrauen. Situatives Vertrauen hingegen bezieht sich auf den spezifischen Kontext einer Interaktion und erlerntes Vertrauen auf vergangene Erfahrungen, die für ein spezifisches System relevant sind (Hoff & Bashir, 2015). Dabei beeinflussen Letztere das situative Vertrauen.

Diese Arbeit ist jedoch an der Theorie von Lee und See (2004) orientiert, da die Trennschärfe der Theorie von Hoff und Bashir (2015) als diffus bzw. die Aufteilung auf die genannten drei Ebenen für den geplanten Versuchsaufbau als nicht geeignet erachtet wird. Das Modell von Lee und See (2004) kann als Weiterentwicklung des Modells von Mayer et al. (1995) für die Anwendung in einem Technologiekontext aufgefasst werden (Körber,



2019). Es gliedert die wahrgenommene Vertrauenswürdigkeit in drei Dimension (siehe Abbildung 3), welche im Folgenden genauer beschrieben werden.

Abbildung 3. Modell zu Vertrauen in Automation (Körber, 2019)

Vertrauenswürdigkeit von Automation. Automatisierte Systeme, welche effizient und zuverlässig arbeiten, können als vertrauenswürdige angesehen werden (Lee & See, 2004). Dabei kann die Vertrauenswürdigkeit auch unabhängig von der Interaktion durch den vorauseilenden Ruf des Systems geprägt werden (Hoff & Bashir, 2015). Gemäss Lee und See (2004) ist die wahrgenommene Vertrauenswürdigkeit in die drei Dimensionen (a) Performance, (b) Process und (c) Purpose aufgegliedert. Diese drei Faktoren können mit jenen aus dem Modell von Mayer et al. (1995) verglichen werden und entsprechen jenen, die in «Vertrauen in KI» präsentiert wurden. Sie beeinflussen ebenfalls nicht nur das Vertrauen, sondern interagieren auch miteinander (Lee & See, 2004). Körber (2019) folgt dem Modell von Lee und See (2004), unterteilt aber die drei Dimension in detailliertere Facetten, welche ursprünglich als fünf Dimensionen konzipiert wurden: (a) *Reliability/Competence*, (b) *Understandability/Predictability* und (c) *Intention of Developers*. Diese drei Faktoren werden nun im Sinn von Lee und See (2004) genauer beschrieben, um ein Verständnis zu schaffen, das unabhängig von den Adaptionen dieses Modells im Kontext von KI ist.

Performance. Performance bezieht sich auf die Handlungen des Systems und umfasst vergangene sowie aktuelle Interaktionen mit diesem (Lee & See, 2004). Dabei beinhaltet dieser Faktor Aspekte wie Zuverlässigkeit, Vorhersagbarkeit sowie Fähigkeit und kann mit Sheridans (1992, zitiert nach Lee & See, 2004, S. 59) Konzept der Robustheit verglichen werden. So basiert das Vertrauen darauf, dass das System die Ziele der nutzenden Person zuverlässig erreicht (Lee & See, 2004).

Process. Hingegen umfasst Process die Wahrnehmung, dass die Algorithmen für die Situation angemessen und in der Lage sind, die Ziele der nutzenden Person zu erreichen (Lee & See, 2004). So basiert dieser Faktor nicht auf spezifischen Handlungen, sondern vielmehr auf der Stabilität und der Integrität des Systems. Hier kann wiederum ein Vergleich mit dem Konzept der Verständlichkeit von Sheridan (1992, zitiert nach Lee & See, 2004, S. 59) gezogen werden. So wird eine Person einem System vertrauen, wenn dessen Algorithmen verständlich und fähig sind, die gesetzten Ziele zu erreichen (Lee & See, 2004). Jedoch sind Algorithmen leistungsfähiger Systeme teils komplex und daher kompliziert zu verstehen.

Purpose. Abschliessend beschreibt Purpose die Wahrnehmung des Wohlwollens des Trustees gegenüber dem Trustor und das Ausmass, in dem das System im Sinn des Entwicklungszwecks eingesetzt wird (Lee & See, 2004). Im Gegensatz zum Menschen kann zwar der Automation, ebenso wie dem KI-System, keine Intentionalität zugeschrieben werden, jedoch den Entwicklern und Entwicklerinnen.

Neigung zum Vertrauen in Automation. Vertrauen unterscheidet sich von Mensch zu Mensch, abhängig von den jeweiligen Entwicklungserfahrungen, Persönlichkeitsmerkmalen sowie kulturellen Hintergründen (Körber, 2019). So sind manche Menschen stärker geneigt zu vertrauen als andere (Lee & See, 2004). Hoff und Bashir (2015) verstehen unter dispositionellem Vertrauen die relativ stabile Neigung eines Individuums, sich auf die Automation zu verlassen. Demnach unterscheidet sich das Vertrauen unabhängig vom Kontext oder einem bestimmten System aufgrund der Nationalität, der Ethnie, der Religion und des Alters. Dabei können individuelle und kulturelle Unterschiede die Interaktion zwischen Mensch und Automation auf unerwartete Weise beeinflussen (Lee & See, 2004). Vertrauen in Automation ist ebenso wie das zwischenmenschliche Vertrauen kulturell geprägt (Hoff & Bashir, 2015; Lee & See, 2004; Razin & Feigh, 2023). Folglich stellt der kulturelle Hintergrund, insbesondere in Bezug auf Machtdistanz sowie individualistische und kollektivistische Prägungen, einen relevanten Faktor dar (Lee & See, 2004).

Zudem können Altersunterschiede sowohl durch Kohorteneffekte als auch anhand altersbedingte kognitive Veränderungen von oder der Kombination dieser beiden Faktoren erklärt werden (Hoff & Bashir, 2015). Menschen unterschiedlichen Alters können bei der Analyse der Vertrauenswürdigkeit abweichende Strategien anwenden.

In Bezug auf das Geschlecht und dessen Einfluss auf das Vertrauen in Automation gibt es keine klaren Erkenntnisse; jedoch zeigt sich, dass das Geschlecht bei der Interaktionen mit anderen Technologien eine Rolle spielen kann (Hoff & Bashir, 2015; Razin & Feigh, 2023).

Hoff und Bashir (2015) erachten Persönlichkeitsmerkmale des Trustors als weitere Komponente des dispositionellen Vertrauens. Im Kontext der Big-Five-Persönlichkeitsmerkmale bestehen Erkenntnisse, nach denen extrovertierte und emotional stabile Personen eher bereit sind, einem automatisierten System zu vertrauen. Letztlich gehen Hoff und Bashir (2015) davon aus, dass Personen mit einer höheren Neigung zu vertrauen eher dazu tendieren, sich auf zuverlässige Systeme zu verlassen, während ihr Vertrauen nach Systemfehlern möglicherweise stärker abnimmt.

Gemäss Hoff und Bashir (2015) spielen Vorerfahrungen, unabhängig von deren spezifischen Auswirkungen, nahezu in allen Fällen eine Rolle bei der HMI. Bestehende Erfahrungen mit einem automatisierten System oder einer ähnlichen Technologie können die Vertrauensbildung erheblich beeinflussen. Lee und See (2004) vertreten die Auffassung, dass Erwartungen an eine bestimmte Situation durch spezifische frühere Erfahrungen mit ähnlichen Situationen bestimmt werden. Hoff und Bashir (2015) unterschieden hier nach

Fachwissen, also dem Verständnis eines bestimmten Fachbereichs, jedoch nicht bezogen auf Automation und Erfahrung mit Automatisierung. So wird das Fachwissen dem situativen Vertrauen zugeschrieben und die Erfahrung mit Automatisierung dem erlernten Vertrauen, da diese beiden Aspekte jeweils unterschiedliche Auswirkungen auf Vertrauen haben.

Nutzungsintention bei Automation. Dass Vertrauen einen Einfluss auf die Nutzungsintention hat, ist naheliegend. Lee und See (2004) gehen, ähnlich wie Tschopp und Ruef (2020), im Kontext von KI davon aus, dass ein System, welches nicht vertrauenswürdig ist, mit hoher Wahrscheinlichkeit nicht genutzt wird. Damit sich Vertrauen bilden kann, muss eine Person sich vorab jedoch auf das System verlassen, da dessen Entwicklung auf den Erfahrungen mit dem System und den Beobachtungen von dessen Verhalten beruht (Lee & See, 2004). Diesbezüglich konstatieren Hoff und Bashir (2015), dass bestehende Einstellungen und Erwartungen auf die Vertrauensbildung sowie die nachfolgenden Nutzungsentscheidungen einwirken können. Razin und Feigh (2023) gehen noch einen Schritt weiter und postulieren, dass Vertrauen zwar nicht vollständig die Nutzungsintention aufklärt, jedoch stark mit dieser korreliert, während Letztere als Entscheidung aufgefasst werden kann, einem System zu vertrauen.

Dynamisches Vertrauen in Automation. Wie bereits angeführt, ist der Status des Vertrauens als dynamisches Konstrukt umstritten (u. a. Lee & See, 2004; Siegrist, 2021). Manche Menschen tendieren dazu, ihr Vertrauen in ein System erheblich zu verändern, wenn sich dessen Fähigkeiten ändern, bei anderen ist dies wiederum kaum der Fall (Lee & See, 2004). Unter anderem kann dieser Aspekt wiederum abhängig vom Systemtyp variieren und dadurch beeinflusst werden, wie die Handlungen des Systems beobachtbar sind. Folglich hängt die Vertrauensbildung von den verfügbaren Informationen ab, was dahingehend kritisch ist, dass die Nutzung relevant für den Informationsfluss ist und gleichzeitig Vertrauen voraussetzt. Hoff und Bashir (2015) gehen in diesem Zusammenhang davon aus, dass Menschen Maschinen zunächst als perfekt erachten und dementsprechend Purpose das anfängliche Vertrauen dominiert. Diese Perspektive wird auch von Lee und See (2004) sowie Razin und Feigh (2023) vertreten, obgleich sich nach Lee und See (2004) das Vertrauen in Automation nicht nach einem festen Muster entwickelt. Vielmehr hängt das Vertrauen in das System, in Ermangelung von Erfahrung vor und während der ersten Interaktionen, zunächst von Purpose ab. Mit fortschreitender Beziehung, Erfahrung und Beobachtung wächst die Relevanz von Performance sowie Process für die Entwicklung des Vertrauens (Hoff & Bashir, 2015; Lee & See, 2004).

Dies deutet darauf hin, dass die anfängliche Leistung eines Systems eine wesentliche Grundlage für Vertrauen ist und anfängliche Erfahrungen sowie frühere Fehler die Vertrauensbildung nachhaltig beeinflussen können (Hoff & Bashir, 2015; Lee & See, 2004). Dabei beeinflusst die Zuverlässigkeit zu Beginn dauerhaft das Vertrauen, selbst wenn sich diese im Verlauf der Zeit verändert (Lee & See, 2004), was für das Vertrauen als statisches Konstrukt spricht. Gleichzeitig erholt sich Letzteres von einer Abnahme bei akuten Fehlern, wohingegen es sich bei chronischen Fehlern erst mit deren Handhabbarkeit wieder regeneriert (Lee & See, 2004). Demnach führen Fehler dazu, dass das Vertrauen mit der Zeit schwindet und nur langsam wiederhergestellt wird. Gemäss Hoff und Bashir (2015) kann die Vertrauensbildung als dynamischer Prozess verstanden werden, wobei die Konfrontation mit neuen Informationen das Gefühl des Vertrauens drastisch verändern kann. Diese Dynamik ist jedoch auch davon abhängig, wie das Vertrauen gemessen wird bzw. ob es sich um unerfahrene oder erfahrene Nutzer und Nutzerinnen handelt (Razin & Feigh, 2023). Zusammenfassend kann sich das Vertrauen in Automation auf Basis einer direkten Beobachtung des Systemverhaltens (Performance), des Verständnisses über die Algorithmen (Process) oder der beabsichtigten Nutzung des Systems (Purpose) verändern (Lee & See, 2004).

Adäquates Vertrauen in Automation. Gemäss Lee und See (2004) ist der Mensch nicht grundsätzlich bereit, der Automatisierung genügend Vertrauen zu schenken. Gleichzeitig kann einem System auch dann vertraut werden, wenn dieses nicht angemessen ist. Wenn dabei das Vertrauen die Systemfähigkeiten übersteigt, wird von Übervertrauen gesprochen, was mit einer missbräuchlichen Nutzung (Misuse) einhergeht. Im umgekehrten Fall ist die Rede von Misstrauen, das wiederum eine Nichtnutzung (Disuse) zur Folge hat. Allgemein hängt adäquates Vertrauen davon ab, wie gut das Vertrauen kalibriert ist, also ob das Vertrauen des Trustors mit der Vertrauenswürdigkeit des Trustees übereinstimmt. Der Begriff der Kalibrierung bezieht sich dabei auf die Übereinstimmung zwischen dem Vertrauen des Menschen in das System und dessen Fähigkeiten (Lee & See, 2004).

Analog hierzu beschreiben Merritt et al. (2019) unter dem Terminus *Complacency* ein selbstgefälliges Verhalten, welches mit der suboptimalen Kontrolle der Automatisierungsleistung einhergeht und dem Vertrauen ähnlich ist. Selbstgefälligkeit tritt meist bei besonders zuverlässigen Systemen auf und ist unabhängig von der Expertise des bedienenden Menschen. Neben der Zuverlässigkeit des Systems haben auch das Vertrauen und die Arbeitsbelastung der bedienenden Person Einfluss auf das selbstgefällige Verhalten,

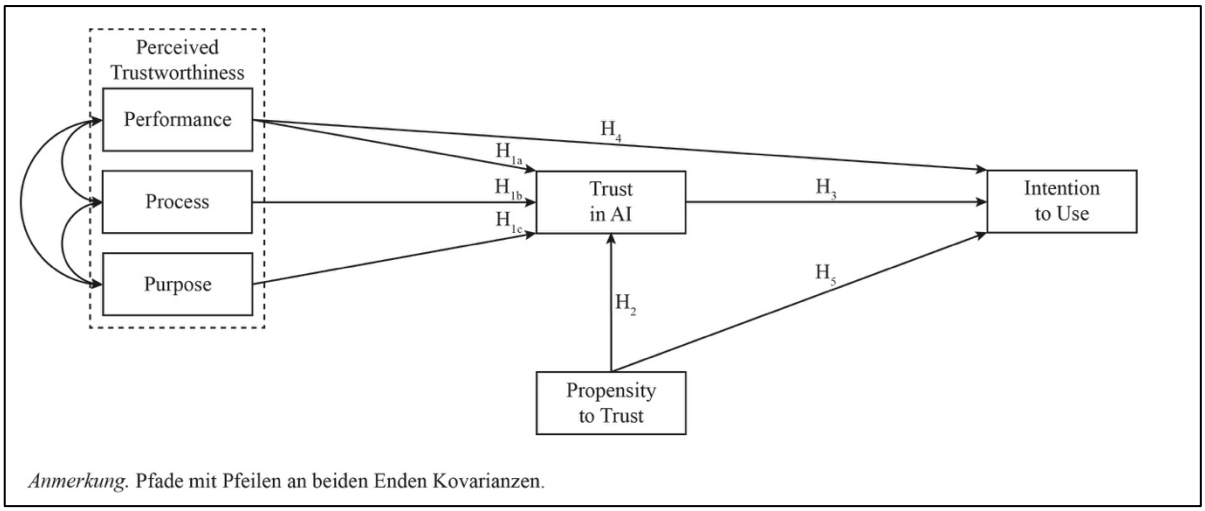
wobei dieses mit Überwachungsaufgaben interferieren kann. Daraus kann geschlossen werden, dass bei einer Kombination manueller und automatisierter Aufgaben die Selbstgefälligkeit eine Verlagerung der Aufmerksamkeit auf die manuelle Aufgabe beinhaltet (Merritt et al., 2019).

Aus diesem Grund ist die Förderung eines adäquaten Vertrauens entscheidend für die Vermeidung von Misuse und Disuse der Automatisierung (Hoff & Bashir, 2015; Lee & See, 2004). Gemäss Gebru et al. (2022) sollte das Vertrauen angemessen kalibriert sein, um eine sichere und effektive HMI zu gewährleisten. Dabei ist diese Kalibrierung eng mit der Robustheit, der Transparenz und der Fairness des automatisierten Systems verbunden. Zudem kann ein adäquates Vertrauen zu einer HMI-Leistung führen, welche jener des Menschen oder der Maschine allein überlegen ist (Lee & See, 2004). Ergänzend führen Hoff und Bashir (2015) an, dass Systeme, die eine genaue Rückmeldung über ihre Zuverlässigkeit oder ihre Funktionsweise geben, ein adäquates Vertrauen begünstigen.

Relevanz des Vertrauens in Automation. Vertrauen ist ein relevantes Konzept, um die HMI zu beschreiben, welche die Produktivität und die Akzeptanz automatisierter Systeme sowie anderer Computer-Technologien beeinflusst (Lee & See, 2004). Das Vertrauen in Automation unterscheidet sich zwar vom zwischenmenschlichen Vertrauen, dennoch existieren auch Parallelen (Hoff & Bashir, 2015). Die Vertrauenswürdigkeit automatisierter Systeme ist in die drei Dimension (a) Performance, (b) Process und (c) Purpose gegliedert (Lee & See, 2004). Diese Faktoren entsprechen den unter «Vertrauen in KI» beschriebenen und sind vergleichbar mit jenen von Mayer et al. (1995). Dabei wird Vertrauen eine entscheidende Rolle in Hinblick auf die Nutzungsintention zugeschrieben (Lee & See, 2004). Entsprechend nimmt die Relevanz des Vertrauens aufgrund der wachsenden Verbreitung automatisierter Systeme für den Privatgebrauch zu und dieses kann für den Erfolg der nächsten Generation von Computer-Technologien entscheidend sein (Lee & See, 2004). Dabei kann ChatGPT als ein solches System für den Privatgebrauch betrachtet werden. Infolgedessen ist es für die Entwicklung von KI-Systemen essenziell zu verstehen, wodurch das Vertrauen beeinflusst wird. Aktuell gibt es bereits zahlreiche konzeptuelle Modelle zu Vertrauen in KI, welche auf Vertrauen in Automation beruhen, jedoch selten empirisch untersucht werden, was das Fehlen einer Theorie zu Vertrauen in KI zur Folge hat (Solberg et al., 2022). Vor diesem Hintergrund ist das Ziel dieser Arbeit herauszufinden, welche Faktoren das Vertrauen in ChatGPT beeinflussen. Dazu werden im Folgenden konkrete Fragestellungen hergeleitet.

Fragestellung

Auf Grundlage des im vorhergehenden Kapitel «Vertrauen» dargelegten theoretischen Hintergrunds und der genannten Ziele wird in dieser Arbeit der Frage nachgegangen, welche Faktoren im Kontext von ChatGPT einen Einfluss auf das Vertrauen in KI haben. Hierzu wurde das in Abbildung 4 ersichtliche Modell zu Vertrauen in KI (Trust in AI) konzipiert. Es basiert auf Lees und Sees (2004) Theorie zu Vertrauen in Automation und ist beeinflusst durch die Modelle von Solberg et al. (2022, siehe Abbildung 1) sowie Körber (2019, siehe Abbildung 3) – welche wiederum von Vertrauen in Automation abgeleitet sind. Dabei besteht nicht der Anspruch, Vertrauen in KI zu generalisieren, da – für eine angemessene Untersuchung (Tschopp & Ruef, 2018) – mit ChatGPT nur ein konkretes KI-Blackbox-



Modell als Forschungsobjekt gewählt wurde.

Abbildung 4. Konzeptuelles Modell zu Vertrauen in KI (eigene Abbildung)

Im Kontrast zu Körber (2019) wurde im konzeptuellen Modell auf den moderierenden Faktor Familiarity verzichtet. Dieser kann mit dem Faktor Previous Experience nach Solberg et al. (2022) verglichen werden, welcher im Kontext von KI auch als Einflussfaktor für Propensity to Trust betrachtet wird. Dabei erachtet der Verfasser dieser Arbeit Previous Experience als systemunabhängige Variable. Zudem sieht Körber (2019) in diesem Faktor eine Moderatorvariable, welche nicht zwangsläufig gemessen werden muss.

Hingegen wurde das konzeptuelle Modell im Sinn von Bedué und Fritzsche (2021) um den Aspekt Intention to Use ergänzt. So ergibt sich aus der Entscheidung zu vertrauen unter

anderem die Nutzungsintention (Razin & Feigh, 2023). Aus dieser Ergänzung resultiert die explorative Unterfrage, welche Faktoren einen Einfluss auf die Nutzungsintention haben.

Aufgrund der uneindeutigen Erkenntnisse der bestehenden Forschung wurde eine weitere explorative Unterfrage verfasst. Konkret soll in dieser Arbeit untersucht werden, welche Faktoren einen Einfluss auf die Neigung zu Vertrauen haben. Wie bereits ausführlich erklärt, bezieht sich Letztere auf das kontext- und systemunabhängige Vertrauen (Solberg et al., 2022), wobei unklar erscheint, welche Faktoren prinzipiell – und wenn ja, in welcher Form – einen Einfluss haben.

Ergänzend wird in einer Längsschnittuntersuchung einer weiteren explorativen Unterfrage nachgegangen. Diese bezieht sich darauf, ob es nach einer Intervention Veränderungen der Zusammenhänge zwischen den Einflussfaktoren und dem Vertrauen in KI gibt. Denn die Vertrauensbildung wird als ein dynamischer Prozess erachtet, welcher durch die Konfrontation mit Informationen signifikant beeinflusst werden kann (Hoff & Bashir, 2015). Inwiefern Vertrauen aber tatsächlich als dynamisches Konstrukt zu verstehen ist, bleibt jedoch umstritten (Siegrist, 2021).

Diese vier Forschungsfragen, zusammen mit der im Kapitel «Vertrauen» aufgeführten Theorie sind Basis der Hypothesenbildung. Dabei ist zu beachten, dass lediglich Hypothesen gebildet wurden, welche zumindest theoretisch untermauert werden können. Auch sind nicht alle theoretischen Erkenntnisse in die Hypothesenbildung eingeflossen. So wurden nur Publikationen berücksichtigt, die zum Zeitpunkt der Hypothesenbildung bereits identifiziert und analysiert worden waren. Die Ableitung dieser Hypothesen folgt nun im Kapitel «Hypothesenbildung».

Hypothesenbildung

Wie bereits erwähnt, sind nicht alle theoretischen Erkenntnisse in die Hypothesenbildung eingeflossen. So wurden im Verlauf dieser Arbeit weitere wissenschaftliche Artikel einbezogen, welche zwar den theoretischen Hintergrund der Hypothesen, jedoch nicht die eigentliche Hypothesenbildung beeinflusst haben. Dieses Vorgehen wurde in Abstimmung mit der Betreuungsperson definiert, um die methodische Integrität zu wahren. So wurden zunächst die Hypothesen gebildet, dann die zur Hypothesenprüfung notwendigen Analyseverfahren bestimmt und abschliessend die Daten erhoben sowie entsprechend den vorab definierten Verfahren analysiert.

Für die Hypothesenbildung wurden zunächst die grundlegende Studie zu Vertrauen in Automation von Lee und See (2004), sowie Metaanalysen und Überblicksartikel dieses Konstrukts berücksichtigt (u. a. Hoff & Bashir; Körber, 2019; Razin & Feigh, 2023). Ergänzend wurden Publikationen, welches sich im Kontext von KI mit dem Vertrauen befassen analysiert (u. a. Choung et al., 2023; Frank et al., 2023; Schepman & Rodway, 2023; Siau & Wang, 2018; Tschopp & Ruef, 2020). Die Literatur wurde dahingehend bewertet, ob diese ein klares Verständnis über des Vertrauenskonstrukt aufweisen und sich dabei von anderen Konstrukten abgrenzen (u. a. Gebru et al., 2022; Kaplan et al., 2023; Langer et al., 2023; Lewis & Marsh, 2022; Sindermann et al., 2022; Solberg et al., 2022). Dementgegen wurden Studien, welche Vertrauen nicht von anderen Konstrukten, wie bspw. Kooperation abgrenzen, als vage in ihrer Aussagekraft bewertet (u. a. Bedué & Fritzsche, 2021; Siegrist, 2021).

Dabei wurden die Hypothesen – analog zu den Forschungsfragen – konzeptuell oder explorativ formuliert. Unter konzeptuellen Hypothesen versteht der Verfasser dieser Arbeit Hypothesen, welche durch empirische Erkenntnisse untermauert werden. Explorative Hypothesen hingegen zeichnen sich dadurch aus, dass zu ihnen keine oder nur einzelne empirische Studien gefunden wurden, ihre theoretische Untermauerung tendenziell vage ist und sich Befunde hierzu aus unterschiedlichen Studien teilweise widersprechen. In den folgenden werden die Hypothesen zu den jeweiligen Forschungsfragen abgeleitet.

Konzeptuelle Hypothesen zum Pfadmodell. Die wahrgenommene Vertrauenswürdigkeit hat, ebenso wie die Neigung zum Vertrauen, sowohl im Kontext von Automation (u. a. Hoff & Bashir, 2015; Körber, 2019; Lee & See, 2004) als auch hinsichtlich der KI (u. a. Bedué & Fritzsche, 2021; Gebru et al., 2022; Kaplan et al., 2023; Lewis & Marsh, 2022; Schepman & Rodway, 2023; Siau & Wang, 2018; Solberg et al., 2022; Tschopp & Ruef, 2020) einen Einfluss auf das Vertrauen. Dabei wird die wahrgenommene Vertrauenswürdigkeit durch die drei im System verorteten Faktoren (a) Performance, (b) Process und (c) Purpose bestimmt (Körber, 2019).

