HMD Praxis der Wirtschaftsinformatik
https://doi.org/10.1365/s40702-026-01314-8

SPEKTRUM

Multimodale Deepfake-Erkennung in der Praxis: Framework und Prototyp für die digitale Forensik

Louis Jarvers · Isabel Bezzaoui · Jonas Fegert

Eingegangen: 12. Februar 2026 / Angenommen: 16. Juli 2026
© The Author(s) 2026

Zusammenfassung Die Verbreitung generativer KI hat zu einem rasanten Anstieg synthetischer Medieninhalte im Internet geführt. Diese Entwicklung verändert die digitale Öffentlichkeit grundlegend, denn hochqualitative Deepfakes lassen sich selbst von geschulten Personen kaum noch sicher von authentischen Medien unterscheiden. Gleichzeitig sind die Folgen KI-manipulierter Inhalte, wie Betrug, Falschinformation oder Vertrauensverlust, umfassend bekannt. Während regulative oder bildungspolitische Maßnahmen primär die Detektionsfähigkeiten der Allgemeinheit stärken können, benötigen besonders digitalforensische Fachkräfte in Sicherheitsbehörden, Presse oder Privatwirtschaft technische Unterstützung bei der Identifikation von Deepfakes. Das Forschungsprojekt „Multi-Modal Deepfake Detection" (MuDDi) entwickelt hierfür Werkzeuge zur multimodalen Erkennung von KI-Inhalten und verbindet dabei bisher getrennte Forschungsansätze zu Inhaltsformaten wie Text, Bild, Video und Audio. Dieser Artikel stellt den aktuellen Forschungsstand von MuDDi auf Basis der Design-Science-Research-Methodologie dar. Nach der Problembeschreibung präsentiert der Artikel umsetzungs- und praxisrelevante Anforderungen aus qualitativen Expert:inneninterviews, die das technische Framework und den Prototypen als Forschungsartefakte leiten. Das Framework beschreibt die konkrete Umsetzung der Design-Anforderungen, während der Prototyp die Analyse von Bildern, Videos und Audio mittels KI-basierter End-to-End-Detection-Module, Wasserzeichenerkennung, der Extraktion pulsbasierter Signale aus Gesichtsregionen, kontextsensitivem Reasoning und dem Abgleich von Lippenbewegungen zu Tonspur demonstriert. Abschließend skizziert der Beitrag die geplante Umsetzung als Proof of Concept in der öffentlichen Verwaltung sowie die Evaluation und Kommunikation des Projekts.

✉ Louis Jarvers · Jonas Fegert
Institut für Wirtschaftsinformatik, Karlsruher Institut für Technologie, Karlsruhe, Deutschland
E-Mail: louis.jarvers@kit.edu

Isabel Bezzaoui
Goethe-Universität Frankfurt, Frankfurt, Deutschland

Published online: 06 August 2026

Schlüsselwörter Deepfakes · KI-generierte Inhalte · Automatische KI-Erkennung · Multimodale Erkennungssoftware · Design Science Research

Abstract The rapid diffusion of generative AI has led to a sharp increase in synthetic content across the internet. This development is fundamentally reshaping the digital spaces, as high-quality deepfakes have become increasingly difficult to distinguish from authentic media—even for trained experts. At the same time, the societal consequences of AI-manipulated content, including fraud, dis- and misinformation, and erosion of trust, are widely recognized. While regulatory interventions and educational initiatives can primarily enhance detection capabilities among the general public, digital forensic professionals in security agencies, journalism, and the private sector request advanced technical tools to reliably identify deepfakes. The research project "Multi-Modal Deepfake Detection" (MuDDi) addresses this need by developing tools for the multimodal detection of AI-generated content, integrating research approaches that have so far largely evolved in isolation across text, image, video, and audio. This article presents the current state of MuDDi's research using a design science research methodology. Following its problem definition, the article presents practice-oriented and implementation-relevant requirements from qualitative expert interviews, which inform the development of both the technical framework and the prototype as research artifacts. The framework operationalizes the identified design principles, while the prototype demonstrates the analysis of images, videos, and audio through AI-based end-to-end detection modules, watermark detection, extraction of pulse-based signals from facial regions, and the synchronization of lip movements with the audio track. The article concludes by outlining the planned implementation of the system as a proof of concept within public administration, as well as the project's evaluation and dissemination strategy.

Keywords Deepfakes · AI-generated Content · Automated AI detection · Multimodal detection software · Design Science Research

1 Einleitung: KI-generierte Inhalte als gesamtgesellschaftliche Herausforderung

Generative KI breitet sich schneller aus als die Einführung des PCs oder des Internets und prägt zunehmend den Alltag (Bick et al. 2025). Hohe Verfügbarkeit, geringe Nutzungskosten und einfache Zugänglichkeit tragen dabei maßgeblich zu einer globalen Zunahme KI-generierter Inhalte auf sozialen Netzwerken, wie X, Instagram oder TikTok bei (Corsi et al. 2024). Zwischen 2022 und 2024 stieg der Anteil von KI-generiertem Text auf den Plattformen wie Medium und Quora auf bis zu 40 %, auf der Fotoplattform Pixiv im gleichen Zeitraum sogar auf knapp 50 % (Senaweera et al. 2018; Wei und Tyson 2024): Mit dieser massiven Ausweitung entstehen neue Risiken: Synthetische Medien dienen dem Identitäts- und Finanzbetrug (Carpenter 2025), verletzen Persönlichkeits-, Gleichstellungs- und Urheberrechte (Sandoval et al. 2024) und tragen zur Verbreitung von Falschinformationen bei (Kharvi 2024). Die Folgen reichen von psychischen Belastungen wie Depressionen, Angststörungen

und Suiziden bei Betroffenen (Berlin und Fehrensen 2025), über volkswirtschaftliche Schäden (Colman 2025), bis hin zu gesellschaftlichem Vertrauensverlust (Vaccari und Chadwick 2020).

Mehr KI-generierte Inhalte und zunehmender Missbrauch sind auch deshalb kritisch, weil Menschen synthetische Inhalte kaum noch zuverlässig von authentischen Medien unterscheiden können (Cooke et al. 2025; Diel et al. 2024). Insbesondere Deepfakes, die in der weiten Definition als manipulierte oder vollständig synthetisierte, hochqualitative Inhalte verstanden werden (Altuncu et al. 2024), fordern selbst Fachkräfte heraus, die sich in Sicherheitsbehörden, Medien, der Finanz- oder Versicherungswirtschaft oder Zivilgesellschaft professionell mit der Verifikation von Informationen beschäftigen. Um sie technisch zu unterstützen, verfolgt das Forschungsprojekt „Multi-Modal Deepfake Detection“ (MuDDi) mithilfe der Design-Science-Research-Methodologie einen nutzerzentrierten Ansatz zur Entwicklung eines Erkennungssystems, bei dem Endanwender:innen Bilder, Videos, Audio und Text in einem System (multimodal) überprüfen können. Dieser Artikel stellt den bisherigen Forschungsstand des Projektes vor und zeigt, wie technische, organisatorische und gesellschaftliche Anforderungen zusammengeführt werden müssen, um Deepfake-Erkennung praxisnah und adaptiv zu gestalten.

