INTERNATIONALE HOCHSCHULE

Bachelorarbeit

Internationale Hochschule

Ethik, Sicherheit und klinische Anforderungen an KI-Systeme in der Psychologie

Peter Wildhaber-Wälchli

28. Februar 2026

INHALTSANGABE

Die zunehmende Integration generativer KI-Systeme in psychologische Beratungs- und Versorgungskontexte wirft grundlegende Fragen nach Verantwortung, Datenschutz und professioneller Rollenabgrenzung auf. Ziel dieser Arbeit war die Entwicklung und Evaluation eines strukturierten, klinisch anschlussfähigen KI-Workflows, der ethische, regulatorische und professionsbezogene Anforderungen systematisch in architektonische Designentscheidungen übersetzt.

Auf Grundlage eines Design-Science-Research-Ansatzes wurde ein integrierter Anforderungskatalog (Kategorien A–D) abgeleitet und in einen klinischen KI-Workflow überführt. Dieser wurde durch eine prototypische Plattformarchitektur sowie zwei exemplarisch instanziierte Artefakte – eine gerätegebundene Passkey-Authentifizierung (D2) und ein Transparenzartefakt zur Begrenzung impliziter Autoritätszuschreibung (B4) – operationalisiert. Die Evaluation erfolgte artefaktbasiert mittels strukturierter Control Sheets auf Prozess-, Architektur- und Instanzebene unter Rückgriff auf DPIA-Logik, LINDDUN und Privacy-by-Design-Prinzipien.

Die Ergebnisse zeigen, dass zentrale delegierbare Pflichten – insbesondere hinsichtlich Datenschutz, Transparenz und struktureller Verantwortungsabsicherung – prinzipiell technisch operationalisierbar sind. Verantwortung wird dabei nicht primär über individuelles Verhalten, sondern über Systemdesign realisiert. Die Arbeit leistet damit einen konzeptionellen Beitrag zur Frage, unter welchen strukturellen Bedingungen KI-Nutzung im Mental-Health-Kontext professionell legitimierbar ist.

Schlüsselwörter: Künstliche Intelligenz, Psychologie, Datenschutz durch Design, KI-Governance, Design Science Research, Transparenz, Professionelle Verantwortung

ABSTRACT

The increasing integration of generative AI systems into psychological counseling and mental health contexts raises fundamental questions regarding responsibility, data protection, and professional role boundaries. The aim of this thesis was to develop and evaluate a structured, clinically compatible AI workflow that systematically translates ethical, regulatory, and professional requirements into architectural design decisions.

Following a Design Science Research approach, an integrated requirements catalogue (categories A–D) was derived and operationalized within a clinical AI workflow. This conceptual artifact was instantiated through a prototype platform architecture and two exemplary technical artifacts: a device-bound passkey authentication mechanism (D2) and a transparency artifact limiting implicit authority attribution (B4). Evaluation was conducted artifact-based using structured control sheets across process, architecture, and implementation levels, drawing on DPIA logic, LINDDUN, and privacy-by-design principles.

The results indicate that key delegable obligations—particularly concerning data protection, transparency, and structural responsibility enforcement—can in principle be technically operationalized. Responsibility thus emerges not primarily from individual user behavior, but from system design. The

thesis contributes a structural framework for determining under which conditions AI use in mental health settings can be considered professionally legitimate.

Keywords: Artificial Intelligence, Psychology, Privacy by Design, AI Governance, Design Science Research, Transparency, Professional Responsibility

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

APA	American Psychological Association
DPIA	Data Protection Impact Assessment
DSG	Datenschutzgesetz
DSR	Design Science Research
DSRM	Design Science Research Methodology
DSGVO	Datenschutz-Grundverordnung
EFPA	European Federation of Psychologists' Associations
EU	Europäische Union
GenAI	Generative Künstliche Intelligenz
GPT	Generative Pre-trained Transformer
KI	Künstliche Intelligenz
LINDUUN	Linking, Identifying, Non-repudiation, Detecting, Data Disclosure, Unawareness, Non-compliance
LLM	Large Language Models
MDR	Medical Device Regulation
OECD	Organization for Economic Cooperation and Development
UI	User Interface
WHO	World Health Organization

1 Einleitung

Die rasante Entwicklung generativer KI-Systeme markiert einen tiefgreifenden Wandel in der Art und Weise, wie Wissen erzeugt, verarbeitet und angewendet wird. Insbesondere im Gesundheitswesen – und hier im sensiblen Feld der psychologischen Beratung und Unterstützung – treffen technologische Innovationspotenziale auf hohe ethische, rechtliche und professionelle Anforderungen. Während KI-Systeme zunehmend Aufgaben wie Dokumentation, Strukturierung komplexer Informationen oder unterstützende Reflexion übernehmen, berühren sie zugleich den Kern psychologischer Arbeit: den verantwortungsvollen Umgang mit vulnerablen Menschen, hochsensiblen Daten und professioneller Deutungsmacht.

