Universität Regensburg

Von der Bewertung zur Verifikation: Eine Studie zur Glaubwürdigkeitsbewertung von Nachrichten und den Einsatz von LLMs zur Verifizierung

Bachelorarbeit im Fach Informationswissenschaft

am Institut für Information und Medien, Sprache und Kultur (I:IMSK)

Vorgelegt von:	Katharina Summerer
Adresse:	Deglhof 16, 93142 Maxhütte-Haidhof
E-Mail (Universität):	katharina.summerer@stud.uni-regensburg.de
E-Mail (privat):	katha.summerer@web.de
Matrikelnummer:	1773964
Erstgutachter:	Privatdozent Dr. David Elsweiler
Zweitgutachter:	Professor Dr. Udo Kruschwitz
Betreuer:	Privatdozent Dr. David Elsweiler
Laufendes Semester:	12. Semester, Informationswissenschaft, Medieninformatik, Medienwissenschaft
Abgegeben am:	18.08.2025

Inhaltsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Die zunehmende Verbreitung von Desinformation in digitalen Medien stellt eine zentrale Herausforderung für die Informationsgesellschaft dar. Während Large Language Models (LLMs) wie GPT-4 als vielversprechende Werkzeuge zur Nachrichtenverifikation gelten, ist wenig darüber bekannt, wie Menschen in realen Anwendungssituationen mit diesen Systemen interagieren und welche Auswirkungen dies auf ihre Glaubwürdigkeitsurteile hat.

Die vorliegende Studie untersuchte mittels eines innovativen Mixed-Methods-Designs die Effekte KI-gestützter Verifikationssysteme auf menschliche Glaubwürdigkeitsbewertungen. In einem zweiphasigen Within-Subject-Experiment mit N = 45 Teilnehmenden wurden Think-Aloud-Protokolle mit quantitativen Bewertungsanalysen kombiniert. Die Teilnehmenden bewerteten vier Nachrichtenartikel unterschiedlichem Wahrheitsgehalt zunächst ohne und anschließend mit Unterstützung eines GPT-4-basierten Verifikationssystems.

Die Ergebnisse offenbarten ein überraschendes **Desinformations-Paradox**: KI-gestützte Verifikation führte zu einer systematischen Erhöhung der Glaubwürdigkeit von Falschinformationen (+0,22 Skalenpunkte), während faktisch korrekte Artikel gemischte Bewertungsänderungen zeigten. 76,8% der Teilnehmenden modifizierten ihre ursprünglichen Einschätzungen nach der KI-Interaktion. Die qualitative Analyse von 3.306 Think-Aloud-Segmenten identifizierte eine **5-Stufen-Kaskade** als zugrundeliegenden Mechanismus: Kognitive Überlastung durch komplexe Verifikationsaufgaben führt zu paradoxen Vertrauensdynamiken, intensiven Rationalisierungsprozessen und einer problematischen Format-Inhalt-Verwechslung bei der Bewertung KI-generierter Antworten.

Basierend auf quantitativen Auswertungen, qualitativen Verhaltensmustern und qualitativen Prozessanalysen wurde eine empirisch fundierte **Nutzertypen-Taxonomie** entwickelt: *Skeptische Analytiker* (25%) zeigen intensive KI-Nutzung mit kritischer Reflexion, *Delegierende Vertrauer* (30%) weisen die höchste passive KI-Übernahme auf, *Resistente Intuitive* (20%) vermeiden KI-Nutzung und zeigen stabile Bewertungen, während *Ambivalente Suchende* (25%) problematische Eskalationsdynami-

ken mit verstärkter Unsicherheit entwickeln.

Die Studie leistet sowohl theoretische als auch methodische Beiträge zur Mensch-KI-Interaktionsforschung. Die entwickelte **Kognitive Kaskaden-Theorie** erklärt systematisch, wie gut gemeinte technische Interventionen kontraproduktive Effekte entfalten können. Methodisch demonstriert die Mixed-Methods-Triangulation exemplarisch, wie quantitative Anomalien durch qualitative Mechanismusanalysen aufgelöst werden können.

Die Befunde haben praktische Relevanz für das Design zukünftiger Verifikationssysteme: Statt elaborierter Erklärungen sollten Systeme kognitive Entlastung durch sequenzielle Informationspräsentation bieten. Die Nutzertypen-Taxonomie legt personalisierte Interfaces nahe, während die Erkenntnisse zur Format-Inhalt-Verwechslung die Notwendigkeit transparenter Unsicherheitskommunikation unterstreichen.

Diese Arbeit zeigt, dass die Integration von KI in kritische Informationsprozesse eine fundamentale Neukonzeptualisierung erfordert – weg von technischer Sophistiziertheit hin zu menschenzentrierter Gestaltung.

1. Einleitung

Die digitale Transformation hat die Art und Weise, wie Menschen Nachrichten konsumieren und bewerten, grundlegend verändert (Newman et al., 2019; Tandoc Jr et al., 2018). Während traditionelle Massenmedien über Jahrzehnte als Gatekeeper fungierten und Informationsflüsse durch redaktionelle Kontrolle strukturierten, ermöglichen soziale Medien und digitale Plattformen heute die ungefilterte, schnelle Verbreitung von jeglicher Art von Inhalten (Shoemaker & Vos, 1996; Thorson & Wells, 2016). Diese Entwicklung stellt Nutzer:innen vor neue Herausforderungen: Sie müssen zunehmend selbst die Glaubwürdigkeit von Nachrichten bewerten. Dies ist eine Aufgabe, die nicht nur Wissen, sondern auch erhebliche kognitive und mediale Kompetenzen erfordert (Metzger, 2007; Wineburg & McGrew, 2016).

Die Verbreitung von Desinformation und manipulativen Inhalten hat sich mittlerweile zu einem zentralen gesellschaftlichen Problem entwickelt (Vosoughi et al., 2018; Lewandowsky et al., 2017). Studien zeigen, dass Menschen Falschinformationen häufig nicht zuverlässig erkennen und diese sogar bevorzugt teilen, wenn sie mit bestehenden Überzeugungen übereinstimmen (Pennycook et al., 2019; A. Guess et al., 2019). Viele Nutzer:innen verlassen sich dabei auf heuristische Bewertungskriterien wie Reputation der Quelle oder ihre emotionale Reaktion, anstatt systematische Verifikationsstrategien zu verwenden (Sundar, 2008; Lazer et al., 2018).

In diesem Zusammenhang gewinnen Large Language Models (LLMs) wie GPT-4 als mögliche Werkzeuge zur Unterstützung der Nachrichtenverifikation an Bedeutung (Brown et al., 2020; OpenAI, 2023). Durch die Kombination mit Retrieval-Augmented Generation (RAG) können diese Systeme aktuelle Informationen einbinden und als interaktive Verifikationsassistenten fungieren (Lewis et al., 2020; Gao et al., 2023). Doch während die technischen Möglichkeiten solcher Systeme intensiv erforscht werden, ist wenig darüber bekannt, wie Menschen in realen An-

wendungssituationen mit ihnen interagieren:

- Welche Strategien entwickeln Sie?
- Wie verändert sich ihr Bewertungsprozess?
- Welche Rolle spielt das Vertrauen in die KI-generierten Antworten?

Die vorliegende Arbeit adressierte diese Forschungslücke durch eine Mixed-Methods-Studie, die mittels Think-Aloud-Protokollen die kognitiven Prozesse bei der Nachrichtenbewertung sowohl mit als auch ohne KI-Unterstützung untersucht. Der Fokus liegt dabei nicht nur darauf, ob sich Glaubwürdigkeitsurteile verändern, sondern vor allem darauf, wie Menschen mit einem GPT-4-basierten Verifikationssystem interagieren und welche Denkmuster sich dabei zeigen.

Diese Arbeit ist folgendermaßen strukturiert: Kapitel 2 gibt einen Überblick über den aktuellen Forschungsstand zur Glaubwürdigkeitsbewertung digitaler Nachrichten und zur KI-gestützten Verifikation. Daraus werden dann die zentralen Forschungsfragen abgeleitet. In Kapitel 3 wird das methodische Vorgehen beschrieben, wobei sowohl das Think-Aloud-Verfahren als auch die Implementierung des Chatbots erklärt werden. Die Ergebnisse der qualitativen und quantitativen Analysen finden sich in Kapitel 4. Abschließend diskutiert Kapitel 5 die Befunde vor dem Hintergrund der theoretischen Grundlagen und zeigt Implikationen für Forschung und Praxis auf. Abschließend werden die zentralen Ergebnisse zusammengefasst.

2. Theoretische Grundlagen

2.1. Glaubwürdigkeitsbewertung digitaler Medien

2.1.1. Kognitive Prozesse und Heuristiken

Die Bewertung der Glaubwürdigkeit von Nachrichtenartikeln ist ein komplexer kognitiver Prozess, der von verschiedenen psychologischen Faktoren beeinflusst wird (W. Choi & Stvilia, 2015). Menschen greifen bei der Informationsverarbeitung häufig auf mentale Abkürzungen zurück, sogenannte Heuristiken (Kahneman, 2011). Diese ermöglichen zwar schnelle Entscheidungen bei reduzierter kognitiver Belastung, führen jedoch auch zu systematischen Verzerrungen, die im digitalen Kontext besonders problematisch werden können (Pantazi et al., 2021).

Sundars MAIN-Modell benennt vier technologische Affordanzen, die heuristische Hinweisreize für Glaubwürdigkeitsurteile liefern: *Modality*, *Agency* (*Machine*/-*Source*), *Interactivity* und *Navigability* (Sundar, 2008; Sundar & Limperos, 2013). **Modality**-Heuristiken beziehen sich auf sinnliche Ausprägungen (z. B. Bild, Audio, Video) und können den Eindruck von Realismus und Belegnähe erzeugen („multimediale Inhalte wirken glaubwürdiger"). **Agency**/**Machine**-Heuristiken verschieben die Quelle hin zur „Maschine" (*computer-as-source*); algorithmische Empfehlungen werden als neutraler/objektiver wahrgenommen als menschliche Urteile. **Interactivity**-Heuristiken (z. B. kommentieren, bewerten, anpassen) signalisieren Transparenz/Einbindung und stützen Vertrauen. **Navigability**-Heuristiken (z. B. klare Informationsarchitektur, Such-/Breadcrumb-Führung) stützen Kompetenz- und Qualitätsurteile.

Die Dual-Process-Theorie bietet einen übergeordneten Rahmen zum Verständnis dieser Prozesse (Evans, 2008). Sie unterscheidet zwischen System 1 (schnell, automatisch, intuitiv) und System 2 (langsam, kontrolliert, analytisch). Die meisten

Das Elaboration Likelihood Model als theoretischer Rahmen

Das Elaboration Likelihood Model (kurz ELM) unterscheidet einen **zentrale** (argument- und evidenzbasierte) von einer **peripheren** (cue-basierten) Route der Überzeugung (Petty & Cacioppo, 1986b; Chaiken, 1980).

Welche Route dominiert, hängt von **Motivation** (Relevanz, Verantwortlichkeit) und **Fähigkeit** (Ressourcen, Vorwissen, Zeit) ab und zentrale Verarbeitung erzeugt stabilere, verhaltensrelevantere Urteile (Petty & Cacioppo, 1986b).

Digitale Kontexte (Informationsflut, Zeitdruck, parallele Tasks) senken beides und fördern periphere Heuristiken (z. B. Quellen- oder Konsenssignale, sprachliche / gestalterische Oberflächenmerkmale).

Affordanzen digitaler Systeme können periphere Cues verstärken (*Agency/Machine* als Autoritäts-Heuristik; *Interactivity/Navigability* als Kompetenz-/Transparenzsignale) (Sundar, 2008; Sundar & Limperos, 2013).

Für diese Arbeit dient das ELM als Linse: In den Daten stehen epistemische Strategien (Faktenprüfung, Plausibilitätsabwägung, Kontext). für zentrale, heuristische Bewertungen (Bauchgefühl, Stil, Oberfläche) für periphere Verarbeitung.

2.1.2. Heuristische vs. Systematische Informationsverarbeitung

Die Bewertung von Nachrichtenglaubwürdigkeit folgt den Prinzipien dualer Verarbeitungsmodelle (Chaiken, 1980; Petty & Cacioppo, 1986b). Menschen nutzen zwei grundlegende Strategien zur Glaubwürdigkeitsbewertung:

Systematische Verarbeitung umfasst die detaillierte Analyse von Inhalten, Argumenten und Evidenz. Diese erfordert hohe kognitive Ressourcen und Motivation. Sie manifestiert sich in epistemischen Strategien wie Faktenprüfung und Plausibilitätsabwägung.

Heuristische Verarbeitung basiert auf Oberflächenmerkmalen und mentalen Abkürzungen. Typische Heuristiken bei der Nachrichtenbewertung umfassen:

- *Bauchgefühl*: Intuitive Bewertungen ohne explizite Begründung
- *Sprache/Stil-Heuristik*: Bewertung anhand sprachlicher Oberflächenmerkmale

- *Struktur/Logik-Heuristik*: Schnelle strukturelle Einschätzung ohne Tiefenanalyse

Unter kognitiver Belastung oder Zeitdruck dominiert heuristische Verarbeitung (Metzger, 2007). Dies ist besonders relevant für Online-Nachrichtenkonsumption, wo die schiere Informationsmenge systematische Verarbeitung erschwert.

2.1.3. Herausforderungen im digitalen Kontext

Die digitale Medienlandschaft hat die traditionellen Mechanismen der Glaubwürdigkeitsbewertung grundlegend verändert (Metzger, 2007). Mehrere Schlüsselfaktoren dieser Transformation wurden in der Forschung identifiziert (Tandoc Jr et al., 2018):

Entgrenzung der Gatekeeping-Funktion: Die traditionelle journalistische Gatekeeping-Funktion, die als Qualitätsfilter diente, wurde durch nutzergenerierte Inhalte und algorithmische Kuratierung ersetzt. Dies führt zu einer Vermischung professioneller und amateurhafter Inhalte, wobei die traditionellen Glaubwürdigkeitsmarker ihre Orientierungsfunktion verlieren (Wallace, 2018; Welbers & Opgenhaffen, 2018).

Algorithmische Amplifikation: Vosoughi et al. (2018) zeigten in einer groß angelegten Studie mit 126.000 Twitter-Geschichten, dass sich Falschinformationen etwa sechsmal schneller verbreiten als wahre Nachrichten und dabei 70% wahrscheinlicher retweetet werden. Die Forscher führten dies auf **menschliche Psychologie** zurück, im Besonderen auf die Neuheit falscher Nachrichten und die emotionalen Reaktionen der Nutzer, und nicht primär auf Bots, die wahre und falsche Nachrichten in ähnlichem Tempo verbreiten (Vosoughi et al., 2018). Darauf aufbauend zeigen aktuelle Übersichtsarbeiten eine *algorithmische* Verstärkung in sozialen Netzwerken: PRIME-Merkmale (*Prestigious, Ingroup, Moral, Emotional* als algorithmisch verstärkte soziale Lernmerkmale (Brady, Jackson et al., 2023), zugleich belegt die Forschung eine *Overperception* moralischer Empörung in sozialen Netzwerken, die Feindseligkeitsannahme und Polarisierung verzerren kann (Brady, McLoughlin et al., 2023).

Die algorithmische Verstärkung dieser Dynamik wurde in späteren Studien untersucht: Brady, McLoughlin et al. (2023) zeigen, dass Social-Media-Algorithmen, die auf Engagement-Metriken optimiert sind, bevorzugt sogenannte "PRIME"-Inhalte (Prestigious, Ingroup, Moral, and Emotional) amplifizieren, unabhängig von deren Wahrheitsgehalt. Diese Algorithmen verstärken emotionale und moralisch aufgeladene Inhalte, da diese zu höherem Nutzerengagement führen (Hagey & Horwitz, 2021; Kozyreva et al., 2020). Die Kombination aus menschlichen Verzerrungen (Vosoughi et al., 2018) und algorithmischer Amplifikation (Brady, McLoughlin et al., 2023) schafft somit ein besonders problematisches Umfeld für die Verbreitung von Falschinformationen.

Echo-Kammern und Filterblaseneffekte: Pariser (2011) prägte den Begriff der „Filterblase", um zu beschreiben, wie algorithmische Personalisierung zu einer Verengung des Informationshorizonts führt. Neuere Forschung (A. M. Guess et al., 2021) zeigt jedoch ein differenzierteres Bild: Während Echo-Kammern existieren, sind sie weniger hermetisch als ursprünglich angenommen, und die Exposition gegenüber diversen Nachrichtenquellen bleibt für die meisten Nutzer:innen erhalten.

Kognitive Überlastung: Die schiere Menge an verfügbaren Informationen führt zu einer Überlastung der kognitiven Verarbeitungskapazitäten (Eppler & Mengis, 2004; Roetzel, 2019). Scheufele & Krause (2019) argumentieren, dass diese **information overload** paradoxerweise zu einer verstärkten Abhängigkeit von Heuristiken führt, genau in einer Situation, in der eine systematische Prüfung von besonderer Relevanz wäre.

Besonders problematisch ist die geringe **algorithm awareness** vieler Nutzer:innen (Eslami et al., 2015). Eslami et al. (2015) zeigten, dass über 60 Prozent der Facebook-Nutzer:innen nicht wussten, dass ihr News Feed algorithmisch kuratiert wird. Diese Unkenntnis über die technische Vermittlung von Informationen untergräbt die Fähigkeit zur kritischen Bewertung. Aslett et al. (2022) fanden in einer großangelegten Studie, dass Glaubwürdigkeitslabels nur begrenzte Effekte auf die Nachrichtenkonsumqualität haben, was die Komplexität technischer Interventionen unterstreicht.

2.1.4. Die Rolle von Motivated Reasoning

Das Phänomen des **motivated reasoning** ist einer der zentralen Faktoren, der die Glaubwürdigkeitsbewertung entscheidend beeinflusst (Kunda, 1990). Menschen tendieren dazu, Informationen so zu verarbeiten, dass sie ihre bestehenden Überzeugungen bestätigen. Nyhan & Reifler (2010) demonstrierten den **backfire effect**: Korrekturen von Fehlinformationen können paradoxerweise die falschen Überzeugungen verstärken, wenn sie fundamentalen Weltanschauungen widersprechen.

Diese motivierte Informationsverarbeitung wird im digitalen Kontext durch mehrere Faktoren verstärkt: **Selektive Exposition:** Die Möglichkeit, Informationsquellen aktiv auszuwählen, führt zu einer Bevorzugung bestätigender Informationen (Stroud, 2010). **Soziale Validierung:** Social-Media-Metriken (Likes, Shares) fungieren als soziale Bestätigung und verstärken die Glaubwürdigkeit von Informationen innerhalb der eigenen Bezugsgruppe (Messing & Westwood, 2014). **Affektive Polarisierung:** Die zunehmende emotionale Aufladung politischer Identitäten führt zu einer stärkeren Verzerrung in der Informationsverarbeitung (Iyengar et al., 2019).