2 Methodisches Vorgehen: Design Science Research

Design Science Research (DSR) zielt darauf ab, „Wissen darüber zu erweitern, wie Dinge konstruiert oder angeordnet werden können und sollten (d. h. designed)“

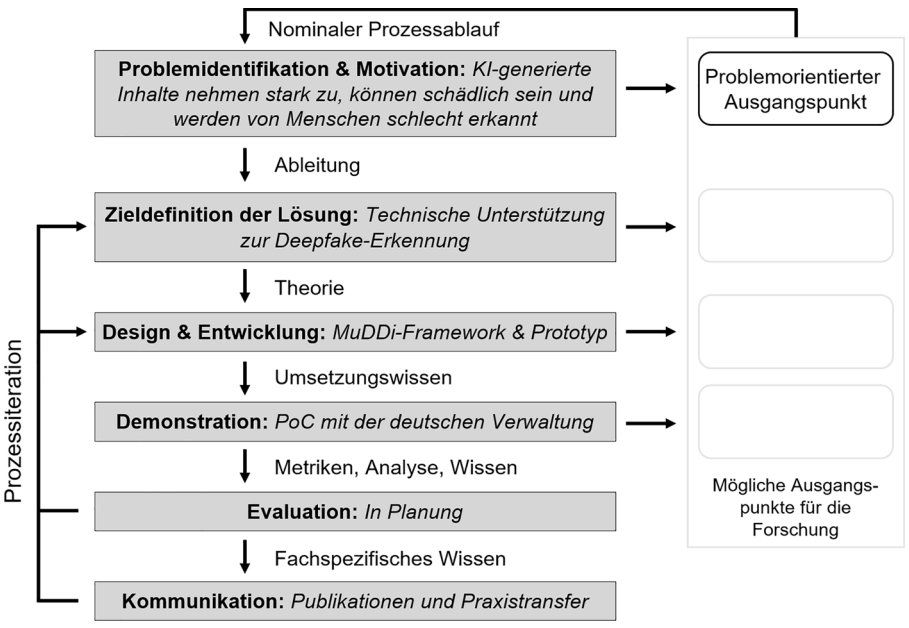

Abb. 1 DSRM Prozess für das MuDDi-Projekt (Adaption nach Peffers et al. 2007)

(A. Hevner et al. 2019, S. 3), um eine „Lösung für ein praxisrelevantes Problem der realen Welt“ zu entwickeln (Kuechler und Vaishnavi 2008, S. 492). Mit diesem Theorieverständnis folgen die konkrete Umsetzung des MuDDi-Projektes und der Aufbau dieses Beitrags den ersten drei von sechs Aktivitäten der Design-Science-Research-Methodologie (DSRM) nach Peffers et al. (2007; vgl. Abb. 1). Anders als Prozessansätze von Walls et al. (1992), Hevner et al. (2004) oder Cole et al. (2005) trennt Peffers et al. DSR-Methodologie die Problemidentifikation, Zieldefinition und Entwicklung der Lösung in drei separate Schritte. Dieses granulare Vorgehen eignet sich besonders für Forschungsvorhaben mit problemorientiertem Ausgangspunkt, da die explizite Zieldefinition samt Anforderungserhebung zusätzliches, übertragbares Design Knowledge jenseits des Prototypen generiert.

Der Anstoß für das MuDDi-Projekt waren zunehmende KI-Missbrauchsvorfälle im US-Präsidentschaftswahlkampf 2024, wo beispielsweise der Deepfakes von Taylor Swift-Fans Unterstützung für Donald Trump suggerierten, Wahlberechtigte KI-generierte, irreführende Anrufe erhielten oder minderwertige KI-Bilder (sog. „KI-Slop“, siehe Jarvers et al. 2026) Kamala Harris als Kommunistin diffamierten (Munoz 2024). Die Vorfälle verdeutlichen die Problemstellung, der sich das Projekt MuDDi widmet: Synthetische Inhalte nehmen stark zu, erzeugen erhebliche Risiken und bleiben für Menschen – insbesondere bei hochqualitativen Deepfakes – schwer erkennbar. Diese Problemidentifikation entspricht DSRM-Aktivität 1 und wird in Kapitel 3 vertieft. Darauf aufbauend umfasst DSRM-Aktivität 2 die Zieldefinition mit Anforderungsanalyse, die im MuDDi-Projekt kontextübergreifend angelegt ist, um die Deepfake-Detektion in Bereichen wie Sicherheitsbehörden, Medien, Industrie und Zivilgesellschaft gleichsam zu unterstützen. Die hierfür durchgeführte Interview-basierte Anforderungserhebung von Bezzaoui et al. (2025) wird in Kapitel 4 dargestellt, deren Erkenntnisse in die Artefakt-basierte Entwicklung des MuDDi-Frameworks und -Prototypen (DSRM-Aktivität 3) überführt werden. Gemäß DSRM kann dieses Artefakt „jedes gestaltete Objekt sein, in dessen Design ein Forschungsbeitrag eingebettet ist“ (Peffers et al. 2007, S. 55) – Artefakte und Forschungsbeiträge werden in Kapitel 5 erläutert. Die Ergebnisse werden im Rahmen eines Proof of Concepts in Kooperation mit der öffentlichen Verwaltung getestet, wobei insbesondere sicherheitsrelevante Anwendungsfälle zur Erkennung von KI-manipulierten Bildern und Videos im Kontext verzerrender Einflussnahme auf den öffentlichen Diskurs priorisiert werden. Parallel dazu erfolgen wissenschaftliche und praxisnahe Publikationen und Transferaktivitäten (DSRM Aktivitäten 4–6), die in Kapitel 6 als Ausblick vorgestellt werden.

3 Problemidentifikation: Technische Unterstützung für menschliche Schwierigkeiten bei der Deepfake-Erkennung

Vergleichende Studien aus den USA, Deutschland und China zeigen, dass Einzelpersonen hochqualitative synthetische Inhalte kaum zuverlässig von authentischen Medien unterscheiden können (Frank et al. 2023). In einer systematischen Literaturrecherche aggregierte Werte belaufen sich auf lediglich 56 % Erkennungswahrscheinlichkeit für Deepfakes (Diel et al. 2024) – ein Wert, der in Experimenten

bestätigt wurde (Cooke et al. 2025). Die Möglichkeit der Erkennung sinkt insbesondere mit zunehmendem Alter, bei fremdsprachigen Inhalten oder bei der Identifikation von künstlich erzeugten Gesichtern (Groh et al. 2024). Besonders kritisch ist die Feststellung, dass größeres Vorwissen über synthetische Medien keinen signifikanten Effekt auf die Fähigkeit zur Erkennung solcher Inhalte zeigen (Cooke et al. 2025), da zahlreiche Maßnahmen auf eben diese Verbesserung abzielen, wie Kozyreva et al. (2024) in ihrer umfassenden Toolbox für individuelle Interventionen gegen Online-Fehlinformationen zeigen.

