

Vertrauen in Künstliche Intelligenz

Wie Framing das Vertrauen in LLM-basierte Applikationen- und Antworten

beeinflusst

Vorstudie Bachelorarbeit

Fabian Ryf

Hochschule Luzern, Wirtschaft

HSLU-W

BSc Business Psychology (BP)

Markt- und Konsumentenpsychologie

Dr. Andreas Hüsser

05.12.2025

Management Summary

Das vorliegende Forschungsdesign bildet die konzeptionelle Grundlage für die geplante Bachelorarbeit, die den Einfluss unterschiedlicher Darstellungsweisen von KI-Konfidenzwerten auf das Vertrauen und die Technologieakzeptanz bei Large Language Model-basierten Assistenzsystemen untersuchen soll. Angesichts der rasanten Verbreitung von KI-Assistenten wie ChatGPT mit über 700 Millionen wöchentlich aktiven Nutzern adressiert die geplante Studie eine zentrale Forschungslücke: Wie beeinflusst die Art der Transparenzkommunikation das Nutzervertrauen in KI-Systeme? Diese Frage gewinnt dadurch an Relevanz, dass nahezu die Hälfte aller KI-Anfragen der Informationsbeschaffung und praktischen Anleitungen dient..

Das vorgeschlagene Forschungsvorhaben stützt sich auf das Artificial Intelligence Technology Acceptance Model (AI-TAM) von Baroni et al., welches das klassische TAM-Modell von Davis um KI-spezifische Konstrukte wie Vertrauen in KI und Kollaborationsintention erweitert. Diese theoretische Fundierung wird durch die Integration des Attribute Framing-Effekts nach Levin und Gaeth ergänzt, um zu untersuchen, wie identische Konfidenzinformationen durch unterschiedliche Valenz-Darstellung die Wahrnehmung und das Vertrauen beeinflussen. Methodisch sieht das Design ein experimentelles Between-Subjects-Verfahren mit drei Bedingungen vor: positives Framing der Konfidenzwerte als Zuverlässigkeit, negatives Framing als Unsicherheit sowie eine Kontrollgruppe ohne Konfidenzanzeige. Die geplante Manipulation soll während der natürlichen Interaktion mit einem digitalen Assistenten erfolgen.

Das Hypothesengerüst umfasst zehn Annahmen, die sowohl direkte Framing-Effekte als auch komplexe Mediations- und Moderationsbeziehungen zwischen den AI-TAM-Konstrukten postulieren.

Die vorgesehene Datenerhebung kombiniert mehrere methodische Zugänge in einer Triangulation. Standardisierte Online-Befragungen mit validierten Skalen sollen in drei Phasen – vor, während und nach der KI-Interaktion – die relevanten Konstrukte wie Explainable AI Trust, Behavioral Intention und die klassischen TAM-Variablen

erfassen. Parallel dazu ist eine automatisierte Verhaltensbeobachtung durch System-Logging geplant, die Nutzungsdaten wie Interaktionshäufigkeit, Session-Dauer und tatsächliche Accuracy Scores dokumentiert.

Die geplante Stichprobe soll aus Nutzern eines digitalen Assistenten rekrutiert werden, wobei eine randomisierte Zuweisung zu den Experimentalbedingungen direkt bei der ersten Interaktion erfolgt. Die Auswertung erfolgt mittels Strukturgleichungsmodellierung (SEM), welche die simultane Prüfung aller angenommenen Beziehungen zwischen den Konstrukten ermöglicht. Dabei werden sowohl direkte Effekte des Framings auf das Vertrauen als auch indirekte, möglicherweise medierte Effekte auf die Nutzungsintention analysiert, unter Kontrolle demografischer und Störvariablen.

Die potenzielle Relevanz dieser Forschung liegt in der praktischen Anwendbarkeit für die Gestaltung transparenter KI-Applikationen. Während KI-Anbieter rechtlich verpflichtet sind, auf mögliche Fehler hinzuweisen, gibt es bislang wenig evidenzbasiertes Wissen darüber, wie diese Kommunikation optimal gestaltet werden sollte.

Inhaltsverzeichnis

Ausgangslage	1
Theoretische Einbettung, Forschungsfrage, Hypothesen und Forschungsdesign	1
Theoretische Einbettung	1
Technology Acceptance Model	1
Artificial Intelligence-Technology Acceptance Model	2
Framing-Effekt	5
Forschungsfrage	6
Latente Konstrukte	6
Hypothesenübersicht	7
Ausformulierte Hypothesen	7
Forschungsdesign	8
Experimentelles Design	8
Between-Subject Design	8
Ablauf Experiment	10
Pre-Interaktionsphase «digitaler Assistent»	10
Interaktionsphase «digitaler Assistent»	10
Post-Interaktionsphase «digitaler Assistent»	10
Stimulus-Konzept	11
Konkrete Stimulus-Beispiele	11
Manipulationscheck Stimulus	11
Methodentriangulation	11
Online-Befragung als Rahmenstruktur	13
Verhaltensbeobachtung durch System-Logging	13
Experimentelle Manipulation im Feldkontext	13
Methodenintegration	13
Abgrenzung des Forschungsdesigns	14

Inhaltlich	14
Methodisch	15
Operationalisierung Konstrukte	15
Stichprobe/Feldzugang	15
Beschreibung Stichprobe	15
Rekrutierung Stichprobe	16
Rekrutierung Stichprobe Pre-Test	16
Pre-Test: Durchführung und Analyse	16
Kritische Fragen und Überlegungen	16
Stimulus-Design	16
Experiment Ablauf	16
Pre-Test	17
Zielsetzung	17
Working Plan	17
Reflection Work	17
Quellenverzeichnis	18