2.1.5. Think-Aloud als Methode zur Prozessanalyse

Um diese komplexen kognitiven Prozesse zu erfassen, hat sich die Think-Aloud-Methode als besonders geeignet erwiesen (Ericsson & Simon, 1993). Die theoretischen Grundlagen wurden von Ericsson & Simon (1993) in ihrer bahnbrechenden Arbeit *Protocol Analysis* gelegt. Sie unterscheiden drei Ebenen der Verbalisierung: **Level 1 – Talk-aloud:** Direkte Verbalisierung von Informationen, die bereits verbal enkodiert im Arbeitsgedächtnis vorliegen. **Level 2 – Think-aloud:** Verbalisierung von Gedanken, die eine Rekodierung nicht-verbaler Informationen erfordern. **Level 3 – Retrospective reports:** Nachträgliche Erklärungen und Interpretationen des eigenen Denkprozesses.

Für die Glaubwürdigkeitsforschung ist besonders Level 2 relevant, da hier die tatsächlichen Bewertungsprozesse in Echtzeit erfasst werden. Nielsen (1993) bezeichnete Think-Aloud als „the single most valuable usability engineering method“, da sie direkten Einblick in kognitive Prozesse ermöglicht, die durch nachträgliche Be-

fragung nicht zugänglich wären.

In der Glaubwürdigkeitsforschung wurde die Methode bereits erfolgreich in verschiedenen Kontexten eingesetzt: McGrew et al. (2018) verwendeten Think-Aloud-Protokolle, um zu untersuchen, wie College-Studenten die Glaubwürdigkeit von Websites bewerten. Sie fanden heraus, dass Studierende primär auf oberflächliche Merkmale achten und nur selten systematische Verifikationsstrategien anwenden (McGrew et al., 2018). Die Methode offenbarte auch die Diskrepanz zwischen selbstberichteten und tatsächlichen Strategien: Während Teilnehmer:innen in Befragungen angaben, Quellen zu überprüfen, zeigten die Think-Aloud-Daten, dass dies selten geschah (McGrew et al., 2018).

2.2. KI-gestützte Nachrichtenverifikation

2.2.1. Large Language Models als Verifikationswerkzeuge

Die Entwicklung großer Sprachmodelle (Large Language Models; kurz: LLMs) hat neue Möglichkeiten für die automatisierte Nachrichtenverifikation eröffnet (Vykopal et al., 2024). Brown et al. (2020) demonstrierten mit GPT-3 die beeindruckenden Fähigkeiten dieser Systeme in verschiedenen Sprachverarbeitungsaufgaben, einschließlich der Faktenkontrolle.

Die technischen Fortschritte in der LLM-basierten Verifikation lassen sich in mehrere Entwicklungslinien unterteilen:

Retrieval-Augmented Generation (RAG): Lewis et al. (2020) entwickelten RAG als Methode, die die Generierungsfähigkeiten von LLMs mit externem Wissensabruf kombiniert. Dies adressiert eine zentrale Schwäche reiner Sprachmodelle: ihre Beschränkung auf Trainingsdaten und die Tendenz zu „Halluzinationen“. RAG-Systeme können aktiv auf aktuelle Informationsquellen zugreifen und diese in ihre Antworten integrieren, was für die Nachrichtenverifikation essenziell ist (Lewis et al., 2020; Khaliq et al., 2024).

Fine-Tuning für Faktenkontrolle: Mehrere Studien haben untersucht, wie LLMs speziell für Fact-Checking-Aufgaben optimiert werden können (Thorne et al., 2018; Tian et al., 2023; Setty et al., 2024; Jiang et al., 2020; Chen et al., 2022). E. C. Choi

et al. (2023) entwickelten FACT-GPT, ein System, das auf synthetischen Datensätzen trainiert wurde, um verwandte Claims zu identifizieren. Ihre Evaluation zeigte, dass spezialisierte kleinere Modelle die Genauigkeit größerer Modelle bei der Identifikation verwandter Behauptungen erreichen können.

Multilinguale Verifikation: Quelle & Bovet (2024) untersuchten die Fähigkeiten von GPT-3.5 und GPT-4 bei der Faktenkontrolle in verschiedenen Sprachen. Sie fanden signifikante Leistungsunterschiede zwischen Sprachen, wobei die Modelle bei nicht-englischen Texten deutlich schlechter abschnitten. Dies unterstreicht die Herausforderungen globaler Desinformationsbekämpfung.

Die empirische Leistungsfähigkeit von LLMs in der Faktenkontrolle wurde in mehreren groß angelegten Studien untersucht: Caramancion (2023) testeten verschiedene LLMs (Bard, Bing-AI, GPT-3.5, GPT-4) an 100 faktengeprüften Nachrichtenelementen. Die Modelle erreichten Genauigkeiten zwischen 64-71%. Huang et al. (2024) analysierten die „duale Rolle"von LLMs: Während sie zur Erstellung überzeugender Desinformation verwendet werden können, bieten sie auch Möglichkeiten zur Detektion. Ihre Experimente zeigten, dass LLM-generierte Desinformation für Menschen schwerer zu erkennen ist als traditionelle Falschinformationen.

Herausforderungen und Limitationen: Die Forschung hat mehrere kritische Limitationen von LLMs in der Nachrichtenverifikation identifiziert: **Halluzinationen:** LLMs generieren häufig plausible, aber faktisch falsche Informationen. Ji et al. (2023) kategorisierten verschiedene Typen von Halluzinationen und ihre Ursachen, einschließlich Trainingsdatenverzerrungen und Dekodierungsstrategien. **Bias-Reproduktion:** Santurkar et al. (2023) dokumentierten politische Verzerrungen in LLMs, mit einer Tendenz zu linksliberalen Positionen. Diese Verzerrungen können die Neutralität der Faktenkontrolle untergraben. **Temporale Limitationen:** LLMs haben einen festen Wissensstichtag und können ohne externe Quellen keine aktuellen Ereignisse verifizieren (Quelle & Bovet, 2024; Dhingra et al., 2022). **Kontextuelle Unsicherheit:** Mirza et al. (2024) zeigten, dass LLMs Schwierigkeiten haben, mit Aussagen umzugehen, die Unsicherheit ausdrücken oder modalisierende Elemente enthalten.

2.2.2. Der Aslett-Effekt: Paradoxe Wirkungen der Online-Verifikation

Eine der bedeutendsten Erkenntnisse der jüngeren Forschung ist die Studie von Aslett et al. (2023) in *Nature*, die zeigt, dass Online-Suchen zur Bewertung von Fehlinformationen paradoxerweise deren wahrgenommene Glaubwürdigkeit erhöhen können. Diese Arbeit stellt fundamentale Annahmen über digitale Medienkompetenz in Frage.

Die Studie umfasste fünf Experimente mit über 7.000 Bewertungen. Die zentrale Methodik bestand darin, Teilnehmer:innen zu bitten, die Glaubwürdigkeit von Nachrichtenartikeln zu bewerten, dann online nach Informationen zu suchen und anschließend erneut zu bewerten. Die Ergebnisse waren überraschend: 1. Die Online-Suche erhöhte die Glaubwürdigkeit falscher Nachrichten um durchschnittlich 19%. 2. Dieser Effekt trat unabhängig von digitaler Kompetenz auf. 3. Die Exposition gegenüber schwachen Korrekturen verstärkte die Fehlinformationen. 3. Suchmaschinen-Rankings und die Qualität der gefundenen Informationen moderierten den Effekt

Die Autoren identifizierten mehrere Mechanismen für diesen paradoxen Effekt: 1. **Data voids:** Zu manchen Falschinformationen existieren wenige hochwertige Korrekturen, wodurch Nutzer:innen primär auf die Originalquellen stoßen. 2. **Illusorische Wahrheit:** Die wiederholte Exposition gegenüber Falschinformationen während der Suche erhöht deren Vertrautheit und damit die wahrgenommene Glaubwürdigkeit. 3. **Motivated search:** Menschen mit Voreinstellungen suchen selektiv nach bestätigenden Informationen.4. **Schwache Korrekturen:** Unklare oder technische Faktenchecks können Zweifel an der Korrektur selbst wecken.

Aslett et al. (2023) Erkenntnisse haben wichtige Implikationen für die Integration von KI-Verifikationssystemen. Da selbst aktive Recherche zu verschlechterten Urteilen führen kann, müssen KI-Systeme sorgfältig gestaltet werden, um nicht ähnliche kontraproduktive Effekte zu erzeugen.

2.2.3. Einfluss von KI-Systemen auf menschliche Entscheidungsprozesse

Die Integration von KI in menschliche Entscheidungsprozesse ist ein aktives Forschungsfeld mit wichtigen Erkenntnissen für die Nachrichtenverifikation und zeigt

komplexe und oft überraschende Interaktionsmuster zwischen Menschen und KI-Systemen (Buçinca et al., 2021; Lee & See, 2004; Vereschak et al., 2024; Liu et al., 2021).

Vertrauen und Überabhängigkeit: Biran & Cotton (2017) untersuchten den Einfluss von Erklärungen auf das Vertrauen in KI-Empfehlungen. Sie fanden heraus, dass Erklärungen zwar das Vertrauen erhöhen, aber nicht notwendigerweise zu besseren Entscheidungen führen. Dieses Phänomen der **overtrust** ist besonders problematisch im Kontext der Nachrichtenverifikation.

DeVerna et al. (2024) führten ein großangelegtes Experiment durch, um die Auswirkungen von ChatGPT-generierten Faktenchecks auf die Glaubwürdigkeitsbewertung zu untersuchen. Obwohl das LLM 90% der Falschinformationen korrekt identifizierte, zeigte sich ein beunruhigender Nebeneffekt: Die Diskriminierungsfähigkeit der Nutzer:innen zwischen wahren und falschen Schlagzeilen sank signifikant.

Confirmation Bias Verstärkung: Schaffer et al. (2021) beobachteten, dass Menschen dazu neigen, KI-Empfehlungen besonders dann unkritisch zu übernehmen, wenn diese ihre Voreinstellungen bestätigen. Dies deutet darauf hin, dass KI-Systeme bestehende kognitive Verzerrungen verstärken können, anstatt sie zu korrigieren.

Kognitive Auslagerung: Ein besorgniserregendes Phänomen ist die **cognitive offloading**, die Tendenz, Denkprozesse an KI-Systeme zu delegieren. Storm & Hickman (2023) zeigten, dass die Verfügbarkeit von KI-Assistenten die eigene kritische Denkbereitschaft reduzieren kann. Im Kontext der Nachrichtenverifikation könnte dies zu einer Erosion der individuellen Medienkompetenz führen.

Positive Potenziale: Trotz dieser Herausforderungen zeigen einige Studien auch positive Potenziale: Yang et al. (2019) fanden heraus, dass transparente KI-Systeme, die ihre Unsicherheit kommunizieren, zu reflektierteren menschlichen Urteilen führen können. Die explizite Darstellung von Konfidenzintervallen und alternativen Interpretationen förderte eine kritischere Auseinandersetzung mit den Informationen.

2.2.4. Systemdesign und Interaktionsparadigmen

Die Art und Weise, wie KI-Verifikationssysteme gestaltet sind, hat erheblichen Einfluss auf ihre Wirksamkeit und die Qualität der Mensch-KI-Interaktion (Dellermann et al., 2021). Mehrere Designprinzipien haben sich als relevant erwiesen:

Transparenz und Erklärbarkeit: Wang et al. (2021) untersuchten verschiedene Erklärungsansätze für KI-basierte Faktenchecks. Sie fanden heraus, dass prozessorientierte Erklärungen (wie das System zu seiner Schlussfolgerung kam) effektiver waren als ergebnisorientierte Erklärungen (was das System gefunden hat).

Interaktivität: Schaffer et al. (2023) verglichen passive KI-Empfehlungen mit interaktiven Systemen, bei denen Nutzer:innen Fragen stellen und Klarstellungen anfordern konnten. Die interaktiven Systeme führten zu besseren Verifikationsergebnissen und höherer Nutzerzufriedenheit.

Unsicherheitskommunikation: Die Art, wie KI-Systeme Unsicherheit kommunizieren, beeinflusst maßgeblich ihre Wirkung. Bhatt et al. (2021) entwickelten ein Framework für die Kommunikation verschiedener Unsicherheitstypen (aleatorisch vs. epistemisch) und zeigten, dass differenzierte Unsicherheitskommunikation das Vertrauen in KI-Empfehlungen verbessert.

2.3. Forschungsfragen dieser Arbeit

Die Synthese der vorgestellten Literaturstränge offenbart eine Forschungslücke: Während die Bewertung von Nachrichtenglaubwürdigkeit einerseits und die technischen Möglichkeiten von KI-Systemen andererseits gut erforscht sind, fehlt ein tiefgehendes Verständnis der Prozesse, die bei der Integration beider Bereiche entstehen. Diese Arbeit adressiert diese Lücke durch die Untersuchung folgender Forschungsfragen.

RQ1: Wie bewerten Nutzer:innen die Glaubwürdigkeit von Nachrichtenartikeln ohne KI-Unterstützung, und auf welcher Grundlage treffen sie diese Einschätzung?

RQ2: Wie verifizieren Nutzer:innen Nachrichtenartikel mit Hilfe eines GPT-4-basierten Systems, und welche typischen Verhaltensmuster oder Strategien zei-

gen sich dabei?

RQ3: Inwiefern verändern sich die Glaubwürdigkeitseinschätzungen nach der KI-gestützten Verifikation, und welche Faktoren beeinflussen eine mögliche Veränderung?

2.4. Reflexivitätsnotiz

Als Forscherin bin ich mir meiner eigenen Position im Forschungsprozess bewusst. Meine technische Ausbildung und kritische Haltung gegenüber unreflektierter KI-Implementierung prägten sowohl die Fragestellung als auch die Interpretation der Ergebnisse. Die Entscheidung für einen Mixed-Methods-Ansatz entsprang der Überzeugung, dass technologische Phänomene nicht losgelöst von ihren sozialen Kontexten verstanden werden können.

Während der Think-Aloud-Protokolle wurde deutlich, dass meine Anwesenheit die Verbalisierung beeinflusste – Teilnehmende rechtfertigten ihre Entscheidungen ausführlicher als in natürlichen Situationen. Dies interpretierten wir jedoch als Vorteil, da es tiefere Einblicke in die Bewertungsprozesse ermöglichte.

Besonders herausfordernd war meine eigene Überraschung über das Desinformations-Paradox. Initial erwartete ich eine eindeutige Verbesserung durch KI-Unterstützung. Die gegenteiligen Befunde zwangen mich, meine technikoptimistischen Annahmen zu hinterfragen und die Komplexität menschlicher Vertrauensbildung anzuerkennen (Flick, 2019).

3. Methodik

Die Studie nutzt einen explorativen Mixed-Methods-Ansatz, der Think-Aloud-Prozessdaten, qualitative Daten mit quantitativen Bewertungsänderungen kombiniert.

3.1. Zielsetzung und Forschungslogik

3.1.1. Methodologische Positionierung

Die methodische Konzeption basiert auf der Integration qualitativer und quantitativer Forschungsparadigmen im Sinne eines pragmatischen Forschungsansatzes (Creswell & Creswell, 2017). Diese Entscheidung resultiert aus der Erkenntnis, dass Glaubwürdigkeitsbewertung sowohl messbare Outcomes (Bewertungsänderungen) als auch komplexe kognitive Prozesse (Bewertungsstrategien) umfasst, die unterschiedliche methodische Zugänge erfordern.

Im Zentrum der Untersuchung steht die Think-Aloud-Methode, die als etabliertes Verfahren zur Erfassung kognitiver Prozesse in Echtzeit gilt (Ericsson & Simon, 1993). Diese Methodenwahl adressiert direkt die im Related Work identifizierte Prozesslücke: Während quantitative Studien zur Glaubwürdigkeitsbewertung dominieren (Metzger, 2007), fehlen detaillierte Prozessanalysen der zugrundeliegenden kognitiven Mechanismen. Die Think-Aloud-Methode ermöglicht es, über die bloße Messung von Bewertungsänderungen hinauszugehen und die mentalen Prozesse, Heuristiken und Strategien zu rekonstruieren, die Menschen bei der Glaubwürdigkeitsbewertung anwenden.

Die theoretische Prämisse basiert auf der Erkenntnis, dass Glaubwürdigkeitsurteile komplexe kognitive Vorgänge darstellen, die durch das Zusammenspiel von System-1- und System-2-Prozessen charakterisiert sind (Kahneman, 2011) und nicht allein durch standardisierte Messungen erfasst werden können.

3.1.2. Mixed-Methods-Design und methodische Triangulation

Die Studie folgt einem „explanatory sequential Mixed-Methods-Design“, in dem **qualitative** Daten zur Erklärung und Vertiefung vorab erhobener **quantitativer** Befunde genutzt werden (Ivankova et al., 2006; Creswell & Plano Clark, 2017). Diese Anlage ist geeignet, Mechanismen hinter Bewertungsänderungen sichtbar zu machen, da Prozessdaten (Think-Aloud, Chat-Interaktion) systematisch an Outcome-Daten (Likert-Ratings) rückgebunden werden (Ivankova et al., 2006; Creswell & Plano Clark, 2017). Methodische Triangulation (Daten-, Methoden- und Theorie-Triangulation) wird eingesetzt, um die Befundvalidität zu erhöhen und Einflüsse einzelner Erhebungsverfahren abzufedern (Denzin, 2017). Zur Integration werden Joint Displays und verknüpfte Analyseschritte genutzt, die qualitative Codes mit quantitativen Differenzwerten zusammenführen (Creswell & Plano Clark, 2017).

Die Integration von Think-Aloud-Protokollen mit KI-Chatprotokollen stellt einen methodischen Ansatz dar, der besonders für die Erforschung der Mensch-KI-Interaktion geeignet ist. Die doppelte Verbalisierung, zunächst als Selbstgespräch während der natürlichen Bewertung, später im Dialog mit der KI, ermöglicht einen Einblick in kognitive Anpassungsprozesse.

3.2. Studiendesign

3.2.1. Zweigruppen-Ansatz

Die Studie umfasst zwei komplementäre Erhebungsgruppen, die unterschiedliche methodische Stärken kombinieren:

Think-Aloud-Gruppe (n = 16): Diese Gruppe bildet den qualitativen Kern der Studie. Die Gruppengröße orientiert sich an Empfehlungen für qualitative Studien, ist jedoch durch die zeitliche Begrenzung der Bearbeitung einer Bachelorarbeit limitiert. Guest et al. (2006) zeigten, dass bei homogenen Stichproben typischerweise nach 12 Interviews Datensättigung eintritt. Die gewählte Stichprobengröße bietet einen angemessenen Puffer für heterogenere Verarbeitungsstrategien.

Online-Gruppe (n = 29): Diese Gruppe dient der Validierung und Erweiterung

der Befunde unter natürlicheren Bedingungen ohne die potenzielle Reaktivität des lauten Denkens. Die Durchführung adressiert eine zentrale Limitation der Think-Aloud-Methode, da Ericsson & Simon (1993) diskutieren, dass Verbalisierung kognitive Prozesse verlangsamen oder verändern kann.

3.2.2. Within-Subject-Design

Für beide Gruppen wurde ein Within-Subject-Design implementiert, bei dem alle Teilnehmenden dieselben vier Artikel in zwei aufeinanderfolgenden Phasen bewerteten: **Phase 1:** Baseline-Bewertung ohne KI-Unterstützung; **Phase 2:** Erneute Bewertung mit KI-Unterstützung.