Die Wirksamkeit dieser Maßnahmen, die auf eine Steigerung der menschlichen Urteilskraft zielen, bleibt daher unterschiedlich zu bewerten und ist nicht vollumfänglich untersucht. Erkennungstrainings mit Deepfake-Beispielen konnten ein nachweisbaren Lerneffekt erzielen, während Maßnahmen zur allgemeinen „Deepfake Awareness“ lediglich minimale Verbesserungen bewirken (Diel et al. 2024). Auch Hinweise oder Ratschläge zur Identifikation synthetischer Inhalte führen zu gemischten Ergebnissen, wobei insbesondere Checklisten kaum Effekte zeigen (Somoray und Miller 2023). Dagegen erweisen sich Ansätze wie die Karikaturisierung typischer Synthetisierungsartefakte oder die Unterstützung durch KI-Systeme und kollaborative Gruppenarbeit als besonders wirksam (Fosco et al. 2025). Warnhinweise und Labels können zwar die Unterscheidung zwischen KI-generierten und menschlich produzierten Inhalten erleichtern, beeinflussen jedoch weder die wahrgenommene Glaubwürdigkeit noch die Weiterverbreitungsabsicht in konsistenter Weise (Li und Yang 2024; Wittenberg et al. 2025). Herkunftsnachweise von Mediendateien erscheinen theoretisch vielversprechend, bergen in der Praxis jedoch erhebliche Herausforderungen: Sie erfordern nicht nur Vertrauen in das Nachweissystem selbst, sondern können paradoxerweise das Vertrauen in authentische Inhalte mindern (Feng et al. 2023). Dieselbe Herausforderung des Theorie-Praxis-Transfers stellt sich auch für die online-basierte Bilderrückwärtssuche: Während die Maßnahme vielfach als Hilfsmittel zur Erkennung von Deepfakes empfohlen wird (Aprin et al. 2022; Postiglione et al. 2025), zeigt die Umsetzung, dass ein Training von Proband:innen zur Rückwärtssuche keine Verbesserung bei der Bewertung echter Medieninhalte entwickelt (Qian et al. 2022). Ein typischer Anwendungsfall für die Bilderrückwärtssuche ist die Überprüfung von älteren, bereits veröffentlichten synthetischen Bildern, die fälschlich als Beleg für ein aktuelles Ereignis verbreitet werden (sogenannte „Out-of-Context“-Nutzung).

Diese Befunde zeigen deutlich: Eine ausschließlich menschliche Deepfake-Erkennung reicht nicht aus. Erforderlich sind technische Systeme, die menschenzentriert gestaltet, multimodal ausgerichtet und in bestehende Arbeitsstrukturen integriert sind. Genau hier setzt das durch die Bundesagentur für Sprunginnovationen (SPRIN-D) gefördert Forschungsprojekt MuDDi des FZI Forschungszentrum Informatik an: In einer qualitativen Interviewstudie wurden praxisnahe Anforderungen an Erkennungstools nach Design-Science-Grundsätzen systematisch erhoben. Diese Ergebnisse (siehe Kapitel 4) stellen sicher, dass der zu entwickelnde technische MuDDi-Prototyp in die Einsatzrealität integriert werden kann. Der Anwendungskontext des MuDDi-Prototyps konzentriert sich auf Personen, die aufgrund ihrer Tätigkeit intensiv mit der Erkennung von synthetischen Inhalten konfrontiert sind,

darunter Journalist:innen, Risikoanalyst:innen, Krisenmanager:innen oder Beschäftigte von Sicherheitsbehörden.

Die Entwicklung des MuDDi-Prototyps adressiert zum einen die Anforderungen der Interview-Studie, zum anderen den aktuellen Stand der Forschung. Grundsätzlich gelten multimodale Methoden zur Deepfake-Erkennung als zentral, um die Grenzen unimodaler Systeme zu überwinden und die Erkennungsleistung zu erhöhen (Tan et al. 2025; M. Wang 2025). Gleichwohl bestehen weiterhin erhebliche Forschungslücken: Abbas und Taeihagh (2024) betonen in ihrer systematischen Literaturrecherche die Notwendigkeit robuster multimodaler Verfahren, die audiovisuelle Fälschungen, Attribut-Manipulationen und synthetische Inhalte auch unter Echtzeitanforderungen zuverlässig detektieren können. Weitere Arbeiten identifizieren Defizite bei Skalierbarkeit, Effizienz, Robustheit, Generalisierbarkeit sowie fehlende Benchmarks für eine systematische Evaluation (Khan et al. 2025; Tan et al. 2025). Darüber hinaus erfordern ethische und rechtliche Fragestellungen, die Integration menschlicher Expertise sowie proaktive Ansätze wie digitale Wasserzeichen verstärkte Aufmerksamkeit (Abbas und Taeihagh 2024). MuDDi verbindet diese Dimensionen, indem es technische Innovation mit Nutzer:innen-Anforderungen und Praxistauglichkeit systematisch verknüpft.

4 Zieldefinition: Drei priorisierte Design Principles verbinden Problemstellung und Umsetzung

Im Forschungsprojekt MuDDi haben Bezzaoui et al. (2025) eine domänenübergreifende Anforderungsanalyse für Deepfake-Detektionssysteme durchgeführt. Hierfür wurden 15 deutschsprachige Fachkräfte für Deepfakeerkennung aus Verwaltung, Journalismus, Industrie, Sicherheitsbehörden und Zivilgesellschaft befragt. Die qualitative Studie folgt methodisch Mullarkey und Hevner (2019); die Interviews wurden in MAXQDA deduktiv-induktiv codiert und in zehn Meta-Requirements und fünf Design Principles überführt. Als Teil der DSRM-Aktivität 2 definieren Bezzaoui et al. (2025) somit Zielanforderungen für die nutzer:innenzentrierte Entwicklung des MuDDi-Frameworks und -Prototypen und verbinden technische Machbarkeit mit nutzungsnaher Gestaltung.