Die Psychologie befindet sich damit an einem Schnittpunkt mehrerer Disziplinen. Erkenntnisse aus Informatik, Ethik, Recht und klinischer Praxis konvergieren in einem multidimensionalen Spannungsfeld, in dem technologische Machbarkeit nicht mit ethischer Zulässigkeit oder professioneller Angemessenheit gleichzusetzen ist. Besonders generative KI-Systeme, deren Ausgaben probabilistisch entstehen und deren interne Entscheidungslogiken nur eingeschränkt nachvollziehbar sind, stellen etablierte Konzepte von Verantwortung, Transparenz und Kontrolle infrage.

Vor diesem Hintergrund stehen Psychologinnen und Psychologen zunehmend vor der Herausforderung, KI-gestützte Werkzeuge in ihre Praxis zu integrieren, ohne fundamentale Prinzipien wie Nichtschaden, Autonomie, Vertraulichkeit und professionelle Integrität zu kompromittieren. Die vorliegende Arbeit setzt genau an dieser Schnittstelle an und adressiert die Frage, wie abstrakte ethische und regulatorische Anforderungen in konkrete, technisch implementierbare und klinisch anschlussfähige Gestaltungsprinzipien übersetzt werden können. Ziel ist es, einen verantwortungsvollen KI-Workflow zu entwickeln, der sowohl den normativen Ansprüchen als auch den praktischen Realitäten psychologischer Versorgung gerecht wird, exemplarisch angewendet auf die Plattform mentalhealth-gpt.ch.

1.1 Mental Health im digitalen Wandel

Die psychologische Versorgung befindet sich in einem tiefgreifenden Transformationsprozess. Steigende Belastungsniveaus, wachsender Versorgungsbedarf und strukturelle Engpässe in vielen Gesundheitssystemen erhöhen den Druck auf psychologische Fachpersonen – nicht nur in klinischer, sondern auch in administrativer und organisatorischer Hinsicht.

Parallel dazu verändert die rasante Entwicklung digitaler Technologien, insbesondere generativer KI, die Bedingungen mental-health-bezogener Unterstützung grundlegend: Informationen, Reflexionsangebote und sprachbasierte „Beratung“ sind über Chatbots jederzeit verfügbar, skalierbar und niederschwellig. Damit verschiebt sich das Feld von einem primär institutionell gerahmten Versorgungssystem hin zu einem hybriden Ökosystem aus professionellen, digitalen und informellen Unterstützungsformen.

Empirisch ist diese Entwicklung bereits messbar – und klinisch ziemlich relevant. Eine national erhobene US-Studie im Frühjahr 2025 zeigt, dass 13,1 % der 12- bis 21-Jährigen schon generative KI für Mental-Health-Ratschläge genutzt haben; bei 18- bis 21-Jährigen lag die Rate bei 22,2 % (McBain et al., 2025, 1–3). Unter den Nutzenden suchten 65,5 % mindestens monatlich Rat, und 92,7 % stuften die erhaltene Unterstützung als zumindest „somewhat helpful" ein. Diese Daten unterstreichen, dass GenAI-Interaktionen für viele Ratsuchende bereits ein Teil der Selbstregulation und der Vorverarbeitung emotionaler Probleme sind – oft vor dem ersten professionellen Kontakt. Für die Praxis bedeutet das, dass sich Anamnese, Erwartungshaltungen und Risikodynamiken verändern, weil Klient:innen nicht „unberührt" in den Kontakt kommen, sondern bereits KI-basierte Deutungen, Ratschläge oder Formulierungen mitbringen.