Die Performance bezieht sich auf die Zuverlässigkeit, die Kompetenz und die Fähigkeit des Systems (Körber, 2019). So hängt das Vertrauen von den Aufgaben sowie den Ergebnissen ab und wird direkt von der Performance beeinflusst (Hoff & Bashir, 2015). Anwendende neigen dazu, Systemen zu vertrauen, welche ihre Ziele zuverlässig erreichen (Lee & See, 2004). So sind die Zuverlässigkeit, die Verlässlichkeit und die Zweckmässigkeit der Funktionen wesentliche Vorläufer des Vertrauens (Hoff & Bashir, 2015). Dabei wird Performance auch als Prädiktor für Vertrauen in KI betrachtet (Bedué & Fritzsche, 2021; Kaplan et al., 2023; Lewis & Marsh, 2022; Siau & Wang, 2018; Solberg et al., 2022).

Der Faktor Process beschreibt das Verständnis der Anwendenden darüber, wie das System arbeitet und ob dieses für die Ziele der Anwendenden geeignet ist (Körber, 2019). Dabei fördern das Verständnis und die Berechenbarkeit des Systems das Vertrauen (Hoff & Bashir, 2015). Konkret neigen Menschen dazu, Systemen zu vertrauen, deren Algorithmen nachvollziehbar sind (Lee & See, 2004). Dieses Verständnis kann auch als wahrgenommene Transparenz verstanden werden (Solberg et al., 2022). So hat sich auch das Verständnis über das KI-System als Prädiktor für das KI-Vertrauen erwiesen (Bedué & Fritzsche, 2021; Kaplan et al., 2023; Lewis & Marsh, 2022; Siau & Wang, 2018; Solberg et al., 2022).

Purpose beschreibt die Wahrnehmung, dass die Entwickler und Entwicklerinnen des Systems den Anwendenden Gutes tun wollen (Körber, 2019). Diese positive Einstellung gegenüber den Anwendenden kann als Wohlwollen verstanden werden (Lee & See, 2004). Dabei ist auch das Wohlwollen der KI-Entwickler und -Entwicklerinnen als Prädiktor für das Vertrauen zu betrachten (Bedué & Fritzsche, 2021; Kaplan et al., 2023; Siau & Wang, 2018; Solberg et al., 2022). Daraus ergibt sich die konzeptuelle Hypothese 1: Es wird erwartet, dass die wahrgenommene Vertrauenswürdigkeit einen positiven Einfluss auf das Vertrauen in KI hat.

Dementgegen wird die Neigung zum Vertrauen als kontext- sowie systemunabhängiges Vertrauen beschrieben und ist beim Trustor verortet (Solberg et al., 2022). Sowohl Einstellungen als auch Erwartungen können die Vertrauensbildung beeinflussen (Hoff & Bashir, 2015), wobei die Neigung zum Vertrauen als Einflussfaktor des Vertrauens aufgefasst wird (Körber, 2019). Es ist feststellbar, dass die Neigung zum Vertrauen unabhängig von der wahrgenommen Vertrauenswürdigkeit des KI-Systems einen Einfluss auf das Vertrauen hat (Kaplan et al., 2023; Schepman & Rodway, 2023; Sindermann et al., 2022; Solberg et al., 2022). In Anlehnung daran wird die konzeptuelle Hypothese 2 gebildet, in deren Rahmen erwartet wird, dass die Neigung zum Vertrauen einen positiven Einfluss auf das Vertrauen in KI hat.

Explorative Hypothesen zum Pfadmodell. Die Entscheidung, einem System zu vertrauen, kann unter anderem als Nutzungsintention bezeichnet werden (Razin & Feigh, 2023). Vertrauen korreliert dabei stark mit Letzterer (Razin & Feigh, 2023). So beeinflusst das Vertrauen in KI auch die Nutzungsintention in Hinblick auf KI-Systeme (Bedué & Fritzsche, 2021; Choung et al., 2023; Frank et al., 2023). Von diesem Zusammenhang wird die explorative Hypothese 3 abgeleitet, die besagt, dass das Vertrauen in KI einen Einfluss auf die Nutzungsintention hat.

Ebenso wird teilweise auch Performance als direkter Prädiktor für die Nutzungsintention von KI-Systemen betrachtet (Choung et al., 2023). Dies resultiert in der Ableitung der explorativen Hypothese 4, bei der davon ausgegangen wird, dass Performance einen Einfluss auf die Nutzungsintention hat.

Für die Faktoren Process und Purpose der wahrgenommenen Vertrauenswürdigkeit wurden in der initialen Literaturrecherche keine Hinweise zu einem Einfluss auf die Nutzungsintention gefunden; dementsprechend wurden hierzu auch keine Hypothesen gebildet. Jedoch wird ergänzend in der Literatur auch der Neigung zum Vertrauen ein direkter Einfluss auf die Nutzungsentscheidungen zugeschrieben (Hoff & Bashir, 2015). Bestehende Einstellungen und Erwartungen können demnach auch ohne spezifische Erfahrungen mit dem System die initiale Nutzungsentscheidung beeinflussen. Dies offenbart sich auch punktuell im Kontext von KI-Systemen (Choung et al., 2023). Hiervon wurde die explorative Hypothese 5 abgeleitet, die besagt, dass die Neigung zum Vertrauen einen Einfluss auf die Nutzungsintention hat.

Explorative Hypothesen bezüglich der Neigung zum Vertrauen. Erwartungen an eine bestimmte Situation werden durch ähnlich wahrgenommene vergangene Situationen bestimmt (Lee & See, 2004). Demnach können Vorerfahrungen mit einer ähnlichen Technologie den Prozess der Vertrauensbildung erheblich verändern (Hoff & Bashir, 2015). In diesem Zusammenhang hat sich gezeigt, dass eine grössere Erfahrung auch mit grösserem Vertrauen in KI einhergeht (Kaplan et al., 2023; Solberg et al., 2022). Jedoch ist die Rolle der Vorerfahrung für die Neigung zum Vertrauen uneindeutig (Solberg et al., 2022). Daher wird mit Hypothese 6 die explorative Hypothese gebildet, dass die Vorerfahrung einen Einfluss auf die Neigung zum Vertrauen hat.

Aus mehrerer Studien geht zudem hervor, dass das Alter in Hinblick auf Vertrauen ein wesentlicher Faktor ist (Hoff & Bashir, 2015). Allerdings werden neben den Alters- auch Kohorteneffekte diskutiert. Im Kontext von KI konnte festgestellt werden, dass jüngere Personen eher dazu neigen zu vertrauen als ältere (Montag, Kraus, et al., 2023). Jedoch gibt es gegensätzliche Erkenntnisse, was das Vertrauen angeht, so zum Beispiel, dass mit dem Alter die Risikobereitschaft sinkt, was wiederum das Vertrauen verringert, während gleichzeitig Alter für Vertrautheit sorgt (Razin & Feigh, 2023). Von diesem Umstand wurde die explorative Hypothese 7 abgeleitet, nach der das Alter einen Einfluss auf die Neigung zum Vertrauen hat.

Zudem kann davon ausgegangen werden, dass Letztere in Abhängigkeit von der Persönlichkeit variiert (Körber, 2019), wobei sie auch als Persönlichkeitsmerkmal betrachtet wird (Lee & See, 2004). Konkret konnten Abhängigkeiten zwischen mehreren spezifischeren Big-Fife-Persönlichkeitsmerkmalen und Vertrauen beobachtet werden (Hoff & Bashir, 2015). Des Weiteren haben sich Persönlichkeitsmerkmale ebenfalls als ein signifikanter Prädiktor im KI-Kontext erwiesen (u. a. Kaplan et al., 2023; Sindermann et al., 2022). Dabei tendieren Menschen mit einer höheren Offenheit eher dazu, einem KI-System zu vertrauen (Kaplan et al., 2023). Auf einen möglichen Zusammenhang mit Gewissenhaftigkeit wird hingegen nur selten hingewiesen (u. a. Schepman & Rodway, 2023). Stattdessen gibt es mehrere Hinweise darauf, dass extrovertierte Menschen eher dazu neigen, KI-Systemen zu vertrauen (Kaplan et al., 2023; Schepman & Rodway, 2023). Darüber hinaus ist festzustellen, dass Verträglichkeit einen Einfluss auf Vertrauen haben kann (Schepman & Rodway, 2023; Sindermann et al., 2022), und auch Neurotizismus weist einen negativen Zusammenhang hiermit auf (Kaplan et al., 2023). Jedoch widersprechen sich die theoretischen und empirischen Erkenntnisse, insbesondere im Kontext unterschiedlicher

Kulturen, teilweise (Sindermann et al., 2022). Infolgedessen wird mit Hypothese 8 eine weitere explorative Hypothese abgeleitet, die besagt, dass Persönlichkeitsmerkmale einen Einfluss auf die Neigung zum Vertrauen haben.

Die Geschlechtsunterschiede bezüglich des Vertrauens in Automation sind ebenfalls inkonsistent, es kann jedoch festgehalten werden, dass das Geschlecht eine Rolle spielen kann (Hoff & Bashir, 2015; Razin & Feigh, 2023). Im Kontext von KI erweist es sich teilweise als einflussreicher Faktor – Männer vertrauen der KI tendenziell mehr als Frauen (Kaplan et al., 2023). Gleichzeitig weisen Ergebnisse anderer Studien darauf hin, dass es keine Geschlechtsunterschiede gibt (Montag, Kraus, et al., 2023). Dies resultiert in der explorativen Unterschiedshypothese 9, der zufolge sich die die Neigung zum Vertrauen abhängig vom Geschlecht unterscheidet.

Explorative Hypothese zum dynamischen Vertrauen. Häufig wird Vertrauensbildung als dynamischer Prozess verstanden (u. a. Hoff & Bashir, 2015; Lee & See, 2004; Razin & Feigh, 2023). Dabei verändert Erfahrung das Vertrauen; genauer gesagt spielen Purpose und die Neigung zum Vertrauen zu einem früheren Zeitpunkt eine grössere Rolle, während im Verlauf der Zeit die Relevanz von Performance sowie Process zunimmt (Lee & See, 2004; Razin & Feigh, 2023). So sind Erwartungen an und das Verständnis über die Fähigkeiten des KI-Systems zu Beginn gering (Langer et al., 2023). Jedoch ist umstritten, ob Vertrauen eher als dynamisches oder als statisches Konstrukt verstanden werden sollte (Siegrist, 2021). Demzufolge wird abschliessend die explorative Hypothese 10 abgeleitet, dass sich der Zusammenhang zwischen den Einflussfaktoren und dem Vertrauen in KI an zwei Messzeitpunkten unterscheidet.

Das folgende Kapitel «Methoden» ist (a) der Operationalisierung, (b) der Stichprobe, (c) der Datenerhebung und (d) der Datenanalyse dieser Arbeit gewidmet. Dabei werden von den in diesem Kapitel gebildeten konzeptuellen und explorativen Hypothesen statistische Hypothesen abgleitet. Für die Prüfung dieser statistischen Hypothesen werden die notwendigen Analyseverfahren diskutiert und bestimmt.

Methoden

Diese Arbeit basiert auf einer Querschnitt- und einer Längsschnittuntersuchung. Dabei wurde Letztere anhand eines Subsamples der Ersteren durchgeführt. Entsprechend weicht die Methodik der Längsschnittuntersuchung nur bezüglich (a) der Datenerhebung des zweiten Messzeitpunktes, (b) der Intervention und (c) der Datenanalyse von der Querschnittuntersuchung ab. Daher werden in diesem Kapitel ausschliesslich die Unterschiede in separaten Unterkapiteln dargestellt.

Die Datenerhebung erfolgte online, auch wenn die Versuchspersonen im Rahmen von Vorlesungen der Hochschule für Wirtschaft FHNW rekrutiert wurden und so eine Teilnahme an der Umfrage vor Ort möglich war. Die Querschnittuntersuchung umfasst zum einen die empirische Prüfung des konzeptuellen Modells (siehe Abbildung 4) anhand einer Pfadanalyse und zum anderen Analyse von Einflussfaktoren des dispositionellen Vertrauens. Bei der Längsschnittuntersuchung handelt es sich um eine Ein-Gruppen-Pretest-Posttest-Untersuchung zur Überprüfung der zeitlichen Veränderung des Zusammenhangs zwischen dem Vertrauen und dessen Einflussfaktoren in Kontext einer instruktionalen Intervention.

Somit ergibt sich eine rein quantitative Untersuchung, auch wenn initial zunächst Überlegungen angestellt wurden, einen qualitativen oder einen Mixed-Methods-Ansatz zu wählen, da Aspekte der Fragestellung explorativer Natur sind. Im Rahmen des vorab durchgeführten und im Kapitel «Einleitung» abgehandelten Literatur-Reviews ergab sich jedoch die Erkenntnis, dass der quantitative Ansatz die geeignete Methode zur Beantwortung der Fragestellung ist. Konkret konnten anhand der bestehenden Konzepte des Vertrauens konzeptuelle und in diesem Kapitel schliesslich empirische Hypothesen abgeleitet werden. Zudem existieren validierte Erhebungsinstrumente, welche eine zuverlässige Messung erlauben und im folgenden Kapitel «Erhebungsinstrumente» genauer beschrieben werden.

Erhebungsinstrumente

Mit Ausnahme des Geschlechts erfolgte die Datenerhebung anhand metrischer Skalen. Dabei wurde das Alter als Single-Item und alle weiteren Variablen anhand reliabler Fragebögen erfasst. Aufgrund der Reliabilität dieser Erhebungsinstrumente wurde auf eine Faktorenanalyse und Messung der internen Konsistenz mittels Cronbachs-Alpha verzichtet. Zum einen wären deren Ergebnisse bei der kleinen vorliegenden Stichprobe nicht aussagekräftig gewesen (Bujang, Omar & Baharum, 2018) und zum anderen hätten diese den Rahmen dieser Masterarbeit überstiegen. Die konkreten Erhebungsinstrumente für die Messung der jeweiligen Konstrukte werden im Folgenden beschrieben.

Messung des Vertrauens. Im Sinn von Hoffman et al. (2023) wurde für die Messung des Vertrauens in KI der deutschsprachige *Trust in Automation Questionnaire (TiA)* von Körber (2019) adaptiert. Der Fragebogen wurde nach der klassischen Testtheorie entwickelt und umfasst sechs konzeptuelle Subskalen: (a) die sechs Items der Subskala *Reliability/Competence* (ω = .92) operationalisieren die Performance, (b) die vier Items der Subskala *Understanding/Predictability* (ω = .81) den Faktor Process und (c) die zwei Items der Subskala *Intention of Developers* (ω = .79) den Aspekt Purpose. Darüber hinaus dienten (d) die zwei Items der Subskala *Familiarity* (ω = .83) zur Operationalisierung der Vorerfahrung, (e) die drei Items der Subskala *Propensity to Trust* (ω = .78) zur Messung der Propensity to Trust und (f) die zwei Items der Subskala *Trust in Automation* (α = .85) sollten den Faktor Trust in AI repräsentieren. Die Subskalen mit den 19 Items und deren Adaption sind in Tabelle A1 in Anhang A ersichtlich.

Messung der Nutzungsintention. Zur Messung der Nutzungsintention wurde die Subskala *Behavioral intention of non-users* (α = .98) des von Choung et al. (2023) im Kontext von Vertrauen in KI adaptierten TAM-2-Fragebogens verwendet. Die vier Items wurden vorab sinngemäss aus dem Englischen übersetzt und entsprechend dem Zweck interpretiert. Die originalen und die übersetzten Items sind in Tabelle A2 in Anhang A ersichtlich.

Messung der Persönlichkeitsmerkmale. Die Big-Five-Persönlichkeitsmerkmale wurden anhand des von Körner et al. (2008) entwickelten NEO-Fünf-Faktoren-Inventars-30 (NEO-FFI-30) – einer ökonomischen Kurzversion des von Borkenau und Ostendorf (1993, zitiert nach Körner et al., 2008) übersetzten NEO-FFI – erfasst. Die fünf Subskalen beziehen sich auf die robusten Persönlichkeitsfaktoren nach Costa und McCrae (1989, zitiert nach Körner et al., 2008), darunter (a) Offenheit (α = .67), (b) Gewissenhaftigkeit (α = .78), (c) Extraversion (α = .72), (d) Verträglichkeit (α = .75) sowie (e) Neurotizismus (α = .81), und beinhalten jeweils sechs Items. Aus Gründen der Ökonomie wurde nur die Hälfte der Items anhand der ausgewiesenen Faktorladung von Körner et al. (2008) selektiert. Infolgedessen wurden lediglich 15 anstelle der eigentlichen 30 Items des NEO-FFI-30 verwendet. Andernfalls könnte der grössere Zeitaufwand für die Umfrage mit einer hohen Abbruchrate und einem Qualitätsverlust einhergehen (Galesic & Bosnjak, 2009). Die für die Umfrage selektierten Items wurden im Originalwortlaut verwendet und sind in Tabelle A3 in Anhang A ersichtlich.

Stichprobe

Das Sample wurde im Rahmen von Vorlesungen des Bachelorstudiengangs Wirtschaftsinformatik des Praxispartners an der Hochschule für Wirtschaft FHNW gezogen. Somit handelt es sich um eine Ad-hoc-Stichprobe, welche die Generalisierbarkeit limitiert. Den Versuchspersonen wurde im Rahmen der Vorlesungen genügend Zeit zur Beantwortung der Umfrage gewährt; die Teilnahme war dennoch freiwillig. Das Subsample der Längsschnittuntersuchung besteht ausschliesslich aus Studierenden des Moduls *Cyber Security Management*. Im Rahmen der Vorlesungen wurde auch die Intervention in Form eines 15-minütigen Inputs durchgeführt. Es folgt nun die Beschreibung der Stichproben für die Querschnittuntersuchung und die Längsschnittuntersuchung getrennt vorgenommen.

Stichprobe der Querschnittuntersuchung. Die Querschnittuntersuchung umfasst eine Ad-hoc-Stichprobe von 114 Versuchspersonen. Eine Versuchsperson wurde aus der Stichprobe entfernt, da von dieser ausnahmslos alle Items mit dem mittleren Skalenwert von 3 beantwortet wurden. Dies ist auch mit der Tendenz zur Mitte nicht zu rechtfertigen und führte zu Zweifeln bezüglich der Ernsthaftigkeit der Auskünfte. Zudem wurden vier Versuchspersonen ausgeschlossen, da hier jeweils mehrere Items zu den Variablen des Pfadmodells unbeantwortet blieben. Bei Personen mit nur einem fehlenden Wert zu den Subskalen wurde dieser anhand des Medians ersetzt. Zwei Versuchspersonen gaben an, keine Erfahrung mit ChatGPT zu haben, und weitere zwei verzichteten bei diesem Item auf eine Antwort. Diese vier Versuchspersonen wurden aus der Analyse ausgeschlossen. So wurde fehlende Erfahrung mit ChatGPT als hartes Ausschlusskriterium definiert, da andernfalls die Qualität der Daten in Bezug auf das gemessene Vertrauen in die Technologie in Frage zu stellen wäre (Siegrist, 2021). Damit wurden 105 Versuchspersonen für die Datenanalyse der Querschnittuntersuchung berücksichtigt.

Teststärkeanalyse der Querschnittuntersuchung. Da sich die Teststärke für eine Pfadanalyse nicht ohne weiteres berechnen lässt, wurde in Abstimmung mit der Betreuungsperson dieser Arbeit auf eine A-priori-Teststärkeanalyse verzichtet. Für die ungerichtete multiple Regression mit fünf Prädiktoren für die Analyse des Einflusses der Persönlichkeitsmerkmale auf Propensity to Trust wurde a priori eine Teststärkeanalyse mittels G*Power (Faul, Erdfelder, Lang & Buchner, 2007) durchgeführt. Diese ergab für einen nach Cohen (1992) signifikanten mittleren Effekt ($f^2 > .15$, $p = .05$, $1\text{-}\beta = .8$) eine Sample-Grösse von 55 Versuchspersonen. Damit bringt die Sample-Grösse von 105 Versuchspersonen für die multiple Regressionsanalyse eine ausreichende Teststärke mit sich.

Stichprobenbeschreibung der Querschnittuntersuchung. Die berücksichtigten Versuchspersonen der Querschnittuntersuchung waren mit zwei Ausnahmen Studierende der Hochschule für Wirtschaft FHNW. Eine Person studierte an einer anderen Hochschule, eine weitere Person gab an, nicht zu studieren. Unter den Versuchspersonen befanden sich 54 Männer und 51 Frauen – keine der Personen wählte die Antwortoption «divers». Ausserdem verzichteten 12 Versuchspersonen auf eine Altersangabe; die übrigen 93 Personen waren zum Zeitpunkt der Erhebung im Durchschnitt 22.09 Jahre alt (*SD* = 3.98). Die Stichprobe der Querschnittuntersuchung beinhaltet zudem ein Subsample von 27 Personen, welches in der Längsschnittuntersuchung berücksichtigt wurde und im Folgenden beschrieben wird.

Stichprobe der Längsschnittuntersuchung. Für das Subsample der Längsschnittuntersuchung waren 30 Versuchsperson der Querschnittuntersuchung qualifiziert. Dies waren die Studierenden, welche im Modul Cyber Security Management der Hochschule für Wirtschaft FHNW an der Umfrage teilgenommen hatten. Im Rahmen dieses Moduls wurden auch die Intervention und die zweite Datenerhebung durchgeführt. Abzüglich der im Unterkapitel «Stichprobe der Querschnittuntersuchung» beschriebenen Ausschlüsse bestand diese Stichprobe anfänglich aus 27 Versuchspersonen. Aufgrund der Drop-outs in der zweiten Datenerhebung und der Personen, welche bei der Intervention nicht anwesend waren, ergab sich eine Stichprobengrösse von 10 Personen.

Teststärkeanalyse der Längsschnittuntersuchung. Das Ergebnis der A-priori-Teststärkeanalyse für eine MANOVA mit vier Prädiktoren mittels G*Power (Faul et al., 2007) wies, für einen nach Cohen (1992) signifikanten mittleren Effekt ($f > .25$, $p = .05$, $1\text{-}\beta = .8$) eine Sample-Grösse von 128 Versuchspersonen aus. Allerdings war bekannt, dass diese Stichprobengrösse nicht erreichbar sein würde, da das Modul von weniger als 50 Personen besucht wurde. Mit der vorliegenden Sample-Grösse von lediglich 10 Versuchspersonen wurde somit keine ausreichende Teststärke erreicht. Entsprechend konnten hier keine signifikanten Ergebnisse erwartet werden. In Ermangelung einer ordinalskalierten Alternative, welche ein konservativeres Vorgehen ermöglicht hätte, wurde in Absprache mit der Betreuungsperson dennoch auf dieses Verfahren zurückgegriffen.

Stichprobenbeschreibung der Längsschnittuntersuchung. Die Versuchspersonen der Längsschnittuntersuchung waren ausschliesslich Bachelor-Studierende des Moduls Cyber Security Management. Vier der Versuchspersonen dieses Subsamples gaben sich als männlich aus und sechs als weiblich. Die Hälfte der Personen teilte kein Alter mit, die übrigen fünf Personen waren durchschnittlich 27.4 Jahre alt (*SD* = 3.44).

Datenerhebung

Die anonyme Umfrage zur Datenerhebung wurde mittels des FHNW-Umfragetools TIVIAN online erstellt. Zu Beginn der Umfrage wurde kurz auf den Zweck der Datenerhebung und Untersuchung hingewiesen. Ergänzend folgte eine Aufklärung über die Freiwilligkeit und die Datenbearbeitung, welche aktiv bestätigt werden mussten. Neben dem gänzlichen Verzicht auf die Teilnahme an der Umfrage war es den Versuchspersonen auch freigestellt, einzelne Items auszulassen und dennoch die Umfrage weiter fortzuführen.

Struktur der Datenerhebung. Im ersten Block der Umfrage wurden die demographischen Daten zu Studierenden-Status, Geschlecht und Alter sowie das Ausschlusskriterium der vorrangigen Benutzung von ChatGPT erhoben.

Der zweite Block der Umfrage bezog sich auf die Variablen zum Pfadmodell. Es wurde auf eine Randomisierung der Items verzichtet und stattdessen auf die originale Reihenfolge von Körber (2019) zurückgegriffen. Die insgesamt 22 Items zum Pfadmodell wurden auf drei Frageseiten mit zweimal sieben und einmal acht Items aufgeteilt, um eine Überladung zugunsten der Lesbarkeit zu vermeiden. Die Seiten wurden so gestaltet, dass bei Variablen mit nur zwei oder drei Items auf jeder Seite lediglich ein Item der jeweiligen Variable platziert wurde, ohne die Reihenfolge von Körber (2019) zu alternieren. Dementsprechend wurden die drei Items der Nutzungsintention an die Reihenfolge von Körber (2019) angepasst, sodass auch diese gleichmässig auf die drei Frageseiten verteilt waren.

Der dritte und letzte Block der Umfrage beinhaltete die Persönlichkeitsmerkmale. Auch hier wurden drei Frageseiten mit in diesem Fall fünf Items pro Seite erstellt. Die von Körner et al. (2008) vorgeschlagene gleichmässige Abwechslung der fünf Faktoren wurde beibehalten, sodass jede Seite ein Item zu jeder Variable beinhaltete. Wie bereits im Unterkapitel Erhebungsinstrumente beschrieben, wurden lediglich 15 der eigentlichen 30 Items des NEO-FFI-30 erhoben. Auf diese Weise wurde das Risiko eines möglichen grösseren Datenverlusts aufgrund eines vorzeitigen Abbrechens der Umfrage kontrolliert, da die Relevanz dieser Variablen gegenüber jenen für das Pfadmodell untergeordnet ist.