Ausgangslage

Theoretische Einbettung, Forschungsfrage, Hypothesen und Forschungsdesign

Theoretische Einbettung

Modelltheoretisch knüpft die vorliegende Arbeit an frühere Studien in den Bereichen Vertrauen in künstliche Intelligenz/technologische Veränderungen, wahrgenommene Nützlichkeit sowie Benutzerfreundlichkeit und die daraus abgeleitete Nutzungsabsicht an. Als theoretische Grundlage dient zunächst das Technology Acceptance Model (TAM) von Fred Davis aus dem Jahr 1987, welches die Rahmenbedingungen zur Analyse von Adoptionsprozessen neuer Technologien schafft (Davis, 1989). Den zweiten Baustein liefert die Erweiterung des TAM-Modells durch Baroni et al. (2022). Diese ergänzt das Modell um zusätzliche Faktoren wie das Vertrauen in KI-gestützte Assistenten und bildet diese im Artificial Intelligence Technology Acceptance Model (AI-TAM) ab (Baroni et al., 2022).

Zuletzt wird der Framing-Effekt theoretisch beleuchtet, da dieser für die gewählte Stimulus-Wahl relevant ist. Konkret wird dabei die Form des Attribute-Framing-Effekts betrachtet (Druckman, 2001; Freling et al., 2014).

Technology Acceptance Model (Davis, 1989)

Das TAM wurde entwickelt, um die mangelnde Nutzerakzeptanz von Informationssystemen zu adressieren, die als Haupthindernis für den Erfolg neuer Technologien identifiziert wurde. Davis untersuchte 112 Angestellte und Manager eines grossen nordamerikanischen Unternehmens, die zwei unterschiedliche Softwaresysteme nutzten - ein elektronisches Mailsystem und einen Texteditor. Das Modell basiert auf der Attitude-Paradigm aus der Psychologie, speziell auf Fishbein und Ajzens Theory of Reasoned Action (Fishbein & Ajzen, 1975). TAM besagt, dass die tatsächliche Systemnutzung durch die Verhaltensintention bestimmt wird, welche von der Einstellung zur Nutzung abhängt.

Diese Einstellung wird durch zwei zentrale Konstrukte geprägt:

- Perceived Usefulness – «the degree to which an individual believes that using a particular system would enhance his or her job performance» – und
- Perceived Ease of Use – «the degree to which an individual believes that using a particular system would be free of physical and mental effort».

Die Studie zeigte, dass Perceived Usefulness etwa 50% einflussreicher auf die Nutzung war als Ease of Use, wobei das Modell 36% der Varianz in der tatsächlichen Nutzung erklären konnte.

Während TAM erfolgreich die Akzeptanz traditioneller Informationssysteme erklärt, erweist es sich für KI-basierte Systeme als unzureichend. KI-Systeme unterscheiden sich durch ihre probabilistische Natur und inhärente Unsicherheit - Eigenschaften, die Vertrauen zu einem Faktor machen, der im ursprünglichen TAM nicht berücksichtigt wird. Zudem werden KI-Systeme nicht nur als Werkzeuge, sondern oft auch als kollaborative Partner wahrgenommen, was neue Dimensionen der Mensch-Maschine-Interaktion eröffnet. Diese Lücke adressieren Baroni et al. (2022) mit ihrer Erweiterung des TAM-Modells (Baroni et al., 2022).

Artificial Intelligence-Technology Acceptance Model (Baroni et al., 2022)

Notwendigkeit der TAM-Erweiterung für KI-Systeme. Das von Davis (1989) entwickelte Technology Acceptance Model (TAM, (Davis, 1989)) basiert auf der Theory of Reasoned Action und erklärt Technologieakzeptanz durch die Faktoren der wahrgenommenen Nützlichkeit (Perceived Usefulness) und der wahrgenommenen Benutzerfreundlichkeit (Perceived Ease of Use). Während das TAM die Adoption traditioneller Informationssysteme bereits untersucht hat, erweist es sich für KI-basierte Systeme als unzureichend. Der Grund dafür liegt in den Unterschieden von KI-Systemen: Ihre probabilistische Natur, die Unsicherheit und ihre Wahrnehmung als kollaborative Partner anstelle reiner Werkzeuge. Diese Eigenschaften machen Vertrauen zu einem Faktor, der im ursprünglichen TAM nicht abgebildet wird. Darüber hinaus erfordern «Human-in-the-Loop»-Ansätze, dass Nutzer aktiv zur Verbesserung der KI beitragen – eine Dimension der Kollaboration, die das klassische Modell ebenfalls nicht

berücksichtigt.

Erweiterungen des TAM zum AI-TAM. Baroni et al. (2022) erweiterten das TAM um drei zusätzliche Konstrukte: «Explainable AI Trust» (Vertrauen in KI) aus der Literatur zu «Explainable AI» (XAI), «Collaborative Intention» (Kollaborationsabsicht) zur Messung der Bereitschaft zur Teilnahme an «Human-in-the-Loop»-Mechanismen sowie die Vertrautheit mit der Technologie und dem Anwendungskontext. Vertrauenskonstrukt basiert auf der Arbeit von Hoffman et al. (Hoffman et al., 2019). Dabei misst «Explainable AI Trust» die Zuversicht in die Ergebnisse der KI. Die «Collaborative Intention» erfasst die Bereitschaft der Nutzer, aktiv zur Verbesserung der KI beizutragen, was ein kritischer Faktor für «Human-in-the-Loop»-Systeme ist. Dieses übt einen Einfluss auf Kernkonstrukte des TAM aus.

