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»)
- Unabhängige Variable: Das Accuracy-Framing mit drei Stufen:
  - Hoch
  - Mittel
  - Niederig

Diese Designstruktur wird in der Literatur als Between-Subjects-Design mit Kovariate bezeichnet, wobei die kontinuierliche Variable als statistische Kontrollvariable dient (Kim, 2018). 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

Tabelle 1

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 dargestellt	Antwortunsicherheit: 20% oder «Diese Antwort hat eine Fehlerwahrscheinlichkeit von 20%»
Kontrollgruppe	Gruppe 3	Kein Score wird angezeigt (Status Quo)	-

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

Tabelle 2

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.

- 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 (Kontrollvariabeln)
  - 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.
- 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, 1987) 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

Pretest-Analyse

Reliabilitätsanalyse (Cronbach's Alpha)

Die Reliabilität aller verwendeten Skalen wird mittels Cronbach's Alpha überprüft. Für jedes Konstrukt wird die interne Konsistenz berechnet: AI Output Trust, Perceived Usefulness, Perceived Ease of Use, Behavioral Intention und Collaborative Intention. Bei unzureichender Reliabilität werden Item-Total-Korrelationen analysiert, um problematische Items zu identifizieren. Items mit ungenügender Trennschärfe werden überarbeitet oder eliminiert. Zusätzlich wird die Veränderung des Alpha-Werts bei Itemausschluss berechnet (Alpha if item deleted), um die optimale Itemkombination zu bestimmen.

Manipulationscheck (t-test unabh. Stichproben)

Die Wirksamkeit der experimentellen Manipulation wird durch ein t-test für unabhängige Stichproben überprüft. Die Antworten auf die Manipulationscheck-Fragen werden nach Experimentalgruppen aufgeschlüsselt analysiert. Bei der Positiv-Frame-Gruppe wird erwartet, dass Teilnehmende «Konfidenz/Sicherheit» als Darstellungsform angeben, bei der Negativ-Frame-Gruppe «Unsicherheit/Fehlerwahrscheinlichkeit». Zusätzlich wird mittels einfaktorieller ANOVA getestet, ob sich die wahrgenommene Vertrauenswürdigkeit zwischen den Gruppen bereits im Pretest signifikant unterscheidet. Falls die Manipulation zu schwach ist ($<$ 80% Erkennungsrate), werden Anpassungen am Stimulus vorgenommen.

Hauptanalyse: Deskriptive Statistik

Stichprobenbeschreibung und Gruppenvergleichbarkeit

Die deskriptive Analyse beginnt mit der Charakterisierung der Stichprobe. Für alle demografischen Variablen (Alter, Geschlecht, Bildung) werden Häufigkeiten, Mittelwerte und Standardabweichungen berichtet, sowohl für die Gesamtstichprobe als auch getrennt nach Experimentalgruppen. Die Vergleichbarkeit der randomisierten Gruppen wird mittels unabhängigem t-test und einfaktorieller ANOVAs (kontinuierliche Variablen) überprüft. Signifikante Unterschiede würden auf Randomisierungsprobleme

hinweisen und müssten in späteren Analysen als Kovariaten berücksichtigt werden. Zusätzlich werden die Verteilungen der AI-Vorerfahrung und Nutzungserfahrung zwischen den Gruppen verglichen (potenzielle Konfundierungsvariablen).

Skaleneigenschaften und Verteilungsanalysen

Für alle erhobenen Konstrukte werden Mittelwerte, Standardabweichungen, Schiefe berichtet und mittels Histogrammen visualisiert. Q-Q-Plots visualisieren Abweichungen von der Normalverteilung. Die finale Reliabilität der Skalen wird erneut mit Cronbach's Alpha bestimmt und mit den Pretest-Werten verglichen. Interkorrelationen zwischen allen Konstrukten werden in einer Korrelationsmatrix dargestellt, um erste Hinweise auf Zusammenhänge zu erhalten und Multikollinearität zu identifizieren. Die deskriptiven Statistiken der Trust-Ratings während der Interaktion werden separat für jede Alva-Antwort berichtet, um mögliche Veränderungen über die Interaktionen hinweg zu identifizieren.

Manipulationscheck

Mittels einfaktorieller ANOVA wird getestet, ob sich die wahrgenommene Vertrauenswürdigkeit zwischen den Gruppen signifikant unterscheidet.

Nutzungsverhalten und Systemvariablen

Die während der LLM-Interaktion automatisch erfassten Variablen werden deskriptiv ausgewertet. Die durchschnittliche Anzahl der Interaktionen pro Person und Gruppe wird berichtet, ebenso die Verteilung der tatsächlichen Accuracy Scores. Die Session-Dauer wird analysiert, um Hinweise auf Engagement oder Frustration zu erhalten. Die Komplexität der gestellten Fragen wird kategorisiert (einfach/mittel/komplex) und deren Verteilung zwischen den Gruppen verglichen. Diese Variablen dienen später als Kontrollvariablen in den Hauptanalysen und ermöglichen es, die ökologische Validität der Ergebnisse zu bewerten.

Hauptanalyse: Strukturgleichungsmodellierung (SEM)

Das verwendete AI-TAM-Modell wird mittels Strukturgleiungsmodell (SEM) analysiert (Bollen, 1989). Die Methode ermöglicht die simultane Schätzung aller Hypothesen

(H1–H9) und berücksichtigt dabei die angenommenen Abhängigkeiten zwischen den Konstrukten.