Skalenniveau der Datenerhebung. Mit Ausnahme des Geschlechts, welches nominal anhand der drei Antwortoptionen (a) «weiblich», (b) «männlich» und (c) «divers» erhoben wurde, sowie dem Alter, welches in einem Textfeld als natürliche Zahl eingegeben wurde, wurden alle Items anhand einer fünfstufigen Likert-Skala von *1 – stimme gar nicht zu* bis *5 – stimme voll zu* gemessen. Abgesehen vom Geschlecht sind somit alle Daten für die Analysen intervallskaliert.

Datenerhebung der Querschnittuntersuchung. Die Datenerhebung der Querschnittuntersuchung fand zwischen dem 28. Februar und dem 19. März 2024 statt. Damit endete die Erhebungsphase vor der Durchführung der Intervention der Längsschnittuntersuchung, um eine mögliche Beeinflussung der Daten durch diese zu vermeiden. Die Umfragen wurden im Kontext unterschiedlicher Vorlesungen durch Dozierende des Praxispartners an der Hochschule für Wirtschaft FHNW bekanntgemacht. Damit ist eine Mehrfachteilnahmen generell nicht ausgeschlossen, jedoch verhindert TIVIAN eine solche im selben Browser. Zudem ist nicht davon auszugehen, dass Teilnehmende dieselbe Umfrage mehrfach ausfüllen.

Datenerhebung der Längsschnittuntersuchung. Die Daten des Pretests der Längsschnittuntersuchung wurden am 28. Februar 2024 zu Beginn der Vorlesung erhoben. Die Umfrage des Pretests ist mit jener der Querschnittuntersuchung weitgehend identisch. Lediglich das Item zur Abfrage des Studierendenstatus wurde entfernt und stattdessen ein Item zur Generierung einer eindeutigen Identifizierung der beiden Datensätze ergänzt.

Die Datenerhebung des Posttests erfolgte am 3. April 2024 gegen Ende der ersten Vorlesungsstunde, um auch möglicherweise verspäteten Personen die Teilnahme zu ermöglichen. Für den Posttest wurde die Umfrage des Pretests durch zwei Items ergänzt, in denen die Teilnahme an der Pretest-Erhebung und an der Intervention abgefragt wurde. Zudem wurden die Items zu den Persönlichkeitsmerkmalen entfernt, da diese für die Ein-Gruppen-Pretest-Posttest-Untersuchung nicht relevant waren und die entsprechenden Daten durch den Pretest bereits vorlagen. So sind Persönlichkeitsmerkmale als stabil bzw. die Zeit überdauernd zu betrachten (u. a. John & Srivastava, 1999).

Der Zeitraum zwischen den beiden Erhebungen betrug sechs Wochen, was der Empfehlung von Körber (2019) für eine längsschnittliche Untersuchung des Vertrauens entspricht. Zwischen diesen beiden Erhebungszeitpunkten wurde mit den Versuchspersonen eine Intervention durchgeführt, welche im Folgenden genauer beschrieben wird.

Intervention der Längsschnittuntersuchung

Mit Hilfe der Intervention sollten die Faktoren Performance, Process und Purpose, welche Trust in AI beeinflussen, manipuliert werden. Sie wurde am 20. März 2024 in Form eines 15-minütigen Inputs im Rahmen Vorlesungen des Moduls Cyber Security Management durchgeführt.

Die Intervention begann mit der Klärung allgemeiner Begrifflichkeiten rund um KI. Anschliessend wurde auf den Faktor Purpose abgezielt. Dies umfasste die Übermittlung von Informationen zum ChatGPT-Entwickler OpenAI, unter anderem zu dessen Mission und dessen Vision, welche frei auf der Website verfügbar sind. Ebenso wurde der Zweck von ChatGPT thematisiert, wobei auch kritische Argumente hinsichtlich der Schädlichkeit dieses KI-Systems anhand wissenschaftlicher Publikationen (Brown et al., 2020; Scheurer et al., 2023) aufgezeigt wurden.

Es folgten Folien zum Faktor Process, mit denen anhand des Bestsellers «Die KI war's» von Prof. Dr. Katharina Zweig (2023) der probabilistische Algorithmus von ChatGPT erklärt wurde. Dabei wurde betont, dass die Texte anhand statistischer Wahrscheinlichkeiten gebildet werden und entsprechend kein Anspruch auf Richtigkeit besteht.

Abschliessend wurden im Kontext des Faktors Performance anhand wissenschaftlicher Publikationen (Brown et al., 2020; Taecharungroj, 2023) die Fähigkeiten und die Unfähigkeit von ChatGPT aufgezeigt. Dabei wurde erklärt, dass es sich bei den Resultaten von ChatGPT um «intelligentes Raten» (Zweig, 2023) handelt und die Genauigkeit meist unter 60 % liegt (Brown et al., 2020).

Insgesamt wurden mit der Intervention alle drei Aspekte der wahrgenommen Vertrauenswürdigkeit manipuliert. Die Inhalte der Intervention sind in der PowerPoint-Präsentation in Anhang B ersichtlich.

Datenanalyse

Die Datenanalyse erfolgte mittels R (R Core Team, 2022). Vor der eigentlichen Analyse wurden die Skalen negativer Items umgepolt. Anschliessend wurden aus den einzelnen Items der jeweiligen Skalen die Mittelwerte als neue Variablen berechnet. Dies betraf alle intervallskalierten Items mit Ausnahme des Alters.

Für die Analyse der Daten wurde, in Übereinstimmung mit den Konventionen nach Cohen (1992), das Signifikanzniveau auf $p = .05$ festgelegt. Folglich werden Effekte mit Signifikanzwerten darunter als vorhanden und mit Signifikanzwerten darüber als nicht vorhanden diskutiert. Es wird also nicht zwischen marginal-, hoch- und höchstsignifikant unterschieden. Bedeutender erscheint die Rolle der Teststärke, denn eine zu hohe Teststärke kann zu einer Signifikanz von Effekten führen, welche jedoch mit einem Alpha-Fehler einhergehen könnte (Cohen, 1992). Des Weiteren wurde die Effektstärke nach Cohen (1992) bewertet.

Im Folgenden werden die empirischen Hypothesen gemeinsam mit den statistischen Methoden zur Analyse dieser Hypothesen aufgeführt. Dabei wird die Wahl des jeweiligen Analyseverfahrens begründet und die Prozesse der Voraussetzungsprüfungen werden beschrieben.

Querschnittanalysen. Die Querschnittuntersuchung umfasst neben der Pfadanalyse zur Analyse des Pfadmodells zwei lineare sowie eine multiple Regressionsanalysen und eine einfaktorielle ANOVA ohne Messwiederholung zur Analyse der Einflussfaktoren der Neigung zum Vertrauen, welche im Folgenden beschrieben werden.

Pfadanalyse des konzeptuellen Pfadmodells. Im Rahmen des Pfadmodells der Querschnittuntersuchung wurden folgende empirische Hypothesen geprüft:

H_1 Die Faktoren der wahrgenommen Vertrauenswürdigkeit, (a) Performance, (b) Process und (c) Purpose, stehen in Zusammenhang mit Trust in AI.

H_2 Propensity to Trust steht in Zusammenhang mit Trust in AI.

H_3 Trust in AI steht in Zusammenhang mit Intention to Use.

H_4 Performance steht in Zusammenhang mit Intention to Use.

H_5 Propensity to Trust steht in Zusammenhang mit Intention to Use.

Begründung der Pfadanalyse. Die Überprüfung dieser Hypothesen erfolgte anhand der Pfadanalyse mithilfe des r-Pakets lavaan (Rosseel, 2012). Obwohl bei der Pfadanalyse im Gegensatz zum Strukturgleichungsmodell (SEM) die interne Konsistenz nicht berücksichtigt wird, wurde dennoch auf diese zurückgegriffen, da bei der vorliegenden Stichprobengrösse für ein SEM keine adäquate Teststärke zu erwarten wäre (Kim, 2005). Zudem beziehen sich die im Kapitel «Vertrauen» aufgeführten empirischen Studien teilweise auf Pfadanalysen.

Voraussetzungsprüfung der Pfadanalyse. Zur Prüfung der Voraussetzungen wurden zunächst die endogenen Variablen mittels QQ-Plott sowie Mardias-Test und Shapiro-Wilk-Test auf ihre multivariate respektive univariate Normalverteilung hin geprüft. Anschliessend erfolgte die Prüfung der Linearität anhand des Streudiagramms.

Modellpassung des Pfadmodells. Die Passung der Modellierung des Pfadmodells anhand des konzeptuellen Modells zu Vertrauen in KI (siehe Abbildung 4) wurde zunächst anhand einer von Hu und Bentler (1999) empfohlenen Kombination der Fit-Indizes (a) Comparative Fit Index (CFI), (b) Standardized Root Mean Squared Residual (SRMR) und (c) Root Mean Squared Error of Approximation (RMSEA) geprüft. Dabei wurden die Cut-off-Werte (CFI > .95, SRMR < .08, RMSEA < .06) von Hu und Bentler (1999) verwendet. Die Fit-Indizes zeigten eine mangelnde Passung der Daten in Hinblick auf das konzeptuelle Pfadmodell, was Modellrespezifikationen notwendig machte.

Für diese Verbesserung wurden zunächst die statistischen Modifikationsindizes aufgerufen, woraufhin eine Pfadergänzung in das Modell integriert wurde. Hierbei wurde jeweils der Pfad gewählt, welcher zum einen konzeptuell sinnvoll erschien und zum anderen die grösste geschätzte signifikante Verbesserung der Modellpassung ($\Delta\ \chi 2$, $\Delta\ p < .05$) mit sich brachte. Daraus ergaben sich drei Modellrespezifikationen bis zum gewünschten Modell-Fit. Diese schrittweise Respezifikation des Pfadmodells ist im Kapitel «Ergebnisse» ausformuliert.

Analyse der Einflussfaktoren der Neigung zum Vertrauen. Die Querschnittsanalyse der Neigung zum Vertrauen umfasste die Prüfung folgender empirischer Hypothesen:

- H_6 Vorerfahrung steht in Zusammenhang mit Propensity to Trust.
- H_7 Alter steht in Zusammenhang mit Propensity to Trust.
- H_8 Die Persönlichkeitsmerkmale (a) Offenheit, (b) Gewissenhaftigkeit, (c) Extraversion, (d) Verträglichkeit und (e) Neurotizismus stehen in Zusammenhang mit Propensity to Trust.
- H_9 Propensity to Trust unterscheidet sich aufgrund des Geschlechts.

Begründung der linearen Regressionsanalysen. Die Untersuchung der Einflussfaktoren der Propensity to Trust erfolgte bei den intervallskalierten Variablen mittels linearer Regressionsanalyse. Aufgrund der Tatsache, dass im Kontext der Hypothesen und der zugrundliegenden Theorie von einer kausalen Beziehung auszugehen ist, wurde die Regressionsanalyse auch bei den einfachen Zusammenhängen der Korrelationsanalyse vorgezogen. Dabei wurde der Einfluss von Vorerfahrung und Alter auf Propensity to Trust jeweils mittels einer einfachen linearen Regression bzw. der Einfluss der Persönlichkeit mittels multipler Regressionsanalyse überprüft.

Voraussetzungsprüfung der linearen Regressionsanalysen. Vor der eigentlichen Analyse wurden zunächst die Linearitätsvoraussetzung mittels Rainbow-Test (Utts, 1982) geprüft. Die Normalverteilung der Residuen wurde mittels Histogramm und QQ-Plot optisch begutachtet. Anhand des Breusch-Pagan- und des Durbin-Watson-Tests wurde die Homoskedastizität respektive die Unabhängigkeit der Residuen geprüft. Bei der multiplen Regressionsanalyse wurde zudem die Multikollinearität mittels Varianzinflationsfaktor-Werten (VIF) untersucht. Ergänzend wurden Ausreisser mittels Cooks-Distance des r-Pakets olsrr (Hebbali, 2024) diagnostiziert und anschliessend ausgeschlossen.

Begründung der einfaktoriellen ANOVA. Die Untersuchung des Geschlechtsunterschieds bei Propensity to Trust wurde, wie bereits angedeutet, mittels einfaktorieller ANOVA ohne Messwiederholung vorgenommen. Bei der Erhebung des Geschlechts hätten sich drei Gruppen ergeben könnten, sodass eine Analyse mittels t-Test unmöglich wäre. Wie der Stichprobenschreibung zu entnehmen ist, ergaben sich zwar nur zwei Geschlechter aus der Umfrage. Da jedoch im Sinne der methodischen Integrität mit der Betreuungsperson vereinbart wurde, die Analyseverfahren vor der Datenerhebung zu definieren und nicht post hoc zu ändern, wurde davon abgesehen, auf einen t-Test auszuweichen.

Voraussetzungsprüfung der einfaktoriellen ANOVA. Zunächst wurde die Voraussetzung der Normalverteilung mittels Shapiro-Wilk-Test geprüft und die Daten wurden mittels Histogramm und QQ-Plot optisch begutachtet. Ergänzend wurde die Verteilung mit Hilfe des D'Agostino- und des Anscombe-Glynn-Signifikanztests auf Schiefe respektive Kurtosis überprüft. Abschliessend folgte die Überprüfung der Varianzhomogenität anhand des Levene-Tests.

Längsschnittanalyse. Im Rahmen der Ein-Gruppen-Pretest-Posttest-Untersuchung der Längsschnittuntersuchung wurde folgende empirische Hypothese geprüft:

H_{10} Der Zusammenhang von (a) Trust in AI mit (b) Performance, (c) Process, (d) Purpose und (e) Propensity to Trust unterscheidet sich zu den beiden Messzeitpunkten.

Begründung der einfaktoriellen MANOVA. Die Wahl der MANOVA als geeignetes Analyseverfahren erfolgte in Abstimmung mit der Methodenberatung der Hochschule für Angewandte Psychologie FHNW, auch wenn der geringe Stichprobenumfang die Möglichkeit für ein signifikantes Ergebnis einschränkte. Mangels einer ordinalskalierten Alternative wurden Optionen wie (a) eine Behelfslösung, bei welcher aus den abhängigen Variablen zunächst der Summenscore als neue Variable gebildet wird, um anschliessend die Unterschiede der beiden Messzeitpunkte mittels Wilcoxon-Test zu vergleichen, (b) eine Diskriminanzanalyse, (c) eine Faktorenanalyse oder (d) die Aufnahme beider Messungen in ein SEM mit der Betreuungsperson evaluiert und verworfen. Diese Verfahren wären entweder mit einer Alphafehler-Kumulierung einhergegangen oder hätten die quantitativen Kompetenzen des Autors überstiegen. Gleichzeitig wurde so die Faustregel zur Stichprobengrösse, dass mindestens so viele Fälle bestehen (hier 10) wie abhängige Variablen (hier fünf) existieren, eingehalten (Hemmerich, 2024).

Voraussetzungsprüfung der einfaktoriellen MANOVA. Zur Prüfung der Voraussetzungen wurde zunächst die multivariate Normalverteilung der abhängigen Variablen mittels Shapiro-Wilk-Test und Mardia-Test geprüft. Anschliessend erfolgte die Prüfung der Homogenität der Kovarianz-Matrizen mittels Box-M-Test, gefolgt vom Levene-Test zur Prüfung der der Varianzenhomogenität. Im nächsten Schritt wurden auf multivariate Ausreisser mittels Mahalanobis-Distance geprüft. Anhand der Streudiagramme wurde die lineare Beziehung zwischen den abhängigen Variablen für jeden Messzeitpunkt untersucht. Abschliessend wurde die Multikollinearität zwischen den abhängigen Variablen getestet.

Ergebnisse

Dieses Kapitel stellt die Ergebnisse der Querschnitt- und der Längsschnittuntersuchung, getrennt voneinander, beginnend mit der Querschnittuntersuchung, vor.

Ergebnisse der Querschnittuntersuchung

Die Querschnittuntersuchung umfasste eine Pfadanalyse, zwei lineare sowie eine multiple Regressionsanalyse und eine einfaktorielle ANOVA ohne Messwiederholung.

Pfadanalyse des konzeptuellen Pfadmodells. Im Folgenden werden zunächst die Voraussetzungsprüfung der Pfadanalyse, dann die Respezifikation des konzeptuellen Pfadmodells und abschliessend die Ergebnisse der Pfadanalyse, welche sich auf die Prüfung der Hypothesen 1 bis 5 beziehen, dargestellt.

Ergebnisse der Voraussetzungsprüfung der Pfadanalyse. Bei der Prüfung der multivariaten Normalverteilung resultierte der Mardias-Test bei der Schiefe in einem signifikanten Ergebnis ($p < .001$), was auf eine Verletzung dieser Voraussetzung hindeutete. Bei der Kurtosis hingegen zeigte sich kein signifikantes Ergebnis ($p < .874$). Des Weiteren lag gemäss dem Shapiro-Wilk-Test eine Verletzung der univariaten Normalverteilung vor. So ergaben sich zu den endogenen Variablen Trust in AI ($p = .002$) und Intention to Use ($p < .001$) signifikante Resultate. Das QQ-Plot der multivariaten Normalverteilung (Abbildung C1) und die Streudiagramme zur Prüfung der Linearität (Abbildung C2) sind in Anhang C ersichtlich; Letztere schien demnach weitestgehend gegeben zu sein. Da jedoch die Normalverteilung nicht gegeben war, wurde die Modellpassung mittels robuster Fit-Indizes anhand des Standardfehlers bewertet.

Tabelle 1
Schrittweise Respezifikation des Pfadmodells

	CFI	SRMR	RMSEA	90 % KI	*df*	Δ $\chi 2$	Δ *p*
Konzeptuelles Modell	.680	.182	.287	[.211, .371]	5		
1. Respezifikation	.895	.128	.184	[.091, .245]	4	30.01	< .001
2. Respezifikation	.975	.077	.103	[.000, .235]	3	12.02	< .001
3. Respezifikation	1.00	.010	< .001	[.000, .105]	2	6.92	< .001
Finales Modell	1.00	.038	< .001	[.000, .153]	4	3.96	.14

Anmerkungen. CFI: Comparative Fit Index, SRMR: Standardized Root Mean Squared Residual, RMSEA: Root Mean Square Error of Approximation, 90 % KI: 90 % Konfidenzintervall für RMSEA.

Respezifikation des Pfadmodells. Für das Erreichen eines angemessenen Modell-Fits war die schrittweise Respezifikation des konzeptuellen Modells notwendig (siehe Tabelle 1). Diese Schritte werden im Folgenden detailliert geschildert.

Erste Respezifikation des Pfadmodells. Im ersten Schritt wurde der Pfad von Propensity to Trust zu Performance hinzugefügt, was den Modell-Fit signifikant verbesserte ($\Delta \chi 2 = 30.014$, $\Delta p < .05$). Jedoch erwies sich die Passung weiterhin als mangelhaft (CFI = .895, SRMR = .128, RMSEA = .184).

Zweite Respezifikation des Pfadmodells. Die Ergänzung des Pfades von Propensity to Trust zu Process in der zweiten Respezifikation führte ebenfalls zu einer signifikanten Verbesserung der Modellpassung ($\Delta \chi 2 = 12.018$, $\Delta p < .05$), der RMSEA (.103) lag jedoch, entgegen den beiden anderen Indizes CFI (.975) und (SRMR) .077) nicht im Bereich des Akzeptablen.

Dritte Respezifikation des Pfadmodells. Bei der dritten Modellrespezifikation (siehe Abbildung 5) wurde der Pfad von Propensity to Trust zu Purpose ergänzt. Dies resultierte abermals in einer signifikanten Verbesserung des Modell-Fits ($\Delta \chi 2 = 6.9215$, $\Delta p < .05$) mit durchgehend guten Fit-Indizes (CFI = 1.0, SRMR = .010, RMSEA < .001). Alternativ dazu wurden die drei neuen Pfade, wie von Mayer et al. (1995) postuliert, durch moderierende Effekte von Propensity to Trust auf die Zusammenhänge zwischen den drei Variablen der Perceived Trustworthiness und Trust in AI ersetzt, was jedoch keinen angemessenen Modell-Fit (CFI = .653, SRMR = .145, RMSEA = .162) mit sich brachte.

Finale Respezifikation des Pfadmodells. Die schrittweise Respezifikation des Pfadmodells offenbarte keine signifikanten Zusammenhänge zur Intention to Use (siehe Ergebnisse der dritten Respezifikation des Pfadmodells). Zudem reduzierten die drei hinzugefügten Pfade die Freiheitsgrade und damit die Erklärungsgüte des Modells. Des Weiteren haben die Hypothesen der Pfade zur Variable Intention to Use, wie im Kapitel «Hypothesenbildung» erörtert, einen explorativen Charakter. Daher wurden in der finalen Modellrespezifikation (siehe Abbildung 6) die direkten Pfade von (a) Performance und (b) Propensity to Trust zu Intention to Use entfernt. Dies stellte keine signifikante Verbesserung ($\Delta \chi 2 = 3.9636$, $\Delta p = .14$), aber auch keine Verschlechterung des statistischen Modells dar, denn es lag ein ausgezeichneter Modell-Fit (CFI = 1.0, SRMR = .038, RMSEA < .001) vor. Jedoch verbesserten sich dadurch die konzeptuelle Passung und die Güte des Modells, da sich die Freiheitsgrade wieder erhöhten ($\Delta df = 2$).

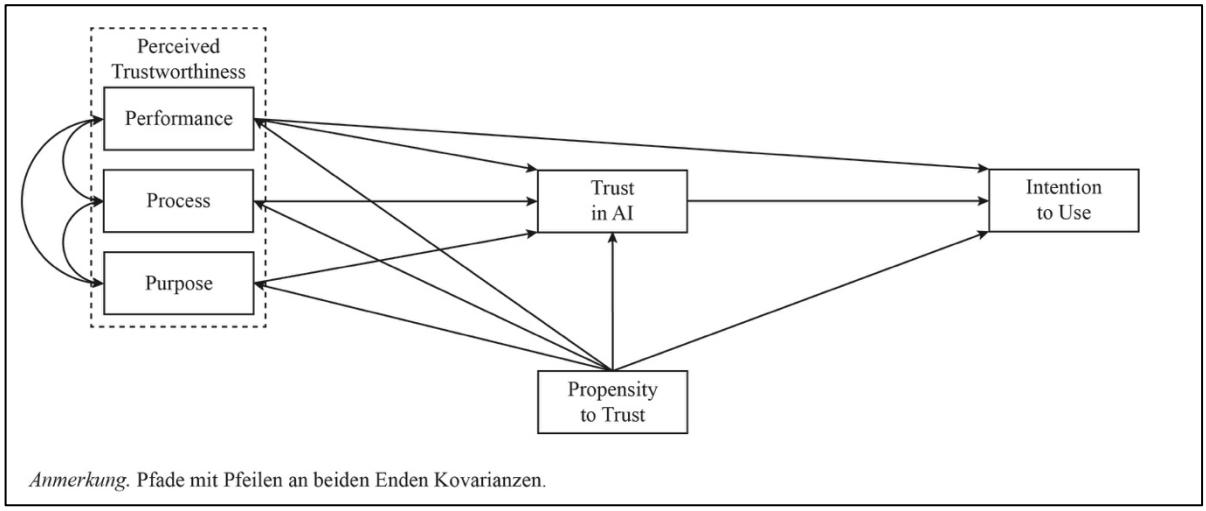


Abbildung 5. Dritte Respezifikation des Pfadmodells (eigene Abbildung)

Ergebnisse der Pfadanalyse des Pfadmodells. Die Mittelwerte und die Standardabweichungen aller im Pfadmodell analysierten Variablen sind in Tabelle 2 ersichtlich. Aus Gründen der Transparenz wird neben dem finalen Modell auch die dritte Modellrespezifikation erläutert. Dadurch wird ein Vergleich der differenzierenden Ergebnisse ermöglicht, was der Verfasser dieser Arbeit aufgrund der in Bezug auf den erneuten Modell-Fit nicht notwendigen Respezifikation angemessen erscheint.

Tabelle 2
Mittelwerte und Standardabweichungen der Variablen des Pfadmodells

Performance	Process	Purpose	Propensity	Trust	Intention
3.17 (.489)	3.12 (.543)	2.81 (.688)	2.87 (.677)	3.09 (.826)	4.32 (792)

Anmerkungen. Propensity: Propensity to Trust; Trust: Trust in AI; Intention: Intention to Use; Standardabweichung in Klammern; $N = 105$.

Ergebnisse der dritten Respezifikation des Pfadmodells. Das Pfadmodell der dritten Respezifikation umfasst 18 der möglichen 20 Parameter ($df = 2$). Dabei erklären die Pfade von (a) Performance, (b) Process, (c) Purpose und (d) Propensity to Trust zum Mediator Trust in AI 49.7 % ($R^2 = .497$) der Varianz. Zudem sind 14.0 % ($R^2 = .140$) der Varianz auf die direkten Pfade von (a) Trust in AI (b) Performance und (c) Propensity to Trust zur endogenen Variable Intention to Use zurückzuführen. Abschliessend erklären die Pfade von Propensity to Trust zu den Variablen der Perceived Trustworthiness (a) Performance 30.2 % ($R^2 = .302$), (b) Process 14.1 % ($R^2 = .141$) und (c) Purpose 6.4 % ($R^2 = .064$) der Varianz.