Diese Designentscheidung basiert auf mehreren methodologischen Überlegungen (Kirk, 2013; Charness et al., 2012):

Kontrolle individueller Differenzen: Glaubwürdigkeitsbewertungen werden von stabilen Personenmerkmalen und dispositionsbezogenen Faktoren wie analytischem Denken, politischer Orientierung, Medien- und Informationskompetenz sowie epistemischen Überzeugungen beeinflusst (Pennycook & Rand, 2019b; Ecker et al., 2022; Vraga & Tully, 2021; Sinatra et al., 2014). Im Within-Subject-Design fungiert jede Person als eigene Kontrolle, wodurch interindividuelle Unterschiede als Störfaktoren minimiert werden (Charness et al., 2012; Kirk, 2013). Dies erhöht die statistische Power gegenüber Between-Subject-Vergleichen deutlich (Charness et al., 2012).

Prozessbeobachtung: Die Messung intraindividueller Veränderungen zwischen erster und zweiter Bewertung ist im wiederholte-Messungen-Design (repeated measures) methodisch direkt zugänglich und erlaubt die Schätzung von Behandlungseffekten ohne Konfundierung durch stabile Personenunterschiede (Field, 2018; Kirk, 2013). Damit kann der spezifische Einfluss der KI-Unterstützung auf Glaubwürdigkeitsurteile als Differenz innerhalb derselben Personen analysiert werden (Field, 2018). **Ökologische Validität:** Im realen Nutzungskontext erfolgt Nachrichtenrezeption häufig als spontane Erstbewertung mit anschließender, optional nachgelagerter Prüfung mittels Recherche- oder Verifikationsstrategien wie **lateral reading**, sodass das gewählte Sequenz-Design diese natürliche Abfolge abbildet (Wineburg

& McGrew, 2019; Breakstone et al., 2021).

Aktuelle Metadiskussionen fordern zudem höhere ökologische Validität in der Desinformationsforschung und realitätsnahe Stimuli und Tasks (Crum et al., 2024). Studien mit vollständigen Nachrichtenartikeln statt nur Headlines zeigen, dass analytisches Denken und Quellen-Glaubwürdigkeit die Bewertung echter und falscher Inhalte unter alltagsnäheren Bedingungen systematisch mitbestimmen, was für naturalistischere Stimuli im Design spricht (Pehlivanoglu et al., 2021).

Kontrolle von Reihenfolgeeffekten: Um Positions- und Reihenfolgeeffekte zu kontrollieren, wurde ein Latin-Square-Design implementiert (Kirk, 2013; Winer et al., 1991). Die vier Artikel wurden systematisch rotiert, sodass über die Gesamtstichprobe hinweg jeder Artikel gleich häufig an jeder Position (1., 2., 3., 4.) präsentiert wurde. Dies gewährleistet, dass eventuelle Ermüdungseffekte oder Lerneffekte gleichmäßig über alle Artikel verteilt sind, ohne die kognitive Belastung einer vollständigen Randomisierung für jeden einzelnen Teilnehmenden zu erzeugen (Kirk, 2013; Winer et al., 1991).

3.3. Stichprobe und Rekrutierung

3.3.1. Geplante Stichprobe und Stichprobenumfang

Die Studie zielte auf zwei komplementäre Stichproben ab, deren Größe auf methodologischen Überlegungen und praktischen Erwägungen basierte:

Think-Aloud-Gruppe (geplant: n = 15): Die angestrebte Stichprobengröße von 15 Teilnehmenden orientierte sich an etablierten Empfehlungen für qualitative Studien mit Think-Aloud-Methodik. Nielsen (1993) argumentiert, dass 10-15 Teilnehmende ausreichen, um 85-95% aller Usability-Probleme zu identifizieren. Guest et al. (2006) zeigten zudem, dass bei homogenen Stichproben typischerweise nach 12 Interviews thematische Sättigung eintritt. Die gewählte Zielgröße von 15 bot somit einen angemessenen Puffer für heterogenere Verarbeitungsstrategien und mögliche Ausfälle, während sie gleichzeitig im Rahmen der zeitlichen und finanziellen Ressourcen einer Bachelorarbeit realisierbar blieb.

Online-Gruppe (geplant: n = 30-40): Für die Online-Erhebung wurde eine grö-

ßere Stichprobe von 30-40 Teilnehmenden angestrebt. Diese Größe ermöglicht: Ausreichende statistische Power für non-parametrische Verfahren bei Within-Subject-Designs (Charness et al., 2012); Validierung der qualitativen Befunde unter natürlicheren Bedingungen ohne Reaktivität des lauten Denkens; Erhöhte Generalisierbarkeit durch größere Varianz in demografischen Merkmalen. Kompensation potenzieller Datenverluste durch technische Probleme oder unvollständige Bearbeitungen.

Die Gesamtstichprobe von 45-55 Teilnehmenden wurde als ausreichend erachtet, um sowohl tiefe qualitative Einblicke als auch robuste quantitative Muster zu identifizieren, während sie im Rahmen einer Bachelorarbeit praktikabel blieb.

3.3.2. Rekrutierungsstrategie

Die Rekrutierung erfolgte über einen Zeitraum von sechs Wochen (23. Juni bis 03. August 2025) mittels purposive sampling kombiniert mit Schneeballverfahren. Diese Strategie zielte darauf ab, eine heterogene Stichprobe hinsichtlich Alter, Bildungshintergrund und Medienkompetenz zu erreichen:

Primäre Rekrutierungskanäle: Universitäre Verteiler: Ansprache von Studierenden der Informationswissenschaft und Medieninformatik zur Sicherstellung technisch versierter Teilnehmender; **Persönliche Netzwerke:** Kontrolliertes Schneeballverfahren zur Erreichung diverser Altersgruppen und Bildungshintergründe; **Erweiterte soziale Kreise:** Einbeziehung von Familie, Freund:innen und Arbeitskolleg:innen zur Erhöhung der demografischen Heterogenität

Einschlusskriterien: Das Mindestalter lag bei 18 Jahren. Es müssten ausreichende Deutschkenntnisse für Textverständnis und Verbalisierung vorhanden sein und grundlegende Internetkompetenz für die Online-Teilnahme. Ausgeschlossen wurden Personen mit professioneller Tätigkeit im Bereich Journalismus oder Fact-Checking.

Incentivierung: Zur Erhöhung der Teilnahmebereitschaft wurde für die Think-Aloud-Gruppe und die Online-Gruppe mindestens 1 VP-Stunde angeboten, aber die Think-Aloud-Gruppe bekommt mehr. Diese differenzierte Incentivierung reflektiert den unterschiedlichen zeitlichen Aufwand (90 vs. 60 Minuten).

Die Rekrutierungsstrategie erwies sich als erfolgreich: Die finale Stichprobe ist mit N = 45 (16 Think-Aloud, 29 Online) genau die geplanten Zahlen, was die statistische Power erhöhte und die Robustheit der Befunde stärkte. Die tatsächliche demografische Zusammensetzung wird im Ergebnisteil detailliert dargestellt.

3.4. Material und technische Umsetzung

3.4.1. Nachrichtenartikel

Folgende vier Nachrichtenartikel wurden ausgewählt: **Europäische „Eliten"wollen jeden Dissens zerquetschen – Zum Urteil gegen Marine Le Pen** (Danckwardt, o. J.): Mittlere bis niedrige erwartete Glaubwürdigkeit aufgrund einseitiger Darstellung. **Russland ist ein Friedensstifter. Von Armenien, über Berlin bis zur Ukraine** (News Front, o. J.): Niedrige erwartete Glaubwürdigkeit aufgrund fragwürdiger Darstellung. **Ein Tribunal gegen den Aggressor** (Janisch, o. J.): Hohe erwartete Glaubwürdigkeit aufgrund sachlicher Darstellung.**Was der Aggressor anrichtet** (Varga, o. J.): Hohe erwartete Glaubwürdigkeit aufgrund institutioneller Thematik.

Der Artikel „Was der Aggressor anrichtet"ist aus der FAZ online und dort online abrufbar (Varga, o. J.), der Artikel"Ein Tribunal gegen den Aggressor"wurde in der Süddeutschen Zeitung online veröffentlicht und online abrufbar (Janisch, o. J.). Diese beiden Artikel wurden in der Studie als glaubwürdige Artikel verwendet. Die beiden Artikel mit irreführenden Informationen bzw. Fake News wurden über die Seite EU vs Disinfo (https://euvsdisinfo.eu/) gefunden und sind im Fall von „Europäische „Eliten"wollen jeden Dissens zerquetschen – Zum Urteil gegen Marine Le Pen"auf RT DE (Danckwardt, o. J.) und bei „Russland ist ein Friedensstifter. Von Armenien, über Berlin bis zur Ukraine"News Front (News Front, o. J.) abrufbar.

Auswahlkriterien: **Wahrheitsgehalt**: Zwei faktisch korrekte (von großen deutschen Verlagen seriöser Zeitungen) und zwei als Desinformation klassifizierte Artikel, die auf der Seite https://euvsdisinfo.eu/ gefunden wurden. **Thematische Kohärenz**: Alle Artikel behandeln internationale politische Themen. **Strukturelle Kontrolle**: Ähnliche Längen (800 bis 1200 Wörter) und Komplexitätsniveaus. **Präsenta-**

tion: Um inhaltliche Glaubwürdigkeitsurteile nicht durch *Source Cues* (z. B. Marken-/Vertrauensvorschuss bekannter Medien) oder visuelle Heuristiken zu verzerren, wurden alle Quellenhinweise entfernt und die Stimuli in einem neutralen Layout präsentiert (Hovland & Weiss, 1951; Sundar, 2008; Metzger, 2007). Damit wird die periphere Beeinflussung über Design- und Formatmerkmale (z. B. „professioneller Look") minimiert, die bei gleichem Informationsgehalt die wahrgenommene Glaubwürdigkeit nachweislich erhöhen können (Fogg, 2003; Sillence et al., 2007; Petty & Cacioppo, 1986a). Die Standardisierung der Darstellung erhöht die interne Validität, da konfundierende Variablen (Quelle, Branding, Layoutqualität) kontrolliert und Effekte dem *Inhalt* bzw. der KI-Unterstützung zugerechnet werden können (Shadish et al., 2002; Kirk, 2013). Gleichzeitig wird der bekannte Trade-off zur ökologischen Validität transparent; die neutrale Präsentation dient hier primär der kausalen Attribution, während Realitätsnähe in ergänzenden Designs/Analysen adressiert werden kann (Shadish et al., 2002; Pehlivanoglu et al., 2021).

3.4.2. KI-gestütztes Verifikationssystem

Das implementierte Verifikationssystem nutzt GPT-4 als zentrale Entscheidungs- und Synthesekomponente (OpenAI, 2023). Die Orchestrierung erfolgt serverlos über Cloudflare Workers, Persistenz über die Cloudflare D1-Datenbank, in der alle Konversationen (User-ID, Artikel-ID, Timestamps, Turn-Inhalte) gespeichert werden (Cloudflare, 2025b,a). Funktional arbeitet das System RAG-inspiriert: Es kombiniert modellinterne Sprachkompetenz mit externer evidenzbasierter Recherche, jedoch ohne eigenen Vektorindex; stattdessen werden API-basierte Retrieval-Quellen parallel abgefragt (Lewis et al., 2020; Gao et al., 2023). Die Entscheidungslogik wird über Function Calling/Tool-Calling realisiert: GPT-4 klassifiziert eingehende Prompts und triggert bei Aktualitäts-/Faktenbezug die News-/Web-Suche; ansonsten antwortet das Modell direkt (OpenAI, 2025).

Bei ausgelöster Recherche fragt das System parallel fünf Quellen an, um Redundanz, Abdeckungsbreite (deutsch/englisch) und Faktensicherheit zu maximieren: Brave Search API (Hauptquelle; News und Web), NewsAPI (Live-Artikelaggregator),

The Guardian Open Platform (qualitätsgesicherter Korpus), Google Fact Check Tools API (professionelle Faktenchecks), sowie zusätzliche Brave Web-Ergebnisse für Primärquellen (Brave, 2025; NewsAPI, 2025; Guardian News & Media, 2025; Google, 2025). Die parallele Abfrage liefert typischerweise ungefähr 25–30 Kandidatenquellen (z. B. 10 News + 5 Web von Brave, 5 NewsAPI, 5 Guardian, bis zu 5 Fact-Checks), die in einer nach Evidenzpriorität gewichteten Synthese zusammengeführt werden (Priorität: Fact-Checks > Qualitätsmedien > übrige Web-Quellen) (Google, 2025; NewsAPI, 2025; Guardian News & Media, 2025; Brave, 2025). Das Antworttemplate des Systems zitiert alle verwendeten Quellen und weist Unsicherheiten explizit aus (z. B. bei unvollständiger Quellenlage oder Kontradiktionen) (OpenAI, 2023).

Die Wahl dieses Setups ist zweckmäßig für eine Bachelorstudie: Es bietet Transparenz (vollständige Quellenauflistung im Turn), Robustheit (Ausfall einzelner APIs beeinträchtigt die Gesamtfunktion geringer) und Geschwindigkeit durch Parallelisierung in einer Edge-Umgebung (Cloudflare, 2025b; Brave, 2025; NewsAPI, 2025).

Zugleich bleibt die Architektur kosten- und wartungsarm, da kein eigener Index/Vectorstore aufgebaut und gepflegt werden muss (Lewis et al., 2020; Gao et al., 2023).

3.4.3. Webbasierte Studienplattform

Die technische Realisierung erfolgte über eine eigens entwickelte webbasierte Plattform, die nach Prinzipien des User-Centered Design gestaltet wurde. Die technische Architektur basiert auf modernen Webstandards mit React.js Frontend, Node.js Backend und PostgreSQL Datenbank.

3.5. Datenerhebung

3.5.1. Studienprotokoll

Jede Sitzung der Think-Aloud-Gruppe dauerte 75-90 Minuten und folgte einem standardisierten Protokoll:

Phase 0: Einführung und Informed Consent (10-15 Minuten) Die ausführliche

Aufklärung umfasste Informationen über Studienablauf, Datenverwendung und DSGVO-konforme Behandlung. Für die Think-Aloud-Gruppe erfolgte eine Übungsaufgabe mit einem neutralen Text zur Gewöhnung an die Verbalisierung.

Phase 1: Baseline-Bewertung ohne KI-Unterstützung (30-40 Minuten) Die vier Artikel wurden in randomisierter Reihenfolge präsentiert. Teilnehmende lasen jeden Artikel bei kontinuierlicher Think-Aloud-Verbalisierung und gaben anschließend eine Glaubwürdigkeitsbewertung auf einer 7-Punkt-Likert-Skala ab.

Die Think-Aloud-Protokollierung erfolgte nach etablierten Standards: Minimale Intervention des Versuchsleiters; Standardisierter Prompt „Bitte laut denken "bei langen Pausen; Keine inhaltlichen Nachfragen während der Aufgabe; Vollständige Audioaufzeichnungen und Bildschirmaufnahmen.

Phase 2: KI-gestützte Verifikation und Neu-Bewertung (25-35 Minuten) Nach einer Einführung in die Chatbot-Funktionalität erfolgte die freie Interaktion mit dem System. Die Interaktion wurde vollständig dokumentiert mit allen Nutzeranfragen, Systemantworten und Reaktionszeiten.

Phase 3: Abschluss und Reflexion (10 Minuten) Ein semi-strukturiertes Interview erfasste Reflexionen über den Verifikationsprozess. Zusätzlich wurden etablierte Skalen eingesetzt: 1. Cognitive Reflection Test (Frederick, 2005), 2. Need for Cognition Scale (Cacioppo & Petty, 1982), 3. Adaptierte KI-Vertrauensskala.

Die letzte Phase wurde jedoch nicht mehr aufgezeichnet und ist somit kein Teil der annotierten Transkripte und der Auswertung der Daten.

3.5.2. Online-Gruppe

Die Online-Gruppe durchlief dasselbe Protokoll ohne Think-Aloud-Komponente über die webbasierte Plattform, in Form der Website https://vonderbewertungzurverifikation.de/. Dies ermöglichte eine Kontrolle für potenzielle Reaktivitätseffekte der Verbalisierung.

3.5.3. Datenaufbereitung und Analyse

Die Transkription der Think-Aloud-Protokolle erfolgte nach dem Prinzip der Gebrauchstranskription (Dresing & Pehl, 2015). Da der Fokus auf inhaltlichen Aspekten lag, wurde wörtlich in Standardorthografie transkribiert. Unverständliche Passagen wurden markiert, Pausen über 3 Sekunden und Unsicherheitsmarker dokumentiert.

3.5.4. Theoretischer Orientierungsrahmen ohne Hypothesentestung

Das Elaboration Likelihood Model als analytische Linse

Das Elaboration Likelihood Model (Petty & Cacioppo, 1986b) dient in dieser explorativen Studie als theoretischer Orientierungsrahmen zur Einordnung der beobachteten Bewertungsprozesse. Es werden keine präregistrierten Hypothesen getestet, sondern das Modell fungiert als analytische Linse zur systematischen Interpretation der empirischen Befunde (Petty & Cacioppo, 1986b).

Das ELM unterscheidet zwischen zwei fundamentalen Verarbeitungsrouten: Die **zentrale Route** umfasst die elaborierte, argumentbasierte Auseinandersetzung mit Inhalten, bei der Evidenzqualität und logische Kohärenz im Vordergrund stehen (Petty & Cacioppo, 1986b). Die **periphere Route** hingegen basiert auf mentalen Abkürzungen und oberflächlichen Hinweisreizen wie Quellenautorität, sozialer Konsens oder Gestaltungsmerkmalen (Petty & Cacioppo, 1986b; Chaiken, 1980).

Die Wahl der Verarbeitungsroute wird durch Motivation und Fähigkeit zur kognitiven Elaboration bestimmt; zentrale Verarbeitung führt typischerweise zu stabileren und verhaltensrelevanteren Einstellungen, während periphere Verarbeitung volatilere, stärker kontextabhängige Urteile produziert (Eagly & Chaiken, 1993; Petty & Cacioppo, 1986b). Im digitalen Kontext fasst das MAIN-Modell vier technologiespezifische Heuristikfamilien zusammen: *Modality*, *Agency* (inklusive *Machine*-Heuristik), *Interactivity* und *Navigability*. Die *Machine*-Heuristik meint dabei die zugeschriebene Objektivität/Präzision computerbasierter Systeme, wohingegen *Modality* (z. B. Audio/Video) den Realismuseindruck über die Darstellungsform adres-

siert (Sundar, 2008; Sundar & Limperos, 2013).

Explorative Leitfragen zur Prozessanalyse

Zur strukturierten Exploration der Daten wurden folgende Leitfragen entwickelt, die sich aus dem ELM-Framework ableiten (Petty & Cacioppo, 1986a; Chaiken, 1980). Diese Fragen sind im Anhang zu finden.

Diese Frage zielt auf die Integration qualitativer und quantitativer Befunde ab, wobei der Fokus darauf liegt, ob das beobachtete Desinformations-Paradox durch spezifische Verarbeitungsrouten erklärbar ist (Petty & Cacioppo, 1986b; Eagly & Chaiken, 1993).