Die zehn Meta-Requirements lassen sich in fünf Design Principles (DP) einordnen: Nutzbarkeit (DP1) fordert, dass die Tools ohne technisches Vorwissen nutzbar, intuitiv gestaltet und ohne verpflichtende Anmeldung oder Datenerhebung anwendbar sind – auch für Personen, die beruflich regelmäßig mit synthetischen Inhalten konfrontiert sind. Denn wer die Anwendung nicht versteht, nutzt das System nicht. Transparenz (DP2) unterstreicht, dass Nutzer:innen die Urteile des Tools nachvollziehen können und Wahrscheinlichkeiten statt binärer Entscheidungen erhalten sollen. Über alle Sektoren lehnen die befragten Expert:innen eine Black-Box-Detektion oder binäre Echt-Fake-Labels ab und fordern stattdessen begründete Wahrscheinlichkeiten mit konkreten Hinweisen auf genutzte Signale. Dadurch soll die Software eine professionelle Beurteilung unterstützen, nicht ersetzen. Semantik (DP3) betont die Notwendigkeit einer multimodalen, kontextsensitiven Analyse, die in mehreren Interviews unterstrichen wird. Eine technische Lösung muss Informationen zwischen

Audio, Bilder, Videos und Text verknüpfen, den Kontext prüfen und auf Kohärenz bewerten. Kompatibilität (DP4) betrifft die Integration in bestehende Infrastrukturen und die Einhaltung rechtlicher sowie ethischer Vorgaben, insbesondere des Datenschutzes. Schließlich fordert Governance (DP5) eine transparente, nachvollziehbare und vertrauenswürdige Entwicklung und Verwaltung der Tools.

Die Interviews bestätigen die Ergebnisse der Problemidentifikation: Während aktuell einfache „cheap fakes“ (technisch wenig aufwendige Manipulationen, wie ein echtes Video mit ausgetauschter Tonspur oder irreführender Beschriftung, siehe z. B. Qian et al. (2022)) in der praktischen Arbeit noch dominieren, nimmt die Qualität synthetischer Medien schnell zu, sodass eine rein menschliche Erkennung bald nicht mehr ausreichen wird. Interviewpartner:innen aus Sicherheitsbehörden und Verwaltung prognostizieren hierfür ein enges Zeitfenster; Praktiker:innen aus Medien und Finanzsektor weisen auf eine wachsende, aber je nach Sektor oder Wirtschaftsbereich unterschiedlich schnelle Entwicklung hin (z. B. Identitätsbetrug im Personalbereich oder CEO-Fraud).

Im Sinne des Information Frameworks von Hevner et al. (2004) bilden die Ergebnisse von Bezzaoui et al. (2025) eine Brücke zwischen Problembeschreibung im Forschungskontext und Artefakt-Design als Kern von IS-Forschung (siehe Abb. 2). Die Design Principles 1 bis 3 bilden dabei den harten Kern einer technischen Erkennungslösung, die den Menschen befähigt. Nur so entsteht ein glaubwürdiger, skalierbarer Beitrag zur Sicherung von „Informationsintegrität“ (OECD 2024) – nicht als Ersatz professioneller Urteilskraft, sondern als deren künftige Bewertungsinfrastruktur.

5 Gestaltung & Entwicklung: Framework und Prototyp

5.1 Framework für die automatisierte Deepfake-Erkennung

Auf Basis der zuvor erhobenen Anforderungen und Design Principles wurde ein multimodales, modulares, nutzer:innenzentriertes Framework konzipiert, das zentrale Arbeitsschritte automatisiert und transparente, nachvollziehbare Erkenntnisse über die analysierten Inhalte bereitstellt (siehe Abb. 3).

Zur Sicherstellung eines konsistenten und skalierbaren Workflows definiert das Framework feste Schnittstellen und Übergabepunkte. Dadurch wird sowohl eine standardisierte, vollständig automatische Ausführung als auch die künftige Integration neuer Analyseverfahren ermöglicht.

Für eine breite Anwendbarkeit wurde eine einheitliche Eingangsschnittstelle für Bilder, Videos und Audio definiert (vgl. DP1). Ein gemeinsamer Pre-Processing-Layer harmonisiert multimodale Eingaben für Bild, Video, Audio und verteilt sie an die zuständigen Analysemodule (siehe Kapitel 5.2). Ein wesentlicher Zwischenschritt ist die Information-Retrieval-Phase, in der Kontextinformationen zu den Inhalten Bild, Video oder Audio eingeholt werden (z. B. Veröffentlichungszeitpunkt, Autor:in, Beschreibungen, vgl. DP3). Diese Zusatzinformationen werden insbesondere bei Reasoning-basierten Detektionsansätzen von Sprachmodellen (Large Language Models, LLMs) und bildverarbeitenden Sprachmodellen (Vision Language

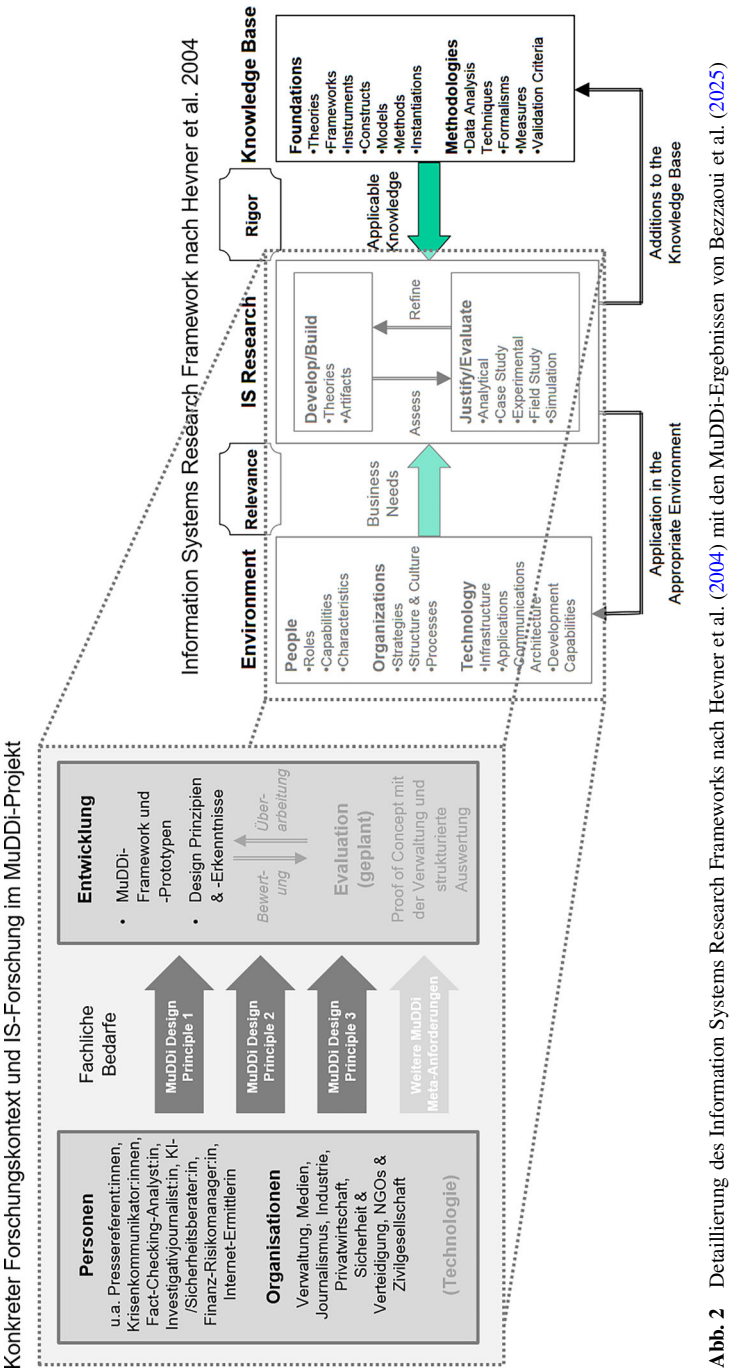

Abb. 2 Detaillierung des Information Systems Research Frameworks nach Hevner et al. (2004) mit den MuDDi-Ergebnissen von Bezzaoui et al. (2025)