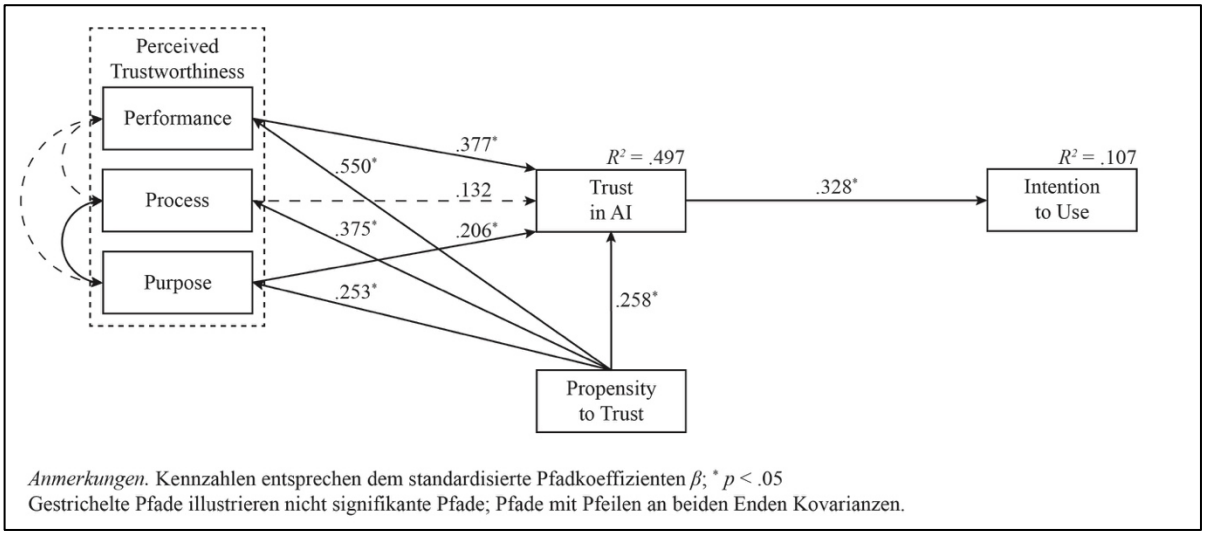


Abbildung 6. Pfadanalyse des finalen Pfadmodells (eigene Abbildung)

Dabei erweisen sich die direkten Zusammenhänge zwischen (a) Performance bei mittlerer Effektstärke (β = .377, $p < .001$), (b) Purpose bei kleiner Effektstärke (β = .206, $p = .022$) sowie (c) Propensity to Trust bei kleiner Effektstärke (β = .258, $p = .048$) und dem Mediator Trust in AI als signifikant. Dies gilt jedoch nicht für den Zusammenhang mit (d) Process (β = .132, $p = .106$). Hingegen sind für die endogene Variable Intention to Use keine signifikanten Zusammenhänge von (a) Performance (β = .165, $p = .151$), (b) Trust in AI (β = .169, $p = .187$) und (c) Propensity to Trust (β = .107, $p = .352$) ersichtlich. Die im Rahmen der Respezifikation des Pfadmodells ergänzten Pfade von Propensity to Trust zu (a) Performance bei grosser Effektstärke (β = .550, $p < .001$), zu (b) Purpose bei mittlerer Effektstärke (β = .375, $p < .001$) und zu Process bei kleiner Effektstärke (β = .253, $p = .039$) zeigen sich alle signifikant. Bei den Koeffizienten der wahrgenommen Vertrauenswürdigkeit liegt lediglich ein signifikanter kleiner Effekt zwischen Process und Purpose (β = .197, $p = .030$) vor. Die Koeffizienten von Performance und Process (β = .045, $p = .674$) bzw. Performance und Purpose (β = .075, $p = .477$) sind hingegen nicht signifikant.

Ergebnisse des finalen Pfadmodells. Das finale Pfadmodell (siehe Abbildung 6) umfasst 16 der möglichen 20 Parameter ($df = 4$). Im Vergleich zum Pfadmodell der dritten Respezifikation gibt es, was die aufgeklärte Varianz angeht, lediglich einen Unterschied. So wurden ausschliesslich Pfade bezüglich der endogenen Variable Intention to Use respezifiziert. Dies resultiert darin, dass nur noch 10.7 % ($R^2 = .107$) der Varianz dieser

endogenen Variable aufgeklärt werden. Alle weiteren Varianzen sind erwartungsgemäss mit jenen der dritten Respezifikation identisch.

Ebenso führte die Modelloptimierung gegenüber der dritten Respezifikation lediglich zur Veränderung eines Effekts. Konkret weist Trust in AI nun einen kleinen signifikanten direkten Zusammenhang (β = .328, $p < .001$) mit Intention to Use auf. Des Weiteren ist der indirekte Zusammenhang von Performance mit Intention to Use bei geringem Effekt signifikant (β = *.123*, $p = .019$). Hingegen ist für den indirekten Pfad zwischen Propensity to Trust und Intention to Use kein signifikanter Effekt (β = *.084*, $p = .068$) erkennbar. Die gesamten Ergebnisse des finalen Pfadmodells sind in Tabelle 3 aufgelistet.

Tabelle 3
Pfadkoeffizienten und Anteile aufgeklärter Varianz des finalen Pfadmodells

	β	*p*	95 % KI	R^2
Trust in AI ~				.497
Performance	.377	< .001*	[.200, .553]	
Process	.132	.106	[-.028, .292]	
Purpose	.206	.022*	[.030, .382]	
Propensity to Trust	.258	.048*	[.003, .512]	
Intention to Use ~				.107
Trust in AI	.328	< .001*	[.138, .517]	
Performance	*.123*	*.019**	[.020, .226]	
Propensity to Trust	*.084*	*.068*	[-.006, .175]	
Performance ~				.302
Propensity to Trust	.550	< .001*	[.376, .723]	
Process ~				.141
Propensity to Trust	.375	< .001*	[.167, .583]	
Purpose ~				.064
Propensity to Trust	.253	.039*	[.013, .492]	
Performance ~~ Process	.045	.674	[-.126, .194]	
Process ~~ Purpose	.197	.030*	[.017, .334]	
Purpose ~~ Performance	.075	.477	[-.106, .227]	

Anmerkungen. 95 % KI = 95 % Konfidenzintervall; ~: AV regressiert auf UV, ~~: UV1 kovariiert mit UV2. Indirekte Pfade in kursiv; $N = 105$. * $p < .05$

Analyse der Einflussfaktoren der Neigung zum Vertrauen. In diesem Unterkapitel werden die Ergebnisse der Prüfung der Hypothesen 6 bis 9 behandelt, anhand derer Einflussfaktoren auf Propensity to Trust untersucht wurden. Dabei werden getrennt nach den Hypothesen zunächst die Voraussetzungsprüfung und anschliessend die Ergebnisse der jeweiligen Analyse dargestellt.

Lineare Regressionsanalyse der Vorerfahrung. Bei der Analyse der Regression zwischen Vorerfahrung und Propensity to Trust wurde zunächst die Prüfung der Voraussetzungen sowie anschliessend die eigentliche Regressionsanalyse vorgenommen.

Voraussetzungsprüfung der linearen Regressionsanalyse der Vorerfahrung. Der Rainbow-Test zur Prüfung der Linearität war nicht signifikant ($F_{(53, 50)} = .713$, $p = .887$). Auch das Streudiagramm (siehe Abbildung D1 in Anhang D) erwies sich als unauffällig und die Voraussetzung der Linearität war gegeben. Das Histogramm (siehe Abbildung D2 in Anhang D) und das QQ-Plot (siehe Abbildung D3 in Anhang D) der Verteilung der Fehlerwerte entsprachen annähernd einer Normalverteilung. Der Breusch-Pagan-Test zur Prüfung der Homoskedastizität fiel ebenfalls nicht signifikant aus ($\chi^2 = 1.33$, $p = .249$) und der Wert von 1.99 (> 1, < 3) des Durbin-Watson-Tests ($p = .954$) liess erkennen, dass keine Autokorrelation vorlag. Die Ausreisserdiagnostik mittels Cooks-Distance offenbarte sieben Ausreisser (siehe Abbildung D4 Anhang D), welche von der Analyse ausgeschlossen wurden, was in einem Umfang von 98 Personen des Subsamples resultierte.

Ergebnisse der linearen Regressionsanalyse der Vorerfahrung. Die deskriptive Statistik dieser Regressionsanalyse ($n = 98$) zeigte bei Propensity to Trust einen Mittelwert von 2.87 ($SD = .616$) und bei Vorerfahrung von 3.24 ($SD = 1.22$). In der Regressionsanalyse konnte hier kein signifikanter Zusammenhang von Vorerfahrung und Propensity to Trust ($F_{(1,98)} = .407$, $R^2 = .004$, korrigiertes $R^2 = -.006$, $p = .525$) aufgedeckt werden.

Lineare Regressionsanalyse des Alters. Abermals wurde vor der Regressionsanalyse zwischen Alter und Propensity to Trust zunächst die Prüfung der Voraussetzungen vorgenommen.

Ergebnisse der Voraussetzungsprüfung der linearen Regressionsanalyse des Alters. Der Rainbow-Test war nicht signifikant ($F_{(47, 44)} = .968$, $p = .544$), das Streudiagramm (siehe Abbildung E1 in Anhang E) fiel ebenfalls unauffällig aus. Somit war die Annahme der Linearität erfüllt. Gemäss Histogramm (siehe Abbildung E2 in Anhang E) und QQ-Plot (siehe Abbildung E3 in Anhang E) sowie dem nicht signifikanten Breusch-Pagan-Test ($\chi^2 = 3.54$, $p = .060$) bzw. dem Wert von 2.06 (> 1, < 3) des Durbin-Watson-Tests ($p = .764$)

war die Voraussetzung der Normalverteilung der Residuen erfüllt. Die Cooks-Distance offenbarte fünf Ausreisser (siehe Abbildung E4 in Anhang E), die von der Analyse ausgeschlossen wurden und zu einem Subsample von 88 Personen führte.

Ergebnisse der lineare Regressionsanalyse des Alters. Bei dieser Regressionsanalyse (n = 88) wies die deskriptive Statistik für Propensity to Trust einen Mittelwert von 2.91 (SD = .627) und für das Alter einen Mittelwert von 21.6 (SD = 2.96) aus. Im Rahmen der Analyse zeigte sich kein signifikanter Zusammenhang zwischen Alter und Propensity to Trust ($F_{(1,86)}$ = .001, R^2 < .001, korrigiertes R^2 = -.012, p = .981).

Multiple lineare Regressionsanalyse der Persönlichkeitsmerkmale. Im Zuge der multiplen Regressionsanalyse zwischen den Persönlichkeitsfaktoren und Propensity to Trust wurde bei der vorangehenden Voraussetzungsprüfung ergänzend die Multikollinearität geprüft.

Ergebnisse der Voraussetzungsprüfung der multiplen linearen Regressionsanalyse. Die Annahme der Linearität war nach der Durchführung des Rainbow-Tests ($F_{(53, 46)}$ = .752, p = .843) und der optischen Prüfung der Streudiagramme (siehe Abbildung F1 in Anhang F) erfüllt. Die VIF-Werte lagen bei den unabhängigen Variablen (a) Offenheit (1.087), (b) Gewissenhaftigkeit (1.107), (c) Extraversion (1.173), (d) Verträglichkeit (1.214) und (e) Neurotizismus (1.109) alle deutlich unter 10, womit nicht von einem Problem hinsichtlich einer Multikollinearität auszugehen war. Die Normalverteilung der Residuen war ebenfalls gegeben. So waren weder in Histogramm (siehe Abbildung F2 in Anhang F) und QQ-Plot (siehe Abbildung F3 in Anhang F) noch beim Breusch-Pagan-Test (χ^2 = 3.25, p = .071) oder am Wert von 1.94 (> 1, < 3) des Durbin-Watson-Tests (p = .670) Auffälligkeiten sichtbar. Die Cooks-Distance offenbarte acht Ausreisser (siehe Abbildung F4 in Anhang F), die von der Analyse (n = 97) ausgeschlossen wurden.

Ergebnisse der multiple lineare Regressionsanalyse. Die Mittelwerte und die Standardabweichungen der Variablen dieser multiplen Regressionsanalyse sind in Tabelle 4 aufgeführt.

Tabelle 4
Mittelwerte und Standardabweichungen der Variablen der multiplen Regressionsanalyse

Propensity	Offenheit	Gewissen.	Extraversion	Verträglich	Neurotizismus
2.83 (.565)	3.26 (.869)	4.04 (.628)	3.62 (.710)	2.53 (.702)	2.48 (765)

Anmerkungen. Standardabweichung in Klammern; Gewissen.: Gewissenhaftigkeit; n = 97.

Aus der Regressionsanalyse ging ein signifikanter mittlerer Zusammenhang zwischen den Persönlichkeitsfaktoren und Propensity to Trust ($F_{(5,91)} = 3.26$, $R^2 = .152$, korrigiertes $R^2 = .105$, $p = .009$) hervor, wobei 15.2 % der Varianz in Propensity to Trust durch diese fünf Faktoren aufgeklärt werden. Dabei stehen jedoch lediglich (a) die Gewissenhaftigkeit (β = -.255, $p = .017$) und (b) der Neurotizismus (β = -.291, $p = .005$) in einem signifikanten kleinen negativen Zusammenhang mit Propensity to Trust. Die Zusammenhänge von (c) Extraversion (β = .041, $p = .696$), (d) Offenheit (β = .167, $p = .104$) und (e) Verträglichkeit (β = .142, $p = .198$), stehen zwar in einem positiven, aber nicht signifikanten Zusammenhang mit Propensity to Trust. Die Ergebnisse der multiplen Regressionsanalyse sind in Tabelle 5 dargestellt.

Tabelle 5
Multiple Regressionsanalyse für Persönlichkeitsmerkmale auf Propensity to Trust

	B	β	*p*
Konstante	3.526		
Offenheit	.109	.167	.104
Gewissenhaftigkeit	-.230	-.255	.017*
Extraversion	.033	.041	.696
Verträglichkeit	.114	.142	.198
Neurotizismus	-.215	-.291	.005*
R^2		.152	.009*

Anmerkungen. $n = 97$.
* $p < .05$

Einfaktorielle ANOVA der Geschlechtsunterschiede. Zur Analyse der Geschlechtsunterschiede bei Propensity to Trust wurde zunächst die Prüfung der Voraussetzungen und anschliessend die Varianzanalyse realisiert.

Ergebnisse der Voraussetzungsprüfung der einfaktoriellen ANOVA. Die Prüfung der Normalverteilung mittels Shapiro-Wilk-Test ($p = .010$) deutete auf eine Verletzung dieser Voraussetzung hin. Zudem offenbarte das Ergebnis des D'Agostino-Signifikanztests ($p = .016$) eine schiefe Verteilung. Diese Ergebnisse sind ebenfalls im Histogramm (siehe Abbildung G1 in Anhang G) und im QQ-Plot (siehe Abbildung G2 in Anhang G) optisch nachvollziehbar. Auch wenn der Anscombe-Glynn-Signifikanztest ($p = .411$) keine gewölbte Verteilung andeutete, schien die Bedingung der Normalverteilung verletzt zu sein. Der

Levene-Test ($F_{(1,103)} = .007$, $p = .933$) gab jedoch keine Hinweise auf eine Verletzung der Varianzhomogenität.

Obwohl die ANOVA als relativ robust gegenüber Verletzungen der Normalverteilungsannahme gilt, sind dennoch Ergebnisverzerrungen möglich (Schmider, Ziegler, Danay, Beyer & Bühner, 2010). Daher wurde zusätzlich zur ANOVA der ordinalskalierte Kruskal-Wallis-Test angewendet, um anschliessend die beiden Ergebnisse miteinander zu vergleichen. Dieser Vergleich diente der Sicherstellung, dass die Verletzung der Normalverteilungsannahme keinen Einfluss auf das Ergebnis der ANOVA hat.

Ergebnisse der einfaktorielle ANOVA. Die deskriptive Analyse der Mittelwerte, der Standardabweichungen und der Mediane von Propensity to Trust wies bereits darauf hin, dass kein signifikanter Geschlechtsunterscheid besteht (siehe Tabelle 6). Mittelwerte, Standardabweichungen und Mediane waren nahezu identisch. Entsprechend konnte auch durch die ANOVA kein signifikanter Geschlechtsunterschied bei Propensity to Trust ($F_{(1,103)} = .602$, $p = .440$) identifiziert werden. Dieses Ergebnis spiegelte auch der Kruskal-Wallis-Tests ($\chi^2 = .629$, $p = .428$) wider.

Tabelle 6
Mittelwerte, Standardabweichungen und Median von Propensity to Trust nach Geschlecht

	n	*M*	*SD*	*Md*
weiblich	51	2.82	.67	2.67
männlich	54	2.92	.68	2.67

Ergebnisse der Längsschnittuntersuchung

Im Rahmen der Längsschnittuntersuchung wurde die zeitliche Veränderung der Einflussfaktoren von Trust in AI untersucht. Für die Analyse dieser Unterschiede zwischen zwei Messzeitpunkten wurde eine einfaktorielle MANOVA mit Messwiederholung berechnet. Vor der eigentlichen Analyse wurden zunächst die Voraussetzungen geprüft.

Ergebnisse der Voraussetzungsprüfung der einfaktoriellen MANOVA. Weder der Shapiro-Wilk-Test (siehe Tabelle H1 in Anhang H) noch der Mardia-Test zeigten bezüglich Schiefe (p = .230) und Kurtosis (p = .244) Signifikanzen auf, welche gegen eine multivariate Normalverteilung der abhängigen Variablen gesprochen hätten. Der Box M-Test erwies sich ebenfalls als nicht signifikant (p = .253), womit von einer Homogenität der Kovarianz-Matrizen ausgegangen werden konnte. Der Levene-Test (siehe Tabelle H2 Anhang H) fiel bei keiner der abhängigen Variablen signifikant aus, was wiederum eine Varianzenhomogenität annehmen liess. Die Analyse mittels Mahalanobis-Distance zeigte keine multivariaten Ausreisser. Durch die optische Begutachtung der Streudiagramme (siehe Abbildung A1 Anhang H) konnte auf eine lineare Beziehung zwischen den abhängigen Variablen geschlossen werden. Die Korrelation zwischen den abhängigen Variablen (siehe Tabelle 7) war zwar teilweise gross, jedoch nicht grösser als r = .669, womit kein Problem mit Multikollinearität anzunehmen war. Damit waren alle Voraussetzungen für eine MANOVA erfüllt.

Tabelle 7
Korrelationen von Trust in AI, Propensity to Trust, Performance, Process und Purpose

	1	2	3	4	5
1. Trust in AI					
2. Propensity to Trust	.628				
3. Performance	.288	.368			
4. Process	.509	.476	.415		
5. Purpose	.187	.669	.174	.288	

Anmerkungen. n = 10.

Ergebnisse der einfaktoriellen MANOVA. Die deskriptive Statistik der abhängigen Variablen (a) Trust in AI, (b) Propensity to Trust, (c) Performance, (d) Process und (e) Purpose (siehe Tabelle 8) deutet darauf hin, dass zwischen den Messzeitpunkten keine signifikanten Unterschiede bestehen. Dementsprechend hat auch die MANOVA keine signifikanten Unterschiede im Zusammenhang dieser Variablen ($F_{(5, 14)} = .104$, $\eta^2 = .04$, $\Lambda = .964$, $p = .990$) offenbart. In Anbetracht dessen wurde auf die Durchführung eines Post-hoc-Tests verzichtet.

Tabelle 8
Mittelwerte und Standardabweichungen von Trust in AI, Propensity to Trust, Performance, Process und Purpose nach Messzeitpunkt

	Trust in AI	Propensity	Performance	Process	Purpose
Pretest	2.7 (.587)	2.5 (.423)	2.9 (.492)	3.2 (.463)	2.8 (.425)
Posttest	2.7 (.675)	2.5 (.281)	2.9 (.387)	3.2 (.583)	2.9 (.530)

Anmerkungen. Standardabweichung in Klammern; $n = 10$.

Diskussion

Im Rahmen der Querschnittuntersuchung wurde den Fragen nachgegangen, welche Faktoren im Kontext von ChatGPT einen Einfluss auf (a) das Vertrauen in KI, (b) die Nutzungsintention und (c) die Neigung zum Vertrauen haben. Ergänzend wurde in der Längsschnittuntersuchung geprüft, ob nach einer instruktionalen Intervention Veränderungen der Zusammenhänge zwischen den Einflussfaktoren und dem Vertrauen in KI auftreten. Im Folgenden werden die Hypothesen zu diesen Fragen unter Bezugnahme auf die Ergebnisse beantwortet.

Interpretation der Ergebnisse der Querschnittuntersuchung

Entgegen der Annahme aus Hypothese 1 und somit auch jener von Solberg et al. (2022) zeigten lediglich Performance und Purpose als Faktoren der wahrgenommenen Vertrauenswürdigkeit, nicht aber Process einen direkten Einfluss auf das Vertrauen in KI. Gleichzeitig stimmen diese Ergebnisse mit denen von Langer et al. (2023) überein.

Dies ist dahingehend interessant, dass insbesondere Process mit den in Bezug dazu stehenden Aspekten Erklärbarkeit, Transparenz und Interpretierbarkeit im Kontext von XAI eine hohe Relevanz für das Vertrauen zugeschrieben wird (Hoffman et al., 2023; Shin, 2021). Es könnte damit erklärt werden, dass es sich bei ChatGPT um ein Blackbox-Modell handelt, dessen Funktionsweise für die Nutzenden nicht nachvollziehbar ist (Rai, 2020). Die statistischen Daten deuten jedoch nicht darauf hin, denn die Mittelwerte und die Standardabweichungen der Variablen Process sind vergleichbar mit jenen von Performance und Purpose. Gleichzeitig gibt es keine Hinweise darauf, dass sich die Nachvollziehbarkeit respektive deren Absenz, wie unter anderem von Gebru et al. (2022), Hoffman et al. (2023), Rai (2020) und Shin (2021) angeführt, auf das Vertrauen auswirkt. Vielmehr stellt sich die Frage, ob sich das Fehlen dieses Zusammenhangs auf eine adäquate Kalibrierung des Vertrauens auswirken könnte. Letzteres wäre dann der Fall, wenn Process, wie theoretisch hergeleitet, in der Realität tatsächlich einen Einfluss auf das Vertrauen hätte, aber so wie in den vorliegenden Ergebnissen bei der Vertrauensbildung nicht berücksichtigt würde. Ob das Beschriebene zutrifft, kann aber nicht anhand der Ergebnisse dieser Arbeit beantwortet werden. An dieser Stelle sollte aber auch bedacht werden, dass Transparenz der Theorie zur Folge allein keine adäquate Vertrauensbildung sicherstellt (Okamura & Yamada, 2020). Ausserdem gehen mehr Informationen nicht automatisch mit einem höheren Vertrauen einher (Mackay et al., 2020). Dies erscheint insbesondere in Bezug auf ChatGPT nachvollziehbar. Zwar könnte erklärt werden, wie genau der Algorithmus zu einem Ergebnis

kommt, jedoch kann anhand der Erklärung nicht bewertet werden, ob dieses tatsächlich richtig ist. Vielmehr ist ChatGPT ein probabilistisches Sprachmodell, welches somit keinen Anspruch auf Richtigkeit erhebt, da es sich lediglich um «intelligentes Raten» handelt (Zweig, 2023). Eine Bewertung der Richtigkeit bedarf in diesem Kontext der Konsultation einer weiteren, externen Quelle.

Dass mit der Variable Purpose das Wohlwollen der Entwickler und Entwicklerinnen ein relevanter Faktor für Vertrauen ist, lässt sich zwar theoretisch gut begründen (u. a. Bedué & Fritzsche, 2021; Kaplan et al., 2023; Siau & Wang, 2018; Solberg et al., 2022). Bei näherer Betrachtung scheint es im Untersuchungskontext jedoch nicht selbstverständlich. So ist OpenAI ausschliesslich für KI-Systeme bekannt, wobei die Organisation selbst möglicherweise zahlreichen Personen kein Begriff ist.

Wie in Hypothese 2 angenommen, beeinflusst auch die Neigung zum Vertrauen das Vertrauen in KI. Es hat sich jedoch gezeigt, dass dieser Einfluss geringer ist als jener von Performance. Dies ist insofern als positiv zu erachten, als die Persönlichkeit der Nutzenden weniger relevant für das Vertrauen zu sein scheint als die Fähigkeiten des KI-Systems. So ist ein bedeutender Faktor für ein adäquates Vertrauen die korrekte Einschätzung der Fähigkeiten des Systems (Gebru et al., 2022; Lee & See, 2004). Jedoch ist anzumerken, dass in der vorliegenden Arbeit nicht die Korrektheit der Einschätzung dieser Fähigkeiten gemessen wurde.

Entsprechend der explorativen Hypothese 3 konnte auch ein direkter Einfluss des Vertrauens in KI auf die Nutzungsintention festgestellt werden. Es zeigte sich jedoch nicht der starke Zusammenhang, der im Kontext von Vertrauen in Automation von Razin und Feigh (2023) postuliert wird. Dies deutet darauf hin, dass neben Vertrauen andere Faktoren, beispielsweise gemäss Choung et al. (2023) die wahrgenommene Nützlichkeit und die Nutzungsfreundlichkeit, ebenfalls – potenziell sogar grössere – Einflüsse auf die Nutzungsintention haben. Entgegen der Empfehlung von Langer et al. (2023) wurde keine spezifische Aufgabe, welche mit ChatGPT erledigt werden sollte, definiert. Folglich könnte ChatGPT für bestimmte Aufgaben als nützlich und für andere als weniger nützlich eingestuft worden sein, was die Messung der Nutzungsintention beeinflusst haben könnte. Möglicherweise wird dennoch Vertrauen eine grössere Relevanz für die erfolgreiche Implantation eines KI-Systems zugeschrieben als tatsächlich vorliegt. Dabei könnten auch die nach Mayer et al. (1995, S. 712–714) von Vertrauen abzugrenzenden Konstrukte Kooperation, Zuversicht und Berechenbarkeit eine Rolle spielen. Entgegen der Annahme

von Tschopp und Ruef (2020) wäre es dennoch denkbar, dass Personen, welche ChatGPT nicht vertrauen, diese trotz dessen nutzen würden.