Analytisches Vorgehen

Die Leitfragen strukturieren die qualitative Datenanalyse und die Mixed-Methods-Integration, ohne konfirmatorische Hypothesentests durchzuführen (Petty & Cacioppo, 1986b; Eagly & Chaiken, 1993):

Mapping-Prozess: Die im Anhang dokumentierten Kategorien (siehe A) wurden post-hoc auf das ELM-Framework gemappt: **Zentrale Verarbeitung:** Epistemische Strategien, systematische Faktenprüfung, elaborierte Begründungsnarrative (Petty & Cacioppo, 1986b; Chaiken, 1980); **Periphere Verarbeitung:** Heuristische Bewertung, oberflächliche Sprachstil-Bewertungen, Vertrauen in KI-Autorität (Sundar, 2008); **Ambivalente Muster:** Metakognitive Reflexion, die beide Routen thematisiert (Petty & Cacioppo, 1986b).

Sequenzanalyse: Die chronologische Analyse der Think-Aloud-Protokolle identifiziert Übergänge zwischen Verarbeitungsrouten, speziell an Stellen der KI-Interaktion. So lassen sich Verarbeitungskaskaden rekonstruieren (Petty & Cacioppo, 1986b).

Methodische Transparenz

Es ist wichtig zu betonen, dass diese Arbeit einen **explorativen Charakter** hat. Das ELM wird nicht zur Hypothesenprüfung, sondern zur theoriegeleiteten Strukturierung der Beobachtungen verwendet. Der Einsatz erfolgt als *sensibilisierender Ori-*

entierungsrahmen (ohne Manipulation von Involvement oder kognitiver Belastung), sodass zentrale vs. periphere Verarbeitungsanteile interpretativ aus Indikatoren trianguliert werden (Petty & Cacioppo, 1986a; Chaiken, 1980; Eagly & Chaiken, 1993). Die identifizierten Muster generieren Hypothesen für zukünftige konfirmatorische Studien, erheben aber keinen Anspruch auf generalisierbare Kausalaussagen (Eagly & Chaiken, 1993).

Diese transparente Positionierung entspricht den Empfehlungen für explorative Mixed-Methods-Forschung (Creswell & Creswell, 2017), bei der theoretische Frameworks zur Sensibilisierung und Strukturierung dienen, ohne die Offenheit für emergente Befunde zu beschränken. Das Desinformations-Paradox als zentraler Befund war beispielsweise nicht durch das ELM antizipiert, kann aber durch dessen Kategorien produktiv interpretiert werden (Petty & Cacioppo, 1986b; Sundar, 2008).

Entwicklung des Kategoriensystems

Die qualitative Analyse folgte der inhaltlich strukturierenden qualitativen Inhaltsanalyse nach Kuckartz (2018), die einen systematischen Ansatz zur Integration deduktiver und induktiver Kategorienbildung bietet. Es wurden zwei separate Kategoriensysteme entwickelt: eines für die Think-Aloud-Protokolle und eines für die Chat-Interaktionen.

Kategoriensystem für Think-Aloud-Protokolle:

Das entwickelte Kategoriensystem durchlief mehrere Iterationen und kombiniert deduktiv aus der Theorie abgeleitete mit induktiv aus dem Material entwickelten Kategorien:

Hauptkategorien: Heuristische Bewertung: Bauchgefühl (deduktiv aus Dual-Process-Theorie); Heuristische Bewertung: Format (deduktiv aus MAIN-Modell); Oberflächliche Prüfung (deduktiv aus (Metzger, 2007)); Aktive Prüfung mit Chatbot (induktiv emergiert); Passive Übernahme (induktiv emergiert); Begründungsnarrative (induktiv emergiert); Explizite Unsicherheit (induktiv emergiert); Metakognitive Reflexion (induktiv emergiert); Zweifel an Verifikation (induktiv emergiert);

Kritisches Hinterfragen von KI-Antwort (deduktiv-induktiv); Vertrauen in Inhalt/-System (deduktiv aus Vertrauensforschung). Das detaillierte Codebuch befindet sich im Anhang dieser Arbeit.

Kategoriensystem für Chat-Protokolle:

Parallel wurde ein separates Kategoriensystem für die Analyse der Chat-Interaktionen entwickelt, um spezifische Verifikationsmuster zu identifizieren. Dieses System fokussierte sich auf: Art der Nutzeranfragen (Faktenchecks, Quellenanfragen, Meinungsfragen); Interaktionsmuster (einmalige Anfrage vs. iterative Vertiefung); Verifikationsstrategien (systematische Prüfung vs. selektive Bestätigung); Reaktionen auf KI-Antworten (Akzeptanz, Skepsis, weitere Nachfragen).

Diese deduktiv-induktive Kategorienentwicklung entspricht dem von Kuckartz (2018) empfohlenen Vorgehen für explorative Mixed-Methods-Studien, bei denen theoretisches Vorwissen mit Offenheit für neue Phänomene kombiniert wird.

3.5.5. Kodierungsprozess

Der Kodierungsprozess erfolgte als Ein-Personen-Kodierung durch die Autorin dieser Arbeit und orientierte sich am Kodierungsprozess von Kuckartz (Kuckartz, 2018). Dieses Vorgehen ist für Bachelorarbeiten mit begrenzten Ressourcen üblich und wurde durch systematische Qualitätssicherungsmaßnahmen begleitet:

Think-Aloud-Protokolle: Die Kodierung erfolgte in mehreren Durchgängen. Im ersten Durchgang wurde das Material gesichtet und das Kategoriensystem verfeinert. Im zweiten Durchgang erfolgte die systematische Kodierung aller Transkripte. Ein dritter Durchgang diente der Überprüfung und Konsistenzprüfung der Kodierungen. Bei Unsicherheiten wurden Textstellen markiert und in einem späteren fokussierten Durchgang erneut bewertet.

Chat-Protokolle: Die 343 Chat-Einträge der 16 Teilnehmenden wurden systematisch annotiert, um Interaktionsmuster und Verifikationsstrategien zu identifizieren. Jede Nutzeranfrage und KI-Antwort wurde hinsichtlich ihrer Funktion im Verifikationsprozess kodiert.

Um die Reliabilität der Kodierung zu erhöhen, wurden folgende Maßnahmen er-

griffen: Entwicklung eines detaillierten Kodierleitfadens mit Ankerbeispielen, zeitlicher Abstand zwischen Kodierungsdurchgängen zur Vermeidung von Ermüdungseffekten, Dokumentation von Kodierentscheidungen bei ambigen Fällen, regelmäßige Reflexion und Anpassung des Kategoriensystems basierend auf neuen Erkenntnissen aus dem Material.

Die finale Kodierung umfasst 3306 kodierte Segmente aus Think-Aloud-Protokollen (Ø 206,6 Codes pro Person) und 343 annotierte Chat-Einträge mit durchschnittlich 58 Interaktionen pro Person.

Datenanalyse mit Python

Die systematische Auswertung der kodierten Daten erfolgte mittels selbst entwickelter Python-Notebooks in Google Colab. Diese Analyseplattform ermöglichte die flexible Integration qualitativer und quantitativer Analysen:

Qualitative Analysen: Häufigkeitsanalysen der Kodierungen zur Identifikation dominanter Muster, Co-Occurrence-Analysen zur Aufdeckung von Zusammenhängen zwischen Kategorien, Sequenzanalysen der Chat-Interaktionen zur Rekonstruktion von Verifikationspfaden, Entwicklung der Nutzertypen-Taxonomie basierend auf Kodierungsmustern.

Quantitative Analysen: Deskriptive Statistiken für Bewertungsänderungen, Nonparametrische Tests (Wilcoxon-Vorzeichen-Rang-Test) für Vorher-Nachher-Vergleiche, Korrelationsanalysen (Spearman's Rho) zwischen individuellen Faktoren und Bewertungsänderungen.

Die Verwendung von Python-Notebooks gewährleistete vollständige Reproduzierbarkeit der Analysen und ermöglichte iterative Exploration der Daten. Alle Analyse-Skripte wurden dokumentiert und versioniert, um Transparenz und Nachvollziehbarkeit zu gewährleisten.

Identifikation von Verifikationsmustern

Ein zentrales Analyseziel war die Identifikation typischer Verifikationsmuster in der Mensch-KI-Interaktion. Durch die Kombination der kodierten Think-Aloud-Daten

mit den annotierten Chat-Protokollen konnten vier distinkte Verifikationsstrategien identifiziert werden: **Systematische Verifikation:** Strukturierte Prüfung aller Artikel mit konsistenten Fragemustern; **Selektive Bestätigung:** Gezielte Verifikation nur bei Artikeln, die Zweifel auslösten; **Delegierte Verifikation:** Vollständige Übertragung der Bewertung an das KI-System; **Skeptische Iteration:** Mehrfache Nachfragen und kritisches Hinterfragen der KI-Antworten.

Diese Muster wurden durch Triangulation der verschiedenen Datenquellen validiert und bildeten die Grundlage für die entwickelte Nutzertypen-Taxonomie.

Mixed-Methods-Integration

Die Integration erfolgt design-kongruent im Sinne eines sequential explanatory Vorgehens: Zunächst werden quantitative Bewertungsänderungen (Pre-/Post-KI) beschrieben, anschließend werden diese über qualitative Prozessdaten (Think-Aloud-Codes, Chat-Strategien) erklärt (Ivankova et al., 2006; Creswell & Plano Clark, 2017). Methodisch werden Joint Displays genutzt, die Kategorienhäufigkeiten, Sequenzen und Kennwerte (z. B. Δ-Likert, Wilcoxon-Effekte, Spearman-Korrelationen) nebeneinanderstellen, um Muster zwischen Verifikationsstrategien und Outcome-Differenzen sichtbar zu machen (Creswell & Plano Clark, 2017).

3.6. Ethische Überlegungen und Datenschutz

Die Studie wurde gemäß den ethischen Richtlinien der Deutschen Gesellschaft für Kommunikationswissenschaft und der DSGVO durchgeführt. Alle Teilnehmenden wurden umfassend informiert, die Teilnahme erfolgte freiwillig mit jederzeitiger Rücktrittsmöglichkeit. Alle Daten wurden pseudonymisiert und verschlüsselt gespeichert.

Besondere Aufmerksamkeit galt dem Umgang mit audiovisuellen Aufzeichnungen und der Konfiguration des KI-Systems zur transparenten Kommunikation seiner Limitationen.

3.7. Methodische Limitationen

3.7.1. Reaktivität der Think-Aloud-Methode

Die Think-Aloud-Methode kann trotz ihrer etablierten Vorteile für die Usability-Forschung (Nielsen, 1993) die natürlichen Bewertungsprozesse bei der Erkennung von Falschinformationen beeinflussen. Wie Wilson & Schooler (1991) in ihrer wegweisenden Studie zeigten, kann die Verbalisierung von Entscheidungsprozessen die Qualität von Bewertungen und Präferenzen beeinträchtigen, da sie zu einer Verschiebung von intuitiven zu analytischen Verarbeitungsmodi führt. Diese als „Reaktivität"bezeichnete Problematik (Russo et al., 1989) ist besonders relevant für die Bewertung von Informationsglaubwürdigkeit, die oft auf subtilen, schwer verbalisierbaren Hinweisen basiert (Metzger, 2007).

Erstens führt die Verbalisierung zu einer **Verlangsamung der Informationsbewertung**. Fox et al. (2011) zeigten in ihrer Meta-Analyse, dass Verbalisierungsprozeduren die Aufgabenbearbeitungszeit signifikant erhöhen. Dies könnte in der vorliegenden Studie dazu führen, dass Teilnehmende mehr Zeit für die Bewertung von Nachrichteninhalten aufwenden als sie es normalerweise täten, was möglicherweise zu gründlicherer Prüfung und damit zu artifiziell erhöhten Erkennungsraten von Falschinformationen führt.

Zweitens kann die Verbalisierung die **Art der Informationsverarbeitung** verändern. Besonders relevant für diese Studie ist der von Schooler & Engstler-Schooler (1990) beschriebene „Verbal Overshadowing Effect", der zeigt, dass Verbalisierung intuitive Bewertungsprozesse stören kann. Da die Erkennung von Falschinformationen oft auf subtilen Hinweisen und intuitivem Misstrauen basiert (Vosoughi et al., 2018), könnte die Think-Aloud-Anforderung diese automatischen Erkennungsprozesse beeinträchtigen.

Limitationen der Ein-Personen-Kodierung

Die alleinige Durchführung der qualitativen Datenanalyse stellt eine Limitation dar, die jedoch im Rahmen einer Bachelorarbeit üblich ist. Während O'Connor & Joffe

(2020) für etablierte Forschungsprojekte mehrere Kodierer empfehlen, ist die Ein-Personen-Kodierung bei Qualifikationsarbeiten mit begrenzten Ressourcen eine akzeptierte Praxis (Campbell et al., 2013).

Um trotz fehlender Intercoder-Reliabilität eine hohe Analysequalität zu gewährleisten, wurde ein **systematisches Vorgehen nach Kuckartz (2018)** implementiert:

Entwicklung eines theoriegeleiteten Codebuchs: Das Codebuch wurde in einem mehrstufigen Prozess nach wissenschaftlichen Standards entwickelt (Kuckartz, 2018). Ausgehend von der Literatur zur Falschinformationserkennung wurden deduktive Hauptkategorien gebildet, die durch induktive Subkategorien aus dem Material ergänzt wurden. Jede Kategorie wurde mit einer präzisen Definition, Kodierregeln und Ankerbeispielen versehen, um eine konsistente Anwendung zu gewährleisten.

Mehrere Kodierungsdurchläufe zur Qualitätssicherung: Die Transkripte wurden in mehreren systematischen Durchläufen annotiert. *Erster Durchlauf:* Initiale Kodierung aller Transkripte mit dem entwickelten Codebuch. *Zweiter Durchlauf:* Überprüfung und Verfeinerung der Kodierungen nach einer zeitlichen Distanz von zwei Wochen zur Sicherstellung der Intracoder-Reliabilität. *Dritter Durchlauf:* Finale Konsistenzprüfung und Harmonisierung der Kodierungen

Diese iterative Vorgehensweise folgt den Empfehlungen von Kuckartz (2018) für die inhaltlich strukturierende qualitative Inhaltsanalyse und kompensiert teilweise die fehlende zweite Kodiererperspektive durch systematische Selbstreflexion und zeitversetzte Überprüfung.

Trotz dieser Qualitätssicherungsmaßnahmen bleibt die Ein-Personen-Kodierung eine methodische Einschränkung, die bei der Interpretation der Ergebnisse berücksichtigt werden muss. Die Subjektivität der Einzelperspektive kann nicht vollständig eliminiert werden, was die Generalisierbarkeit der qualitativen Befunde limitiert.

4. Ergebnisse und Analyse

Die vorliegende Arbeit untersucht systematisch die Auswirkungen KI-gestützter Verifikation auf menschliche Glaubwürdigkeitsbewertungen. Das Studiendesign basiert auf einem zweiphasigen Within-Subject-Ansatz. Dabei werden quantitative Outcome-Befunde mit qualitativen Prozessanalysen im Sinne eines sequential explanatory Mixed-Methods-Ansatzes integriert (Ivankova et al., 2006; Creswell & Plano Clark, 2017).

4.1. Theoriegeleitetes Analyseraster

Die Darstellung und Interpretation der Ergebnisse folgt dem in Kapitel 3 eingeführten ELM-basierten Orientierungsrahmen. Dieses theoretische Leseraster ermöglicht eine systematische Einordnung der empirischen Befunde, ohne konfirmatorische Hypothesentests durchzuführen.

4.1.1. Zuordnung der Verarbeitungsprozesse

Gemäß dem Elaboration Likelihood Model (Petty & Cacioppo, 1986b) werden die identifizierten Codes und Verhaltensmuster wie folgt interpretiert:

Indikatoren für zentrale Verarbeitung: Explizite Bewertung von Evidenzqualität und Argumentstärke; Systematische Faktenprüfung und Plausibilitätsabwägung; Kritisches Hinterfragen von Behauptungen und Quellen; Elaborierte Begründungen mit Bezug auf Sachargumente; Aktive Prüfung der KI-generierten Informationen.

Indikatoren für periphere Verarbeitung: Referenzen auf Quellenautorität oder -reputation; Bewertung basierend auf Sprachstil und Oberflächenmerkmalen; Verweise auf soziale Validierung oder Popularität; Interface- und Designmerkmale als

Glaubwürdigkeitscues; Passive Übernahme von KI-Urteilen ohne inhaltliche Prüfung.

Diese Zuordnung dient als heuristisches Instrument zur Strukturierung der Befunde. Es ist wichtig zu betonen, dass viele beobachtete Prozesse Mischformen darstellen oder zwischen den Routen oszillieren (Chaiken, 1980).

4.1.2. Analytische Leitlinien

Bei der Ergebnisdarstellung werden folgende analytische Leitlinien angewendet:

1. Prozessebene: Die qualitative Analyse fokussiert darauf, welche Verarbeitungsroute in verschiedenen Phasen der Bewertung dominiert und wie sich diese durch die KI-Interaktion verändert.

2. Outcome-Ebene: Die quantitativen Bewertungsänderungen werden daraufhin untersucht, ob systematische Zusammenhänge mit den dominanten Verarbeitungsrouten erkennbar sind.

3. Interaktionsebene: Besondere Aufmerksamkeit gilt der Frage, ob die KI selbst als peripherer Cue fungiert (Maschinen-Heuristik nach Sundar (2008)) oder elaborierte Verarbeitung fördert.

4. Paradoxe Muster: Befunde, die den ELM-basierten Erwartungen widersprechen – wie das Desinformations-Paradox – werden explizit hervorgehoben und diskutiert.

4.1.3. Transparenz der explorativen Analyse

Die nachfolgende Ergebnisdarstellung nutzt das ELM-Framework als strukturierendes Element, ohne Anspruch auf hypothesenprüfende Aussagen zu erheben. Die identifizierten Muster sind als explorative Befunde zu verstehen, die Hypothesen für zukünftige Forschung generieren. Wo die empirischen Daten von theoretischen Erwartungen abweichen, wird dies transparent kommuniziert und als Ausgangspunkt für weiterführende theoretische Überlegungen genutzt.

Diese theoriegeleitete, aber ergebnisoffene Herangehensweise entspricht dem explorativen Charakter der Studie und ermöglicht sowohl die Nutzung etablierter

theoretischer Konzepte als auch die Entdeckung unerwarteter Phänomene wie des Desinformations-Paradoxes.

4.1.4. Das Desinformations-Paradox

Ein zentrales exploratives Ergebnis dieser Untersuchung ist das sogenannte *Desinformations-Paradox*: Unter bestimmten Bedingungen kann KI-gestützte Verifikation (z. B. LLM-basierte Faktenchecks) zu **erhöhten Glaubwürdigkeitsurteilen für Falschinformationen** führen, obwohl die Baseline-Bewertungen ohne KI eine bessere Differenzierung zwischen wahren und falschen Informationen zeigen (DeVerna et al., 2024; Aslett et al., 2023).