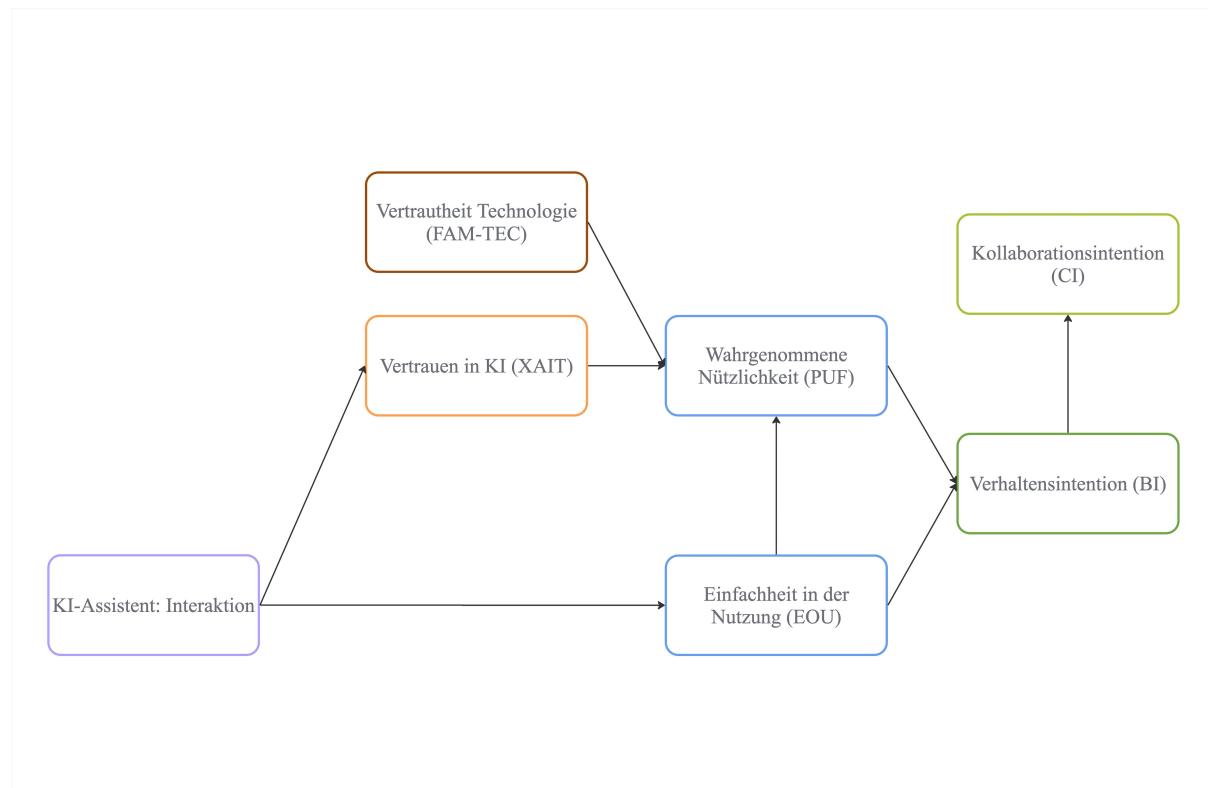


Abbildung 1

Erweitertes TAM-Modell: Artificial Intelligence-Technology Acceptance Model (Baroni et al., 2022)

Validierung durch die BumpOut-Studie (Baroni et al., 2022). Das AI-TAM wurde mit der Anwendung «BumpOut» validiert, einer KI-gestützten App zur Schadensmeldung bei Autounfällen. Die Studie umfasste 400 Teilnehmende in zwei Crowdsourcing-Kampagnen unter unterschiedlichen experimentellen Bedingungen: einer fehlerfreien KI versus einer teilweise fehlerhaften KI. Die App analysiert dabei Schadensbilder automatisch, wobei die Nutzer die von der KI getroffenen Identifikationen bestätigen oder korrigieren können. Die Ergebnisse zeigten hohe Werte für die wahrgenommene Nützlichkeit und Benutzerfreundlichkeit, während die Funktionsfähigkeit der KI nur einen minimalen Einfluss hatte. Besonders bedeutsam war die starke Korrelation zwischen der Nutzungsabsicht («Behavioral Intention») und der Kollaborationsabsicht («Collaborative Intention»). Dies bestätigt, dass Nutzer, die bereit sind, die App zu verwenden, auch bereit sind, zur Verbesserung der KI beizutragen. Das AI-TAM eignet sich daher auch für die Untersuchung der Akzeptanz von Large Language Models (LLMs), da diese Systeme die gleichen kritischen Charakteristika aufweisen: probabilistische Ausgaben, inhärente Unsicherheit und die Notwendigkeit von Nutzervertrauen. LLMs werden zunehmend als kollaborative Partner wahrgenommen, deren Ergebnisse oft Nutzerfeedback erfordern. Insbesondere die XAI-Konstrukte sind hier relevant, da Nutzer nachvollziehen müssen, warum ein LLM eine bestimmte Antwort generiert.

Die Verbindung des AI-TAM mit dem Konzept des Attribute Framings eröffnet neue Forschungsperspektiven. Framing-Effekte könnten insbesondere das XAIT-Konstrukt beeinflussen. So dürfte die Präsentation von KI-Fähigkeiten als Gewinn («95 % Genauigkeit») im Vergleich zu einer Darstellung als Unsicherheit («5 % Fehlerrate») das Vertrauen in die KI («Explainable AI Trust») direkt verändern. Gemäss dem AI-TAM-Modell beeinflusst dieser Faktor wiederum die Nutzungsabsicht («Behavioral Intention»). Für Experimente mit Large Language Models (LLMs) bedeutet dies, dass die Art der Leistungsdarstellung, beispielsweise durch das Attribut Framing der Modellfähigkeiten, die Nutzerakzeptanz beeinflussen könnte. Das AI-TAM bietet hierbei den methodischen Rahmen, um diese Effekte auf den relevanten

Dimensionen präzise zu messen.