Im Widerspruch zu Hypothese 4 steht, dass lediglich ein durch das Vertrauen in KI mediierter Zusammenhang zwischen Performance und der Nutzungsintention festzustellen war. Dies könnte wiederum gegen die von Choung et al. (2023) angeführte Nützlichkeit als Einflussfaktor sprechen. So wurde mit dem Faktor Performance implizit die Nützlichkeit von ChatGPT durch Items zur Zuverlässigkeit und zur Fähigkeit des Systems erfragt. Zudem lässt sich eine hohe Leistung bezüglich der zu erledigende Aufgabe als nützlich interpretieren.

Der in Hypothese 5 angenommene Zusammenhang zwischen der Neigung zum Vertrauen und der Nutzungsintention zeigte sich auch nicht indirekt. Dies kann ebenfalls als positiv erachtet werden, da somit die Persönlichkeit der Nutzenden keinen direkten Einfluss hat und stattdessen vermutet werden kann, dass dem System zuschreibbare Aspekte eine höhere Relevanz für die Nutzungsintention haben.

Die Respezifikation, welche aufgrund der ungenügenden Passung des konzeptuellen Modells zu den Daten notwendig war, legt zudem einen Einfluss der Neigung zum Vertrauen auf die wahrgenommene Vertrauenswürdigkeit nahe. Dies ist insofern überraschend, als Lee und See (2004), aber auch Körber (2019) lediglich auf Moderationen der Zusammenhänge zwischen der wahrgenommenen Vertrauenswürdigkeit und dem Vertrauen durch die Neigung zum Vertrauen hinweisen. Dabei suggeriert die Bezeichnung «wahrgenommene Vertrauenswürdigkeit» bereits, dass es sich um eine subjektive Einschätzung handelt, welche nicht nur durch das System, sondern auch durch die nutzende Person bestimmt wird. Dies ist dahingehend bedenklich, dass so das Vertrauen indirekt auch von systemunabhängigen Faktoren beeinflusst wird, was eine adäquate Vertrauensbildung erschweren könnte.

Die Frage, wie genau sich die Neigung zum Vertrauen zusammensetzt, kann jedoch in dieser Arbeit nicht beantwortet werden. Weder für die Vorerfahrung (Hypothese 6) noch für das Alter (Hypothese 7) oder das Geschlecht (Hypothese 9) konnte ein Einfluss auf die Neigung zum Vertrauen aufgedeckt werden. Diese Hypothesen sind jedoch als explorativ zu erachten und es gibt hierzu keine übereinstimmenden Ergebnisse. So weisen Montag, Kraus, et al. (2023) entgegen den Erkenntnissen von Kaplan et al. (2023) beispielsweise darauf hin, dass es keine Geschlechtsunterschiede gibt. Gleichzeitig zeigen sich weder Alters- noch Kohorteneffekte, wie von Hoff und Bashir (2015) postuliert – wobei dies auch auf die Stichprobe und das Studiendesign zurückzuführen ist. So bringt diese Stichprobe ein hohes

Mass an Homogenität bezüglich des Alters mit sich und die Messung von Kohorteneffekte ist nahezu unmöglich, da es sich bei allen Versuchspersonen um junge Erwachsene handelte, die somit der gleichen Kohorte zuzuordnen sind. Möglicherweise ist Vorerfahrung zudem, entgegen der Annahme von Solberg et al. (2022), nicht dem kontext- und systemunabhängigen Vertrauen zuzuordnen.

Auch diese Arbeit trägt zur Unklarheit bezüglich des Einflusses der Persönlichkeitsmerkmale auf die Neigung zum Vertrauen bei. Die Big-Five-Persönlichkeitsfaktoren kläre gemäss den Ergebnissen nur eine geringe Varianz in der Neigung zum Vertrauen auf. Dabei offenbaren die Ergebnisse der Prüfung von Hypothese 8 lediglich kleine negative Zusammenhänge von Gewissenhaftigkeit und Neurotizismus mit der Neigung zum Vertrauen. Dies ist dahingehend interessant, dass der Verfasser dieser Arbeit mit der Studie von Schepman und Rodway (2023) nur einen Hinweis auf einen möglichen Zusammenhang mit Gewissenhaftigkeit bekannt ist. Zudem konnte der von Kaplan et al. (2023) aufgezeigte Zusammenhang bezüglich Neurotizismus hier ebenfalls nachgewiesen werden.

Dementgegen zeigte sich der Zusammenhang von Offenheit nach Kaplan et al. (2023) und in der chinesischen Stichprobe von Sindermann et al. (2022) nicht. Dies ist möglicherweise auch mit der Operationalisierung dieses Persönlichkeitsmerkmals zu erklären; so ist die Messung der Offenheit anhand des NEO-FFI-30 umstritten, da diese Skala eine relativ geringe Kommunalität und geringe interne Konsistenzen aufweist (Körner et al., 2008). Ebenfalls konnte der von Schepman und Rodway (2023) sowie Kaplan et al. (2023) angeführte Zusammenhang mit Extraversion nicht bestätigt werden. Des Weiteren konnte trotz der entsprechenden Annahmen von Schepman und Rodway (2023) sowie der chinesischen Stichprobe von Sindermann et al. (2022) kein Zusammenhang mit Verträglichkeit festgestellt werden. Die geringe aufgeklärte Varianz deutet darauf hin, dass die Neigung zum Vertrauen vor allem durch andere, in dieser Arbeit nicht berücksichtigte Faktoren bestimmt wird. Dies deutet eine klare Forschungslücke an, da der Einfluss der Neigung zum Vertrauen auf das Vertrauen selbst, aber auch auf die wahrgenommene Vertrauenswürdigkeit beachtlich zu sein scheint.

Interpretation der Ergebnisse der Längsschnittuntersuchung

Die Mittelwerte aller gemessenen Faktoren der Längsschnittanalyse, welche der Analyse von Hypothese 10 zuzuordnen sind, fielen nahezu identisch aus. Damit waren auch keine signifikanten Unterschiede in der multivariaten Varianzanalyse festzustellen. In diesem Kontext ist klar zu betonen, dass aufgrund des geringen Stichprobenumfangs von lediglich 10 Versuchspersonen keine signifikanten Ergebnisse zu erwarten waren. So lässt sich aus einer derart kleinen Stichprobe auch keine klare Schlussfolgerung ziehen. Anhand der nahezu identischen Mittelwerte lässt sich lediglich vermuten, dass, wie von Siegrist (2021) postuliert, das Vertrauen möglicherweise weniger dynamisch ist als häufig angenommen. Entsprechend könnte auch der Prozess der Vertrauensbildung weniger dynamisch sein als von Hoff und Bashir (2015) konstatiert.

Limitationen

Auch wenn der Stichprobenumfang entsprechend der benötigten Teststärke für die Querschnittuntersuchung gross genug war, handelte es sich um eine relativ homogene Stichprobe, welche sich nahezu ausschliesslich aus Studierenden der Wirtschaftsinformatik zusammensetzte. Ergänzend sollte angemerkt werden, dass die Homogenität der Stichprobe möglicherweise auch einen Einfluss auf die Diversität der Persönlichkeitsmerkmale und das Technologievertrauen hatte. Zumindest lässt sich Personen, welche sich für ein Studium mit Informatikbezug entscheiden, eine Affinität zu Technik zuschreiben. Somit ist eine Generalisierung auf die Gesamtbevölkerung nicht zulässig, auch wenn die Stichprobe für ein erstes Verständnis des Konstrukts Vertrauen in KI geeignet ist.

Da bei der Studie kein experimentelles Design verwendet wurde, handelte es sich lediglich um eine korrelative Untersuchung. Entsprechend können keine klaren Aussagen zu einer möglichen Kausalität getroffen werden. Diese kann lediglich anhand der theoretischen Grundlagen diskutiert werden. Des Weiteren hat die schrittweise Respezifikation des Pfadmodells dazu geführt, dass die Signifikanz der Pfade nicht mehr eindeutig zu interpretieren ist. Da mit ChatGPT lediglich die Vertrauenswürdigkeit eines KI-Systems beurteilt wurde, ist ausserdem eine Generalisierung auf andere Systeme ausgeschlossen. Dies lässt auch keine Aussage zum Einfluss der objektiven Vertrauenswürdigkeit auf das Pfadmodell zu.

Die Ergebnisse der Längsschnittanalyse sind aufgrund des geringen Stichprobenumfangs nur begrenzt interpretierbar. Zudem handelte es sich lediglich um eine Ein-Gruppen-Pretest-Posttest-Untersuchung ohne Kontrollgruppe, wobei zwar spezifische

Informationen zum System, aber nicht die Erfahrung mit diesem im gemessenen Zeitraum kontrolliert wurde. Dennoch wurde der von Körber (2019) definierte Zeitraum von sechs Wochen zur Messung der Veränderung des Vertrauens eingehalten.

Es ist auch denkbar, dass das Risiko, das gemäss Mayer et al. (1995) im Rahmen von Vertrauen eingegangen werden muss, bei ChatGPT und dessen Nutzung zu gering ist, als dass Vertrauen tatsächlich zum Tragen kommt respektive notwendig wäre. Dies könnte wiederum den geringen Einfluss des Vertrauens auf die Nutzungsintention erklären. Zudem wurde mit ChatGPT zwar das KI-System definiert, jedoch nicht, wie von Langer et al. (2023) empfohlen, dessen Aufgabenfeld spezifiziert, womit die Umfrage möglicherweise zu abstrakt für Selbstauskünfte über das Vertrauen und die Nutzungsintention in dieses ist. Folglich ist infrage zu stellen, ob Vertrauen in KI als globales Konstrukt für ein spezifisches KI-System oder gar KI als gesamthafte Technologie dienen kann.

Implikationen

Ob der fehlende Einfluss von Process auf Trust in AI einen Einfluss auf die adäquate Vertrauensbildung hat, kann anhand dieser Studie nicht gesagt werden. Daher sollte weitere Forschung betrieben werden, bei der die adäquate Vertrauenswürdigkeit mehrerer KI-Systeme untersucht und verglichen wird, um so festzustellen, ob die Erklärbarkeit, welche bei Blackbox-Modellen nicht gegeben ist, dennoch eine hohe Relevanz diesbezüglich mitbringt.

Um darüber hinaus zu eruieren, ob sich das Modell auch auf andere Blackbox-Modelle übertragen lässt, sollten weitere solche KI-Systeme anhand des in dieser Arbeit postulierten Pfadmodells untersucht werden.

Des Weiteren sollten auch Whitebox- respektive Glassbox-Modelle untersucht werden, um festzustellen, ob es beim Vertrauen einen Unterschied zu Blackbox-Modellen gibt. Auf diese Weise wäre es möglich, potenzielle Unterschiede transparent aufzuzeigen, was Einfluss auf die Gestaltung entsprechender Richtlinien für diese unterschiedlichen Modelle haben könnte.

Die Ergebnisse dieser Arbeit legen nahe, dass im Rahmen von Richtlinien zur Gestaltung vertrauenswürdiger KI-Systeme der Fokus weg von der Erklärbarkeit und hin auf die Faktoren Performance und Purpose gerichtet werden sollte. Speziell sollte das Verständnis über die Zuverlässigkeit und die Leistungsfähigkeit des KI-Systems in den Fokus rücken.

Zudem sollten diese Aspekte mit der Reputation anderer Produkte und Services der Unternehmung übereinstimmen, da die Ergebnisse dieser Studie darauf hindeuten, dass die generelle Wahrnehmung der Organisation unabhängig vom spezifischen KI-System Einfluss auf das Vertrauen hat. So kann ein positiver Ruf der Unternehmung zur erfolgreichen Implementierung eines KI-Systems beitragen, aber auch ausgenutzt werden. Entsprechend kann die Reputation der Organisation nicht nur als Vorteil, sondern auch als Massstab erachtet werden, welchem es gerecht zu werden gilt.

Abschliessendes Fazit

Zum aktuellen Zeitpunkt gibt es kein generelles Modell zu Vertrauen in KI (Solberg et al., 2022). Vielmehr bestehen bereits bei der Definition und dem Verständnis von Vertrauen Schwierigkeiten (Lewis & Marsh, 2022; Tschopp & Ruef, 2020). In dieser Arbeit erfolgte hinsichtlich der Definition von Vertrauen eine Orientierung an Mayer et al. (1995) und an dem von Körber (2019) auf dieses Verständnis zugeschnittenen Fragebogen. Damit hat diese Arbeit einen empirischen Einblick in das Vertrauen in KI ermöglicht und kann als erster Schritt zur Entwicklung eines empirischen Modells dienen.

Die Resultate legen nahe, dass sich, wie von Chi et al. (2021) postuliert, etablierte Konstrukte des Vertrauens nicht vollständig auf Vertrauen in KI – zumindest nicht im Kontext von Blackbox-Modellen – übertragen lassen. So zeigte sich für das Verständnis über die Funktionalität des KI-Systems, im Kontrast zu früheren Befunden (Gebru et al., 2022; Hoffman et al., 2023; Rai, 2020; Shin, 2021) kein Einfluss auf das Vertrauen. Dies kann positiv als Coping-Mechanismus interpretiert werden, sollten sich die Personen bewusst sein, dass sie die Funktionalität des Systems nicht gänzlich nachvollziehen können. Sofern jedoch das Verständnis über diese Funktionalität für die adäquate Vertrauensbildung relevant wäre, würde es darauf hindeuten, dass das Vertrauen weniger gut kalibriert ist, da es auf lediglich zwei der drei relevanten Vertrauensfaktoren basiert. Inwiefern ein fehlendes Verständnis ein Über- oder Misstrauen begünstigt, kann anhand dieser Untersuchung allerdings nicht dargelegt werden.

Weiterhin deuten die Ergebnisse darauf hin, dass die wahrgenommene Vertrauenswürdigkeit des KI-Systems ebenso wie das Vertrauen in KI direkt von der Neigung zu Vertrauen der nutzenden Person beeinflusst wird. Dies weist darauf hin, dass der direkte Einfluss des Systems und damit der entwickelnden Organisation auf dessen Vertrauenswürdigkeit möglicherweise weniger relevant für die Wahrnehmung der Vertrauenswürdigkeit ist als die Disposition der vertrauenden Person. Gleichzeitig weisen

die Ergebnisse darauf hin, dass Vertrauen weniger relevant für den Erfolg von KI-Systemen sein könnte als angenommen. Trotz der momentanen Tendenz zum Trustwashing (Tschopp & Ruef, 2020) sollte das nach Mayer et al. (1995) von Vertrauen abzugrenzende Konstrukt der Kooperation nicht vernachlässigt werden. Gemäss diesem kann Kontrolle die Notwendigkeit von Vertrauen ersetzen, was wiederum zur Kooperation führen kann (Mayer et al., 1995; Tschopp & Ruef, 2020).

Auch wenn Kontrolle eine kontraproduktive Wirkung auf das Vertrauen zugesagt wird (Tschopp & Ruef, 2020), betrachtet der Autor dieser Arbeit die beiden Konstrukte nicht als konkurrenzierend. Wo es möglich ist, Kontrolle auszuüben, beispielsweise in Form von XAI, kann dies grundsätzlich legitim sein. Die Komplexität von KI – insbesondere bei Blackbox-Modellen – lässt Kontrolle aber nicht in allen Fällen zu (Choung et al., 2023; Frank et al., 2023). Gleichzeitig hat der Grad der Autonomie des KI-Systems einen hohen Einfluss auf die Kontrollmöglichkeiten (Choung et al., 2023). Hierbei kann jedoch nicht davon ausgegangen werden, dass ChatGPT vollständig ohne die Kontrolle der Nutzenden verwendet wird.

Literatur

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv:1610.07997 [cs.AI]. https://doi.org/10.48550/arXiv.1606.06565

Bedué, P. & Fritzsche, A. (2021). Can we trust AI? An empirical investigation of trust requirements and guide tosSuccessful AI adoption. *Journal of Enterprise Information Management*, *35*(2), 530–549. https://doi.org/10.1108/JEIM-06-2020-0233

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. et al. (2020). Language models are few-shot learners (NIPS'20). In H. Larochelle, M. Ranzato & R. Hadsell (Hrsg.), *Proceedings of the 34th International Conference on Neural Information Processing Systems* (S. 1877–1901). Red Hook, NY, USA: Curran Associates.

Bujang, M. A., Omar, E. D. & Baharum, N. A. (2018). A Review on Sample Size Determination for Cronbach's Alpha Test: A Simple Guide for Researchers. *The Malaysian Journal of Medical Sciences*, *25*(6), 85–99. https://doi.org/10.21315/mjms2018.25.6.9

Chi, O. H., Jia, S., Li, Y. & Gursoy, D. (2021). Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior*, *118*, 106700. https://doi.org/10.1016/j.chb.2021.106700

Choung, H., David, P. & Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human-Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Conte, N. (2024). Ranked: The Most Popular AI Tools. *Visual Capitalist*. Verfügbar unter: https://www.visualcapitalist.com/ranked-the-most-popular-ai-tools/

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Frank, D.-A., Jacobsen, L. F., Søndergaard, H. A. & Otterbring, T. (2023). In companies we trust: consumer adoption of artificial intelligence services and the role of trust in companies and AI autonomy. *Information Technology & People*, *36*(8), 155–173. https://doi.org/10.1108/ITP-09-2022-0721

Galesic, M. & Bosnjak, M. (2009). Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Gebru, B., Zeleke, L., Blankson, D., Nabil, M., Nateghi, S., Homaifar, A. et al. (2022). A Review on Human-Machine Trust Evaluation: Human-Centric and Machine-Centric Perspectives. *IEEE Transactions on Human-Machine Systems*, *52*(5), 952–962. https://doi.org/10.1109/THMS.2022.3144956

Hebbali, A. (2024). *olsrr: Tools for Building OLS Regression Models.* Verfügbar unter: https://olsrr.rsquaredacademy.com

Hemmerich, W. A. (2024). *Einfaktorielle MANOVA: Voraussetzungen*. StatistikGuru.de. Verfügbar unter: https://statistikguru.de/spss/einfaktorielle-manova/voraussetzungen-11.html

Hendrycks, D., Carlini, N., Schulman, J. & Steinhardt, J. (2022). *Unsolved Problems in ML Safety*. arXiv:2109.13916 [cs.LG]. https://doi.org/10.48550/arXiv.2109.13916

Hendrycks, D. & Mazeika, M. (2022). *X-Risk Analysis for AI Research*. arXiv:2206.05862 [cs.CY]. https://doi.org/10.48550/arXiv.2206.05862

Hendrycks, D., Mazeika, M. & Woodside, T. (2023). *An Overview of Catastrophic AI Risks*. arXiv:2306.12001 [cs.CY]. https://doi.org/10.48550/arXiv.2306.12001

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

Hoffman, R. R., Mueller, S. T., Klein, G. & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*, 1096257. https://doi.org/10.3389/fcomp.2023.1096257

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

John, O. P. & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Hrsg.), *Handbook of personality: Theory and research* (Band 2, S. 102–138). New York: Guilford Press.

Jurafsky, D. & Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3. Auflage, Vorab-Onlinepublikation). Verfügbar unter: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kaplan, A. D., Kessler, T. T., Brill, J. C. & Hancock, P. A. (2023). Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors*, *65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Kim, K. H. (2005). The Relation Among Fit Indexes, Power, and Sample Size in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 368–390. https://doi.org/10.1207/s15328007sem1203_2

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander & Y. Fujita, *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Band 4, S. 13–30.

Körner, A., Geyer, M., Roth, M., Drapeau, M., Schmutzer, G., Albani, C. et al. (2008). Persönlichkeitsdiagnostik mit dem NEO-Fünf-Faktoren-Inventar: Die 30-Item-Kurzversion (NEO−FFI−30). *Psychotherapie, Psychosomatik, Medizinische Psychologie*, *58*, 238–45. https://doi.org/10.1055/s-2007-986199

Langer, M., König, C. J., Back, C. & Hemsing, V. (2023). Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of

Unfair Bias. *Journal of Business and Psychology*, *38*(3), 493–508. https://doi.org/10.1007/s10869-022-09829-9

Lee, J. D. & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lewis, P. R. & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, *72*, 33–49. https://doi.org/10.1016/j.cogsys.2021.11.001

Mackay, A., Fortes, I., Santos, C., Machado, D., Barbosa, P., Boas, V. V. et al. (2020). The Impact of Autonomous Vehicles' Active Feedback on Trust. In P.M. Arezes (Hrsg.), *Advances in Safety Management and Human Factors* (S. 342–352). Cham: Springer. https://doi.org/10.1007/978-3-030-20497-6_32

Mayer, R. C., Davis, J. H. & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. Academy of Management. https://doi.org/10.2307/258792

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A. et al. (2019). Automation-Induced Complacency Potential: Development and Validation of a New Scale. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00225

Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J. et al. (2023). Trust toward humans and trust toward artificial intelligence are not associated: Initial insights from self-report and neurostructural brain imaging. *Personality Neuroscience*, *6*, e3. https://doi.org/10.1017/pen.2022.5

Montag, C., Kraus, J., Baumann, M. & Rozgonjuk, D. (2023). The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. *Computers in Human Behavior Reports*, *11*, 100315. https://doi.org/10.1016/j.chbr.2023.100315

OECD. (2019). *Artificial Intelligence in Society*. Paris: Organisation for Economic Co-operation and Development. https://doi.org/10.1787/eedfee77-en

Okamura, K. & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLOS ONE*, *15*(2), e0229132. https://doi.org/10.1371/journal.pone.0229132

OpenAI. (2024a). *About*. Verfügbar unter: https://openai.com/about

OpenAI. (2024b). *ChatGPT*. Verfügbar unter: https://openai.com/chatgpt/

Parasuraman, R., Sheridan, T. B. & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

R Core Team (2022) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna.

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

Razin, Y. S. & Feigh, K. M. (2023). *Converging Measures and an Emergent Model: A Meta-Analysis of Human-Automation Trust Questionnaires*. arXiv:2303.13799 [cs.HC]. https://doi.org/10.48550/arXiv.2303.13799

Reason, J. (1990). *Human Error.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139062367

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rousseau, D. M., Sitkin, S. B., Burt, R. S. & Camerer, C. (1998). Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review*, *23*(3), 393–404. https://doi.org/10.5465/amr.1998.926617

Schepman, A. & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human-Computer Interaction*, *39*(13), 2724–2741. https://doi.org/10.1080/10447318.2022.2085400

Scheurer, J., Balesni, M. & Hobbhahn, M. (2023, 27. November). *Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure.* arXiv:2311.07590 [cs.CL]. https://doi.org/10.48550/arXiv.2311.07590

Schmider, E., Ziegler, M., Danay, E., Beyer, L. & Bühner, M. (2010). Is It Really Robust? *Methodology*, *6*(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Sheridan, T. B. & Hennessy, R. T. (1984). *Research and Modeling of Supervisory Control Behavior. Report of a Workshop*. Washington, D.C.: National Academy.

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

Siau, K. & Wang, W. (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal*, *31*(2), 47–53. Verfügbar unter: https://www.researchgate.net/profile/Keng-Siau-2/publication/324006061_Building_Trust_in_Artificial_Intelligence_Machine_Learning_and_Robotics/links/5ab8744baca2722b97cf9d33/Building-Trust-in-Artificial-Intelligence-Machine-Learning-and-Robotics.pdf

Siegrist, M. (2021). Trust and Risk Perception: A Critical Review of the Literature. *Risk Analysis*, *41*(3), 480–490. https://doi.org/10.1111/risa.13325

Sindermann, C., Yang, H., Elhai, J. D., Yang, S., Quan, L., Li, M. et al. (2022). Acceptance and Fear of Artificial Intelligence: associations with personality in a German and a Chinese sample. *Discover Psychology*, *2*(1), 8. https://doi.org/10.1007/s44202-022-00020-y

Solberg, E., Kaarstad, M., Eitrheim, M. H. R., Bisio, R., Reegård, K. & Bloch, M. (2022). A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids. *Group & Organization Management*, *47*(2), 187–222. https://doi.org/10.1177/10596011221081238

Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing*, *7*(1), 35. https://doi.org/10.3390/bdcc7010035

Tschopp, M. & Ruef, M. (2018). On Trust in AI – A Systemic Approach. Verfügbar unter: https://www.researchgate.net/publication/336850230_On_Trust_in_AI_-_A_Systemic_Approach

Tschopp, M. & Ruef, M. (2020). AI & Trust – Stop asking how to increase trust in AI. Verfügbar unter: https://www.researchgate.net/publication/339530999_AI_Trust_-Stop_asking_how_to_increase_trust_in_AI

Utts, J. M. (1982). The rainbow test for lack of fit in regression. *Communications in Statistics – Theory and Methods*, *11*(24), 2801–2815. https://doi.org/10.1080/03610928208828423

Yampolskiy, R. V. & Spellchecker, M. S. (2016). *Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures*. arXiv:1610.07997 [cs.AI]. https://doi.org/10.48550/arXiv.1610.07997

Zweig, K. A. (2023). *Die KI war's!: von absurd bis tödlich: Die Tücken der künstlichen Intelligenz* (Originalausgabe.). München: Heyne.