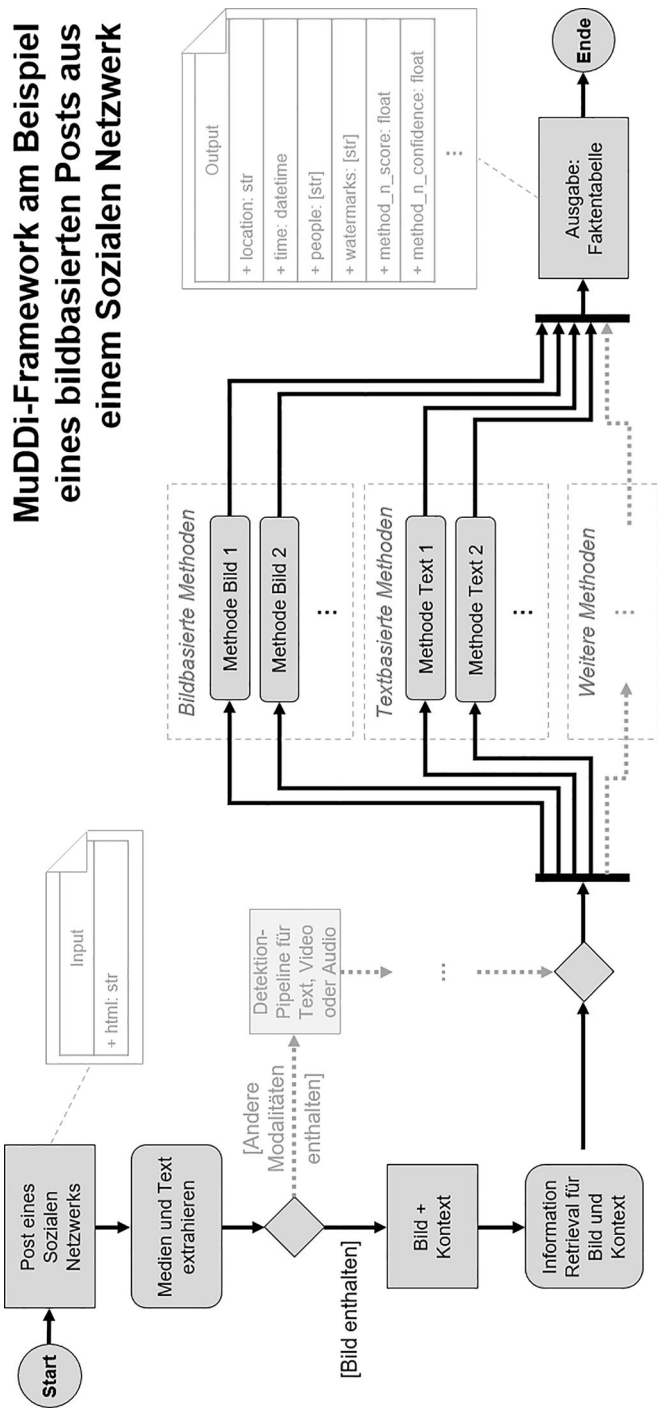

Abb. 3 MuDDi-Framework für die Deepfake-Detection-Pipeline am Beispiel eines bildbasierten Posts aus einem Sozialen Netzwerk

Models, VLMs) genutzt. Die zentralen Ergebnisse des Frameworks werden in einer Faktentabelle basierend auf den Analysemodulen zusammengeführt und erklärt, ohne dabei binäre „Richtig-Falsch"-Kategorien zu vergeben (vgl. DP2). Die Ergebnistabellen können an bestehende Fact-Checking-Plattformen (z. B. CORRECTIV Faktenforum) angebunden oder direkt im System für die Anwender:innen bereitgestellt werden.

Kompatibilität (DP4), also die Anschlussfähigkeit des Prototypen an bestehende Infrastrukturen und rechtliche Vorgaben, und Governance (DP5) im Sinne eines transparenten und überprüfbaren Betriebs der Anwendung sind in der Grundkonzeption des Frameworks und der Prototypen-Umsetzung berücksichtigt. So wurden im Sinne einer offenen und überprüfbaren Code-Basis ausschließlich auf Open Source-Zugänge und -Modelle zurückgegriffen (vgl. DP5) und Datenschutz-Anforderungen von Beginn an integriert („Data Protection by Design", vgl. DP4). Konkrete organisatorische oder rechtliche Anforderungen an ein Deepfake-Detektionssystem ergeben sich jedoch erst im konkreten Anwendungsfall, der ab DSRM-Aktivität 4 (Demonstration) und 5 (Evaluation) im Projekt realisiert werden kann (siehe Kapitel 6).

5.2 Prototyp zur Deepfake-Erkennung in Bildern und Videos

Der MuDDi-Prototyp zeigt die beispielhafte Implementierung wichtiger, multimodaler Analysemodule des Frameworks in Form eines funktionsfähigen Artefakts und realisiert damit die im Framework verankerten Design Principles 1–3.

Aufgrund zahlreicher Missbrauchsmöglichkeiten steht die Gesichtsmanipulation im Mittelpunkt der Deepfake-Generierung (Gong und Li 2024). Da rein Feature-basierte Erkennungsverfahren bei hochqualitativen Gesichtsfakes nicht mehr zuverlässig genug sind, priorisiert MuDDi eine Ende-zu-Ende-Gesichtsbild-Detektion (CNN- und Transformer-basierte Architekturen), die Gesichtsmerkmale implizit mituntersucht und ohne manuell definierte Einzelmerkmale direkt vom Bild zur Echt-/Fälschungs-Einschätzung gelangt. Für das Modell-Training und die Evaluation wurde ein umfassender, divers aggregierter und ausbalancierter Datensatz aus echten und gefälschten Bildern von Gesichtern erstellt (rund 212.000 Trainings-, 20.000 Validierungs- und 20.000 Testbilder basierend u. a. auf DF40, FFHQ, FairFaces und selbstgenerierten Fakes). Für die Analyse im Prototypen wurden Modelle für drei Testszenarien erstellt: Echte vs. synthetisierte Gesichter, originale Gesichter vs. Face Swap, und reale Gesichter aus Image-to-Text-Datensätzen vs. Text-zu-Bild-generierte Gesichtsfakes. Die Modelle erzielten sehr gute Accuracy-, Precision- und Recall-Ergebnisse[1] (zwischen 0,92 bis 0,99), wobei auf Generalisierung ausgerichtete, robustere Varianten schlechter performten (zwischen 0,85 und 0,92). Die im Prototyp integrierte Daten-Pipeline extrahiert Gesichter, passt diese für eine Bewertung an und analysiert ihre Authentizität mithilfe von zwei CLIP-basierten Vision-Transformer-Modellen, einer neuronale Netzarchitektur, die ein Bild in kleine Bild-