Auch auf Seiten der Profession zeigt sich eine rasch wachsende – zugleich aber heterogene – Nutzung. Eine systematische Erhebungen speziell im mental-health-ärztlichen Kontext fand, dass 44 % der befragten Psychiater:innen ChatGPT-3.5 und 33 % GPT-4 genutzt hatten, u. a. zur Unterstützung bei klinischen Fragen; 70 % erwarteten bzw. erlebten Effizienzgewinne in der Dokumentation (Blease et al., 2024, S. 2–5). Ergänzend liefert eine weitere empirische Studie Hinweise darauf, dass generative KI-Tools inzwischen auch bei psychologischen Fachpersonen substantiell angekommen sind (Aral et al., 2025, S. 4–7). In einer Befragung von Kinder- und Jugendpsychiater:innen sowie Psycholog:innen berichteten 47,9 % der Psychiater:innen und 40 % der Psycholog:innen eine vorherige Nutzung von ChatGPT-4o im beruflichen Kontext. Gleichzeitig zeigte sich, dass der Einsatz bislang vor allem auf sprachlich-administrative und strukturierende Aufgaben fokussiert ist, etwa Zusammenfassungen, Entwürfe oder die Organisation klinischer Informationen.

Bemerkenswert ist jedoch auch, dass diese Nutzung von ausgeprägten Vorbehalten begleitet wird. Beide Berufsgruppen bewerteten insbesondere ethische Fragen, Datenschutz, Transparenz sowie fehlende institutionelle Leitlinien kritisch und äußerten Zurückhaltung gegenüber patientennahen oder quasi-therapeutischen Einsatzformen. Die Studie beschreibt diese Haltung als vorsichtigen Optimismus, der mit einem klaren Bedarf an professioneller Aufsicht, ethischen Leitplanken und rollenspezifischen Rahmenbedingungen einhergeht (Aral et al., 2025, S. 2). Solche Befunde stützen die Annahme, dass generative KI nicht mehr nur experimentell genutzt wird, ihre Integration in die Mental-Health-Praxis jedoch maßgeblich von der Lösung grundlegender ethischer und sicherheitsbezogener Fragen abhängt.

Diese Ergebnisse werden durch berufsständische Datenerhebungen gestützt, die sich explizit auf psychologische Praxis beziehen: Laut APA[1] Practitioner Pulse Survey (2024) berichteten in den USA 71 % der Psycholog:innen im Jahr 2024, nie KI zur Unterstützung ihrer Praxis genutzt zu haben – gleichzeitig nutzte „etwa 1 von 10" KI mindestens monatlich, vor allem für Notizen und administrative

[1] American Psychological Association

Tätigkeiten. Im Folgejahr zeigte derselbe Survey eine starke Dynamik (APA, 2025). Der Anteil, der angab, noch nie KI in der Praxis genutzt zu haben, sank auf 44 % (2025).

Gleichzeitig macht derselbe Survey deutlich, dass die zunehmende Nutzung generativer KI mit wachsenden Vorbehalten einhergeht. Mit zunehmender Vertrautheit nannten Psycholog:innen insbesondere Datenschutzrisiken (67%), soziale Schäden (64 %), Bias in Ein- und Ausgaben (63%) sowie unzureichende Transparenz (52%) als zentrale Bedenken (APA, 2025, S. 5).

So berührt GenAI im Mental-Health-Kontext den Kern professioneller Verantwortung. Mental Health ist kein rein informationsbasiertes Feld, sondern beruht auf Beziehung, Kontextsensitivität und einer ethisch gebundenen Deutungspraxis. Gerade deshalb warnt die Fachliteratur zu generativer KI in Mental Health davor, die „glatte“ Sprachfähigkeit generativer Modelle mit klinischer Angemessenheit zu verwechseln. Systeme können überzeugend formulieren und dennoch inhaltlich falsch, unpassend oder risikoreich sein – besonders in vulnerablen Situationen und bei komplexen Verläufen (Blease & Rodman, 2025, S. 2–6).

1.2 Interdisziplinäre Herausforderungen generativer KI im Mental-Health-Kontext

Die Integration künstlicher Intelligenz in den Bereich der psychologischen Versorgung und mentaler Gesundheit ist nicht nur eine technologische Entwicklung, sondern spiegelt einen tieferen strukturellen Wandel in den Professionen, Praktiken und Erwartungen wider. Der Einsatz von KI-Technologien wie generativen Sprachmodellen (GenAI, Large Language Models, LLMs) eröffnet neue Möglichkeiten, beispielsweise in der Unterstützung bei Routineaufgaben, Datenverarbeitung oder der Skalierung niederschwelliger Angebote. Gleichzeitig wirft ihre Anwendung im sensiblen Feld der mentalen Gesundheit grundlegende Fragen auf, die weit über reine Leistungsfähigkeit hinausgehen und interdisziplinäre Betrachtungsweisen erfordern.