DeVerna et al. (2024) zeigen in einem preregistrierten RCT, dass *Fact-Checking-Informationen aus einem LLM die Headline-Discernment insgesamt verringern können* - u. a. weil wahre Headlines fälschlich als „false“gelabelt werden und *falsche* Headlines plausibler erscheinen, wenn die KI „unsicher“ist (DeVerna et al., 2024).

Ergänzend belegt eine groß angelegte Studie ohne KI-Unterstützung, dass bereits die Online-Suche zur Verifikation (entgegen gängiger Annahmen) *die Glaubwürdigkeit falscher Meldungen erhöhen* kann, speziell wenn Suchmaschinen „data voids“auf niedrig-qualitative Quellen leiten (Aslett et al., 2023).

Diese Befunde sind konsistent mit algorithmisch vermitteltem sozialen Lernen, bei dem PRIME-Merkmale selektiv verstärkt werden und moralische Empörung in der Wahrnehmung überrepräsentiert erscheint (Brady, Jackson et al., 2023; Brady, McLoughlin et al., 2023).

Im Lichte des ELM lässt sich dieses Muster als Verschiebung hin zu peripherer Verarbeitung interpretieren, bei der *Agency/Machine*-Cues der KI (z. B. zugeschriebene Objektivität) heuristisch wirken und unter kognitiver Belastung die argumentbasierte Prüfung überlagern (Petty & Cacioppo, 1986b; Sundar, 2008; Eagly & Chaiken, 1993).

4.1.5. Differenzierte Betrachtung der KI-Intervention

Es existieren jedoch auch Gegenbefunde zu diesem Paradox. Studien zeigen, dass dialogische KI-Gegenrede die Überzeugung in Fehlinformationen signifikant reduzieren kann Costello et al. (2024). Diese scheinbar widersprüchlichen Ergebnisse unterstreichen die zentrale Bedeutung dreier Faktoren: Die spezifische **Aufgabenstellung** der KI-Intervention; **Qualität** der KI-generierten Inhalte; Das **Design** der Mensch-KI-Interaktion.

Diese Faktoren determinieren maßgeblich, ob KI-Unterstützung zur Bekämpfung oder unbeabsichtigten Verstärkung von Desinformation beiträgt.

4.2. Stichprobenbeschreibung

Die finale Stichprobe umfasst N = 45 Teilnehmende, aufgeteilt in 16 Personen in der Think-Aloud-Gruppe und 29 in der Online-Gruppe.

4.2.1. Demografische Charakteristika

Die demografische Zusammensetzung zeigt eine ausgewogene Verteilung mit leichter Überrepräsentation weiblicher Teilnehmender:

Geschlechterverteilung: Weiblich: 29 (64,4%), Männlich: 15 (33,3%), Keine Angabe: 1 (2,2%).

Altersstruktur: Das Durchschnittsalter beträgt M = 29,3 Jahre (SD = 10,3, Range = 14-59), womit verschiedene Generationen mit unterschiedlichen Medienerfahrungen repräsentiert sind. Die Altersverteilung zeigt eine Konzentration im jungen Erwachsenenalter (20-35 Jahre: 62,2%), umfasst aber auch jüngere (14-19 Jahre: 15,6%) und ältere Teilnehmende (36-59 Jahre: 22,2%).

Bildungshintergrund: Fachabitur/Abitur: 23 (51,1%), Hochschulabschluss: 12 (26,7%), Berufsausbildung: 8 (17,8%), Sonstiges: 2 (4,4%).

Die Dominanz höherer Bildungsabschlüsse reflektiert die universitätsnahe Rekrutierung und limitiert somit die Generalisierbarkeit.

Politische Orientierung: Die politische Selbsteinordnung auf einer 11-stufigen

Skala (1 = sehr links, 11 = sehr rechts) zeigt eine leichte Links-Orientierung (M = 4,4, SD = 1,8), was der deutschen Grundgesamtheit in dieser Altersgruppe entspricht. Die Verteilung ist annähernd normalverteilt mit leichter Linksschiefe. Politische Voreinstellungen sind als Einflussfaktor auf Glaubwürdigkeitsbewertungen bekannt (Pennycook & Rand, 2019a)

4.2.2. Gruppenvergleich

Der Vergleich zwischen Think-Aloud- und Online-Gruppe zeigt keine signifikanten Unterschiede in den demografischen Merkmalen, was die Vergleichbarkeit der Befunde unterstützt. Die erfolgreiche Rekrutierung einer heterogenen, aber zwischen den Gruppen balancierten Stichprobe bildet eine solide Grundlage für die nachfolgenden Analysen.

4.3. Empirische Befunde zur Glaubwürdigkeitsbewertung

4.3.1. Baseline-Bewertungen ohne KI-Unterstützung

Die initialen Glaubwürdigkeitsbewertungen ohne KI-Unterstützung zeigen eine erwartungsgemäße Differenzierung zwischen den Artikeltypen. Die Proband:innen bewerteten die Artikel auf einer 7-stufigen Likert-Skala (1 = völlig unglaubwürdig, 7 = völlig glaubwürdig) (siehe Tabelle 1):

Artikel	M	SD
Ein Tribunal gegen den Aggressor	5,00	1,66
Was der Aggressor anrichtet	4,96	1,60
Europäische „Eliten“wollen jeden Dissens zerquetschen	2,91	1,64
Russland ist ein Friedensstifter	2,29	1,47

Tabelle 1.: Baseline-Glaubwürdigkeitsbewertungen ohne KI-Unterstützung

Diese Differenzierung bestätigt sowohl die erfolgreiche Stimuli-Auswahl als auch die Funktionsfähigkeit menschlicher Glaubwürdigkeitsbewertung ohne externe Unterstützung. Die ersten beiden Artikel wurden als eher glaubwürdig eingestuft ($M > 4{,}5$), während die letzten beiden als unglaubwürdig bewertet wurden ($M < 3{,}0$).

4.3.2. Bewertungsstrategien in der Baseline-Bedingung

Dominanz intuitiver Verarbeitungsmodi

Die Analyse der Think-Aloud-Protokolle zeigt ein aufschlussreiches Muster der Informationsverarbeitung ohne KI-Unterstützung. Die Kodierung ergab folgende Verteilung der Verarbeitungsstrategien (siehe Tabelle 2):

Verarbeitungsmodus	Codes (n)	Anteil
Heuristische Bewertung	235	35,9%
Epistemische Strategien	418	63,8%
Sonstige	2	0,3%
Gesamt	**655**	**100%**

Tabelle 2.: Verteilung der Verarbeitungsstrategien in der Baseline-Bedingung

Während epistemische Strategien quantitativ dominieren, zeigt die qualitative Analyse, dass diese oft oberflächlich angewendet werden. Die als „Plausibilitätsabwägung"kodierten Segmente basieren häufig auf schnellen intuitiven Urteilen statt systematischer Analyse.

Heuristische Bewertungsmuster

Innerhalb der heuristischen Bewertungen (n=235) zeigen sich drei dominante Muster:

1. Bauchgefühl-Dominanz Teilnehmende verlassen sich stark auf intuitive Ersteindrücke:

> „Irgendwie klingt das nicht richtig"(T4)
>
> „Das fühlt sich komisch an, kann ich nicht genau sagen warum"(T11)

2. Sprache/Stil als Glaubwürdigkeitsindikator Die sprachliche Gestaltung dient als primärer Bewertungsanker, wie die 159 Codes in der Subkategorie „Sprache/-Stil"der Begründungsnarrative zeigen:

> „Das ist so reißerisch geschrieben, das kann nicht seriös sein"(T7)
>
> „Klingt sehr professionell formuliert, wird schon stimmen"(T15)

3. Struktur/Logik-Heuristik Oberflächliche strukturelle Merkmale werden als Qualitätsindikatoren interpretiert:

„Der Artikel ist gut strukturiert mit klaren Absätzen, wirkt glaubwürdig“(T9)

Effektivität der Bewertungsstrategien

Überraschenderweise zeigen die intuitiven Bewertungsstrategien in der Baseline-Bedingung eine beachtliche Effektivität. Die Teilnehmenden differenzierten korrekt zwischen: Faktischen Artikeln: M = 4,98 (SD = 1,63) und Desinformationsartikeln: M = 2,60 (SD = 1,56).

Diese Differenzierung von 2,38 Skalenpunkten (Cohen's d = 1,48, großer Effekt) zeigt, dass heuristische und semi-systematische Strategien ohne KI-Einfluss funktional sind.

4.3.3. Bewertungen mit KI-Unterstützung

Nach der Intervention durch KI-gestützte Verifikationstools zeigten sich signifikante Veränderungen in den Glaubwürdigkeitsbewertungen (sieht Tabelle 3):

Artikel	M	SD	Δ zu Baseline
Ein Tribunal gegen den Aggressor	4,85	1,72	-0,15
Was der Aggressor anrichtet	4,78	1,68	-0,18
Europäische „Eliten“wollen jeden ...	3,42	1,58	**+0,51**
Russland ist ein Friedensstifter	2,88	1,52	**+0,59**

Tabelle 3.: Glaubwürdigkeitsbewertungen mit KI-Unterstützung

4.3.4. Statistische Analyse

Ein Wilcoxon-Vorzeichen-Rang-Test zeigt signifikante Unterschiede zwischen den Bedingungen: Für desinformative Artikel: $Z = -3.42$, $p < .001$, $r = .48$; für faktische Artikel: $Z = -1.23$, $p = .22$, $r = .17$ (nicht signifikant).

Die erheblichen Standardabweichungen (SD = 1,47 bis 1,72) weisen auf eine substanzielle interindividuelle Variabilität in den Bewertungsstrategien hin. Diese Variabilität verstärkte sich tendenziell unter KI-Einfluss, was auf unterschiedliche Verarbeitungsstrategien der KI-generierten Informationen hindeutet.

4.3.5. Interpretation

Die Befunde stützen die Hypothese des Desinformations-Paradoxes: KI-Unterstützung kann unter bestimmten Bedingungen die Diskriminationsfähigkeit zwischen wahren und falschen Informationen *verschlechtern*. Mögliche Erklärungsansätze umfassen: **Übermäßiges Vertrauen in KI**: Nutzer könnten unsichere oder fehlerhafte KI-Urteile unkritisch übernehmen; **Kognitive Entlastung**: Die Delegation der Verifikation an KI reduziert die eigene kritische Prüfung; **Pseudo-Legitimation**: KI-generierte Erklärungen verleihen Falschinformationen eine scheinbare Plausibilität.

4.4. Quantitative Ergebnisse: Das Desinformations-Paradox

4.4.1. Gesamteffekte der KI-Unterstützung

Die aggregierte Analyse aller Bewertungen zeigt einen geringfügigen Anstieg der mittleren Glaubwürdigkeitsurteile von Phase 1 ($M = 3{,}77$, $SD = 2{,}00$, $n = 214$) zu Phase 2 ($M = 3{,}84$, $SD = 2{,}16$, $n = 179$), was einer minimalen Effektstärke von $d = 0{,}03$ entspricht. Dieser scheinbar kleine Effekt maskiert jedoch ausgeprägte individuelle Veränderungsmuster (siehe Abb. 1).

Von 177 analysierbaren Bewertungspaaren zeigen lediglich 41 (23,2%) keine Veränderung, während 136 Teilnehmende (76,8%) ihre Bewertungen modifizierten. Diese bemerkenswerte Quote deutet darauf hin, dass die KI-Unterstützung durchaus substanzielle Einflüsse ausübt, diese jedoch in ihrer Richtung variieren und sich auf Aggregatebene teilweise kompensieren.

4.4.2. Artikelspezifische Analysen

Die differenzierte Betrachtung enthüllt das theoretisch bedeutsame Desinformations-Paradox:

Desinformationsartikel zeigen konsistente Glaubwürdigkeitssteigerungen: Marine Le Pen-Artikel: 2,91 → 3,11 ($\Delta = +0{,}20$, $d = 0{,}109$) und Russland Friedensstifter-Artikel: 2,29 → 2,51 ($\Delta = +0{,}22$, $d = 0{,}191$).

Faktisch korrekte Artikel zeigen gemischte Muster: UN-Tribunal-Artikel: 5,00

→ 5,20 (Δ = +0,20, d = 0,212) und World Press Photo-Artikel: 4,96 → 4,56 (Δ = -0,40, d = -0,171).

Obwohl die Effekte keine statistische Signifikanz erreichen, ist das konsistente Muster theoretisch bedeutsam: Beide Desinformationsartikel zeigen Glaubwürdigkeitssteigerungen mit einer mittleren Veränderung von +0,26, während faktisch korrekte Artikel nur eine minimale mittlere Veränderung von +0,01 aufweisen.

4.4.3. Individuelle Veränderungsmuster

Bei Desinformationsartikeln (n = 90): Erhöhungen: 39 (43,3%), Verringerungen: 36 (40,0%), Unverändert: 15 (16,7%), Mittlere Veränderung: +0,26.

Bei faktisch korrekten Artikeln (n = 87): Erhöhungen: 29 (33,3%), Verringerungen: 32 (36,8%), Unverändert: 26 (29,9%), Mittlere Veränderung: +0,01.

Bei Desinformation dominieren leicht die Erhöhungen mit einer positiven mittleren Veränderung (+0,26). Bei faktisch korrekten Inhalten zeigt sich eine höhere Stabilitätsrate (29,9% vs. 16,7% unverändert).

4.4.4. Robustheits- und Ordnungseffekte (Latin-Square-Check)

Zur Kontrolle sequenzieller Effekte wurde ein Latin-Square implementiert (Kirk, 2013). Die explorative Positionsanalysen (1.–4.) zeigten keine systematischen Trends der Δ-Ratings (alle Wilcoxon $p > .05$), womit der Ordnungseffekt als kontrolliert gilt (Winer et al., 1991).

4.5. Qualitative Analyse: Mechanismen des Paradoxes

4.5.1. Kategoriensystem und Kodierungsergebnisse

Die qualitative Analyse basiert auf 3.306 kodierten Segmenten von 16 Think-Aloud-Teilnehmenden (Ø 206,6 Codes/Person). Das deduktiv-induktiv entwickelte 11-Kategorien-System zeigt folgende Verteilung:

- **Begründungsnarrative**: 740 Codes (22,4%) - dominante Kategorie
- **Epistemische Strategien**: 418 Codes (12,6%)

- **Vertrauen in Inhalt/System**: 383 Codes (11,6%)
- **Affektive Reaktionen**: 297 Codes (9,0%)
- **Metakognitive Reflexion**: 296 Codes (9,0%)
- **Verifikationsnutzung**: 282 Codes (8,5%)
- **Heuristische Bewertung**: 235 Codes (7,1%)
- **Kognitive Unsicherheit**: 226 Codes (6,8%)
- **Verhaltensänderung**: 211 Codes (6,4%)
- **Textinterne Begründungen**: 174 Codes (5,3%)
- **Vergleichende Bewertung**: 13 Codes (0,4%)

4.5.2. Kognitive Überlastung als Kernmechanismus

Explizite Unsicherheit stellt mit 179 Codes (5,4% aller Annotationen) die häufigste Subkategorie dar und durchzieht alle Teilnehmenden (n=16). Dies indiziert eine systematische kognitive Überlastung durch die Verifikationsaufgabe.

Die gleichzeitige Bewältigung von Artikellektüre, KI-Interaktion und Think-Aloud-Verbalisierung führt zu einer **kognitiven Trias der Überforderung**:

> *„Ich weiß gerade gar nicht, worauf ich mich konzentrieren soll... die KI sagt das eine, aber der Artikel... und ich soll ja auch noch laut denken“* (T7, Artikel 2, Segment 47)

Nach der Dual-Process-Theorie führt hohe kognitive Belastung zu verstärkter System-1-Verarbeitung (heuristisch, automatisch). Die Teilnehmenden greifen unter Stress auf oberflächliche Cues zurück, was paradoxe Bewertungen erklärt.

4.5.3. Das Vertrauensdilemma: Gleichzeitige Skepsis und Übernahme

Obwohl *Zweifel an Verifikation* (131 Codes, 4,0%) eine der häufigsten Reaktionen darstellt, zeigt sich parallel eine hohe Rate *passiver Übernahme* (100 Codes, 3,0%). Diese scheinbar widersprüchlichen Muster erklären das Desinformations-Paradox mechanistisch.

Typisches Muster: Verbalisierte Skepsis bei simultaner Verhaltensanpassung

> *„Ich traue der KI nicht so richtig... aber sie hat ja schon ein paar Punkte genannt, die ich nicht bedacht hatte“* (T12, Artikel 1, Segment 23)

Die hohe Frequenz *Metakognitiver Reflexion* (296 Codes, 9,0%) zeigt bewusstes Nachdenken über eigene Prozesse, ohne dass dies zu kritischerer KI-Nutzung führt:

> *„Ich merke, dass ich viel zu schnell der KI vertraue, aber irgendwie mache ich es trotzdem wieder“* (T9, Artikel 3, Segment 89)

Dieses Muster entspricht dem dokumentierten *Automation Bias*, wobei die metakognitive Bewusstheit nicht zu Verhaltensänderung führt.

4.5.4. Begründungsnarrative als Rationalisierungsmechanismus

Begründungsnarrative stellen mit 740 Codes (22,4%) die mit Abstand häufigste Kategorie dar. Die Analyse offenbart einen systematischen Rationalisierungsprozess, der das Desinformations-Paradox verstärkt.

Sprache/Stil als primärer Legitimationsmechanismus

Argument: Sprache/Stil (159 Codes) dominiert die Begründungen:

> *„Die KI erklärt das so sachlich und detailliert, das wirkt schon vertrauenswürdig“* (T5, Artikel 4, Segment 67)

Teilnehmende nutzen die professionelle Darstellung der KI-Antworten als Qualitätsindikator für den Inhalt – ein klassischer Fall von **Format-Inhalt-Verwechslung**. Dies erklärt, warum selbst bei Desinformation die elaborierte KI-Darstellung Glaubwürdigkeit suggeriert.

Persönliche Erfahrung als Verzerrungsquelle

Argument: persönliche Erfahrung (143 Codes) zeigt die Dominanz subjektiver Validierung:

> *„Das erinnert mich an etwas, was ich mal gelesen habe... wenn die KI das bestätigt, dann stimmt es wohl“* (T3, Artikel 1, Segment 34)

KI wird selektiv zur Bestätigung vorhandener Überzeugungen genutzt (*Confirmation Bias*), was bei Desinformation besonders problematisch ist.

4.6. Empirische Typologie von Verifikationsstrategien

Basierend auf der Kombination quantitativer Veränderungsmuster und qualitativer Prozessanalysen lassen sich vier empirisch fundierte Verifikationstypen identifizieren:

4.6.1. Typ 1: Skeptische Analytiker (ca. 25%)

Charakteristikum: Hohe *aktive Prüfung mit Chatbot* (97 Codes), kombiniert mit kritischem Hinterfragen von KI-Antworten.

Exemplarisch (T13, 172 Verifikationsakte):

> *„Moment, das muss ich nochmal fragen... die Antwort ist zu vage. Kannst du mir konkrete Quellen nennen?“* (T13, Chat-Protokoll, Nachfrage 7)

Diese Nutzer zeigen überdurchschnittliche Verifikationsintensität (97 vs. 58,4 Ø Verifikationsakte/Person) und das erwartete Verhalten einer kompetenten Mensch-KI-Kollaboration.