Das Modell umfasst zwei Komponenten: Das Messmodell spezifiziert die Beziehungen zwischen latenten Konstrukten und ihren manifesten Indikatoren. Das Strukturmodell (Abbildung 2) bildet die theoretischen Pfade zwischen den Konstrukten ab. Als exogene Variablen fungieren die Framing-Manipulation (kodiert durch zwei Dummy-Variablen: Dummy_Pos = 1 für positives Framing, 0 sonst; Dummy_Neg = 1 für negatives Framing, 0 sonst; mit der Kontrollgruppe ohne Score-Anzeige als Referenzkategorie), der Accuracy Score (ACTS) sowie die Vertrautheit mit Technologie (FAM-TEC). Als endogene latente Variablen werden AI Output Trust (XAIT), Perceived Usefulness (PUF), Perceived Ease of Use (EOU), Behavioral Intention (BI) und Collaborative Intention (CI) spezifiziert. Tabelle 2 zeigt die Zuordnung der Hypothesen zu den entsprechenden Modellpfaden.

Parameterschätzung und Modell-Fit

Die Parameterschätzung erfolgt mittels Maximum-Likelihood-Verfahren (Bollen, 1989; Kano et al., 1997). Zur Beurteilung des Modellfits werden mehrere Fit-Indizes herangezogen:

- Comparative Fit Index (CFI): Cutoff $\geq .95$ für sehr guten, $\geq .90$ für akzeptablen Fit (Bentler, 1990)
- Tucker-Lewis Index (TLI): Cutoff $\geq .95$ für sehr guten, $\geq .90$ für akzeptablen Fit (Tucker & Lewis, 1973)
- Root Mean Square Error of Approximation (RMSEA): Cutoff $\leq .06$ für sehr guten, $\leq .08$ für akzeptablen Fit (Browne & Cudeck, 1992; Hu & Bentler, 1999; Steiger & Lind, 1980)
- Standardized Root Mean Square Residual (SRMR): Cutoff $\leq .08$ (Bentler, 1995; Jöreskog & Sörbom, 1981)

Evaluation des Messmodells

Zunächst wird das Messmodell mittels konfirmatorischer Faktorenanalyse (CFA) geprüft. Dabei werden die fünf latenten Konstrukte (XAIT, PUF, EOU, BI, CI) ohne strukturelle Pfade analysiert, jedoch mit freien Kovarianzen zwischen allen Konstrukten. Dieses zweistufige Vorgehen (Anderson & Gerbing, 1988) ermöglicht die separate Evaluation der Messqualität, bevor die theoretischen Hypothesen im Strukturmodell getestet werden. Dazu werden drei Kriterien betrachtet.

- Faktorladungen: Items mit standardisierten Ladungen $< .60$ werden für eine theoriegeleitete Modellmodifikation in Betracht gezogen (Gäde et al., 2020)
- Reliabilität: Für jedes Konstrukt wird Cronbach's Alpha berechnet ($\alpha \geq .80$ als akzeptabel; (Nunnally & Bernstein, 1994))
- Validität: Die Korrelationen zwischen den latenten Konstrukten sollten hoch sein ($< .85$)

Strukturmodell und Hypothesentests

Nach Bestätigung des Messmodells wird das vollständige Strukturmodell geschätzt. Die strukturellen Pfade testen die Hypothesen H2–H8 (AI-TAM- und TAM-Standardbeziehungen) sowie H1a und H1b (Framing-Effekte auf XAIT). Alle Pfade werden simultan geschätzt, wodurch die gegenseitigen Abhängigkeiten zwischen den endogenen Variablen berücksichtigt werden. Für jede endogene Variable wird R^2 berichtet, welches den durch die Prädiktoren erklärten Varianzanteil angibt. Die Signifikanz der Pfadkoeffizienten wird mittels z-Test geprüft ($\alpha = .05$, zweiseitig).

Indirekte Effekte und Mediation (H9)

Die Mediationshypothese (H9) besagt, dass der Effekt des Framings auf «Behavioral Intention» durch «Explainable AI Trust» und die nachgelagerten TAM-Konstrukte vermittelt wird. Im vorliegenden Modell existieren multiple indirekte Pfade von Framing zu BI:

- Via PUF: Framing $\rightarrow$ XAIT $\rightarrow$ PUF $\rightarrow$ BI

- Via EOU: Framing → XAIT → EOU → BI
- Via EOU und PUF (seriell): Framing → XAIT → EOU → PUF → BI

Diese indirekten Effekte werden simultan im SEM berechnet. Die Signifikanz wird mittels Bias-Corrected Bootstrap-Verfahren mit 5000 Ziehungen und 95%-Konfidenzintervallen getestet. Separate Effekte werden für beide Framing-Bedingungen (Dummy_Pos, Dummy_Neg) relativ zur Kontrollgruppe berechnet.
Berichtet werden:

- Die spezifischen indirekten Effekte über jeden einzelnen Mediationspfad
- Der totale indirekte Effekt (Summe aller indirekten Pfade)
- Der totale Effekt (Summe aus direktem und indirektem Effekt, sofern ein direkter Pfad Framing → BI spezifiziert wird)

Eine vollständige Mediation liegt vor, wenn kein signifikanter direkter Effekt von Framing auf BI besteht; eine partielle Mediation liegt vor, wenn zusätzlich zu den indirekten Effekten ein signifikanter direkter Effekt verbleibt (Baron & Kenny, 1986).

Kontrolle für Störvariablen

Demografische Variablen (Alter, Geschlecht, Bildungsabschlusswerden als Kovariaten im Modell berücksichtigt.

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 Möglichkeitendie Proband*innen in das Experiment zu holen

  - Footer
  - Button
  - Suchresultate

- Soll das Onboarding zum Experiment vor der errsten 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