Framing-Effekt

Der Framing-Effekt, erstmals von Kahneman und Tversky in ihrer Prospect Theory beschrieben, zeigt, dass Entscheidungen davon beeinflusst werden, wie Informationen präsentiert werden (Tversky & Kahneman, 1986). Der Framing-Effekt zeigt unter anderem, wie identische Szenarien zu unterschiedlichen Präferenzen führen, je nachdem ob sie in Gewinn- oder Verlustbegriffen formuliert werden. Während sich die frühe Forschung auf riskante Entscheidungen konzentrierte, erweiterte sich das Konzept auf verschiedene Framing-Typen wie Risky Choice Framing, Goal Framing und Attribute Framing.

Zusätzlich untersuchte Freling et al. (2014) in ihrer Meta-Analyse 107 Studien zum Attribute Framing und entwickelten dabei eine theoretische Integration mittels Construal Level Theory (CLT). Ihre zentrale Erkenntnis: Die Effektivität von Attribute Framing hängt von der Kongruenz zwischen dem Abstraktionsniveau (Construal Level) des Frames und der psychologischen Distanz des Bewertenden zum geframten Event ab (Freling et al., 2014).

Attribute Framing nach Levin & Gaeth (1988) und Dolgopolova et al. (2022). Attribute Framing unterscheidet sich von anderen Framing-Typen, da hier ein einzelnes Attribut in äquivalenten aber unterschiedlich valenten Begriffen beschrieben wird. Levin und Gaeth demonstrierten dies mit Hackfleisch, das entweder als «75% mager» oder «25% fett» beschrieben wurde (Levin & Gaeth, 1988). Der Attribute Framing-Effekt manifestiert sich in einer valenz-konsistenten Verschiebung: Positive Frames führen zu günstigeren Bewertungen als negative. Ihre Studie zeigte zudem, dass direkte Produkterfahrung den Framing-Effekt abschwächt - ein Befund, der durch ein Averaging-Modell erklärt wird, bei dem zusätzliche Informationsquellen den relativen Einfluss einzelner Frames reduzieren.

Dolgopolova et al. (2022) spezifisch auf Lebensmittelentscheidungen und fanden Effekte für Einstellungen versus Intentionen. Während Gain-Frames signifikant positivere Einstellungen erzeugten, war der Effekt auf Kaufintentionen nahe null und

nicht signifikant. Mehrere Moderatoren wurden identifiziert: Gain-Frames, Interaktionsterme, spezifische Produkte und Studentenstichproben beeinflussten signifikant die Ergebnisse. Diese Befunde unterstreichen die Komplexität des Attribute Framing bei Lebensmitteln, wo zeitliche Diskontierung und die Verzögerung zwischen Konsum und Gesundheitskonsequenzen eine Rolle spielen (Dolgopolova et al., 2022).

Der Attribute Framing-Effekt ist für die Untersuchung der KI-Akzeptanz relevant, da KI-Systeme durch ihre Fähigkeiten (Gain-Frame: «95% Genauigkeit») oder Limitationen (Loss-Frame: «5% Fehlerrate») charakterisiert werden können. Im Kontext des erweiterten TAM-Modells (Davis, 1989) könnte Attribute Framing die Wahrnehmung von Perceived Usefulness und Vertrauen in KI beeinflussen. Die Präsentation von KI-Funktionen als Gewinne («erhöht Produktivität um 30%») versus Verluste («30% manuelle Arbeit bleibt erforderlich») könnte unterschiedliche Akzeptanzmuster erzeugen.

Forschungsfrage

Wie beeinflusst die Framingdarstellung von KI-Konfidenzwerten (positiv vs. negativ) das Vertrauen in KI-generierte Antworten und die daraus resultierende Technologieakzeptanz in LLM-basierten Assistenzsystemen?

Latente Konstrukte

Die latenten Konstrukte werden mittels einer Online-Befragung vor, während und nach der Nutzung der KI-Assistenz erhoben. Die verwendeten Konstrukte stammen grossteils aus dem Technology Acceptance Model (TAM) von Davis sowie aus der Erweiterung dieses Modells durch Baroni et al. (2022). Diese Erweiterung ergänzt das bestehende TAM um KI-relevante Faktoren wie das Vertrauen in erklärbare KI («Explainable AI Trust», XAIT), die Kollaborationsabsicht («Collaborative Intention, CI») und die technologische Vorerfahrung («Familiarity with Technology», FAM-TEC») (Baroni et al., 2022).

Hypothesenübersicht

Ausformulierte Hypothesen

Haupthypothesen (Framing-Effekte).

- H1a: Die positive Darstellung des Accuracy Scores (z.B. «Diese Antwort ist zu 80% korrekt») führt zu einem höheren AI Output Trust als die Kontrollbedingung ohne Score-Anzeige.
- H1b: Die negative Darstellung des Accuracy Scores (z.B. «Diese Antwort hat eine 20% Fehlerwahrscheinlichkeit») führt zu einem niedrigeren AI Output Trust als die Kontrollbedingung ohne Score-Anzeige.

AI-TAM Kernbeziehungen.

- H2: AI Output Trust hat einen positiven Einfluss auf die Perceived Usefulness. Nutzer, die den AI-Ausgaben vertrauen, bewerten das System als nützlicher für ihre Aufgaben.
- H3: AI Output Trust hat einen positiven Einfluss auf die Perceived Ease of Use. Vertrauen in die AI-Ausgaben reduziert die wahrgenommene kognitive Belastung bei der Systemnutzung.

TAM-Standardbeziehungen.