Anhang A

Fragebogen Items der Umfrage

Tabelle A1
Items zum Vertrauen nach Körber (2019)

Variable	Item	Originalwortlaut	Adaptierter Wortlaut
Performance	1	Das System ist imstande, Situationen richtig einzuschätzen.	ChatGPT ist imstande, Prompts richtig einzuschätzen.
	7	Das System arbeitet zuverlässig.	ChatGPT arbeitet zuverlässig.
	12	*Ein Ausfall des Systems ist wahrscheinlich.*	*Ein Ausfall von ChatGPT ist wahrscheinlich.*
	16	Das System kann wirklich komplizierte Aufgaben übernehmen.	ChatGPT kann wirklich komplizierte Aufgaben übernehmen.
	18	*Das System könnte stellenweise einen Fehler machen.*	*ChatGPT könnte stellenweise einen Fehler machen.*
	22	Ich bin überzeugt von den Fähigkeiten des Systems.	Ich bin überzeugt von den Fähigkeiten von ChatGPT.
Process	2	Mir war durchgehend klar, in welchem Zustand sich das System befindet.	Mir war durchgehend klar, in welchem Zustand sich ChatGPT befindet.
	8	*Das System reagiert unvorhersehbar.*	*ChatGPT reagiert unvorhersehbar.*
	13	Ich konnte nachvollziehen, warum etwas passiert ist.	Ich konnte nachvollziehen, warum etwas in ChatGPT passiert ist.
	19	*Zu erkennen, was das System als Nächstes macht, ist schwer.*	*Zu erkennen, was ChatGPT als Nächstes macht, ist schwer.*

Purpose	4	Die Entwickler sind vertrauenswürdig.	OpenAI (Entwickler von ChatGPT) ist vertrauenswürdig.
	9	Die Entwickler nehmen mein Wohlergehen ernst.	OpenAI (Entwickler von ChatGPT) nimmt mein Wohlergehen ernst.
Propensity to Trust	5	*Bei unbekannten automatisierten Systemen sollte man eher vorsichtig sein.*	*Bei unbekannten KI-Systemen sollte man eher vorsichtig sein.*
	14	Ich vertraue einem System eher, als dass ich ihm misstraue.	Ich vertraue einem KI-Systeme eher, als dass ich ihm misstraue.
	21	Automatisierte Systeme funktionieren generell gut.	KI-Systeme funktionieren generell gut.
Trust in AI	10	Ich vertraue dem System.	Ich vertraue ChatGPT.
	17	Ich kann mich auf das System verlassen.	Ich kann mich auf ChatGPT verlassen.
Previous Experience	3	Ich kenne bereits ähnliche Systeme.	Ich kenne bereits ähnliche Systeme wie ChatGPT.
	20	Ich habe ähnliche Systeme bereits genutzt.	Ich habe ähnliche Systeme wie ChatGPT bereits genutzt.

Anmerkungen. Negativ formulierte Items in kursiv.

Tabelle A2
Items zur Nutzungsintention nach Choung et al. (2023)

Variable	Item	Originalwortlaut	Adaptierter Wortlaut
Intention to Use	6	I intend to use [AI smart technologies] in a future	Ich beabsichtige, ChatGPT in Zukunft zu nutzen.
	11	I predict that I would use [AI smart technologies]	Ich gehe davon aus, dass ich ChatGPT nutzen werde.
	15	Using [AI smart technologies] is something I would do in a future.	Die Verwendung von ChatGPT ist etwas, das ich in Zukunft tun werde.

Tabelle A3
Items zu den Big-Five Persönlichkeitsmerkmalen nach Körner et al. (2008)

Variable	Item	Originalwortlaut	Adaptierter Wortlaut
Neurotizismus	23	Wenn ich unter starkem Stress stehe, fühle ich mich manchmal, als ob ich zusammenbräche.	Wenn ich unter starkem Stress stehe, fühle ich mich manchmal, als ob ich zusammenbräche.
	28	Manchmal fühle ich mich völlig wertlos.	Manchmal fühle ich mich völlig wertlos.
	33	Zu häufig bin ich entmutigt und will aufgeben, wenn etwas schief geht.	Zu häufig bin ich entmutigt und will aufgeben, wenn etwas schief geht.
Extraversion	24	Ich habe gern viele Leute um mich herum.	Ich habe gern viele Leute um mich herum.
	29	Ich bin leicht zum Lachen zu bringen.	Ich bin leicht zum Lachen zu bringen.
	34	Ich bin ein fröhlicher, gutgelaunter Mensch.	Ich bin ein fröhlicher, gutgelaunter Mensch.
Offenheit für Erfahrung	25	*Ich finde philosophische Diskussionen langweilig.*	*Ich finde philosophische Diskussionen langweilig.*
	30	*Poesie beeindruckt mich wenig oder gar nicht.*	*Poesie beeindruckt mich wenig oder gar nicht.*
	35	Ich habe oft Spaß daran, mit Theorien oder abstrakten Ideen zu spielen.	Ich habe oft Spaß daran, mit Theorien oder abstrakten Ideen zu spielen.
Verträglichkeit	26	Manche Leute halten mich für selbstsüchtig und selbstgefällig.	Manche Leute halten mich für selbstsüchtig und selbstgefällig.
	31	Im Hinblick auf die Absichten anderer bin ich eher zynisch und skeptisch.	Im Hinblick auf die Absichten anderer bin ich eher zynisch und skeptisch.
	36	Manche Leute halten mich für kalt und berechnend.	Manche Leute halten mich für kalt und berechnend.

Gewissen-haftigkeit	27	Ich versuche, alle mir übertragenen Aufgaben sehr gewissenhaft zu erledigen.	Ich versuche, alle mir übertragenen Aufgaben sehr gewissenhaft zu erledigen.
	32	Wenn ich eine Verpflichtung eingehe, so kann man sich auf mich bestimmt verlassen.	Wenn ich eine Verpflichtung eingehe, so kann man sich auf mich bestimmt verlassen.
	37	Ich bin eine tüchtige Person, die ihre Arbeit immer erledigt.	Ich bin eine tüchtige Person, die ihre Arbeit immer erledigt.

Anmerkungen. Negativ formulierte Items in kursiv.

Anhang B

PowerPoint-Präsentation der Intervention

Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Exkurs

DIGITAL TRUST

Vertrauen in KI

Untersuchung im Kontext dieser Vorlesung für meine Masterarbeit an der Hochschule für Angewandte Psychologie in Zusammenarbeit mit dem Kompetenzzentrum Digital Trust (HSW IWI).

28.02. – 1. Datenerhebung
20.03. – Exkurs: ChatGPT
03.04. – 2. Datenerhebung

APS & Kompetenzzentrum Digital Trust 20.03.24 1

Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Exkurs

DIGITAL TRUST

Künstliche Intelligenz (KI)

- Technologien, die computergestützt das menschliche Denken sowie Entscheidungs- und Problemlösungsverhalten reproduzieren (Bendel, 2022)
- Selbstlernende Systeme, welche grosse Datenmengen selbständig verarbeiten (Bendel, 2022)
- Anwendungsfelder von KI (Bendel, 2018; Bendel, 2019)
  - soziale Roboter
  - Suchmaschinen
  - Chatbots, wie beispielsweise ChatGPT (Taecharungroj, 2023)
    - textbasierte Konversationen mit technischen Systemen
    - anhand Large Language Models (LLM)
    - nutzen Natural Language Processing (NLP) um menschliche Sprache zu verstehen

APS & Kompetenzzentrum Digital Trust 20.03.24 2

Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Exkurs

DIGITAL TRUST

OpenAI

- Non-Profit-Organisation, mit gewinnorientierter Tochtergesellschaft
  - grösster Investor: Microsoft
- KI-Forschung und -Entwicklung seit Ende 2015
  - u.a. ChatGPT (**G**enerative **P**re-trained **T**ransformer)
- Fernziel: Entwicklung einer der Menschheit zugutekommenden Artificial General Intelligence (AGI)

«Our mission is to ensure that artificial general intelligence – AI systems that are generally smarter than humans – benefits all of humanity.»

(OpenAI, 2024a)

Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Exkurs

DIGITAL TRUST

… einer der Menschheit zugutekommenden KI …

«Artificial general intelligence has the potential to benefit nearly every aspect of our lives – so it must be developed and deployed responsibly.»

(OpenAI, 2024b)

- daher ist ChatGPT darauf trainiert, hilfreich, unschädlich und ehrlich zu sein (Scheurer, Balesni & Hobbhahn, 2023)
- **aber** bietet dennoch potenziell schädliche Einsatzmöglichkeiten (Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal et al., 2020)
  - Desinformation, Spam, Phishing, Missbrauch von Rechtsverfahren und Regierungsvorgängen, betrügerisches Verfassen von akademischen Arbeiten und Social Engineering
  - Diskriminierung aufgrund von Biases von Geschlecht, Ethnie, Religion und weitere
- zudem kann ChatGPT Fehlverhalten zeigen und dieses verschleiern, ohne dazu aufgefordert zu werden
  - Bspw. Insiderhandel (Scheurer, Balesni & Hobbhahn, 2023)

Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Exkurs

DIGITAL TRUST

Algorithmus, oder doch nicht

Algorithmus

- die **korrekte Lösung** für ein Problem
- Bspw. Navigationsgeräte

Heuristik

- statistische Wahrscheinlichkeit
- eine oftmals **gute Lösung** für ein Problem
- Bspw. Large Language Models (LLM)

(Zweig, 2023)

APS & Kompetenzzentrum Digital Trust 20.03.24 5

Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Exkurs

DIGITAL TRUST

Was kann ChatGPT, was nicht?

Fähigkeiten

- Kreatives Schreiben (Taecharungroj, 2023)
- Essays (Sekundarstufe) (Taecharungroj, 2023)
- Promptes (Taecharungroj, 2023)
- Programmieren (Taecharungroj, 2023)
- Übersetzen (Brown et al., 2020)
- Lückentextaufgaben (Brown et al., 2020)
- Fragebeantwortung (Brown et al., 2020)

Aber

- kein Anspruch auf Richtigkeit, da LLM (heute) auf Heuristiken basieren müssen (Zweig, 2023)
  - denn trotz Big-Date sind nicht alle Informationen vorhanden
  - es handelt sich um intelligentes Raten anhand eines statistischen Korrelationsmodells
    - Korrelation ≠ Kausalität
    - Raten > 46 % (Brown et al., 2020)
    - GPT-3 ⌀ > 60 % (Brown et al., 2020)

APS & Kompetenzzentrum Digital Trust 20.03.24 6

Quellen

Bendel, O. (Hrsg.). (2018). *Pflegeroboter* (1. Aufl.). Springer Nature. https://doi.org/10.1007/978-3-658-22698-5

Bendel, O. (Hrsg.). (2019) *Handbuch Maschinenethik*. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-17483-5

Bendel, O. (2022). *450 Keywords Digitalisierung* (2. Aufl.). Springer Gabler, Wiesbaden. https://doi.org/10.1007/978-3-658-37492-1

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. et al. (2020). Language models are few-shot learners (NIPS'20). *Proceedings of the 34th International Conference on Neural Information Processing Systems* (S. 1877–1901). Red Hook, NY, USA: Curran Associates Inc.

OpenAI. (2024a). *About*. Verfügbar unter: https://openai.com/about [Zugriff: 08.03.2024]

OpenAI. (2024b). *Safety*. Verfügbar unter: https://openai.com/safety [Zugriff: 08.03.2024]

Scheurer, J., Balesni, M. & Hobbhahn, M. (2023). Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure. https://doi.org/10.48550/arXiv.2311.07590

Taecharungroj, V. (2023). "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing*, 7(1), 35. https://doi.org/10.3390/bdcc7010035

Zweig, K. A. (2023). *Die KI war's!: von absurd bis tödlich: Die Tücken der künstlichen Intelligenz* (Originalausgabe.). München: Heyne.

Anhang C

Prüfung der Voraussetzungen der Pfadanalyse

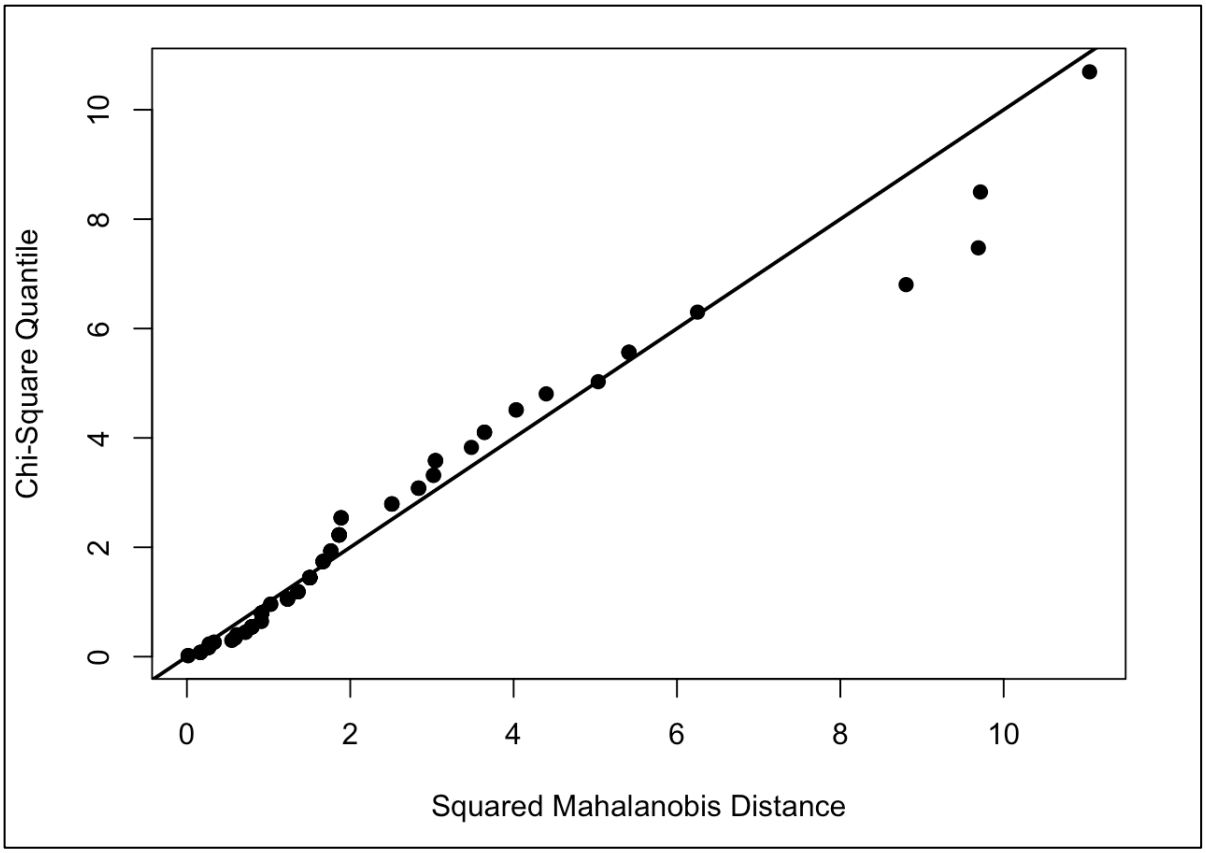


Abbildung C1. QQ-Plot der multivariaten Normalverteilung (eigene Abbildung)

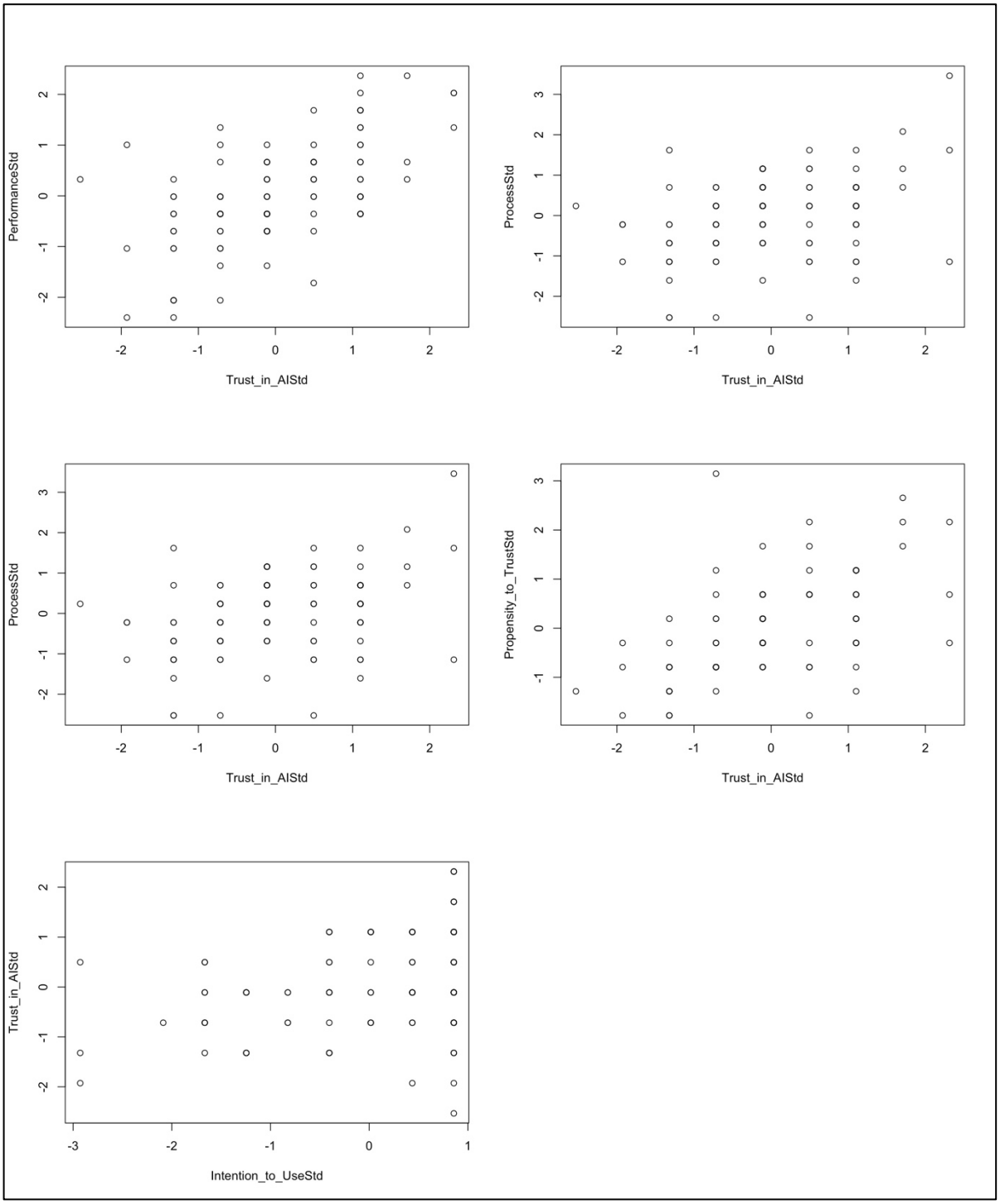


Abbildung C2. Streudiagramme zur Prüfung der Linearität (eigene Abbildung)

Anhang D

Voraussetzungsprüfung der Regressionsanalyse der Vorerfahrung

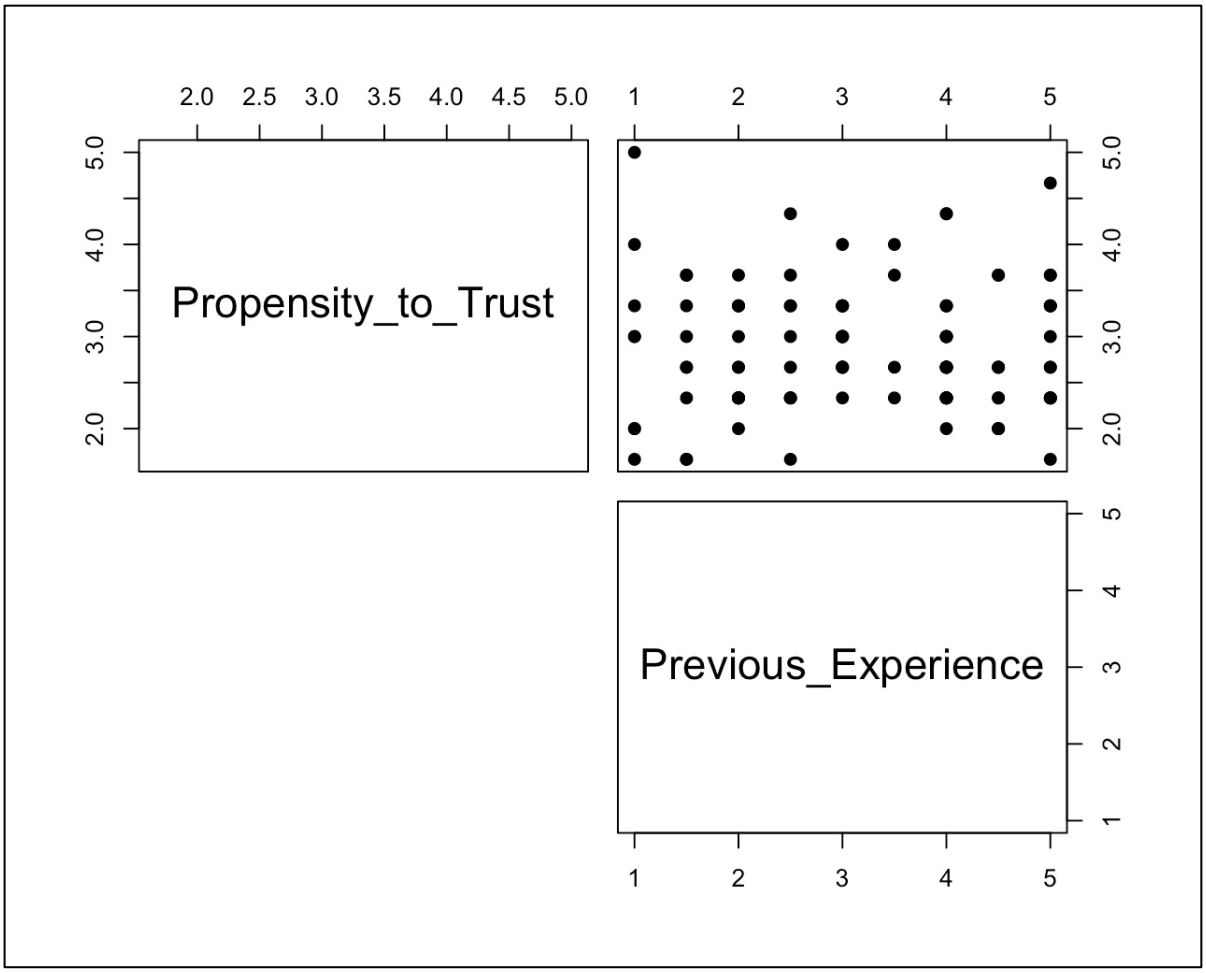


Abbildung D1. Streudiagramme zur Prüfung der Linearität (eigene Abbildung)

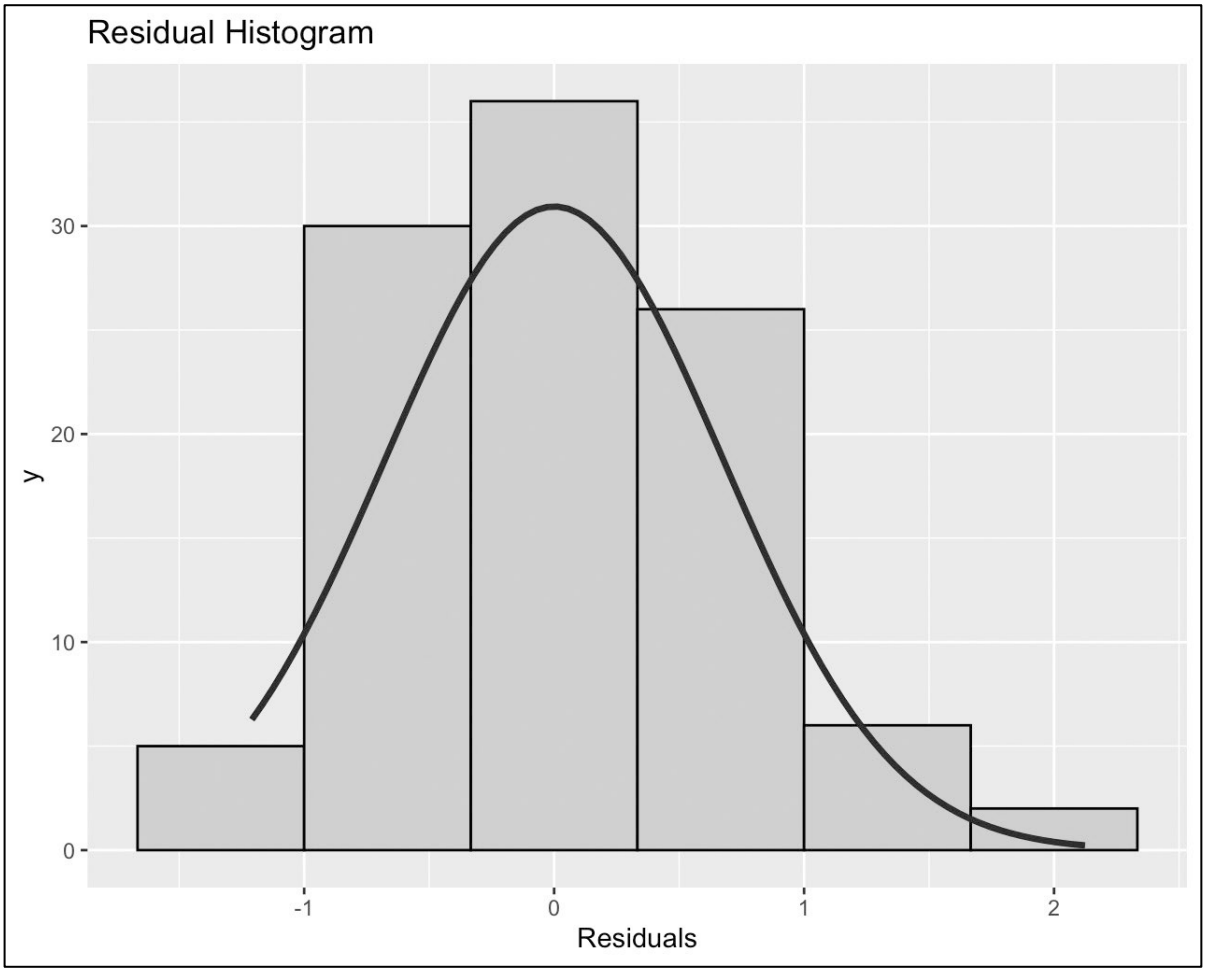


Abbildung D2. Histogramm der Verteilung der Fehlerwerte (eigene Abbildung)