[1] Accuracy bezeichnet dabei den Anteil insgesamt korrekter Einordnungen, Precision den Anteil tatsächlicher Fälschungen unter den als Fälschung erkannten Bildern und Recall den Anteil der erkannten unter allen tatsächlichen Fälschungen.

ausschnitte zerlegt und mit dem aus der Sprachverarbeitung bekannten Transformer-Mechanismus auswertet. Falls ein Bild untersucht werden soll, in dem kein Gesicht erkannt wird, nutzt der Prototyp das vortrainierte OpenSight Community Forensic-Modell, ein allgemein einsetzbares Modell zur binären Echt- vs. KI-Erkennung von Bildern (Park und Owens 2025).

Für die Überprüfung von Deepfakes mit Vision-Language-Modellen hat das MuDDi-Projekt ein VLM mithilfe der Reinforcement Learning Technik „Group Relative Policy Optimization“ (GRPO) verfeinert. Dabei handelt es sich um ein bestärkendes Lernverfahren, das pro Eingabe mehrere Antworten erzeugt und sie relativ zum Gruppendurchschnitt bewertet, ohne ein separates Bewertungsmodell zu benötigen (Shao et al. 2024). GRPO bietet hohe Zielgenauigkeit und Effizienz beim Training und nutzt die Vorteile des Logical Reasoning-Ansatzes von Reasoning-Sprachmodellen, wie DeepSeek R1. Für das Training wurde ein spezialisierter Datensatz angepasst (ca. 70.000 Bilder), in dem definierte Gesichtsareale (Augen, Nase, Mund) des FFHQ Datensatzes (Karras et al. 2019) gezielt verändert wurden. Die in den Prototypen integrierte Untersuchungsmethode für Deepfakes zeigt eine hohe Sensibilität des Modells auf kleine visuelle Veränderung (z. B. bei besonders hochqualitativen Deepfakes oder lokal begrenzte Manipulationen einzelner Gesichtsregionen wie Augen, Nase oder Mund bei Gesichtsvertauschungen („Face-Swaps“)), sowie eine hohe Erklärbarkeit und Interpretation von Nicht-Gesichtsbildern.

Wasserzeichen und Metadaten von Bildern sind grundsätzlich aussagekräftige Indikatoren für die Authentizität bzw. den Ursprung eines Bildes – sowohl für echte als auch generierte Inhalte, da sie die Herkunft unmittelbar kennzeichnen (Korus 2017). Während Wasserzeichen, beispielsweise auf KI-Bildern von Google, Meta und Microsoft, schwerer zu entfernen sind, lassen sich Metadaten auf Ergebnissen von OpenAI, Midjourney oder Flux leicht verändern und werden beim Hochladen in soziale Netzwerke häufig automatisch entfernt. Gleichzeitig wird auch die Wirksamkeit von KI-Wasserzeichen kritisiert: Ihre Verifikation ist stark modellabhängig und kann in der Regel nur beim Anbieter selbst erfolgen, wobei dies für automatisierte Überprüfungen zumeist kostenpflichtig ist (Fernandez et al. 2023; Gowal et al. 2025). Der aktuelle Prototyp enthält daher lediglich eine Erkennung für Metadaten sowie einen Open-Source Wasserzeichen-Standard, der bisher wenig verbreitet ist. Erweiterungen für sichtbare und unsichtbare Wasserzeichen sind möglich.

Zeitliche Inkonsistenzen und ungleichmäßige Veränderung in Gesichtsregionen können wesentliche Hinweise auf die synthetische Generierung von Videos sein. Zu deren Detektion hat das MuDDi-Projekt Stirn- und Wangenregionen in zu untersuchenden Videos in 14 Zonen unterteilt und Signale der sogenannte „Remote Photoplethysmography“ (rPPG) extrahiert. Bei echten Videos lassen sich durch eine Zeit-Frequenz-Analyse Informationen zum Herzschlag herauslesen (W. Wang et al. 2017); bei künstlich erzeugten Gesichtern kann dieser Puls jedoch gestört sein und diese damit als künstlich erzeugt ausweisen. Ein speziell trainiertes neuronales Netzwerk lernt solche Unterschiede in Mustern und Bewegungen zu erkennen und authentische von gefälschten Videos zu unterscheiden (Fawaz et al. 2020).

Zusätzlich zur Bild- und Videoanalyse integriert der Prototyp auch ein Modul zur Überprüfung von Audiospuren und Lippenbewegungen. Nach der Transkription mit offenen Speech-to-Text-Modellen werden die Lippenbewegungen im Video mit dem

erkannten Text abgeglichen und Unstimmigkeiten als „Lip Sync Errors“ detektiert. Der Prototyp detektiert, wenn die Ton- bzw. Stimmspur eines Videos durch eine andere ersetzt wurde – eine besonders einfach generierbare Fälschung mit großem Schadenspotenzial (Karaboga et al. 2024).

6 Fazit, Limitationen und Ausblick

Ausgehend von der Forschungsfrage, wie sich technische, multimodale und nutzerzentrierte Unterstützungssysteme für die Deepfake-Detektion gestalten lassen hat das MuDDi-Projekt anwendungsorientierte Anforderungen, ein darauf aufbauendes Framework und einen technischen Prototypen entwickelt. Entlang der DSR-Methodologie zeigt das Projekt, dass eine belastbare Lösung durch die Verbindung von Praxisanforderungen, einem modularen Design und der Integration multimodaler Analyseverfahren möglich ist. Die im Rahmen der Zieldefinition herausgearbeiteten Design Principles auf Basis von Bezzaoui et al. (2025) bestätigen zentrale Befunde der Literatur, wonach rein menschliche Detektionsleistungen angesichts der wachsenden Qualität synthetischer Inhalte unzureichend bleiben und technische Unterstützungssysteme erklärbar und kontextsensitiv ausgestaltet sein müssen. Die entwickelten Artefakte materialisieren die Erkenntnisse der soziotechnischen Design Science-Forschung für die praktische Anwendung in sensiblen Arbeitskontexten. Aus praktischer Perspektive leistet MuDDi damit einen zweifachen Beitrag: Erstens bietet der Prototyp eine unmittelbar nutzbare Grundlage für Fachkräfte, die täglich mit potenziell manipulierten Inhalten konfrontiert sind. Zweitens schafft das Framework eine modulare technische Architektur, die dynamische Weiterentwicklungen – etwa neue Erkennungsverfahren oder zusätzliche Datenmodalitäten – flexibel integrieren kann.