- H4: Die Perceived Usefulness hat einen positiven Einfluss auf die Behavioral Intention. Je nützlicher Nutzer Alva einschätzen, desto höher ist ihre Absicht, das System zukünftig zu nutzen.
- H5: Die Perceived Ease of Use hat einen positiven Einfluss auf die Behavioral Intention. Eine als einfach wahrgenommene Nutzung erhöht die Intention zur zukünftigen Systemnutzung.
- H6: Die Perceived Ease of Use hat einen positiven Einfluss auf die Perceived Usefulness. Systeme, die einfach zu nutzen sind, werden als nützlicher wahrgenommen.

- H7: Die Behavioral Intention hat einen positiven Einfluss auf die Collaborative Intention. Nutzer, die beabsichtigen Alva zu nutzen, zeigen auch eine höhere Bereitschaft zur kollaborativen Zusammenarbeit mit dem AI-System.
- H8: Familiarity with Technology hat einen positiven Einfluss auf die Perceived Usefulness. Nutzer, die mit KI-Technologie vertraut sind, schätzen die Nützlichkeit der digitalen Assistenz höher ein.

Mediation.

- H9: Der Effekt des Framings auf die Behavioral Intention wird durch Explainable AI Trust partiell oder vollständig mediert. Bei niedrigen Explainable AI Trust-Werten ist der Unterschied zwischen positivem und negativem Framing grösser als bei hohen Explainable AI Trust-Werten.

Forschungsdesign

Das Experiment untersucht den Einfluss von Accuracy-Framing auf die Technologieakzeptanz von KI-Systemen, basierend auf dem AI-TAM-Modell. In einem 3x2 Between-Subjects-Design wird die Darstellung von Konfidenzwerten (positiver Frame vs. negativer Frame) bei variierenden Accuracy-Scores (hoch, mittel, niedrig) manipuliert. Die experimentelle Manipulation erfolgt während der realen Interaktion mit einem KI-Assistenten.

Experimentelles Design

Between-Subject Design

Das Experiment nutzt ein Between-Subjects-Design mit drei Experimentalgruppen, um den Einfluss des Accuracy-Framings auf das Vertrauen in KI-Systeme zu untersuchen.

- Unabhängige Variable: Das Accuracy-Framing mit zwei Ausprägungen
 - Positiver Frame: Gain-Darstellung (positive «Zuversicht»)
 - Negative Frame: Loss-Darstellung (negative «Zuversicht»)

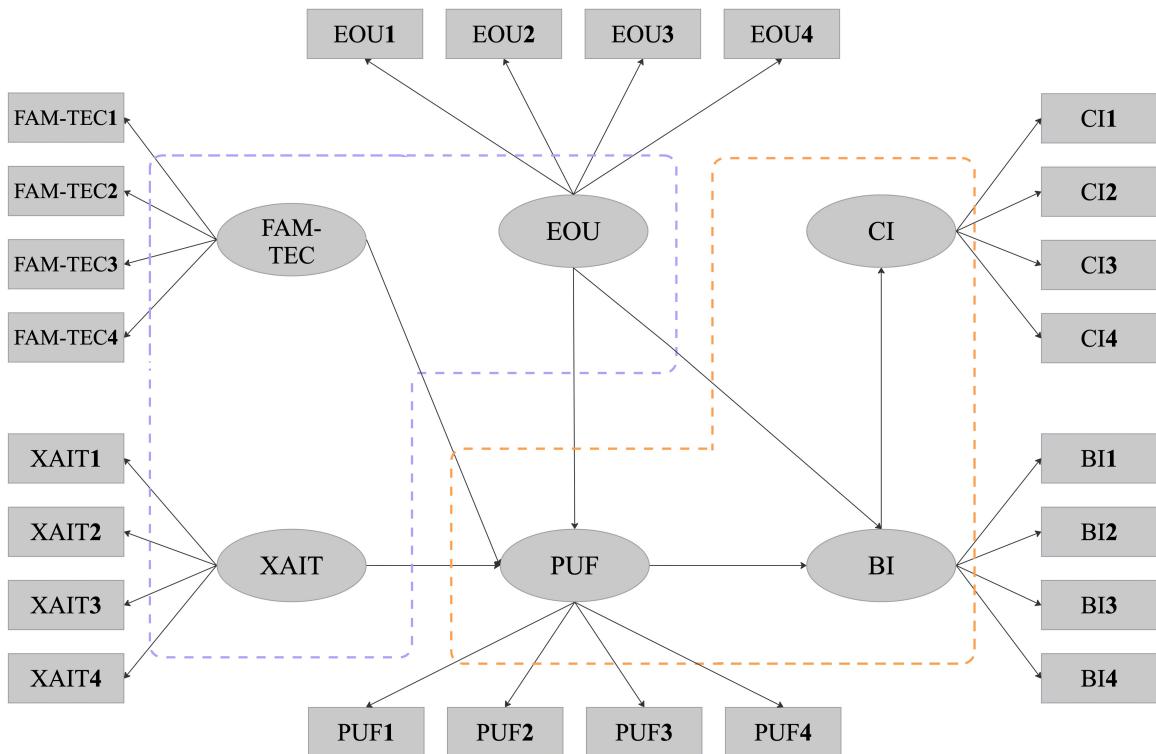


Abbildung 2

Strukturmodell für AI-TAM, Teil des Strukturgleichungsmodell, inkl. Hypothesen, inkl. Stimulus

- Unabhängige Variable: Das Accuracy-Framing mit drei Stufen:
 - Hoch
 - Mittel
 - Niedrig

Diese Designstruktur wird in der Literatur als Between-Subjects-Design mit Kovariate bezeichnet, wobei die kontinuierliche Variable als statistische Kontrollvariable dient (Kim & Song, 2022). Die Kombination eines kategorialen Faktors mit einer kontinuierlichen Variable ermöglicht die Untersuchung von Haupt- und Interaktionseffekten, während gleichzeitig die natürliche Varianz der KI-Performance kontrolliert wird

Ablauf Experiment

Das geplante Experiment ist in drei Phasen gegliedert: eine Phase vor, eine während und eine nach der Interaktion mit dem KI-Assistenten. In jeder dieser Phasen werden die benötigten Daten erhoben. Dies geschieht entweder durch direkte Nutzerbefragung mittels Bewertungsfragen oder durch automatische Berechnung und Speicherung im Hintergrund, wie es beispielsweise beim «Accuracy Score» der Fall ist.