Ein zentraler Aspekt dieses Wandels ist die breite, aber fragmentierte wissenschaftliche Auseinandersetzung mit KI-Anwendungen im Kontext mentaler Gesundheit. Systematische Übersichtsarbeiten betonen, dass KI-Modelle in der Forschung bislang vor allem zur Diagnose, Überwachung und Intervention bei Störungen wie Depression und Angst eingesetzt werden (88%), jedoch sowohl methodologische Limitationen als auch nicht-technische Herausforderungen bestehen. Dazu zählen Datenschutzrisiken, Datenungleichgewichte, mangelnde Robustheit und fehlende klinische Validierung, die die Übertragbarkeit in realweltliche Settings einschränken (Wajid et al., 2025, S. 16–19).

Diese Befunde setzen sich in der Praxis fort. KI-gestützte Systeme werden zwar zunehmend als potenzielle Werkzeuge zur Effizienzsteigerung im Mental-Health-Kontext diskutiert, ihre Rolle in direkten Interaktions- und Versorgungssituationen bleibt jedoch umstritten. Die systematische Übersichtsarbeit von Wang et al. (2025, 5–7) zeigt, dass die Forschung zu generativer KI in mental-health-bezogenen Anwendungen bislang überwiegend experimentell und aufgabenbasiert angelegt ist und einzelne Fähigkeiten isoliert untersucht. In der Folge entsteht eine fragmentierte Wissens-

basis, in der Fragen der kontextsensitiven Anwendung, der ethischen Einbettung und der Übertragbarkeit in reale klinische Arbeitsprozesse bislang nur unzureichend systematisch adressiert werden.

Vor diesem Hintergrund werden in der Fachliteratur zunehmend normative und praxisbezogene Vorbehalte diskutiert. Hillebrand und Baumeister (2025, S. 158–159) zeigen in ihrem fachlichen Kommentar, dass KI-basierte Tools zwar als potenzielle Unterstützung wahrgenommen werden, zugleich jedoch klare Grenzen gesehen werden – insbesondere dort, wo kontextuelles Verstehen, situative Einschätzung und professionelle Verantwortung erforderlich sind. Die Autor:innen betonen, dass generative KI menschliche Urteilskraft nicht ersetzen kann, und warnen vor einer unkritischen Übertragung experimenteller Ergebnisse auf reale psychotherapeutische Settings.

Diese Ambivalenz zeigt sich auch in der Debatte über die Rolle von KI-gestützten Chatbots und digitalen Assistenten. So betonen etwa Torous und Topol (2025, S. 683), dass generative KI zwar neue Möglichkeiten für skalierbare Unterstützung eröffnet und erste Wirksamkeitshinweise vorliegen, die Evidenz jedoch überwiegend frühphasig ist und ihre Übertragbarkeit in reale Versorgungskontexte begrenzt bleibt. Insbesondere weisen die Autoren darauf hin, dass KI-Systeme keine klinische Verantwortung oder Haftung übernehmen können und daher menschliche Fachpersonen weiterhin eine zentrale Rolle bei Sicherheit, Kontextualisierung und professioneller Entscheidungsfindung einnehmen müssen.

Zudem weist die Literatur darauf hin, dass die ethischen und sozialen Implikationen generativer KI im Mental-Health-Kontext über rein technische Fragestellungen hinausgehen. Pandey (2024, S. 7–8) hebt hervor, dass Large Language Models Verzerrungen aus Trainingsdaten reproduzieren, anthropomorphe Zuschreibungen begünstigen und in bestimmten Situationen irreführende oder kontextuell unzureichende Ausgaben erzeugen können. Angesichts der sensiblen und häufig vulnerablen Nutzungssituationen im Bereich mentaler Gesundheit unterstreicht der Autor die Notwendigkeit klarer ethischer Leitplanken, robuster Evaluationsmechanismen und kontinuierlicher Überwachung, um Vertrauen, Verantwortlichkeit und Fairness zu gewährleisten.