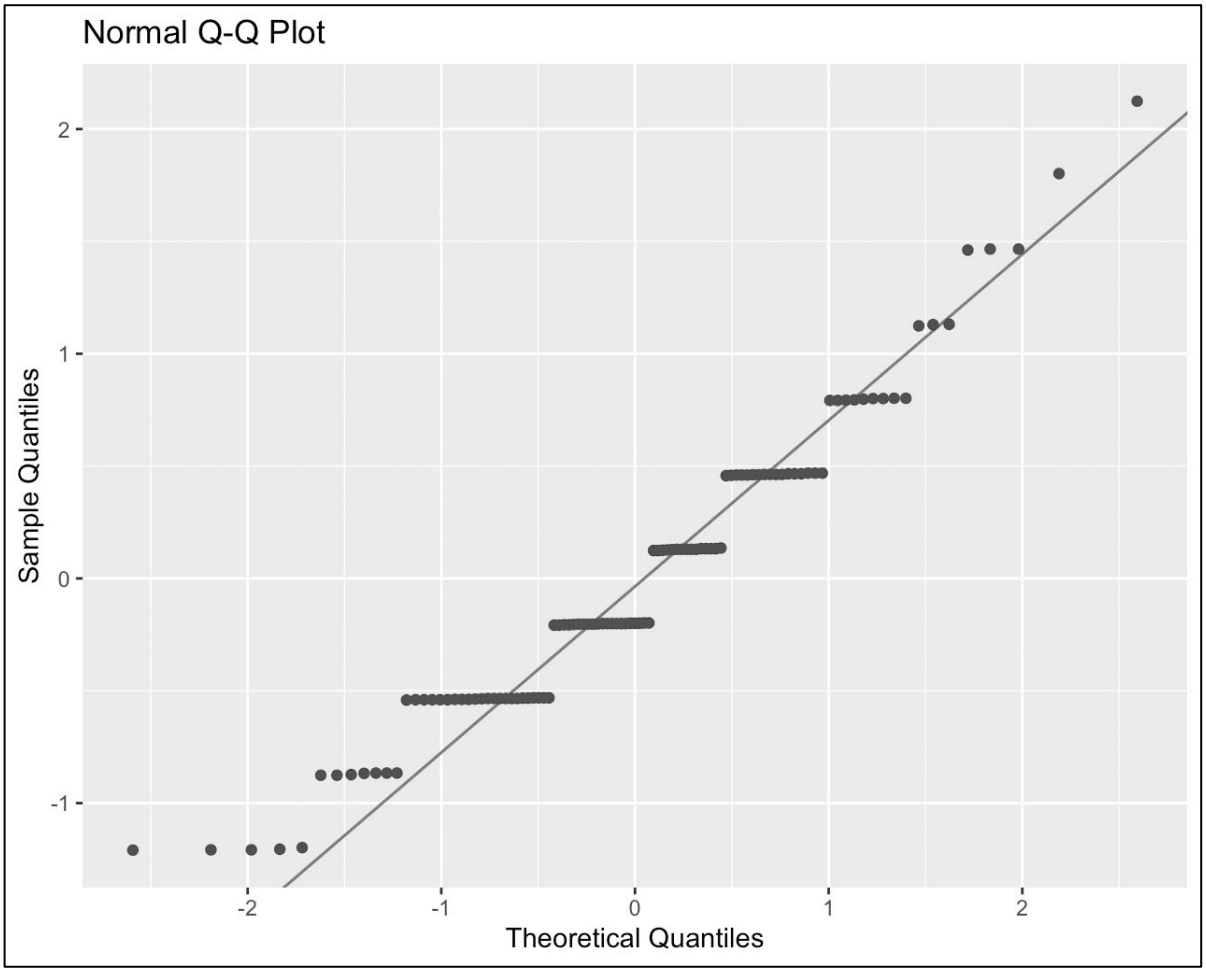


Abbildung D3. QQ-Plot der Verteilung der Fehlerwerte (eigene Abbildung)

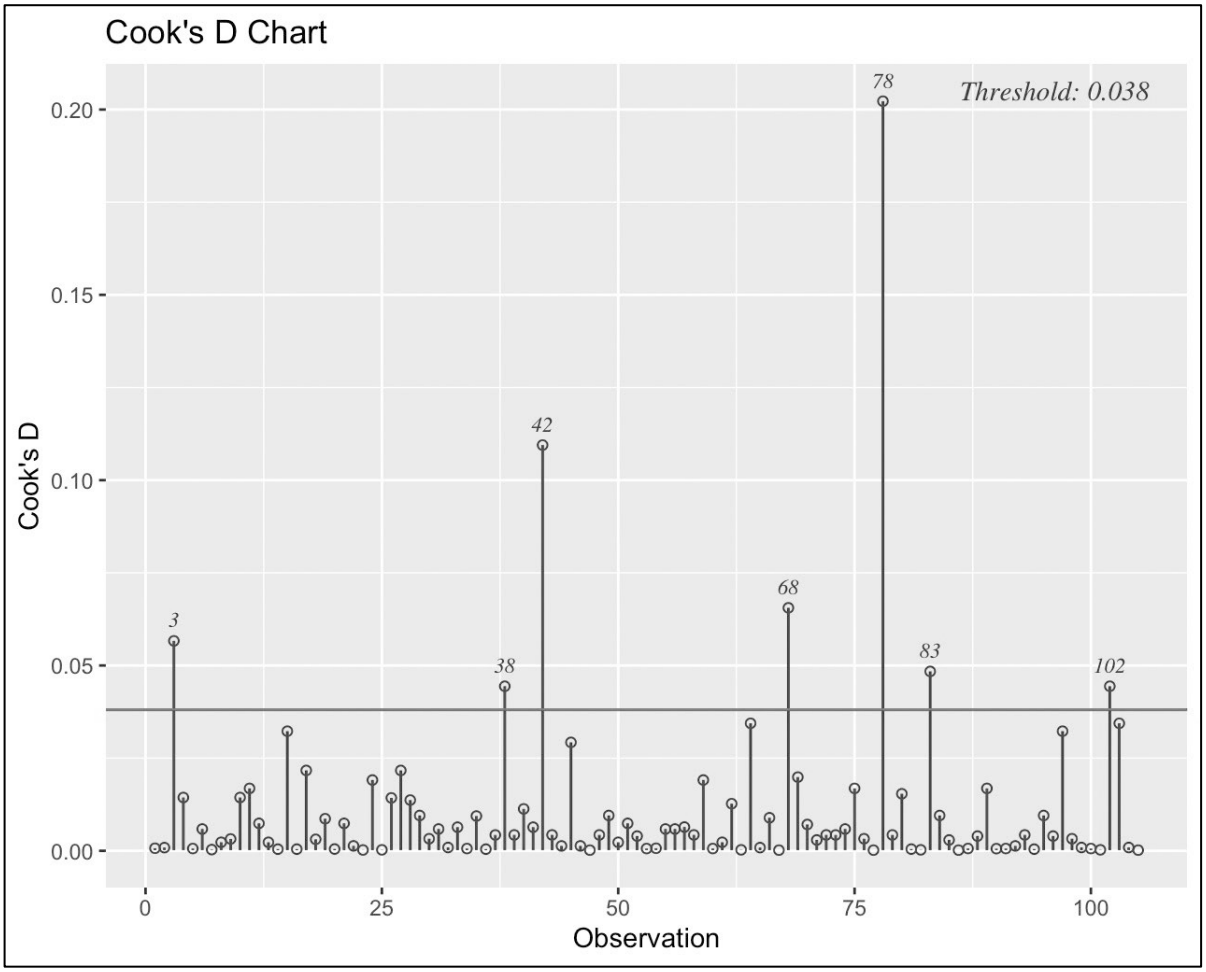


Abbildung D4. Ausreisserdiagnostik mittels Cooks-Distance (eigene Abbildung)

Anhang E

Voraussetzungsprüfung der Regressionsanalyse des Alters

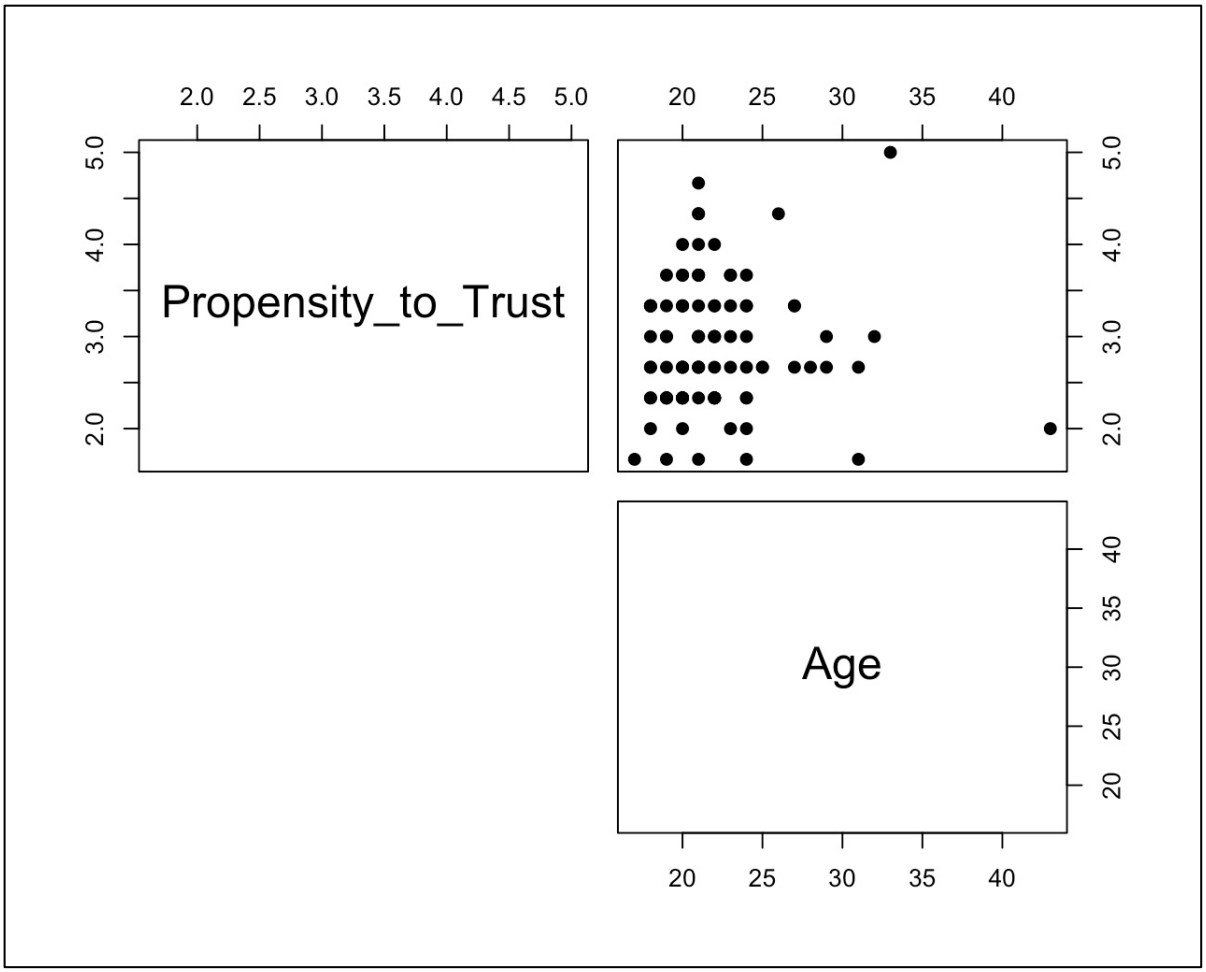


Abbildung E1. Streudiagramme zur Prüfung der Linearität (eigene Abbildung)

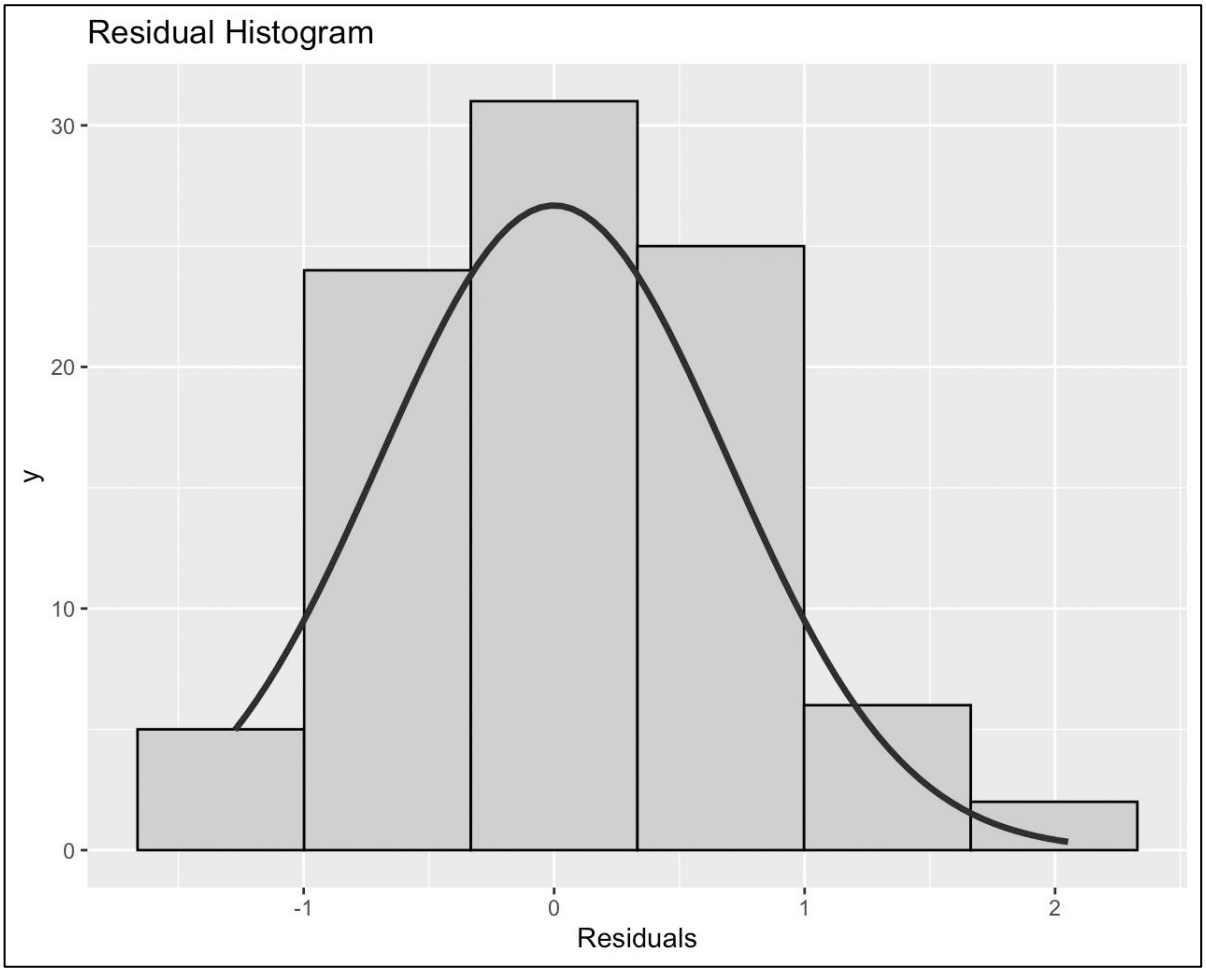


Abbildung E2. Histogramm der Verteilung der Fehlerwerte (eigene Abbildung)

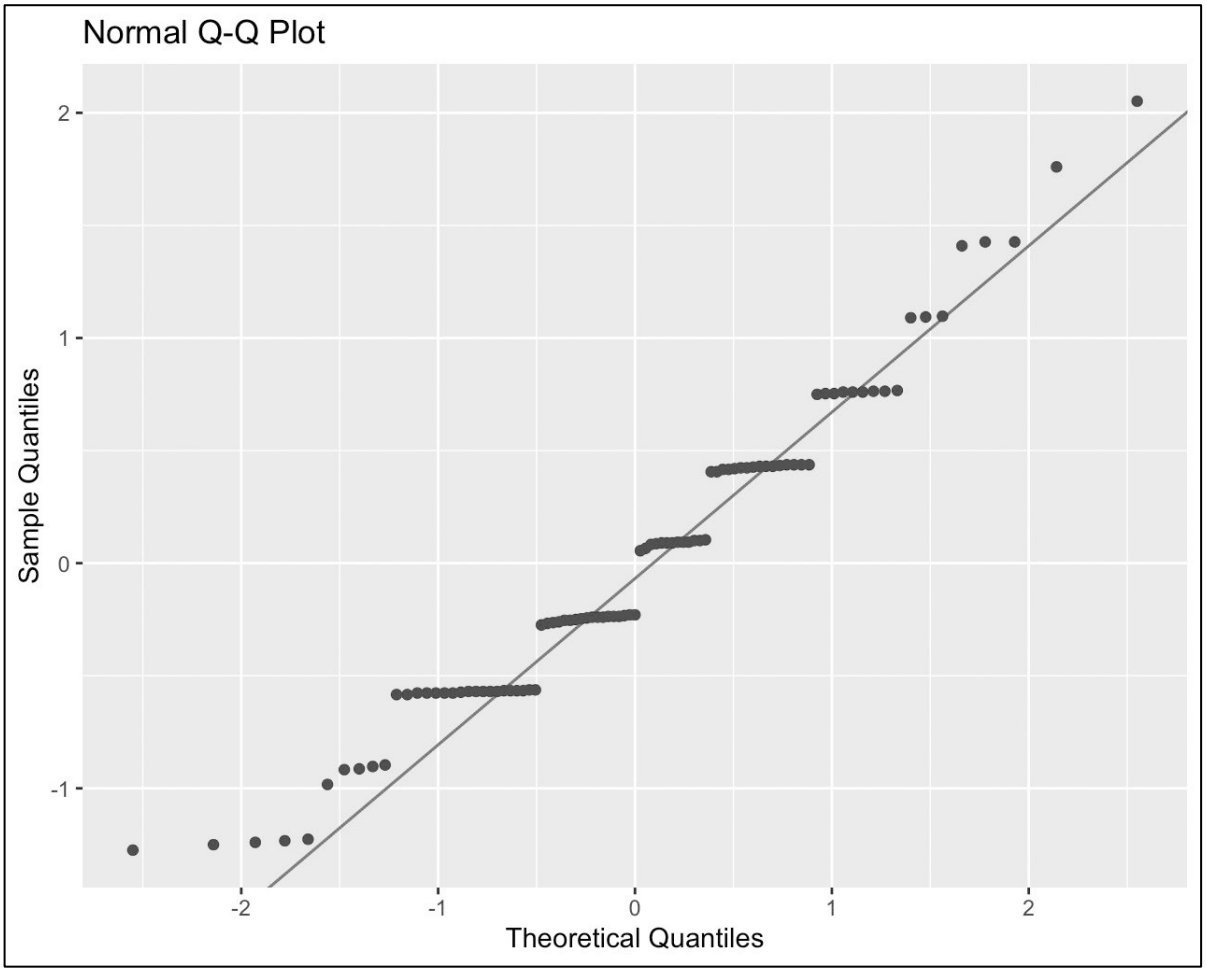


Abbildung E3. QQ-Plot der Verteilung der Fehlerwerte (eigene Abbildung)

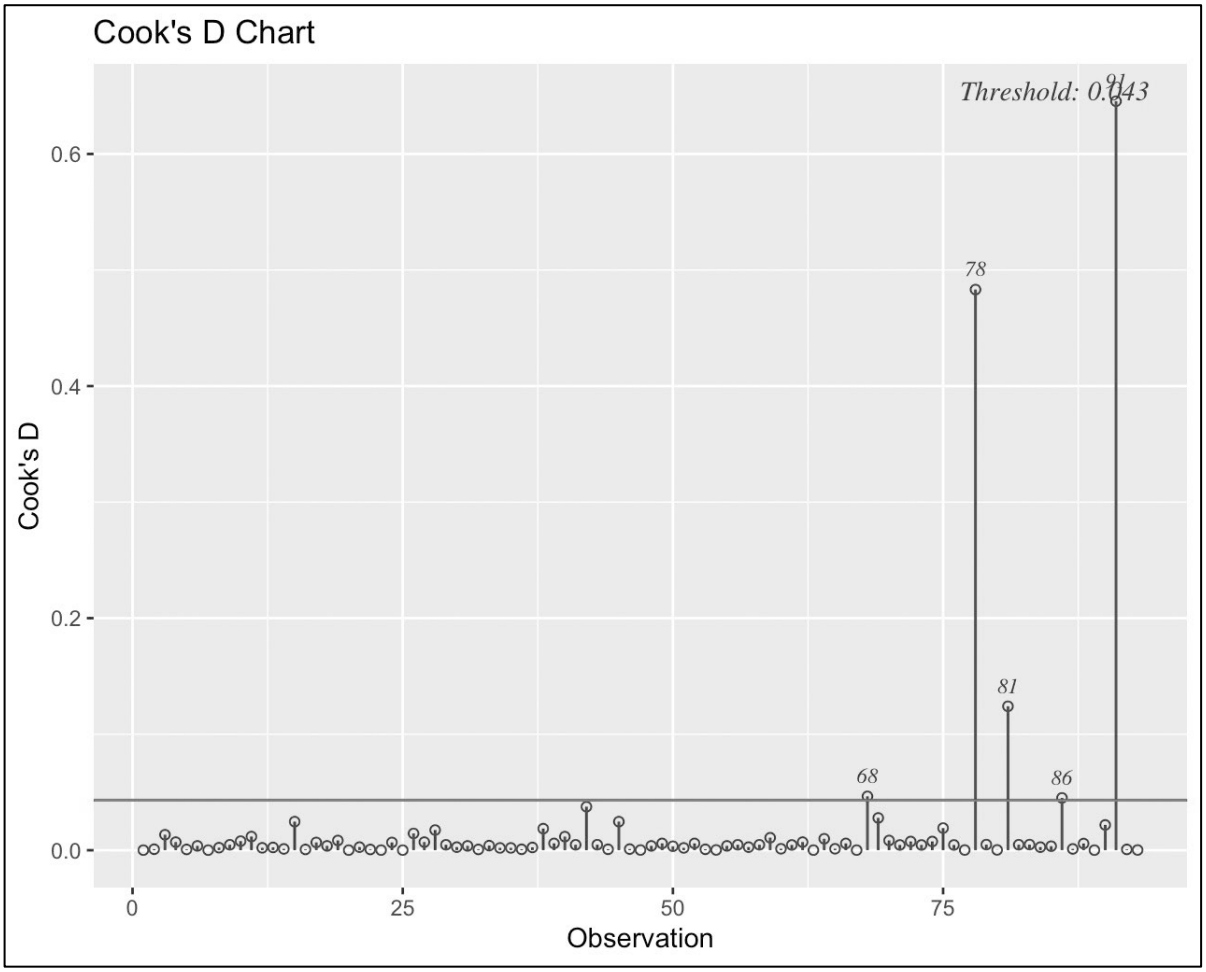


Abbildung E4. Ausreisserdiagnostik mittels Cooks-Distance (eigene Abbildung)