Da die initiale Befragung bereits stattfindet, bevor die Teilnehmenden mit dem System interagieren, wird diese bewusst kurz gehalten. Ziel ist es, die Abbruchrate zu minimieren und eine hohe Abschlussrate des gesamten Experiments zu fördern.

Pre-Interaktionsphase «digitaler Assistent»

- Einführung & Einwilligung
- Erhebung «FAM-TEC»-Konstrukt

Interaktionsphase «digitaler Assistent»

- Anzahl Interaktionen
- Tatsächlicher Accuracy-Score von Antworten
- Dauer der Sitzungen
- Anonymisierte, qualitative Inhalte wie Prompts und Antworten
- Bewertung («Rating»)
- LLM-Einschätzung («Faithfulness Score»)

Post-Interaktionsphase «digitaler Assistent»

- Erhebung «AI-TAM-Konstrukte»
- Erhebung «XAIT»-Konstrukt
- Erhebung «TAM-Konstrukte»
- Erhebung «PUF»-Konstrukt, Teil 2

- Erhebung «EOU»-Konstrukt
- Erhebung «BI»-Konstrukt
- Erhebung «CI»-Konstrukt
- Erhebung soziodemografischer Daten (Kontrollvariablen)
 - Alter
 - Geschlecht
 - Höchster Bildungsabschluss

Stimulus-Konzept

Der Stimulus besteht aus der visuellen und textlichen Darstellung des AI-Accuracy Scores direkt nach jeder LLM-Antwort. Die Manipulation erfolgt in Echtzeit während der natürlichen Interaktion mit dem digitalen Assistenten.

Konkrete Stimulus-Beispiele

Manipulationscheck Stimulus

Als Manipulationscheck werden die Probanden post-experimentell gefragt, ob und in welcher Form ihnen Informationen zur Zuverlässigkeit der Antworten angezeigt wurden, um sicherzustellen, dass die experimentelle Manipulation wahrgenommen wurde.

Methodentriangulation

Das vorliegende Forschungsdesign kombiniert verschiedene Methoden in einer Methodentriangulation, um die Forschungsfrage nach dem Einfluss von KI-Transparenz auf Vertrauen zu beantworten.

Die Triangulation erfolgt auf drei Ebenen:

- experimentelle Manipulation des Accuracy Framings als Between-Subject-Design gewährleistet die interne Validität durch randomisierte Zuweisung.

--	--

(a) *Positiver Frame / Hohe Sicherheit*(b) *Negativer Frame / Hohe Sicherheit*

--	--

(c) *Positiver Frame / Mittlere Sicherheit*(d) *Negativer Frame / Mittlere Sicherheit*

--	--

(e) *Positiver Frame / Niedrige Sicherheit*(f) *Negativer Frame / Niedrige Sicherheit***Abbildung 3***Übersicht der Stimulus-Designs für das 3x2 Experiment*

- Die natürliche Beobachtung während der tatsächlichen LLM-Nutzung erhöht die ökologische Validität, da Nutzer eigene Fragen in realistischen Anwendungskontexten stellen.
- Die standardisierte Befragung mittels validierter Skalen aus dem AI-TAM-Modell ermöglicht die reliable Messung latenter Konstrukte.

Online-Befragung als Rahmenstruktur

Die gesamte Datenerhebung erfolgt über eine webbasierte Plattform, die Pre-Interaction-Messungen (demografische Daten, AI-Vorerfahrung), Post-Interaction-Messungen (AI-TAM-Konstrukte) und die experimentelle Randomisierung steuert. Die Verwendung etablierter Skalen aus dem TAM (Davis, 1989) und AI-TAM (Baroni et al., 2022) gewährleistet die Vergleichbarkeit mit bestehender Forschung. Die standardisierten Items werden auf 5-stufigen Likert-Skalen gemessen.

Verhaltensbeobachtung durch System-Logging

Während der Interaktion mit dem digitalen Assistenten werden automatisiert Verhaltensdaten erfasst: Anzahl der Interaktionen, Session-Dauer(, Fragentypen). Diese Messung liefert weitere Verhaltensindikatoren. Der tatsächliche Accuracy Score wird vom LLM-System für jede Antwort berechnet und protokolliert, wodurch eine kontinuierliche, objektive Performanz-Metrik entsteht, die als Kovariate in die Analysen eingeht.

Experimentelle Manipulation im Feldkontext

Die Framing-Manipulation wird während der natürlichen Nutzung implementiert. Diese Einbettung des Experiments in den realen Anwendungskontext entspricht einem natürlichen Experiment das externe bei ausreichender interner Validität bietet.

Methodenintegration

Die Integration der verschiedenen Datenquellen erfolgt auf Analyseebene: Die experimentelle Gruppenzugehörigkeit (Framing) wird mit den Befragungsdaten (Trust,

TAM-Konstrukte) und den Systemdaten (Accuracy Score) in einem gemeinsamen Datensatz zusammengeführt. Diese Triangulation ermöglicht:

- Konvergenz-Validierung: Trust-Ratings während der Interaktion (Single-Item) werden mit Post-Interaction Trust-Skalen (Multi-Item) korreliert
- Komplementäre Erkenntnisse: Subjektive Wahrnehmungen (Befragung) werden mit objektiven Metriken (System-Logs) kontrastiert
- Moderation/Mediation: Die Interaktion zwischen experimenteller Manipulation und natürlicher Variation kann analysiert werden

Abgrenzung des Forschungsdesigns

Die vorliegende Studie fokussiert auf die valenzorientierte Darstellung von KI-Leistungsmetriken (Attribute Framing) und deren Einfluss auf Vertrauen und Technologie-akzeptanz im Kontext des AI-TAM-Modells.