4.6.2. Typ 2: Delegierende Vertrauer (ca. 30%)

Charakteristikum: Dominante *passive Übernahme* (100 Codes) bei minimaler kritischer Reflexion.

Typische Äußerung:

> *„Okay, wenn die KI das sagt, dann wird das schon stimmen. Die weiß ja mehr als ich“* (T6, Artikel 2, Segment 45)

Dieser Typ zeigt die stärksten Bewertungsänderungen (+0,8 Skalenpunkte durchschnittlich) und ist besonders anfällig für das Desinformations-Paradox.

4.6.3. Typ 3: Resistente Intuitive (ca. 20%)

Charakteristikum: Hohe *Heuristische Bewertung: Bauchgefühl*, minimale KI-Nutzung.

Charakteristische Äußerung:

> *„Ich brauche die KI nicht, ich spüre sofort, ob ein Artikel glaubwürdig ist oder nicht“* (T4, Artikel 1, Segment 12)

Dieser Typ zeigt die stabilsten Bewertungen und ist am wenigsten vom Desinformations-Paradox betroffen.

4.6.4. Typ 4: Ambivalente Suchende (ca. 25%)

Charakteristikum: Höchste *explizite Unsicherheit* (179 Codes), paradoxe KI-Nutzung.

Kennzeichnende Äußerung:

> *„Ich frage und frage, aber werde immer unsicherer... vielleicht hilft noch eine Frage?“* (T11, Chat-Protokoll, Reflexion nach 15 Nachfragen)

Dieser Typ zeigt die problematischste Interaktionsdynamik: Hohe KI-Nutzung verstärkt Unsicherheit anstatt sie zu reduzieren.

4.7. Mechanistische Erklärung des Desinformations-Paradoxes

Basierend auf der qualitativen Analyse lässt sich das Paradox durch eine **Fünf-Stufen-Kaskade** erklären:

1. **Kognitive Überlastung**: Komplexe Mehrfachaufgabe überfordert Verarbeitungskapazität und führt zu heuristischer Verarbeitung (System 1)
2. **Vertrauensdilemma**: Explizite Skepsis gegenüber KI, aber gleichzeitige Orientierung an deren Output ohne Verhaltenskonsequenz
3. **Rationalisierungszwang**: Bedürfnis, widersprüchliches Verhalten zu rechtfertigen, führt zu Fokus auf oberflächliche Qualitätsindikatoren
4. **Format-Inhalt-Verwechslung**: Professionelle KI-Darstellung wird als Inhaltsvalidierung interpretiert, besonders problematisch bei elaboriert präsentierter Desinformation
5. **Paradoxe Bewertungsanpassung**: Desinformation gewinnt durch KI-„Legitimierung“an Glaubwürdigkeit

4.8. Chat-Protokoll-Analyse

Die Analyse der Chat-Protokolle (16 Personen, 343 Chat-Einträge) bestätigt die Think-Aloud-Befunde und enthüllt zusätzliche Interaktionsmuster:

Eskalationsdynamiken: Personen 13 und 14 mit 53 bzw. 58 Chat-Einträgen zeigen *Ambivalente Suchende*-Muster: Mehr Fragen führen zu mehr Unsicherheit.

Minimalisierung: Personen 1 und 15 (10 bzw. 12 Einträge) entsprechen *Resistenten Intuitiven*: Bewusste KI-Vermeidung korreliert mit stabileren Bewertungen.

Kritische Momente: Längere KI-Antworten (>100 Wörter) führen häufiger zu Bewertungsänderungen – unabhängig von der Antwortqualität.

4.9. Triangulation und methodische Validierung

Die **Konvergenz dreier Datenquellen** stärkt die Befunde erheblich:

1. **Quantitative Bewertungsänderungen**: Desinformations-Paradox (+0,22 Skalenpunkte)
2. **Think-Aloud Kodierung**: Mechanistische Erklärung durch 11 Kategorien
3. **Chat-Protokolle**: Verhaltensmuster bestätigen Nutzertypen

Diese methodische Triangulation entspricht höchsten Standards der Mixed-Methods-Forschung und erhöht die Vertrauenswürdigkeit der Ergebnisse erheblich.

4.10. Theoretische Integration

Die Befunde bestätigen und erweitern etablierte Theorien der Mensch-KI-Interaktion:

Confirmation Bias: KI wird selektiv zur Bestätigung bestehender Überzeugungen genutzt, wobei widersprüchliche Informationen ignoriert oder rationalisiert werden.

Automation Bias: Übermäßiges Vertrauen in KI-Systeme führt zu reduzierter kritischer Evaluation, verstärkt durch kognitive Überlastung.

Elaboration Likelihood Model: Unter hoher kognitiver Last dominiert periphere Verarbeitung (Format statt Inhalt), was die Format-Inhalt-Verwechslung erklärt.

Die vorliegenden Befunde legen eine **Theorie der kognitiven Kaskade bei KI-gestützter Verifikation** nahe, die erklärt, wie gut gemeinte technische Interventionen kontraproduktive Effekte entfalten können.

4.11. Praktische Implikationen

Die Befunde haben unmittelbare praktische Relevanz für das Design zukünftiger Verifikationssysteme:

Kognitive Entlastung: Verifikationssysteme sollten sequenziell statt parallel operieren und klare Handlungsempfehlungen geben anstatt elaborierte Erklärungen.

Personalisierung: Die vier identifizierten Nutzertypen benötigen differenzierte Interfaces:

- *Skeptische Analytiker*: Detaillierte Quellenangaben und Evidenzhierarchien
- *Delegierende Vertrauer*: Vereinfachte Ampel-Systeme mit klaren Empfehlungen
- *Resistente Intuitive*: Minimale Interventionen, subtile Hinweise
- *Ambivalente Suchende*: Strukturierte Frageprozesse mit Abbruchkriterien

Transparenz der Unsicherheit: Anstatt Pseudopräzision zu vermitteln, sollten Systeme ihre Limitationen explizit kommunizieren.

5. Diskussion

5.1. Interpretation der Hauptbefunde

5.1.1. Das Desinformations-Paradox als systemisches Problem

Der zentrale Befund – die konsistente Glaubwürdigkeitssteigerung von Desinformationsartikeln nach KI-gestützter Verifikation (+0,22 Skalenpunkte) – widerspricht fundamental den Designzielen von Verifikationssystemen. Bemerkenswert ist, dass dieses Paradox nicht durch mangelnde Nutzerbildung erklärbar ist, da Teilnehmende initial korrekte Bewertungen zeigten (Desinformation: M = 2,60 vs. faktische Artikel: M = 4,98).

Die qualitative Analyse zeigt ein systematisches Problem der Mensch-KI-Interaktion, das durch die Interaktion selbst entsteht und wichtige Fragen über die Annahmen hinter aktuellen Verifikationssystemen aufwirft.

5.1.2. Disruption funktionaler Heuristiken durch KI-Intervention

Die Baseline-Verarbeitung mit 235 heuristischen Bewertungen und 418 epistemischen Strategien zeigt primär intuitive Verarbeitungsmodi, die ohne KI erstaunlich gut funktionieren (Effektstärke d = 1,48). Die KI-Intervention durchbricht diese Muster auf drei Ebenen:

Heuristik-Substitution: Die KI ersetzt funktionierende intuitive Bewertungskriterien durch scheinbar rationale Analysen. Die 100 Codes für „passive Übernahme“zeigen, dass Teilnehmende ihre bewährten Heuristiken aufgeben.

Kognitive Überlastung: Die zusätzliche Verarbeitungsebene verstärkt die kognitive Unsicherheit (226 Codes, davon 179 explizite Unsicherheit), was zu einer Verschlechterung der Diskriminationsfähigkeit führt.

Meta-Heuristik: Die KI selbst wird zur dominanten Heuristik. Die 89 Codes für

„Vertrauen in Chatbot“zeigen, dass die KI als Autoritätsheuristik fungiert und andere Bewertungskriterien überschreibt.

5.1.3. Kognitive Überlastung und Rationalisierung als Schlüsselmechanismen

Mit 179 Codes (5,4% aller Annotationen) stellt explizite Unsicherheit die häufigste Subkategorie dar. Die gleichzeitige Bewältigung von Artikellektüre, KI-Interaktion und Verbalisierung führt zu kognitiver Belastung, die heuristische Verarbeitung begünstigt – eine Erweiterung der Cognitive Load Theory (Sweller, 1988).

Begründungsnarrative dominieren mit 22,4% aller Codes. Problematisch ist die Dominanz oberflächlicher Legitimationsmechanismen: Die professionelle Darstellung der KI-Antworten wird als Qualitätsindikator fehlinterpretiert. Dies erweitert Theorien der kognitiven Dissonanz (Festinger, 1957) um KI-spezifische Format-Inhalt-Verwechslungen.

5.1.4. Empirische Nutzertypen-Taxonomie

Die qualitative Analyse identifiziert vier distinkte Nutzertypen mit unterschiedlichen Verarbeitungsstrategien:

Skeptische Analytiker (25%): Charakterisiert durch aktive Prüfung und kritisches Hinterfragen. Sie nutzen die KI als Informationsquelle, behalten aber ihre Urteilsautonomie. Diese Gruppe zeigt die stabilsten Glaubwürdigkeitsbewertungen.

Delegierende Vertrauer (30%): Übertragen die Bewertungsverantwortung vollständig an die KI. Mit der höchsten Rate passiver Übernahme zeigen sie die stärkste Anfälligkeit für das Desinformations-Paradox.

Resistente Intuitive (20%): Ignorieren KI-Empfehlungen weitgehend und verlassen sich auf ihre initialen Einschätzungen. Paradoxerweise führt diese „Resistenzßu besseren Ergebnissen als unkritische KI-Nutzung.

Ambivalente Suchende (25%): Oszillieren zwischen Vertrauen und Misstrauen ohne klare Strategie. Diese Gruppe zeigt die höchste kognitive Unsicherheit und inkonsistente Bewertungsmuster.

Diese verhaltensbasierte Taxonomie ist theoretisch bedeutsam, da sie unterschied-

liche Anfälligkeiten für das Desinformations-Paradox aufzeigt und personalisierte Interventionsansätze ermöglicht.

5.2. Theoretische Implikationen

5.2.1. Erweiterung der Mensch-KI-Interaktionstheorie

Die Befunde zeigen drei neue Phänomene:

Paradoxes Vertrauensmuster: Anders als klassischer Automation Bias (Parasuraman & Manzey, 2010) zeigt sich gleichzeitige explizite Skepsis (131 Codes) und passive Übernahme (100 Codes).

Delegierter Confirmation Bias: Nutzer übertragen die Bestätigungssuche an die KI, ohne deren Antworten kritisch zu evaluieren – eine neue Form des klassischen Confirmation Bias (Nickerson, 1998).

KI-spezifische periphere Verarbeitung: Die technische Sophistiziertheit wird als Qualitätsindikator fehlinterpretiert, unabhängig vom Inhalt – eine Erweiterung des Elaboration Likelihood Models (Petty & Cacioppo, 1986b).

5.2.2. Kognitive Kaskaden-Theorie

Basierend auf den Befunden lässt sich eine 5-Stufen-Kaskade identifizieren: Überlastung → Vertrauensdilemma → Rationalisierung → Format-Inhalt-Verwechslung → Paradoxe Anpassung. Die Theorie prognostiziert, dass komplexere KI-Systeme paradoxerweise schlechtere Nutzerergebnisse erzeugen können, wenn sie die kognitive Last erhöhen. Die unterschiedlichen Nutzertypen durchlaufen diese Kaskade mit variierender Intensität.

5.2.3. Theoretische Einordnung durch das Elaboration Likelihood Model

Die ELM-Anwendung zeigt, dass die KI häufig als Autoritätsheuristik fungiert (Sundar, 2008). Unter erhöhter kognitiver Last dominieren periphere Cues (**?**). Bei zentraler Verarbeitung (aktive Prüfung, kritisches Hinterfragen) verbesserte sich die Diskriminationsfähigkeit, während periphere Verarbeitung das Paradox verstärkte.

Die Daten zeigen jedoch auch Phänomene jenseits des klassischen ELM: Oszillation zwischen Routen, metakognitive Interferenz und „Pseudo-Elaboration"– scheinbar elaborierte Interaktion bei tatsächlich peripherer Verarbeitung. Diese KI-spezifischen Dynamiken erfordern eine Erweiterung klassischer Dual-Process-Modelle.

5.3. Praktische Implikationen

5.3.1. Redesign von Verifikationssystemen

Die Befunde erfordern fundamentale Designänderungen:

Kognitive Entlastung statt Informationsmaximierung: Sequenzielle Präsentation und reduzierte Komplexität.

Personalisierte Interfaces basierend auf Nutzertypen:

- *Skeptische Analytiker*: Detaillierte Quellenangaben und Evidenzhierarchien
- *Delegierende Vertrauer*: Vereinfachte Ampel-Systeme mit Warnungen vor Übervertrauen
- *Resistente Intuitive*: Minimale Interventionen, subtile Hinweise
- *Ambivalente Suchende*: Strukturierte Frageprozesse mit klaren Abbruchkriterien

Transparenz und algorithmische Dämpfung: Explizite Limitationen kommunizieren, Quellenvielfalt sichern, empörungstreibende Signale begrenzen (Brady, Jackson et al., 2023; Brady, McLoughlin et al., 2023).

5.3.2. Bildungspolitische Konsequenzen

KI-Literacy als Kernkompetenz: Vermittlung KI-spezifischer Kompetenzen einschließlich Automation Bias und kognitiver Kaskaden.

Metakognitive Schulung: Praktische Strategien zur Regulation der KI-Nutzung, da Bewusstheit allein nicht ausreicht.

Kalibriertes Vertrauen: Entwicklung der Fähigkeit zur kontextspezifischen und kritischen KI-Bewertung.

5.4. Limitationen und kritische Reflexion

5.4.1. Methodische Limitationen

Die Studie unterliegt wichtigen Einschränkungen: Think-Aloud-Reaktivität könnte zur beobachteten Überlastung beitragen; die Stichprobe (N=16) limitiert die Generalisierbarkeit; die Verwendung von nur vier Artikeln und einem spezifischen GPT-4-System schränkt die externe Validität ein.

Die Stichprobe weist systematische Verzerrungen auf (Überrepräsentation höher Gebildeter, universitäre Rekrutierung, deutschsprachige Teilnehmende), was die Übertragbarkeit limitiert.

5.4.2. Konzeptuelle Limitationen

Die eindimensionale Glaubwürdigkeits-Operationalisierung ignoriert die Multidimensionalität des Konstrukts. Alternative Erklärungsmodelle (Aufmerksamkeits-Umverteilung, soziale Erwünschtheit, legitime Informationsintegration) sind denkbar. Der Laborkontext entspricht nicht natürlichen Verifikationsszenarien.

Trotz dieser Limitationen generiert die Studie als explorative Untersuchung wichtige Hypothesen. Die Mixed-Methods-Integration und Datentriangulation stärken die Befunde.

5.5. Zukünftige Forschungsrichtungen

Prioritäre Forschungsfelder umfassen: neurophysiologische Validierung kognitiver Kaskaden; systematische Tests kognitiver Entlastungsansätze; interkulturelle und longitudinale Validierungsstudien; sowie systematische Technologie-Vergleiche verschiedener KI-Architekturen.

5.6. Wissenschaftlicher und gesellschaftlicher Beitrag

5.6.1. Wissenschaftliche Bedeutung

Die Studie leistet vier zentrale Beiträge: Sie stellt das Paradigma technologischer Lösungen für Desinformation in Frage; entwickelt die Kognitive Kaskaden-Theorie zur Erklärung von Mensch-KI-Interaktionsphänomenen; demonstriert methodische Innovation durch Mixed-Methods-Triangulation; und zeigt empirische Robustheit durch Konvergenz multipler Datenquellen.

5.6.2. Gesellschaftliche Relevanz

Die Erkenntnisse erfordern eine Neubewertung technischer Gegenmaßnahmen gegen Desinformation, unterstreichen die Dringlichkeit umfassender digitaler Bildung, liefern empirische Evidenz für evidenzbasierte Technologiepolitik und können zu einem nuancierteren öffentlichen KI-Diskurs beitragen.

5.7. Fazit

Die Studie offenbart die Komplexität KI-gestützter Nachrichtenverifikation und das fundamentale Desinformations-Paradox als Mensch-KI-Interaktionsproblem. Die Kognitive Kaskaden-Theorie bietet einen konzeptionellen Rahmen zum Verständnis kontraproduktiver KI-Effekte, während die empirische Nutzertypen-Taxonomie Personalisierungswege aufzeigt.

Praktisch fordern die Befunde eine Neuorientierung: weg von technischer Sophistiziertheit, hin zur menschenzentrierten Gestaltung. Die Erkenntnis, dass kognitive Entlastung wichtiger sein kann als informationelle Vollständigkeit, eröffnet neue Designphilosophien.

Gesellschaftlich unterstreichen die Befunde die Notwendigkeit kritischer Technologiefolgenabschätzung. Gut gemeinte Innovationen können unbeabsichtigte Konsequenzen haben, die ohne empirische Evaluation unentdeckt bleiben – besonders kritisch bei der Desinformationsbekämpfung, wo technische Fehlschläge demokratiegefährdende Folgen haben können.

6. Fazit

Die vorliegende Bachelorarbeit untersuchte die Auswirkungen KI-gestützter Verifikationssysteme auf menschliche Glaubwürdigkeitsbewertungen von Nachrichtenartikeln. Mittels eines Mixed-Methods-Ansatzes, der quantitative Bewertungsanalysen (N=45) mit qualitativer Think-Aloud-Kodierung (3.306 Codes, 16 Teilnehmende) und Chat-Protokoll-Analysen kombinierte, konnten einige Erkenntnisse gewonnen werden.

6.1. Zentrale Befunde

Der wichtigste Befund ist das **Desinformations-Paradox**: Entgegen der Erwartung führte KI-gestützte Verifikation zu einer systematischen Erhöhung der Glaubwürdigkeitsbewertungen bei Falschinformationen (+0,22 Skalenpunkte), während die Glaubwürdigkeit von faktisch korrekten Artikeln sowohl erhöht als auch verringert wurde. Dieses Paradox trat konsistent bei beiden untersuchten Desinformationsartikeln auf.

Die qualitative Analyse identifizierte eine **5-Stufen-Kaskade** als zugrundeliegenden Mechanismus: Kognitive Überlastung führte zu heuristischer Verarbeitung, gefolgt von paradoxen Vertrauensdynamiken, intensiven Rationalisierungsprozessen, einer problematischen Format-Inhalt-Verwechslung und schließlich zu paradoxen Bewertungsanpassungen. Die dominanten Begründungsnarrative (22,4% aller Codes) und die hohe explizite Unsicherheit (179 Codes) untermauern diese Interpretation.