Die Erkenntnisse von MuDDi sind insbesondere durch die bisher fehlende Feldvalidierung beschränkt. Trotz modularer Architektur ist die Stabilität der Detektionsverfahren angesichts der schnellen Weiterentwicklung generativer KI nie vollständig gesichert. Auch fokussiert der bisherige Prototyp primär auf eine technische Audio-, Bild- und Videoanalyse, während komplexere, kontextbasierte Reasoning-Mechanismen noch am Anfang stehen. Ethische, rechtliche und organisatorische Dimensionen können erst in zukünftigen Projektphasen belastbar bewertet werden. Der Ausblick verweist daher auf diese wichtigen, nächsten Schritte: Ein bereits geplanter Proof of Concept in der öffentlichen Verwaltung zur Verifikation verdächtiger Medieninhalte im Kontext der Abwehr von Desinformation ist wesentlich, um die Nutzbarkeit, Robustheit und Detektionsqualität des Systems zu demonstrieren und zu evaluieren (DSRM-Aktivität 4 und 5). Zusätzlich zu diesen Evaluationsergebnissen sind für die iterative Weiterentwicklung besonders Skalierbarkeit, umfassendere LLM-basierte Kontext- und Konsistenzanalysen sowie erweiterte Quellen für Information Retrieval im Fokus. Parallel hierzu erfolgt eine wissenschaftliche und praxisnahe Kommunikation der Forschungs- und Umsetzungsergebnisse, wie in diesem Beitrag (DSRM-Aktivität 6).

Insgesamt zeigt das Projekt, dass technische Deepfake-Detektion nur dann zukunftsfähig ist, wenn sie multimodal, erklärbar und menschenzentriert ausgestal-

tet wird. Die kommenden Forschungsphasen bilden daher nicht nur die nächsten Schritte in der DSR-Methodologie, sondern tragen wesentlich dazu bei, Deepfake-Erkennung als Bestandteil einer breiteren Infrastruktur für Informationsintegrität in digitalen Räumen mit massenhaft KI-generierten Inhalten weiterzuentwickeln.

Danksagung Die Autorinnen und Autoren danken dem gesamten MuDDi-Team am FZI Forschungszentrum Informatik, insbesondere Felix Hertlein, Julia Hofmann, Jacob Langner, Jin Liu, Philipp Reis, Achim Rettinger, Philipp Rigoll, Steffen Thoma, Wilhelm Stork, Martin Zehetner und Christoph Zimmermann.

Funding Open Access funding enabled and organized by Projekt DEAL.

Datenverfügbarkeit Alle dieser Arbeit zugrunde liegenden Daten sind in diesem Artikel enthalten.

Einhaltung ethischer Richtlinien

Interessenkonflikt L. Jarvers, I. Bezzaoui und J. Fegert geben an, dass kein Interessenkonflikt besteht.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden. Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen. Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf http://creativecommons.org/licenses/by/4.0/deed.de.

Literatur

Abbas F, Taeihagh A (2024) Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications 252:124260. https://doi.org/10.1016/j.eswa.2024.124260

Altuncu, E., Franqueira, V. N. L., & Li, S. (2024). Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. Frontiers in Big Data, 7. https://doi.org/10.3389/fdata.2024.1400024

Aprin, F., Chounta, I.-A., & Hoppe, H. U. (2022). "See the Image in Different Contexts": Using Reverse Image Search to Support the Identification of Fake News in Instagram-Like Social Media (S. 264–275). https://doi.org/10.1007/978-3-031-09680-8_25

Berlin, S., & Fehrensen, M. (2025, August 12). Wahn, Psychosen, Suizid: Die dunkelste Seite von KI. Social Media Watch Blog.

Bezzaoui, I., Jarvers, L., Weinhardt, C., & Fegert, J. (2025). Designing deepfake detection systems: Practitioner requirements across sectors. ICIS 2025 Proceedings. https://aisel.aisnet.org/icis2025/hti/hti/20

Bick, A., Blandin, A., & Deming, D. J. (2025). The rapid adoption of generative AI (NBER Working Paper Nr. 32966). National Bureau of Economic Research. https://doi.org/10.3386/w32966

Carpenter, P. (2025). AI, Deepfakes, and the Future of Financial Deception. SEC Investor Advisory Committee.

Cole, R., Purao, S., Rossi, M., & Sein, M. K. (2005). Being Proactive: Where Action Research Meets Design Research. In D. Avison, D. Galletta, & J. I. DeGross (Hrsg.), Proceedings of the International Conference on Information Systems (ICIS) 2005. Association for Information Systems. https://aisel.aisnet.org/icis2005/27