Inhaltlich

- Andere Framing-Typen: Die Studie beschränkt sich auf Attribute Framing und untersucht nicht Risky Choice Framing oder Goal Framing.
- Langzeiteffekte (keine Längsschnittstudie): Gemessen wird die Nutzungsabsicht (Behavioral Intention), nicht die tatsächliche Systemnutzung über längere Zeiträume. Der Intention-Behavior-Gap wird nicht untersucht.
- Alternative Transparenzmechanismen: Neben der Score-Darstellung existieren weitere Transparenzmöglichkeiten (Erklärungen, Quellenangaben, Visualisierungen), die nicht Gegenstand dieser Arbeit sind.
- Kontextübergreifende Generalisierung: Die Untersuchung findet im Verwaltungskontext statt. Ob die Ergebnisse auf medizinische, kreative oder andere Anwendungsbereiche übertragbar sind, bleibt offen.

Methodisch

- Between-Subjects-Design: Jede Person erfährt nur eine Framing-Bedingung.
Intraindividuelle Vergleiche sind nicht möglich.
- Quantitative Fokussierung: Die Studie nutzt standardisierte Skalen, verzichtet jedoch auf qualitative Vertiefungen wie Interviews zur Exploration der zugrunde liegenden kognitiven Prozesse.
- Natürliche Variation des Accuracy Scores: Der tatsächliche Score wird nicht experimentell manipuliert, sondern variiert basierend auf den Nutzeranfragen. Er dient als Kovariate, nicht als unabhängige Variable.

Operationalisierung Konstrukte

Die Operationalisierung der Konstrukte erfolgt in Form von Bewertungsfragen mit einer 5-Punkt Likert-Skalenbewertung. Die Befragung wird mittels Onlinebefragung vor- während- und nach der Verwendung der LLM-Lösung durchgeführt. Die Likert-Skalen sind so skaliert, dass 1 jeweils die negativste Bewertung des jeweiligen Items darstellt und 5 die positivste Bewertung. Eine Item-Batterie beinhaltet zwischen ein bis sechs Items, welche das gewünschte Konstrukt erfassen sollen. Einzelne Item-Batterien beinhalten negativformulierte Items als Kontrollfragen.

Das Operationalisierungsverfahren ist theoriegeleitet, da ein Grossteil der bestehenden Items aus vorherigen Studien (Baroni et al., 2022; Davis, 1987) teilweise übernommen werden kann. Die Items aus den verschiedenen Item-Batterien (latente Konstrukte) müssen jedoch für den geplanten Anwendungsfall überarbeitet und übersetzt werden. Dies betrifft sämtliche definierten Konstrukte des definierten AI-TAM-Modells (Baroni et al., 2022).

Stichprobe/Feldzugang

Beschreibung Stichprobe

Die Stichprobe ist als Gelegenheitsstichprobe zu bezeichnen, da nur potenziell Benutzer mit dem digitalen Assistenten interagieren, welche bereits wissen, dass es diesen gibt. Einschlusskriterien für die zu erhebende Stichprobe sind wie folgt ausgelegt.

Rekrutierung Stichprobe

Die Rekrutierung der Proband*innen geschieht direkt auf der Plattform, wo der digitale Assistent integriert ist. Die Rekrutierung der Benutzer*innen findet somit ausschliesslich digital statt. Wenn sich Benutzer*innen entscheiden mit der digitalen Assistenz auf der Plattform zu interagieren, werden Benutzer*innen nach Akzeptieren der Bestimmungen einer Experimentalbedingung zugewiesen.

Rekrutierung Stichprobe Pre-Test

Pre-Test: Durchführung und Analyse

Kritische Fragen und Überlegungen

Stimulus-Design

- Stimuli unterscheiden sich in drei Dimensionen gleichzeitig: Text (Konfidenz/Unsicherheit), Farbe (grün/orange-rot) UND Icon (✓/☒)
- Manipulationscheck ausreichend bei prominenter Platzierung im Assistenten-Interface?

Experiment Ablauf

- Es gibt mehrere Möglichkeiten die Proband*innen in das Experiment zu holen
 - Footer
 - Button
 - Suchresultate
- Soll das Onboarding zum Experiment vor der ersten LLM-Interkation geschehen oder nachher?
 - Sprich, ich (Proband*in) kann zuerst eine Frage an den digitalen Assistenten richten, danach Onboarding zum Experiment
 - JA/NEIN

Pre-Test

- Framing muss auch irgendwie gepretestet werden
- Wie sieht die Experimentalbedingung aus im Pre-Test
 - Hoch / Mittel / Tief
 - Verschiedene Sicherheitslevel
 - Alles unter 50% ist tief