Anhang F

Voraussetzungsprüfung der multiplen Regressionsanalyse Persönlichkeitsmerkmale

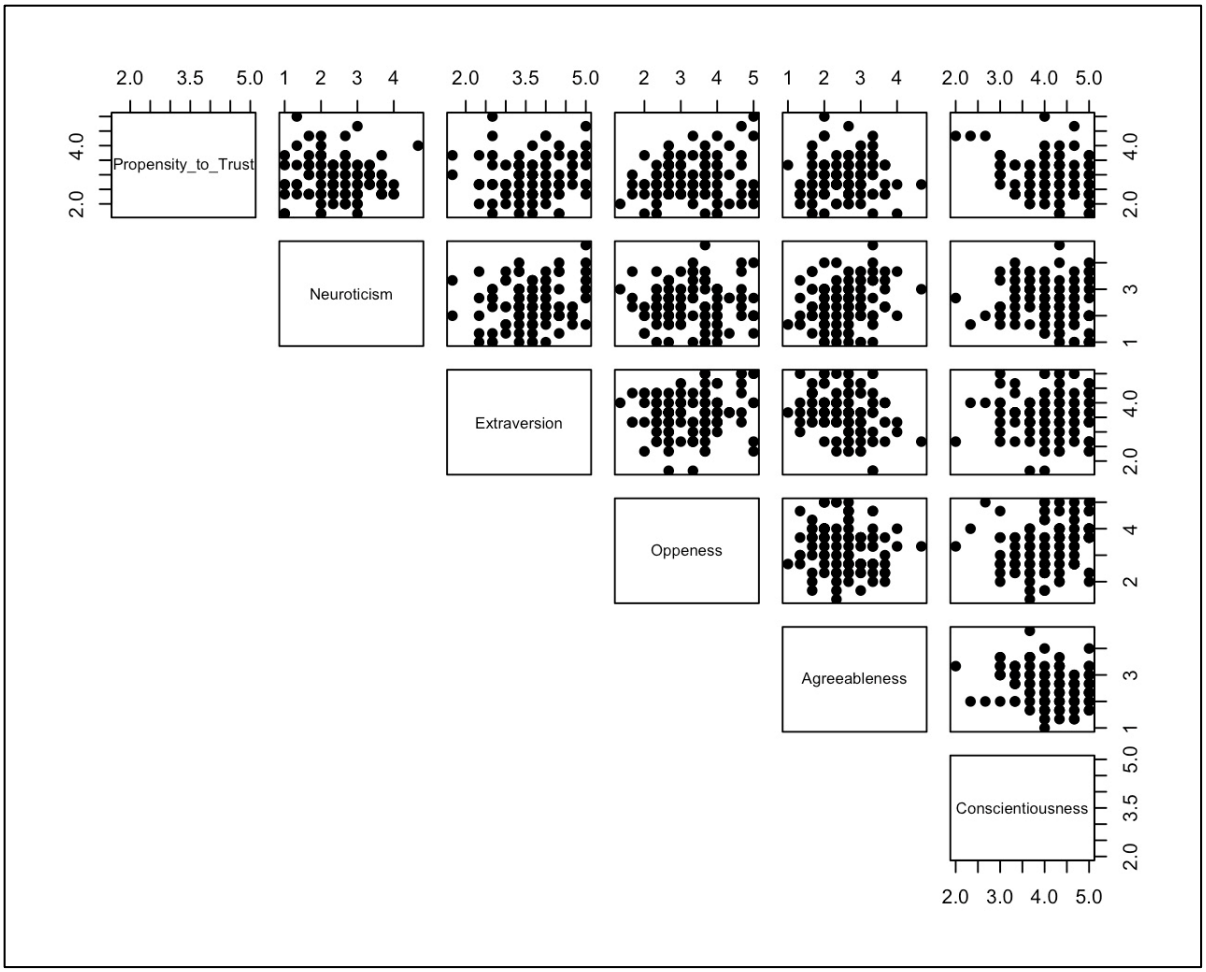


Abbildung F1. Streudiagramme zur Prüfung der Linearität *(eigene Abbildung)*

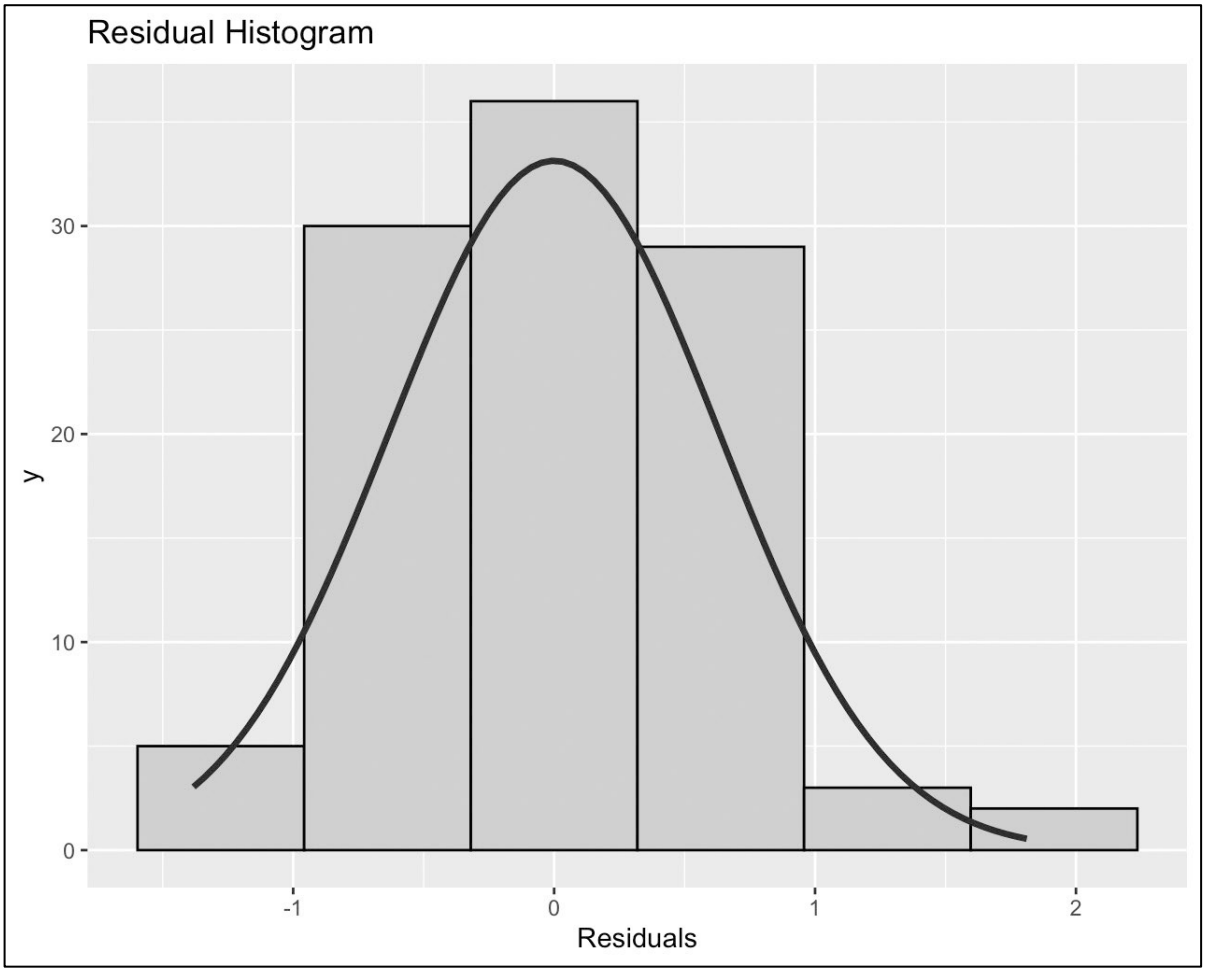


Abbildung F2. Histogramm der Verteilung der Fehlerwerte (eigene Abbildung)

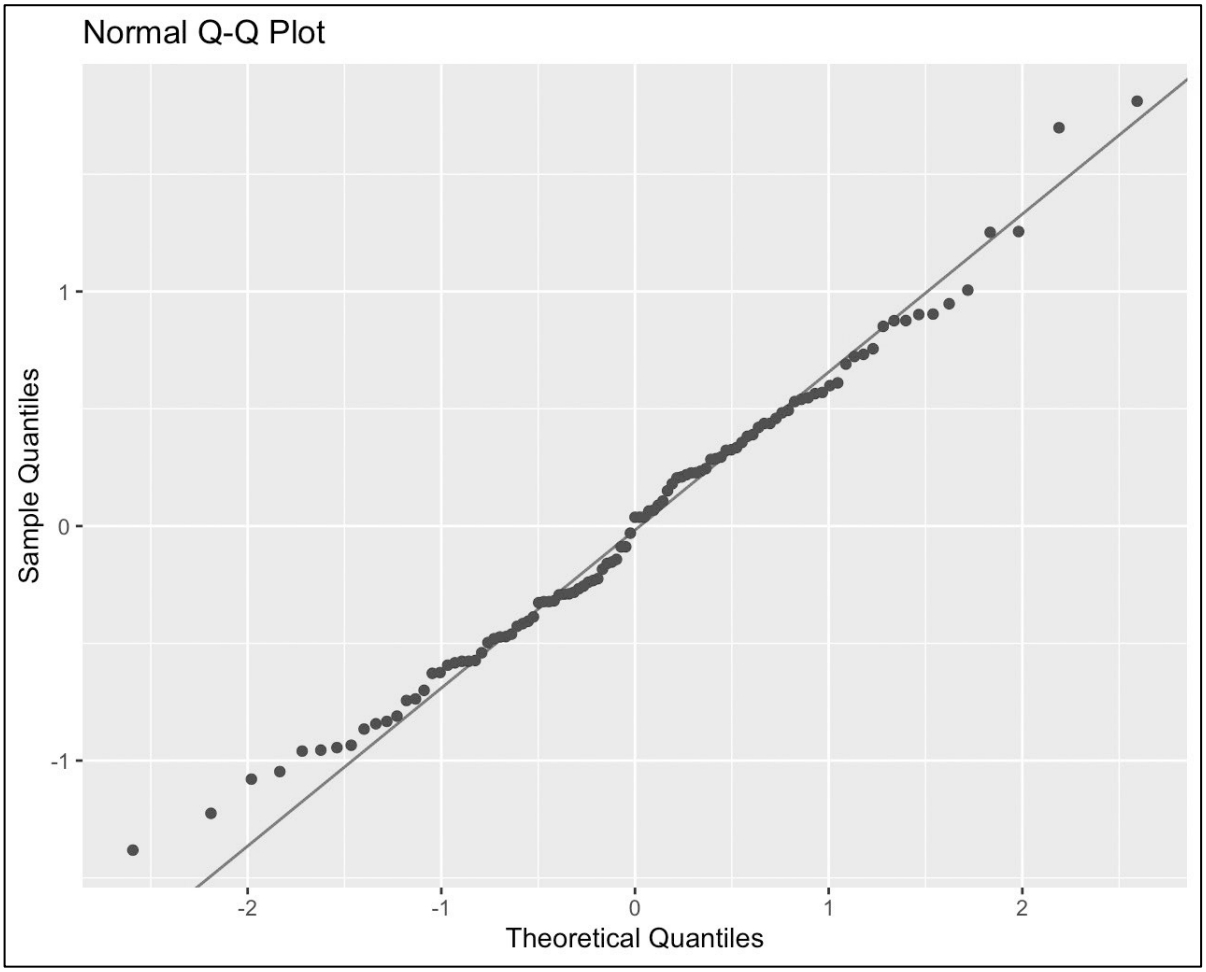


Abbildung F3. QQ-Plot der Verteilung der Fehlerwerte (eigene Abbildung)

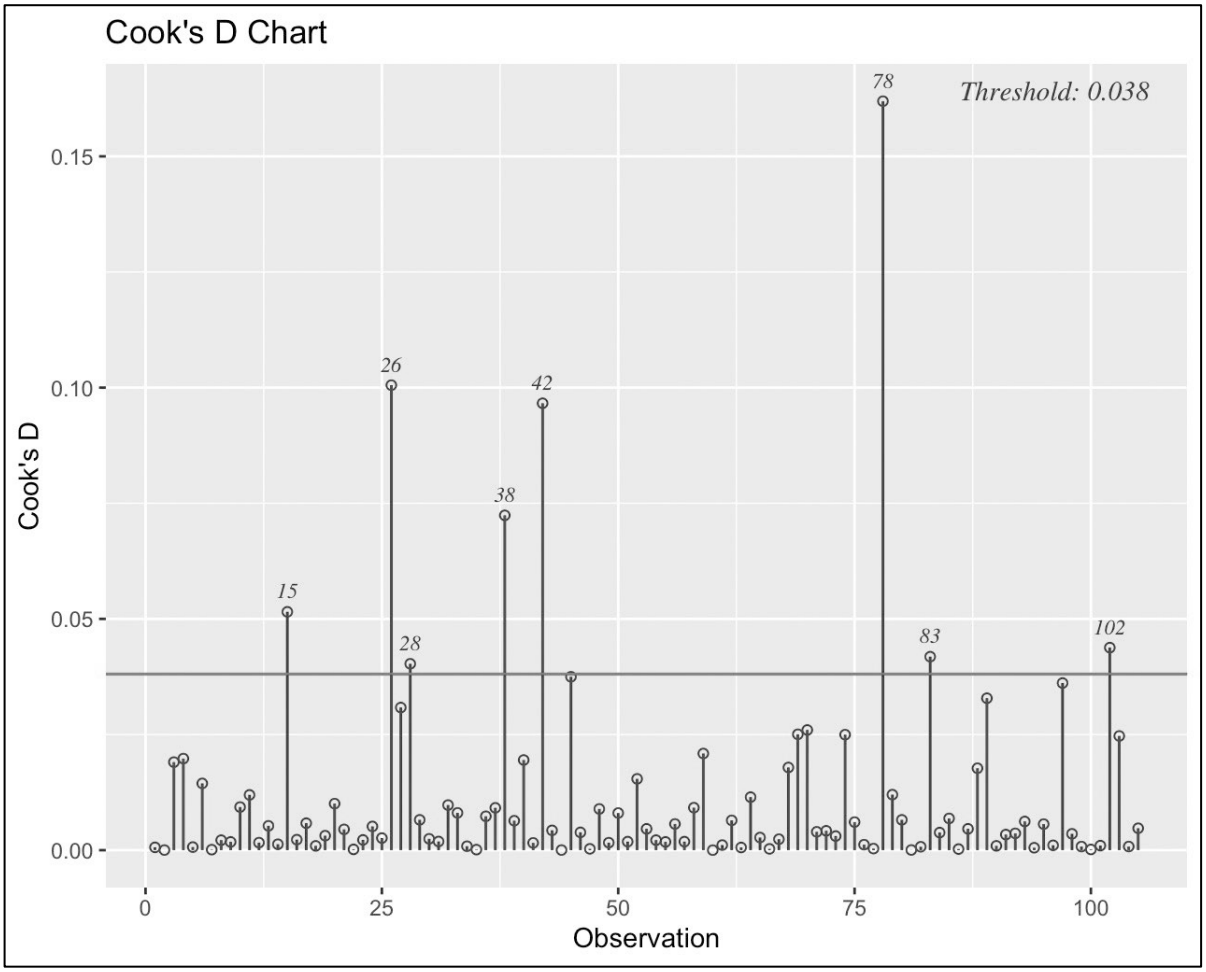


Abbildung F4. Ausreisserdiagnostik mittels Cooks-Distance (eigene Abbildung)

Anhang G

Voraussetzungsprüfung der ANOVA der Geschlechtsunterschiede

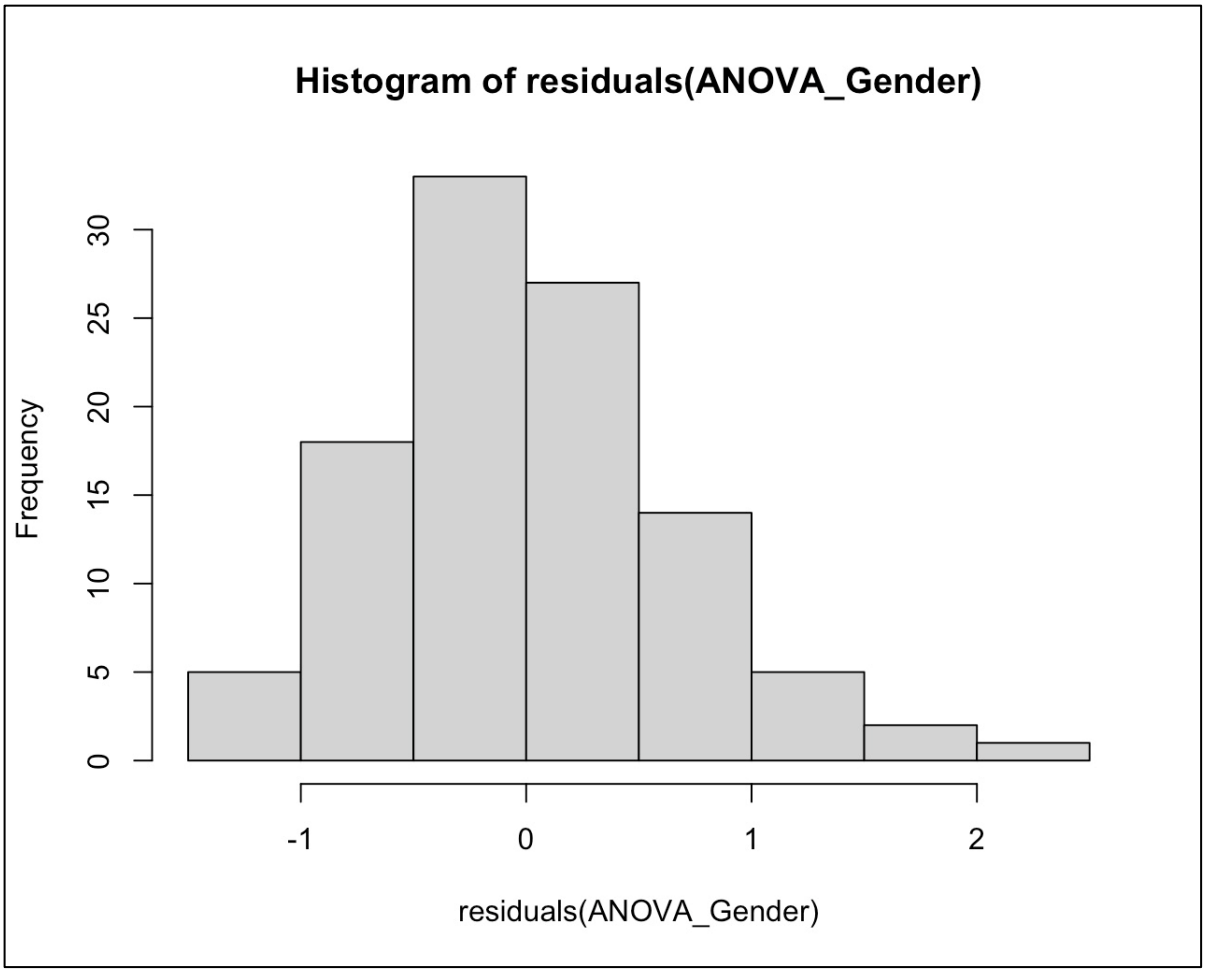


Abbildung G1. Histogramm der Verteilung der Fehlerwerte (eigene Abbildung)

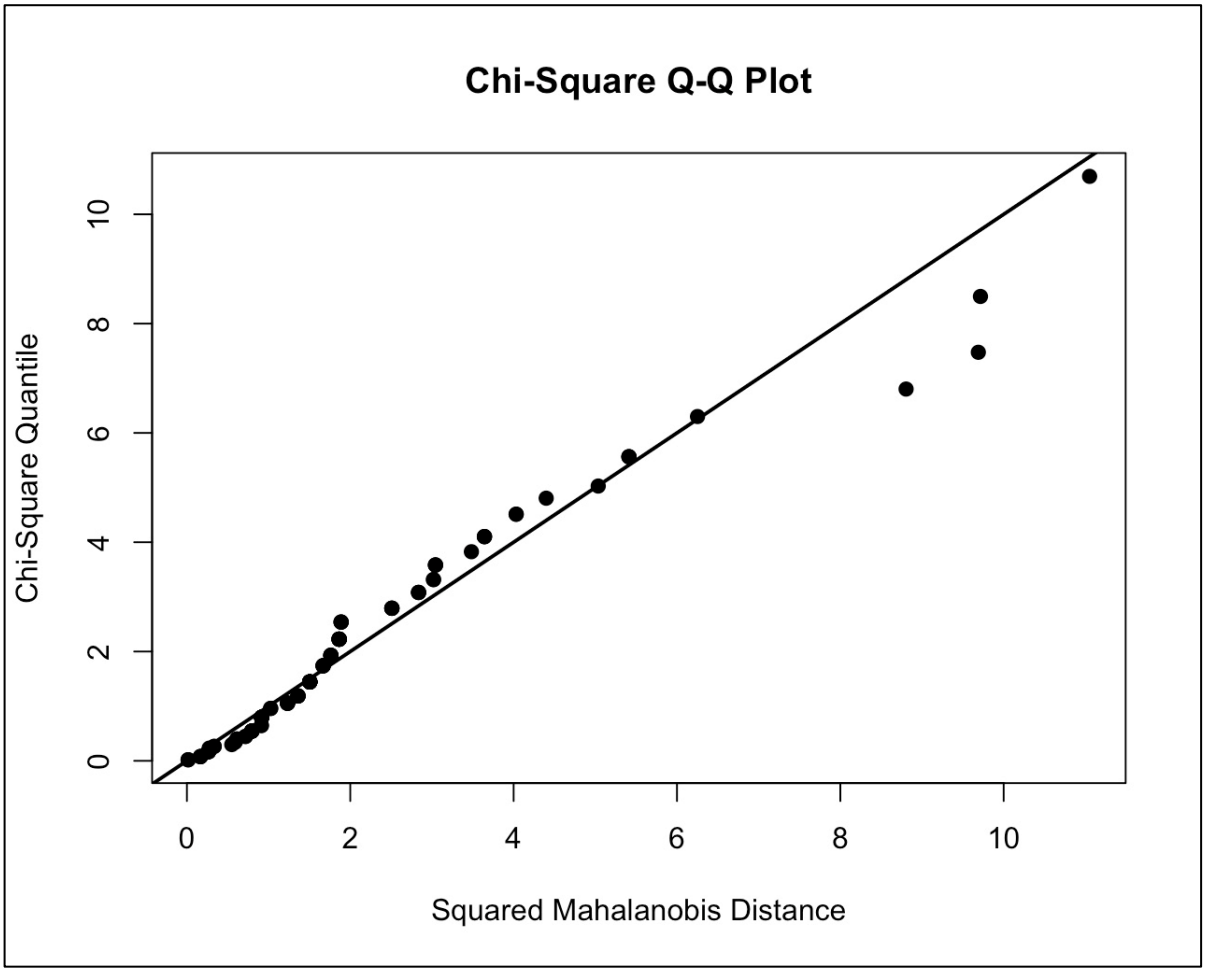


Abbildung G2. QQ-Plot der Verteilung der Fehlerwerte (eigene Abbildung)

Anhang H

Voraussetzungsprüfung der MANOVA der Längsschnittanalyse

Tabelle H1
Shapiro-Wilk-Test

Messzeitpunkt	Variable	Kenngrösse	*p*
Post	Performance	.920	.354
	Process	.944	.598
	Purpose	.868	.095
	Propensity to Trust	.890	.172
	Trust in AI	.929	.440
Pre	Performance	.864	.086
	Process	.905	.247
	Purpose	.906	.258
	Propensity to Trust	.903	.238
	Trust in AI	.873	.108

Tabelle H2
Levene-Test

Variable	*F*	*p*
Performance	.224	.642
Process	.126	.727
Purpose	0.0	1.0
Propensity to Trust	.890	.445
Trust in AI	.224	.642

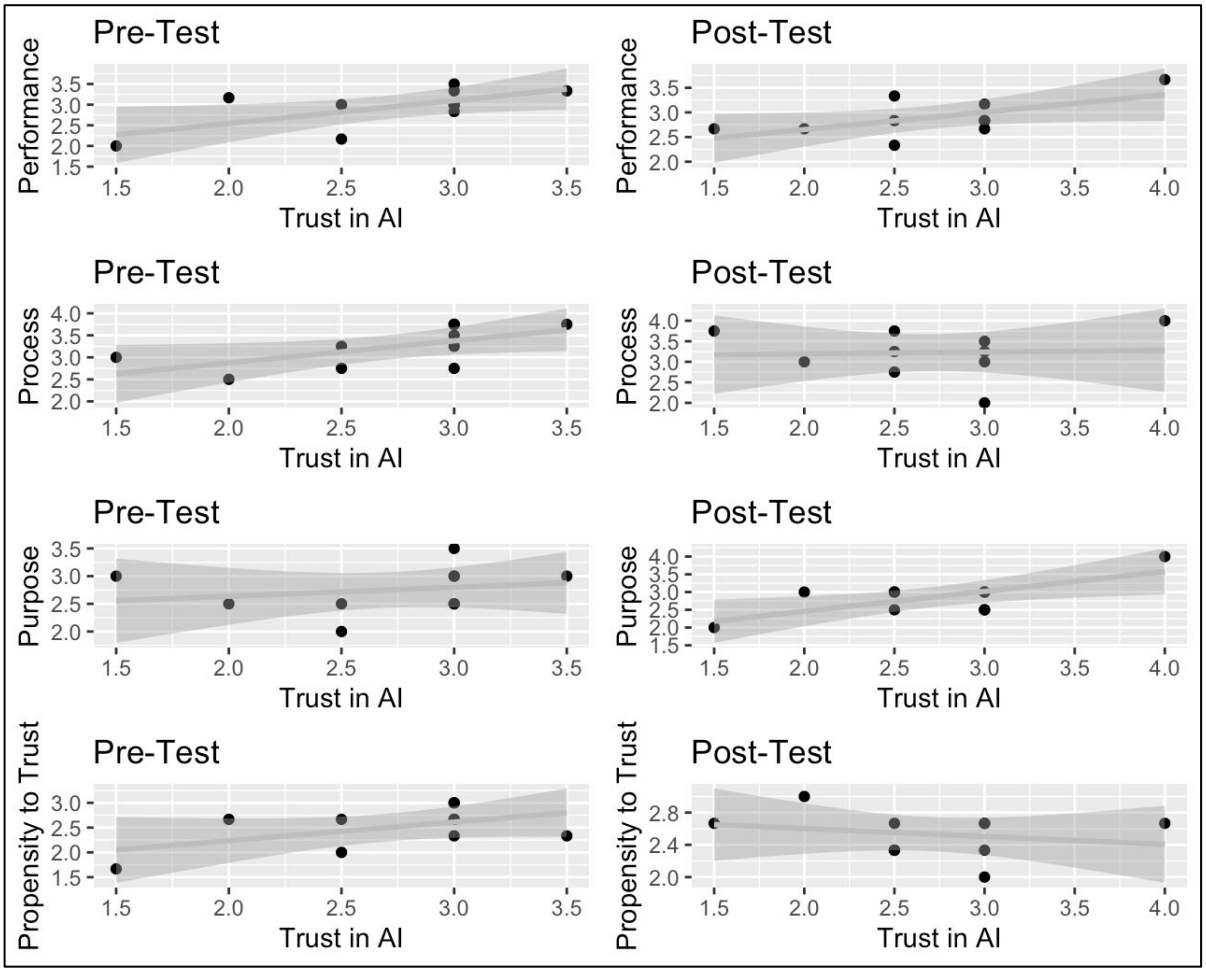


Abbildung H1. Streudiagramme zur Prüfung der Linearität (eigene Abbildung