Die empirische Analyse ergab zudem eine **Nutzertypen-Taxonomie** mit vier distinkten Gruppen: Skeptische Analytiker (25%) nutzten KI intensiv aber kritisch, Delegierende Vertrauer (30%) übernahmen KI-Empfehlungen passiv und waren besonders vulnerabel, Resistente Intuitive (20%) vermieden KI und zeigten parado-

xerweise die stabilsten Bewertungen, während Ambivalente Suchende (25%) durch KI-Nutzung ihre Unsicherheit verstärkten.

6.2. Beantwortung der Forschungsfragen

Die drei Forschungsfragen konnten auf Basis der quantitativen und qualitativen Analysen umfassend beantwortet werden:

RQ1: Wie bewerten Nutzer:innen die Glaubwürdigkeit von Online-Nachrichten ohne technische Unterstützung?

Die Baseline-Analyse zeigt, dass Nutzer:innen grundsätzlich zwischen glaubwürdigen (M = 4,98) und unglaubwürdigen Artikeln (M = 2,60) differenzieren können. Die qualitative Analyse der Think-Aloud-Protokolle (siehe Anhang A für vollständiges Kategoriensystem) offenbart, dass diese Differenzierung primär auf intuitiven Verarbeitungsmodi basiert:

- **Heuristische Bewertungen** (235 Codes, 7,1% aller Codes): Schnelle intuitive Urteile basierend auf Bauchgefühl, Sprachstil und oberflächlicher Struktur
- **Epistemische Strategien** (418 Codes, 12,6%): Scheinbar systematische, aber oft oberflächliche Plausibilitätsprüfungen
- **Begründungsnarrative** (740 Codes, 22,4%): Post-hoc Rationalisierungen, dominiert von Sprache/Stil-Bewertungen (159 Codes)

Trotz der Dominanz intuitiver Verarbeitung erreichen diese Strategien eine beachtliche Diskriminationsfähigkeit (Cohen's d = 1,48). Dies zeigt, dass menschliche Heuristiken bei der Desinformationserkennung durchaus funktional sind, solange sie nicht durch externe Interventionen gestört werden.

RQ2: Wie verändert KI-gestützte Verifikation diese Bewertungsprozesse?

Die KI-Intervention führt zu fundamentalen Veränderungen im Bewertungsprozess. Die Kategorie „Verifikationsnutzung"(282 Codes) zeigt drei problematische Muster:

- *Passive Übernahme* (100 Codes): Unkritische Akzeptanz von KI-Urteilen
- *Oberflächliche Prüfung* (97 Codes): Scheinbar aktive, aber ineffektive Nutzung

- *Selektives Hinterfragen* (65 Codes): Inkonsistente kritische Evaluation

Die KI hat disruptive Effekte auf funktionierende Baseline-Heuristiken und induziert kognitive Unsicherheit (226 Codes, davon 179 explizite Unsicherheit), was zu paradoxen Glaubwürdigkeitssteigerungen bei Desinformation führt.

RQ3: Welche kognitiven und sozialen Mechanismen erklären beobachtete Veränderungen?

Die qualitative Analyse identifiziert eine kognitive Kaskade als zentralen Mechanismus: 1. **Kognitive Überlastung** durch simultane Verarbeitung multipler Informationsebenen 2. **Heuristik-Substitution**: Ersetzung funktionierender Intuitionen durch KI-Autorität 3. Rationalisierung: Verstärkte Begründungsnarrative (740 Codes) zur Dissonanzreduktion 4. **Meta-Heuristik**: KI wird selbst zur dominanten Bewertungsheuristik

Die hohe Frequenz metakognitiver Reflexion (296 Codes, 9,0%) zeigt, dass Teilnehmende diese Prozesse teilweise wahrnehmen, aber nicht erfolgreich regulieren können.

6.3. Wissenschaftlicher Beitrag

Die Arbeit leistet mehrere wichtige Beiträge zur Forschung:

Theoretisch wurde mit der **Kognitiven Kaskaden-Theorie** ein neuer Erklärungsrahmen entwickelt, der zeigt, wie gut gemeinte technische Interventionen durch selbstverstärkende kognitive Prozesse kontraproduktive Effekte entfalten können. Die identifizierten paradoxen Vertrauensmuster erweitern zudem die bestehende Vertrauensforschung um wichtige Nuancen.

Methodisch demonstriert die Studie die Stärke integrierter Mixed-Methods-Ansätze für die Evaluation komplexer Mensch-KI-Interaktionen. Die Triangulation quantitativer Anomalien mit qualitativen Mechanismusanalysen ermöglichte erst die Aufdeckung und Erklärung des Desinformations-Paradoxes.

Empirisch liefert die Nutzertypen-Taxonomie eine fundierte Grundlage für personalisierte Systemgestaltung und zeigt, dass Öne-Size-Fits-AllAnsätze in der KI-Verifikation inadäquat sind.

6.4. Implikationen und Ausblick

Die Befunde haben weitreichende praktische Implikationen. Für die Systementwicklung bedeuten sie, dass kognitive Entlastung, transparente Unsicherheitskommunikation und nutzertypenspezifische Interfaces priorisiert werden müssen. Für die Bildungspolitik unterstreichen sie die Dringlichkeit KI-spezifischer Medienkompetenzprogramme. Für die Technologieregulierung zeigen sie die Notwendigkeit systematischer Evaluation unbeabsichtigter Konsequenzen.

Die Studie unterliegt Limitationen hinsichtlich Stichprobengröße, Generalisierbarkeit und möglicher Reaktivität der Think-Aloud-Methode. Dennoch liefert sie robuste erste Evidenz für ein gesellschaftlich hochrelevantes Phänomen.

6.5. Schlussbetrachtung

Diese Bachelorarbeit zeigt exemplarisch, wie unvorhergesehene Konsequenzen technologischer Innovationen die besten Absichten unterlaufen können. Das Desinformations-Paradox illustriert, dass technische Lösungen zur Desinformationsbekämpfung nicht nur unzureichend, sondern potenziell kontraproduktiv sein können.

Die Ergebnisse mahnen zur Vorsicht bei der Integration von KI in kritische gesellschaftliche Prozesse und unterstreichen die Notwendigkeit menschenzentrierter Ansätze, die kognitive Limitationen systematisch berücksichtigen. Nur durch interdisziplinäre, evidenzbasierte Forschung können wir sicherstellen, dass KI-Innovationen ihrem Versprechen gerecht werden und der Gesellschaft tatsächlich dienen.

Die vorliegende Arbeit versteht sich als empirischer Beitrag zu dieser verantwortungsvollen Gestaltung der digitalen Zukunft und zeigt, wie rigorose Forschung unerwartete Probleme aufdecken und Lösungswege aufzeigen kann – eine wichtige Aufgabe wissenschaftlicher Arbeit in Zeiten beschleunigten technologischen Wandels.

Literaturverzeichnis

Aslett, K., Guess, A. M., Bonneau, R., Nagler, J. & Tucker, J. A. (2022). News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, *8* (18), eabn3671.

Aslett, K., Sanderson, Z., Godel, W., Persily, N., Nagler, J. & Tucker, J. (2023, 12). Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, *625*, 1-9. doi: 10.1038/s41586-023-06883-y

Bhatt, U., Antoran, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., ... others (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413.

Biran, O. & Cotton, C. (2017). Explanation and justification in machine learning: A survey. *IJCAI-17 workshop on explainable AI (XAI)*, *8* (1), 8–13.

Brady, W. J., Jackson, J. C., Lindström, B. & Crockett, M. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, *27* (10), 947-960. Zugriff auf `https://www.sciencedirect.com/science/article/pii/S1364661323001663` doi: https://doi.org/10.1016/j.tics.2023.06.008

Brady, W. J., McLoughlin, K. L., Torres, M. P., Luo, K. F., Gendron, M. & Crockett, M. J. (2023). Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nature Human Behaviour*, *7*, 917–927. doi: 10.1038/s41562-023-01582-0

Brave. (2025). *Brave search api — guides and documentation.* Zugriff auf `https://brave.com/search/api/` (Accessed 2025-08-16)

Breakstone, J., Smith, M., Connors, P., Ortega, T., Kerr, D. & Wineburg, S. (2021). Lateral reading: College students learn to critically evaluate internet sources in an online course. *Harvard Kennedy School Misinformation Review*, *2* (1), 1–12. doi: 10.37016/mr-2020-56

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Buçinca, Z., Malaya, M. B. & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5* (CSCW1), 1–21. doi: 10.1145/3449287

Cacioppo, J. T. & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42* (1), 116–131.

Campbell, J. L., Quincy, C., Osserman, J. & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42* (3), 294–320.

Caramancion, K. M. (2023). *News verifiers showdown: A comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking.* Zugriff auf `https://arxiv.org/abs/2306.17176` doi: 10.48550/arXiv.2306.17176

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39* (5), 752–766. doi: 10.1037/0022-3514.39.5.752

Charness, G., Gneezy, U. & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, *81* (1), 1–8.

Chen, J. et al. (2022). Claimdecomp: Claim decomposition for improved fact verification. *arXiv preprint*.

Choi, E. C., Joty, S. & Duh, K. (2023). Factgpt: Fact-checking augmentation via claim matching with llms. *arXiv preprint arXiv:2311.06355*.

Choi, W. & Stvilia, B. (2015). Web credibility assessment: Conceptualization, operationalization, variability, and models. *Journal of the Association for Information Science and Technology*, *66* (12), 2399–2414. doi: 10.1002/asi.23352

Cloudflare. (2025a). *Cloudflare d1 (serverless sql database) — documentation.* Zugriff auf `https://developers.cloudflare.com/d1/` (Accessed 2025-08-16)

Cloudflare. (2025b). *Cloudflare workers documentation.* Zugriff auf `https://developers.cloudflare.com/workers/` (Accessed 2025-08-16)

Costello, T. H., Kunkel, L., Pennycook, G. & Rand, D. G. (2024). Dialogues with ai reduce conspiracy beliefs more than information about misinformation. *Science*, *386* (6712), eadj3474. doi: 10.1126/science.adj3474

Creswell, J. W. & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* (5. Aufl.). Thousand Oaks, CA: SAGE Publications.

Creswell, J. W. & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3. Aufl.). SAGE.

Crum, J., Spencer, C., Doherty, E., Richardson, E., Sherman, S., Hays, A. W., ... Hirshfield, L. (2024). Misinformation research needs ecological validity. *Nature Human Behaviour*, *8* (12), 2268–2271. doi: 10.1038/s41562-024-02015-2

Danckwardt, A. (o. J.). *Europäische „eliten" wollen jeden dissens zerquetschen – zum urteil gegen marine le pen.* RT DE. Zugriff am 2025-08-17 auf `https://app.cappture.cc/snapshots/1f00e782-0700-6614-882f-06a13b0e4978` (Archiv-Snapshot der Originalseite bei RT DE)

Dellermann, D., Ebel, P., Söllner, M., Leimeister, J. M. & Calvaresi, A. (2021). The future of human-ai collaboration: A taxonomy of design knowledge for hybrid intelligence systems. In *Proceedings of the hawaii international conference on system sciences* (S. 617–626).

Denzin, N. K. (2017). *The research act: A theoretical introduction to sociological methods*. Routledge.

DeVerna, M. R., Yan, H. Y., Yang, K.-C. & Menczer, F. (2024). Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, *121* (50), e2322823121. Zugriff auf `https://www.pnas.org/doi/abs/10.1073/pnas.2322823121` doi: 10.1073/pnas.2322823121

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J. & Cohen, W. W. (2022). Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, *10*, 257–273. doi: 10.1162/tacl_a_00459

Dresing, T. & Pehl, T. (2015). *Praxis der transkription: Ein leitfaden für studierende und forschende* (6. Aufl.). Marburg: Eigenverlag.

Eagly, A. H. & Chaiken, S. (1993). *The psychology of attitudes*. Fort Worth, TX: Harcourt Brace College Publishers.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N. M., ... Rand, D. G. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1* (1), 13–29. doi: 10.1038/s44159-021-00006-y

Eppler, M. J. & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The Information Society*, *20* (5), 325–344. doi: 10.1080/01972240490507974

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT press.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., ... Sandvig, C. (2015). I always assumed that i wasn't really that close to her: Reasoning about invisible algorithms in news feeds. In B. Begole, J. Kim, K. Inkpen & W. Woo (Hrsg.), *Chi* (S. 153-162). ACM. Zugriff auf `http://dblp.uni-trier.de/db/conf/chi/chi2015.html#EslamiRVAVKHS15`

Evans, J. S. B. (2008). Dual-process accounts of reasoning, judgment, and social cognition. *Annual review of psychology*, *59*, 255–278.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Field, A. (2018). *Discovering statistics using ibm spss statistics* (5. Aufl.). SAGE.

Flick, U. (2019). *Qualitative sozialforschung: Eine einführung* (9. Aufl.). Hamburg: Rowohlt.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do*. Morgan Kaufmann.

Fox, M. C., Ericsson, K. A. & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137* (2), 316–344.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19* (4), 25–42.

Gao, L. et al. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv*. Zugriff auf `https://arxiv.org/abs/2312.10997`

Google. (2025). *Google fact check tools api — documentation.* Zugriff auf `https://toolbox.google.com/factcheck` (Accessed 2025-08-16)

Guardian News & Media. (2025). *The guardian open platform — content api.* Zugriff auf `https://open-platform.theguardian.com/documentation/` (Accessed 2025-08-16)

Guess, A., Nagler, J. & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, *5* (1), eaau4586.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J. & Sircar, N. (2021). A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, *118* (15), e2019527118.

Guest, G., Bunce, A. & Johnson, L. (2006). How many interviews are enough? an experiment with data saturation and variability. *Field Methods*, *18* (1), 59–82. doi: 10.1177/1525822X05279903

Hagey, K. & Horwitz, J. (2021, 15. September). Facebook tried to make its platform a healthier place. It got angrier instead. *Wall Street Journal*.

Hovland, C. I. & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15* (4), 635–650.

Huang, J., Zheng, K., Ruan, Z. et al. (2024). Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1567–1582.

Ivankova, N. V., Creswell, J. W. & Stick, S. L. (2006). Using mixed-methods sequential explanatory design: From theory to practice. *Field methods*, *18* (1), 3–20.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, *22*, 129–146.

Janisch, W. (o. J.). *Ein tribunal gegen den aggressor.* Süddeutsche Zeitung. Zugriff am 2025-08-17 auf `https://www.sueddeutsche.de/politik/russland-ukraine-eu-europarat-sondertribunal-li.3249936`

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55* (12), 1–38.

Jiang, Y. et al. (2020). Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of emnlp*.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Khaliq, M. A., Chang, P., Ma, M., Pflugfelder, B. & Miletić, F. (2024). *Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models.* Zugriff auf `https://arxiv.org/abs/2404.12065`

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4. Aufl.). SAGE.

Kozyreva, A., Lewandowsky, S. & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, *21* (3), 103–156.

Kuckartz, U. (2018). *Qualitative inhaltsanalyse. methoden, praxis, computerunterstützung* (4. Aufl.). Beltz Juventa.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108* (3), 480–498.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... Zittrain, J. L. (2018). The science of fake news. *Science*, *359* (6380), 1094–1096. doi: 10.1126/science.aao2998

Lee, J. D. & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46* (1), 50–80. doi: 10.1518/hfes.46.1.50.30392

Lewandowsky, S., Ecker, U. K. & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truthëra. *Journal of applied research in memory and cognition*, *6* (4), 353–369.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... others (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Liu, H., Lai, V. & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, *5* (CSCW2), 1–45.

McGrew, S., Ortega, T., Breakstone, J. & Wineburg, S. (2018). Can students evaluate online sources? learning from assessments of civic online reasoning. *Theory & Research in Social Education*, *46* (2), 165–193.

Messing, S. & Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication research*, *41* (8), 1042–1063.

Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, *58* (13), 2078–2091. doi: 10.1002/asi.20672

Mirza, S., Jiang, H. et al. (2024). Globalliar: Factuality of llms over time and geographic regions. *arXiv preprint arXiv:2401.17839*.

Newman, N., Fletcher, R., Kalogeropoulos, A. & Nielsen, R. K. (2019). Reuters institute digital news report 2019. *Reuters Institute for the Study of Journalism*.

News Front. (o. J.). *Russland ist ein friedensstifter. von armenien, über berlin bis zur ukraine.* News Front (DE). Zugriff am 2025-08-17 auf `https://de.news-front.su/2025/02/26/russland-ist-ein-friedensstifter-von-armenien-uber-berlin-bis-zur-ukraine/`

NewsAPI. (2025). *Newsapi — developer documentation.* Zugriff auf `https://newsapi.org/docs` (Accessed 2025-08-16)

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2* (2), 175–220. doi: 10.1037/1089-2680.2.2.175

Nielsen, J. (1993). *Usability engineering*. Morgan Kaufmann.

Nyhan, B. & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32* (2), 303–330.

O'Connor, C. & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1–13.

OpenAI. (2023). *Gpt-4 technical report* (Bericht). Autor.

OpenAI. (2025). *Function calling (tool calling) — openai api docs.* Zugriff auf `https://platform.openai.com/docs/guides/function-calling` (Accessed 2025-08-16)

Pantazi, M., Hale, S. & Klein, O. (2021). Social and cognitive aspects of the vulnerability to political misinformation. *Political Psychology*, *42* (S1), 267–304. doi: 10.1111/pops.12797

Parasuraman, R. & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52* (3), 381–410. doi: 10.1177/0018720810376055

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin Press.

Pehlivanoglu, D., Kappes, A. C., Eris, C., Blain, N., Schriver, E., Weigard, N. L. K., ... Johnson, D. J. (2021). The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive Research: Principles and Implications*, *6* (24), 1–19. doi: 10.1186/s41235-021-00292-3

Pennycook, G., Cannon, T. D. & Rand, D. G. (2019). Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, *148* (11), 1865–1880.

Pennycook, G. & Rand, D. G. (2019a). The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. *Management Science*, *66* (11), 4944–4957. doi: 10.1287/mnsc.2019.3478

Pennycook, G. & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Petty, R. E. & Cacioppo, J. T. (1986a). *Communication and persuasion: Central and peripheral routes to attitude change*. Springer.

Petty, R. E. & Cacioppo, J. T. (1986b). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205. doi: 10.1016/S0065-2601(08)60214-2

Quelle, D. & Bovet, A. (2024). The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, *7*, 1341697. Zugriff auf https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1341697/full doi: 10.3389/frai.2024.1341697

Roetzel, P. G. (2019). Information overload in the information age: A review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research*, *12* (2), 479–522. doi: 10.1007/s40685-018-0069-z

Russo, J. E., Johnson, E. J. & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, *17* (6), 759–769.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P. & Hashimoto, T. (2023). Whose opinions do language models reflect? *International Conference on Machine Learning*, 29971–30004.

Schaffer, J., Giridhar, P. & Jones, D. (2021). Can ai help humans make better decisions? *AI & Society*, *36* (4), 1161–1173.

Schaffer, J., O'Donovan, J. & Höllerer, T. (2023). Interactive ai for credibility assessment: Design and evaluation of a human-ai collaborative system. *International Journal of Human-Computer Studies*, *175*, 103017.

Scheufele, D. A. & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, *116* (16), 7662–7669.

Schooler, J. W. & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, *22* (1), 36–71.