Zielsetzung

Working Plan

Reflection Work

Quellenverzeichnis

- Baroni, I., Calegari, G. R., Scandolari, D., & Celino, I. (2022). AI-TAM: a model to investigate user acceptance and collaborative intention in human-in-the-loop AI applications. *Human Computation*, 9(1), 1–21.
<https://doi.org/10.15346/hc.v9i1.134>
- Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology [Publisher: Management Information Systems Research Center, University of Minnesota]. *MIS Quarterly*, 13(3), 319–340.
<https://doi.org/10.2307/249008>
- Dolgopolova, I., Li, B., Pirhonen, H., & Roosen, J. (2022). The effect of attribute framing on consumers' attitudes and intentions toward food: A Meta-analysis. *Bio-based and Applied Economics*, 10, 253–264.
<https://doi.org/10.36253/bae-11511>
- Druckman, J. N. (2001). Evaluating framing effects. *Journal of Economic Psychology*, 22(1), 91–101. [https://doi.org/10.1016/S0167-4870\(00\)00032-5](https://doi.org/10.1016/S0167-4870(00)00032-5)
- Fishbein, M., & Ajzen, I. (1975, Mai). *Belief, attitude, intention and behaviour: An introduction to theory and research* (Bd. 27).
- Freling, T. H., Vincent, L. H., & Henard, D. H. (2014). When not to accentuate the positive: Re-examining valence effects in attribute framing. *Organizational Behavior and Human Decision Processes*, 124(2), 95–109.
<https://doi.org/10.1016/j.obhdp.2013.12.007>
- Kim, T., & Song, H. (2022). Communicating the Limitations of AI: The Effect of Message Framing and Ownership on Trust in Artificial Intelligence. *International Journal of Human-Computer Interaction*, 39, 1–11.
<https://doi.org/10.1080/10447318.2022.2049134>
- Tversky, A., & Kahneman, D. (1986, Januar). The Framing of Decisions and the Evaluation of Prospects. In R. Barcan Marcus, G. J. W. Dorn & P. Weingartner (Hrsg.), *Studies in Logic and the Foundations of Mathematics* (S. 503–520, Bd. 114). Elsevier. [https://doi.org/10.1016/S0049-237X\(09\)70710-4](https://doi.org/10.1016/S0049-237X(09)70710-4)

Abbildungsverzeichnis

1	Erweitertes TAM-Modell: Artificial Intelligence-Technology Acceptance Mo-	
	del (Baroni et al., 2022)	3
2	Strukturmodell für AI-TAM, Teil des Strukturgleichungsmodell, inkl. Hy-	
	pothesen, inkl. Stimulus	9
3	Übersicht der Stimulus-Designs für das 3x2 Experiment	12

Tabellenverzeichnis

1	Identifizierte latente Konstrukte zum Einsatz von AI-TAM	20
2	Aufgelistete Hypothesen im Rahmen der Bachelor-Arbeit-Vorstudie . . .	21
3	Experiment-Design, vorhandene Experimentalbedingungen	22
4	Experimentelles 3x2 Design: Manipulation von Framing und Accuracy .	22

Tabelle 1

Identifizierte latente Konstrukte zum Einsatz von AI-TAM

Abkürzung	Name	Definition	Quelle
XAIT	Explainable AI trust	Vertrauen in die generierte Antwort und die LLM-Lösung	(Baroni et al., 2022)
BI	Behaviorial Intention	Verhaltensintention die LLM-Lösung zu Nutzen	(Baroni et al., 2022; Davis, 1989)
CI	Collaborative Intenti- on	Kollaborationsintention zur digitalen Assis- tenz	(Baroni et al., 2022; Davis, 1989)
PUF	Perceived Usefulness	Wahrgenommene Nützlichkeit der generierten Antwort	(Baroni et al., 2022; Davis, 1989)
EOU	Ease of Use	Einfachheit in der Nutzung der KI- Applikation	(Baroni et al., 2022; Davis, 1989)
FAM-TEC	Familiarity with Technology	Vertrautheit in der Nutzung von KI- Technologie	(Baroni et al., 2022; Davis, 1989)

Tabelle 2*Aufgelistete Hypothesen im Rahmen der Bachelor-Arbeit-Vorstudie*

Hypothese	Pfad	Richtung	Theorie
H1a	Dummy_Pos (Stimulus) → XAIT	+	Attribute Frame
H1b	Dummy_Neg (Stimulus) → XAIT	-	Attribute Frame
H2	XAIT → PUF	+	AI-TAM-Modell
H3	XAIT → EOU	+	AI-TAM-Modell
H4	PUF → BI	+	AI-TAM-Modell
H5	EOU → BI	+	AI-TAM-Modell
H6	EOU → PUF	+	AI-TAM-Modell
H7	BI → CI	+	AI-TAM-Modell
H8	FAMTEC → PUF	+	AI-TAM-Modell

Kovariate

ACTS	ACTS → XAIT	kontrolliert	-
------	-------------	--------------	---

Mediation

H9	Framing → XAIT → (PUF, EOU) → BI	Indirekt	-
----	----------------------------------	----------	---

Anmerkung: Alle Pfade werden simultan im Strukturgleichungsmodell (SEM) geschätzt

Tabelle 3*Experiment-Design, vorhandene Experimentalbedingungen*

Bedingung	Gruppe	Manipulation	Beispiel
Positive-Frame	Gruppe 1	Score wird als Konfidenz/Zuverlässigkeit dargestellt	«Antwortsicherheit: 80%» oder «Antwortsicherheit zu 80% zuverlässig»
Negative-Frame	Gruppe 2	Score wird als Unsicherheit/Fehlerwahrscheinlichkeit oder dargestellt	Antwortunsicherheit: 20% oder «Diese Antwort hat eine Fehlerwahrscheinlichkeit von 20%»
Kontrollgruppe	Gruppe 3	Kein Score wird angezeigt (Status Quo)	-

Tabelle 4*Experimentelles 3x2 Design: Manipulation von Framing und Accuracy*

Framing (Unabhängige Variable 1)		
Accuracy (UV 2)	Positiver Frame	Negativer Frame
Hoch	Sicherheit: Hoch	Unsicherheit: Tief
Mittel	Sicherheit: Mittel	Unsicherheit: Mittel
Niedrig	Sicherheit: Tief	Unsicherheit: Hoch

Anmerkung. Die Kontrollgruppe (kein Score) ist in diesem 3x2 Design nicht abgebildet.