Colman, B. (2025, Juli 7). Why detecting dangerous AI is key to keeping trust alive in the deepfake era. World Economic Forum.
Cooke D, Edwards A, Barkoff S, Kelly K (2025) As good as a coin toss: Human detection of AI-generated content. Communications of the ACM 68(10):100–109. https://doi.org/10.1145/3729417
Corsi, G., Marino, B., & Wong, W. (2024). The spread of synthetic media on X. Harvard Kennedy School Misinformation Review. https://doi.org/10.37016/mr-2020-140
Diel A, Lalgi T, Schröter IC, MacDorman KF, Teufel M, Bäuerle A (2024) Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. Computers in Human Behavior Reports 16:100538. https://doi.org/10.1016/j.chbr.2024.100538
Fawaz HI, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, Webb GI, Idoumghar L, Muller PA, Petitjean F (2020) InceptionTime: Finding AlexNet for time series classification. Data Mining and Knowledge Discovery 34(6):1936–1962. https://doi.org/10.1007/s10618-020-00710-y
Feng KJK, Ritchie N, Blumenthal P, Parsons A, Zhang AX (2023) Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions. Proceedings of the ACM on Human-Computer Interaction 7(CSCW2):1–42. https://doi.org/10.1145/3610061
Fernandez, P., Couairon, G., Jégou, H., Douze, M., & Furon, T. (2023). The stable signature: Rooting watermarks in latent diffusion models. 22409–22420. https://doi.org/10.1109/iccv51070.2023.02053
Fosco, C., Josephs, E., Andonian, A., & Oliva, A. (2025). Deepfake caricatures: Amplifying attention to artifacts increases deepfake detection by humans and machines. https://doi.org/10.48550/arXiv.2206.00535
Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M., & Holz, T. (2023). A representative study on human detection of artificially generated media across countries. https://doi.org/10.48550/arxiv.2312.05976
Gong LY, Li XJ (2024) A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges. Electronics 13(3):585. https://doi.org/10.3390/electronics13030585
Gowal, S., Bunel, R., Stimberg, F., Stutz, D., Ortiz-Jimenez, G., Kouridi, C., Vecerik, M., Hayes, J., Rebuffi, S.-A., Bernard, P., Gamble, C., Horváth, M. Z., Kaczmarczyck, F., Kaskasoli, A., Petrov, A., Shumailov, I., Thotakuri, M., Wiles, O., Yung, J., … Kohli, P. (2025). SynthID-image: Image watermarking at internet scale. arXiv. https://doi.org/10.48550/arxiv.2510.09263
Groh M, Sankaranarayanan A, Singh N, Kim DY, Lippman A, Picard R (2024) Human detection of political speech deepfakes across transcripts, audio, and video. Nature Communications 15(1):7629. https://doi.org/10.1038/s41467-024-51998-z
Hevner A, vom Brocke J, Maedche A (2019) Roles of Digital Innovation in Design Science Research. Business & Information Systems Engineering 61(1):3–8. https://doi.org/10.1007/s12599-018-0571-z
Hevner AR, March ST, Park J, Ram S (2004) Design Science in Information Systems Research1. MIS Quarterly 28(1):75–106. https://doi.org/10.2307/25148625
Jarvers, L., Krämer, J., & Fegert, J. (2026). „AI slop in my feed“: An information systems research agenda for high-volume, low-effort AI-generated content and its influence on online platforms. SSRN. https://doi.org/10.2139/ssrn.6946998
Karaboga M, Rovelli S, Frei N, Puppis M, Vogler D, Raemy P, Ebbers F, Runge G, Rauchfleisch A, de Seta G, Gurr G, Friedewald M (2024) Deepfakes und manipulierte Realitäten – Technologiefolgenabschätzung und Handlungsempfehlungen für die Schweiz
Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks (arXiv:1812.04948). arXiv. https://doi.org/10.48550/arxiv.1812.04948
Khan AA, Laghari AA, Inam SA, Ullah S, Shahzad M, Syed D (2025) A survey on multimedia-enabled deepfake detection: State-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. Discover Computing 28(1):48. https://doi.org/10.1007/s10791-025-09550-0
Kharvi PL (2024) Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media. IEEE Security & Privacy 22(4):115–122. https://doi.org/10.1109/msec.2024.3405963
Korus P (2017) Digital image integrity—a survey of protection and verification techniques. Digit Signal Process 71:1–26. https://doi.org/10.1016/j.dsp.2017.08.009
Kozyreva A, Lorenz-Spreen P, Herzog SM, Ecker UKH, Lewandowsky S, Hertwig R, Ali A, Bak-Coleman J, Barzilai S, Basol M, Berinsky AJ, Betsch C, Cook J, Fazio LK, Geers M, Guess AM, Huang H, Larreguy H, Maertens R, Wineburg S et al (2024) Toolbox of individual-level interventions against online misinformation. Nature Human Behaviour 8(6):1044–1052. https://doi.org/10.1038/s41562-024-01881-0

Kuechler B, Vaishnavi V (2008) On theory development in design science research: Anatomy of a research project. European Journal of Information Systems 17(5):489–504. https://doi.org/10.1057/ejis.2008.40

Li F, Yang Y (2024) Impact of Artificial Intelligence—Generated Content Labels On Perceived Accuracy, Message Credibility, and Sharing Intentions for Misinformation: Web-Based, Randomized, Controlled Experiment. JMIR Formative Research 8:e60024. https://doi.org/10.2196/60024

Mullarkey MT, Hevner AR (2019) An elaborated action design research process model. European Journal of Information Systems 28(1):6–20. https://doi.org/10.1080/0960085x.2018.1451811

Munoz, K. (2024). KI und Wahlen: Brennpunkt USA. Bundeszentrale für politische. https://www.bpb.de/themen/nordamerika/usa/552975/ki-und-wahlen-brennpunkt-usa/

OECD (2024) Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing https://doi.org/10.1787/d909ff7a-en

Park, J., & Owens, A. (2024). Community forensics: Using thousands of generators to train fake image detectors. https://doi.org/10.48550/arxiv.2411.04125

Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S (2007) A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems 24(3):45–77. https://doi.org/10.2753/mis0742-1222240302

Postiglione M, Baldwin J, Denisenko N, Fosdick L, Gao C, Gortner I, Pulice C, Kraus S, Subrahmanian VS (2025) GODDS: The Global Online Deepfake Detection System. Proceedings of the AAAI Conference on Artificial Intelligence 39(28):29685–29687. https://doi.org/10.1609/aaai.v39i28.35367

Qian, S., Shen, C., & Zhang, J. (2022). Fighting cheapfakes: Using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. Journal of Computer-Mediated Communication, 28(1). https://doi.org/10.1093/jcmc/zmac024

Sandoval, M.-P., de Almeida Vau, M., Solaas, J., & Rodrigues, L. (2024). Threat of deepfakes to the criminal justice system: A systematic review. Crime Science, 13(1), 41. https://doi.org/10.1186/s40163-024-00239-1

Senaweera, M., Dissanayake, R., Chamindi, N., Shyamalal, A., Elvitigala, C., Horawalavithana, S., Wijesekara, P., Gunawardana, K., Wickramasinghe, M., & Keppitiyagama, C. (2018). A Weighted Network Analysis of User Migrations in a Social Network. 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), 357–362. https://doi.org/10.1109/icter.2018.8615571

Shao Z, Wang P, Zhu Q, Xu R, Song J, Bi X, Zhang H, Zhang M, Li YK, Wu Y, Guo D (2024) DeepSeekMath: Pushing the limits of mathematical reasoning in open language models (arXiv:2402.03300). arXiv. https://doi.org/10.48550/arxiv.2402.03300

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. Computers in Human Behavior, 149, 107917. https://doi.org/10.1016/j.chb.2023.107917

Tan, D., Yang, Y., Niu, C., Li, S., Yang, D., & Tan, B. (2025). A review of deep learning based multimodal forgery detection for video and audio. Discover Applied Sciences, 7(9), 987. https://doi.org/10.1007/s42452-025-07629-3

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social Media + Society, 6(1). https://doi.org/10.1177/2056305120903408

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an Information System Design Theory for Vigilant EIS. Information Systems Research, 3(1), 36–59. https://doi.org/10.1287/isre.3.1.36

Wang, M. (2025). Deepfake detection: A multimodal survey. ITM Web of Conferences, 78, 02027. https://doi.org/10.1051/itmconf/20257802027

Wang, W., den Brinker, A. C., Stuijk, S., & de Haan, G. (2017). Algorithmic Principles of Remote PPG. IEEE Transactions on Biomedical Engineering, 64(7), 1479–1491. https://doi.org/10.1109/tbme.2016.2609282

Wei, Y., & Tyson, G. (2024). Understanding the Impact of AI-Generated Content on Social Media: The Pixiv Case. Proceedings of the 32nd ACM International Conference on Multimedia, 6813–6822. https://doi.org/10.1145/3664647.3680631

Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. PNAS Nexus, 4(6). https://doi.org/10.1093/pnasnexus/pgaf170

Hinweis des Verlags Der Verlag bleibt in Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutsadressen neutral.