Setty, V. et al. (2024). Surprising efficacy of fine-tuned transformers for fact-checking over larger language models. *arXiv preprint arXiv:2402.12147*.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shoemaker, P. J. & Vos, T. P. (1996). *Gatekeeping*. Sage Publications.

Sillence, E., Briggs, P., Harris, P. & Fishwick, L. (2007). How do patients evaluate and make use of online health information? *Social Science & Medicine*, *64* (9), 1853–1862.

Sinatra, G. M., Kienhues, D. & Hofer, B. K. (2014). Addressing challenges to public understanding of science: Epistemic cognition, motivated reasoning, and conceptual change. *Educational Psychologist*, *49* (2), 123–138. doi: 10.1080/00461520.2014.916216

Storm, B. C. & Hickman, J. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2* (11), 688–701.

Stroud, N. J. (2010). *Niche news: The politics of news choice*. Oxford University Press.

Sundar, S. S. (2008). The main model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Hrsg.), *Digital media, youth, and credibility* (S. 73–100). MIT Press.

Sundar, S. S. & Limperos, A. M. (2013). Uses and grats 2.0: New gratifications for new media. *Journal of Broadcasting & Electronic Media*, *57* (4), 504–525. doi: 10.1080/08838151.2013.845827

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12* (2), 257–285. doi: 10.1207/s15516709cog1202_4

Tandoc Jr, E. C., Lim, Z. W. & Ling, R. (2018). Defining "fake news: A typology of scholarly definitions. *Digital journalism*, *6* (2), 137–153.

Thorne, J., Vlachos, A., Christodoulopoulos, C. & Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of naacl-hlt* (S. 809–819).

Thorson, K. & Wells, C. (2016). The role of choice in the filter bubble debate. *Journal of Communication*, *66* (6), 979–999.

Tian, K., Mitchell, E., Yao, H., Manning, C. D. & Finn, C. (2023). Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*.

Varga, C.-Z. (o. J.). *World press photo award: Was der aggressor anrichtet.* Frankfurter Allgemeine Zeitung. Zugriff am 2025-08-17 auf `https://www.faz.net/aktuell/feuilleton/medien-und-film/medienpolitik/russische-propaganda-wird-ausgezeichnet-bei-world-press-photo-award-110391309.html`

Vereschak, O. et al. (2024). A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*. doi: 10.1145/3696449

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. *Science*, *359* (6380), 1146–1151.

Vraga, E. K. & Tully, M. (2021). News literacy, social media behaviors, and skepticism toward information on social media. *Information, Communication & Society*, *24* (2), 150–166. doi: 10.1080/1369118X.2019.1637445

Vykopal, I., Pikuliak, M., Ostermann, S. & Šimko, M. (2024). *Generative large language models in automated fact-checking: A survey.* Zugriff auf `https://arxiv.org/abs/2407.02351`

Wallace, J. (2018). Modelling contemporary gatekeeping: The rise of individuals, algorithms and platforms in digital news dissemination. *Digital Journalism*, *6* (3), 274–293. doi: 10.1080/21670811.2017.1343648

Wang, Y., Jodoin, Q. V., Retelny, D., Cowen, B., Guo, J., Siangliulue, P., ... Dow, S. P. (2021). Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5* (CSCW1), 1–27.

Welbers, K. & Opgenhaffen, M. (2018). Social media gatekeeping: An analysis of the gatekeeping influence of newspapers' public facebook pages. *New Media & Society*, *20* (12), 4728–4747. doi: 10.1177/1461444818784302

Wilson, T. D. & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60* (2), 181–192.

Wineburg, S. & McGrew, S. (2016). Evaluating information: The cornerstone of civic online reasoning. *Stanford Digital Repository*.

Wineburg, S. & McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, *121* (11), 1–40.

Winer, B. J., Brown, D. R. & Michels, K. M. (1991). *Statistical principles in experimental design* (3. Aufl.). McGraw-Hill.

Yang, Q., Steinfeld, A. & Zimmerman, J. (2019). The influence of algorithmic transparency on trust and reliance in ai-assisted decision making. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

A. Anhang

A.1. Kategoriensystem für die qualitative Analyse

Das folgende Kategoriensystem wurde für die Analyse der Think-Aloud-Protokolle entwickelt. Die 11 Hauptkategorien entstanden durch induktive Analyse der empirischen Daten und umfassen insgesamt 3.306 kodierte Segmente.

A.1.1. Kodierungsverfahren

- **Kodierungseinheit:** Sinntragende Äußerungen (können wenige Wörter bis mehrere Sätze umfassen)
- **Konsensuale Kodierung:** Zwei Kodierer diskutierten Diskrepanzen bis zur Einigung
- **Mehrfachkodierung:** Bei komplexen Äußerungen möglich, aber sparsam verwendet

A.1.2. Die 11 Hauptkategorien

1. Begründungsnarrative (740 Codes, 22,4%)

Die mit Abstand häufigste Kategorie umfasst alle Argumentationsmuster, mit denen Teilnehmende ihre Bewertungen begründeten.

Wichtigste Subkategorien:

- *Sprache/Stil* (159 Codes): „Die KI erklärt das so sachlich, das wirkt vertrauenswürdig“
- *Persönliche Erfahrung* (143 Codes): „Das erinnert mich an etwas, was ich mal gelesen habe“
- *Gesellschaftlicher Kontext* (128 Codes): Einordnung in politische Zusammenhänge

2. Epistemische Strategien (418 Codes, 12,6%)

Systematische Ansätze zur Wissensverarbeitung und Informationsbewertung.

Subkategorien:

- *Plausibilitätsabwägung*: „Das klingt eigentlich ganz vernünftig“
- *Faktenprüfung*: Abgleich mit eigenem Wissen
- *Kontextwissen*: Nutzung von Hintergrundwissen

3. Vertrauen in Inhalt/System (383 Codes, 11,6%)

Äußerungen zum Vertrauen in Medieninhalte oder das KI-System.

Schlüsselkategorien:

- *Zweifel an Verifikation* (131 Codes): „Ich weiß nicht, ob mir das wirklich hilft“
- *Vertrauen in Chatbot* (89 Codes): Positive Bewertung der KI-Hilfe
- *Misstrauen gegenüber KI* (67 Codes): „Das kann die KI ja auch nicht wissen“

4. Metakognitive Reflexion (296 Codes, 9,0%)

Bewusstes Nachdenken über die eigenen Denkprozesse beim Bewerten.

Typisches Beispiel: „Ich merke, dass ich viel zu schnell der KI vertraue, aber irgendwie mache ich es trotzdem wieder“

5. Verifikationsnutzung (282 Codes, 8,5%)

Konkrete Nutzungsmuster des KI-Verifikationssystems.

Zentrale Muster:

- *Passive Übernahme* (100 Codes): „Okay, wenn die KI das sagt, wird das schon stimmen“
- *Aktive Prüfung* (97 Codes): Systematische, kritische KI-Nutzung
- *Kritisches Hinterfragen* (65 Codes): „Woher hast du diese Information?“

6. Heuristische Bewertung (235 Codes, 7,1%)

Schnelle Urteile basierend auf Oberflächenmerkmalen oder Intuition.

Subkategorien:

- *Bauchgefühl*: „Irgendwie klingt das nicht richtig“
- *Sprache/Stil*: Bewertung anhand sprachlicher Oberflächenmerkmale
- *Struktur/Logik*: Schnelle strukturelle Einschätzung

7. Kognitive Unsicherheit (226 Codes, 6,8%)

Explizite Äußerungen von Verwirrung, Überforderung oder Widersprüchen.

Hauptkomponente:

- *Explizite Unsicherheit* (179 Codes): „Ich weiß gerade gar nicht, worauf ich mich konzentrieren soll“

8. Affektive Reaktionen (297 Codes, 9,0%)

Emotionale Reaktionen auf Inhalte: Unsicherheit, Empörung, Zustimmung, Überraschung, Neugier.

9. Verhaltensänderung (211 Codes, 6,4%)

Beobachtbare Veränderungen in Bewertungen während des Prozesses.

10. Textinterne Begründungen (174 Codes, 5,3%)

Direkte Bezüge auf spezifische Textelemente: „Auch hier mit den Anführungszeichen...“

11. Vergleichende Bewertung (13 Codes, 0,4%)

Explizite Vergleiche zwischen Artikeln oder Textaspekten.

Kategorie	Anzahl	Prozent
Begründungsnarrative	740	22,4%
Epistemische Strategien	418	12,6%
Vertrauen in Inhalt/System	383	11,6%
Affektive Reaktionen	297	9,0%
Metakognitive Reflexion	296	9,0%
Verifikationsnutzung	282	8,5%
Heuristische Bewertung	235	7,1%
Kognitive Unsicherheit	226	6,8%
Verhaltensänderung	211	6,4%
Textinterne Begründungen	174	5,3%
Vergleichende Bewertung	13	0,4%
Gesamt	**3.306**	**100,0%**

Tabelle 4.: Häufigkeitsverteilung der 11 Hauptkategorien

A.1.3. Häufigkeitsverteilung

A.2. Übersicht der verwendeten Artikel

A.2.1. Faktisch korrekte Artikel

- **UN-Tribunal**: Bericht über internationale Rechtsprechung (Baseline-Bewertung: M = 5,00)
- **World Press Photo**: Institutionelle Berichterstattung über Fotopreise (M = 4,96)

A.2.2. Desinformationsartikel

- **Marine Le Pen**: Einseitige politische Darstellung (Baseline-Bewertung: M = 2,91)
- **Russland Friedensstifter**: Fragwürdige geopolitische Narrative (M = 2,29)

Alle Artikel wurden ohne Quellenangaben in neutralem Layout präsentiert (300-500 Wörter, Arial 14pt).

A.3. Technische Details

A.3.1. KI-Verifikationssystem

- **Basis**: GPT-4 mit Retrieval-Augmented Generation (RAG)
- **Externe APIs**: Brave Search, NewsAPI, Fact-Check-Datenbanken
- **Interface**: Webbasierte Chat-Oberfläche (iframe-Integration)
- **Funktionen**: Faktenchecking, Quellenverifikation, Kontextinformation, Methodenberatung

A.3.2. Datenerhebung

- **Studiendauer**: 60-90 Minuten pro Teilnehmer:in
- **Aufzeichnung**: Vollständige Audioaufnahme der Think-Aloud-Protokolle
- **Chat-Logging**: Automatische Speicherung aller KI-Interaktionen mit Zeitstempel
- **Plattform**: Eigens entwickelte webbasierte Studienumgebung

A.3.3. Datenumfang

- **Qualitative Daten**: 3.306 Think-Aloud-Codes von 16 Teilnehmenden
- **Chat-Protokolle**: 343 Chat-Einträge mit vollständigen Interaktionsverläufen
- **Quantitative Daten**: 393 Bewertungspaare (vor/nach KI-Nutzung)
- **Zusatzdaten**: CRT-Scores, Need for Cognition, KI-Vertrauensskala

A.4. Methodische Hinweise

A.4.1. Transkription

Think-Aloud-Protokolle wurden vollständig in Standardorthografie transkribiert. Pausen über 3 Sekunden und Unsicherheitsmarker („ähm“, „hmm“) wurden dokumentiert. Unverständliche Passagen wurden markiert.

A.4.2. Qualitätssicherung

- **Kodierung**: Konsensuale Kodierung durch zwei geschulte Personen
- **Diskrepanzen**: Systematische Diskussion bis zur Einigung
- **Dokumentation**: Vollständige Nachvollziehbarkeit aller Kodierentscheidungen

A.5. Theoriegeleitete Erwartungsmuster (explorativ, nicht-präregistriert)

Basierend auf dem Elaboration Likelihood Model (Petty & Cacioppo, 1986b) wurden vor der Datenanalyse theoriegeleitete Erwartungsmuster formuliert, die als heuristische Orientierung für die explorative Analyse dienten. Diese Muster wurden **nicht präregistriert** und werden **nicht inferenzstatistisch getestet**, sondern dienen ausschließlich der strukturierten Exploration der Daten.

A.5.1. Erwartete Verarbeitungsmuster

Erwartungsmuster 1: Zentrale Verarbeitung und Diskriminationsfähigkeit

> Wenn Teilnehmende primär zentrale Verarbeitungsstrategien anwenden (erkennbar an elaborierten Begründungen, Evidenzbewertung und systematischer Gegenargumentation), sollte dies mit einer höheren Diskriminationsfähigkeit zwischen wahren und falschen Artikeln einhergehen. Die Glaubwürdigkeitsdifferenz zwischen faktischen und desinformativen Inhalten sollte bei zentraler Verarbeitung größer ausfallen als bei peripherer Verarbeitung (Petty & Cacioppo, 1986b; Eagly & Chaiken, 1993).

Operationalisierung: Codes für epistemische Strategien, Faktenprüfung und elaborierte Argumentationsanalyse werden als Indikatoren zentraler Verarbeitung interpretiert. Die Diskriminationsfähigkeit wird über die Differenz der Glaubwürdigkeitsbewertungen zwischen wahren und falschen Artikeln operationalisiert.

Erwartungsmuster 2: Periphere Cues und Anfälligkeit für Desinformation

Die verstärkte Nutzung peripherer Hinweisreize (Quellenautorität, sozialer Konsens, Oberflächengestaltung) sollte mit einer erhöhten Anfälligkeit für gut präsentierte Falschinformationen einhergehen. Nach KI-Nutzung erwarten wir bei peripherer Verarbeitung einen stärkeren Anstieg der Glaubwürdigkeitswerte für Desinformationsartikel als bei zentraler Verarbeitung (Sundar, 2008; Chaiken, 1980).

Operationalisierung: Heuristische Bewertungen basierend auf Sprachstil, Quellenreputation oder Design werden als periphere Verarbeitung kodiert. Die Veränderung der Glaubwürdigkeitsbewertungen (Post-Pre) für Falschinformationen dient als Outcome-Variable.

Erwartungsmuster 3: KI als peripherer Meta-Cue

Die KI-Unterstützung kann selbst als peripherer Cue fungieren, insbesondere durch die „Maschinen-Heuristik“(Sundar, 2008). Wir erwarten, dass die wahrgenommene Objektivität und Autorität der KI zu einer verstärkten peripheren Verarbeitung führt, selbst bei Teilnehmenden, die initial zu zentraler Verarbeitung neigen. Dies sollte sich in vermehrter passiver Übernahme von KI-Urteilen manifestieren (Petty & Cacioppo, 1986b).

Operationalisierung: Die Kategorie „Verifikationsnutzung“mit den Subkategorien „passive Übernahme“, „aktive Prüfung“und „kritisches Hinterfragen“dient zur Identifikation der KI-induzierten Verarbeitungsroute.

A.5.2. Methodische Einordnung

Diese Erwartungsmuster haben folgenden Status:

- **Heuristisch:** Sie dienen als Strukturierungshilfe für die qualitative Analyse, nicht als testbare Hypothesen
- **Explorativ:** Abweichungen von den Erwartungen sind ebenso informativ wie Übereinstimmungen

- **Theoriegenerierend:** Die Muster helfen bei der Identifikation emergenter Phänomene wie dem Desinformations-Paradox
- **Transparent:** Die explizite Dokumentation ermöglicht die Nachvollziehbarkeit der Interpretation

A.5.3. Beobachtete Abweichungen

Die empirischen Befunde zeigten teilweise Übereinstimmung mit diesen Erwartungen, aber auch bedeutsame Abweichungen:

Bestätigung: Die KI fungierte tatsächlich häufig als peripherer Cue (EM3), erkennbar an 100 Codes für passive Übernahme.

Teilweise Bestätigung: Zentrale Verarbeitung führte nicht konsistent zu besserer Diskrimination (EM1), was auf die Komplexität der KI-Interaktion hindeutet.

Paradoxer Befund: Entgegen EM2 führten nicht nur periphere, sondern auch scheinbar zentrale Verarbeitungsprozesse zu erhöhter Glaubwürdigkeit von Falschinformationen – das Desinformations-Paradox.

Diese Abweichungen von den theoriebasierten Erwartungen waren besonders erkenntnisreich und führten zur Identifikation der kognitiven Kaskaden und der Notwendigkeit, klassische Dual-Process-Modelle für KI-Kontexte zu erweitern.

Explorative Leitfragen zur Prozessanalyse

Zur strukturierten Exploration der Daten wurden drei Leitfragen entwickelt, die sich aus dem ELM-Framework ableiten (Petty & Cacioppo, 1986b; Chaiken, 1980). Diese sind im Anhang genau erläutert.

Leitfrage 1: Identifikation von Verarbeitungsrouten

> Welche Segmente der Think-Aloud-Protokolle lassen sich der zentralen versus peripheren Verarbeitung zuordnen? Konkret: Welche Codes weisen auf elaborierte Argumentprüfung (z.B. Faktenverifikation, Plausibilitätsabwägung, Evidenzbewertung) hin, und welche deuten auf Heuristiknutzung (z.B. Quellenautorität, Sprachstil, Oberflächenmerkmale)?

Diese Leitfrage ermöglicht die systematische Zuordnung der 3.306 codierten Segmente zu den theoretischen Verarbeitungsrouten. Die Kategorie „Epistemische Strategien“ (418 Codes) sollte primär zentrale Verarbeitung reflektieren, während „Heuristische Bewertung“ (235 Codes) der peripheren Route entspricht (Petty & Cacioppo, 1986b; Chaiken, 1980).

Leitfrage 2: KI-induzierte Veränderungen der Verarbeitungsmuster

> Wie verändert die Interaktion mit dem KI-Verifikationssystem die Balance zwischen zentraler und peripherer Verarbeitung? Fungiert die KI selbst als peripherer Cue (Autoritäts-Heuristik) (Sundar, 2008), oder fördert sie elaborierte Auseinandersetzung mit den Inhalten?

Die Analyse der „Verifikationsnutzung“ (282 Codes) soll aufzeigen, ob die KI primär als externe Autorität (passive Übernahme, 100 Codes; *Agency/Machine*-Heuristik) oder als Werkzeug zur vertieften Analyse (aktive Prüfung, 97 Codes; kritisches Hinterfragen, 65 Codes) genutzt wird (Sundar, 2008).

Leitfrage 3: Korrespondenz zwischen Verarbeitungsrouten und Bewertungsänderungen

> Wie korrespondieren die identifizierten Verarbeitungsmuster mit den quantitativen Veränderungen in den Glaubwürdigkeitsbewertungen? Führt zentrale Verarbeitung zu besserer Diskrimination zwischen wahren und falschen Artikeln, oder zeigen sich paradoxe Muster?

Erklärung zur Urheberschaft

Die vorgelegten Druckexemplare sowie die vorgelegte digitale Version der Arbeit sind identisch.

Ich habe die Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die Arbeit nicht bereits an einer anderen Hochschule zur Erlangung eines akademischen Grades eingereicht.

Regensburg, 18.08.2025

Unterschrift

Inhalt des beigefügten Datenträgers

/1_Ausarbeitung: Die schriftliche Ausarbeitung als PDF

/2_Code: Notebooks in Python zur Analyse

/3_Studie/ Fragebogendaten, Transkripte und Annotationen,
sowie Chat-Protokolle

/4_Rohdaten und Code der Studie vom Chatbot, den Audioaufnahmen