Schließlich zeigt sich, dass die Herausforderungen beim Einsatz von KI im Mental-Health-Kontext auch in hohem Maße organisatorische und strukturelle Dimensionen betreffen (Nair et al., 2025, S. 8–10). Empirische Implementierungsstudien aus dem Gesundheitswesen verdeutlichen demnach, dass die Einführung KI-basierter Systeme bestehende Arbeitsprozesse, Rollenverteilungen und Verantwortlichkeiten verändert und ohne strukturierte Einbettung erhebliche Risiken birgt. Insbesondere werden das Fehlen klarer Implementierungsleitlinien, unzureichende Governance-Strukturen sowie mangelnde Einbindung relevanter Akteur:innen als zentrale Faktoren identifiziert, die zu Unsicherheit, zusätzlicher Belastung und potenzieller Verantwortungsdiffusion bei Fachpersonen führen können.

In einer groß angelegten Erhebung unter US-Gesundheitssystemen zeigen Poon et al. (2025, S. 1097–1099), dass KI-Anwendungen zwar zunehmend erprobt werden, ihre Integration jedoch

durch unreife Werkzeuge, regulatorische Unsicherheit und fehlende institutionelle Governance erheblich erschwert wird. Die Autoren betonen daher die Notwendigkeit robuster Bewertungs- und Implementierungsstrukturen, um KI nachhaltig und sicher in den klinischen Alltag zu integrieren. Zugleich weist die Implementierungsforschung darauf hin, dass technische, ethische und organisationale Aspekte der KI-Nutzung im Gesundheitswesen bislang nicht ausreichend integriert betrachtet werden. Reddy (2024, S. 13–14) argumentiert, dass technologische Leistungsfähigkeit allein komplexe Versorgungssysteme nicht transformieren kann und fordert daher, KI-Anwendungen konsequent als sozio-technische Systeme zu analysieren, um ihre Auswirkungen auf klinische Praxis, Verantwortungsstrukturen und professionelle Rollen angemessen zu erfassen.

1.3 Zielsetzung und Forschungsansatz: Bridging the Principles-to-Practice Gap

Während Prinzipien wie Transparenz, Verantwortlichkeit oder menschliche Aufsicht breit anerkannt sind, bleibt häufig unklar, wie sie in konkrete technische, organisatorische und professionelle Strukturen übersetzt werden können, die im Arbeitsalltag tatsächlich handlungsleitend sind (Ibáñez & Olmeda, 2022, S. 1670–1677).

Diese Problematik wird etwa bei Herzog und Blank (2024, S. 15–16) als «Principles-to-Practice Gap» beschrieben. Sie argumentieren dabei aus einer systemischen Perspektive, dass dominante Diskurse zur „ethischen KI“ dazu neigen, normative Anforderungen zu abstrahieren, ohne die sozio-technischen Kontexte zu berücksichtigen, in denen KI-Systeme real eingesetzt werden. Dadurch entstehen Leitlinien, die zwar normativ überzeugend sind, jedoch kaum Orientierung für konkrete Gestaltungs- und Integrationsentscheidungen bieten. Die Autor:innen plädieren daher für Ansätze, die ethische Anforderungen nicht isoliert formulieren, sondern sie als integralen Bestandteil kollaborativer, organisations- und prozessbezogener Gestaltung verstehen.

Gerade im Mental-Health-Kontext ist diese Umsetzungslücke besonders relevant. Hier treffen hohe Anforderungen an Vertraulichkeit, professionelle Verantwortung und den Schutz vulnerabler Personen auf den Wunsch nach praktikabler Unterstützung durch KI-Systeme. Qualitative Befunde zeigen, dass psychologische Fachpersonen zwar potenzielle Entlastung durch KI erkennen, jedoch häufig über keine klaren, organisational verankerten und operationalisierten Strukturen verfügen, um KI verantwortungsvoll, regelkonform und professionell einzusetzen. Verantwortungsvolle Nutzung bleibt damit vielfach eine individuelle Aushandlungsleistung und ist bislang nur unzureichend systemisch unterstützt (Zhang et al., 2023).

Vor diesem Hintergrund zielt die vorliegende Arbeit darauf ab, einen Beitrag zur Schließung dieser «Prinzipien-zu-Praxis-Lücke» zu leisten.

Im Zentrum steht dabei nicht die Entwicklung oder Bewertung einzelner KI-Modelle, sondern die exemplarische und partielle Operationalisierung ethischer und regulatorischer Anforderungen in einer strukturierten, praxisnahen Form.

Zu diesem Zweck verfolgt die Arbeit einen gestaltungsorientierten Forschungsansatz. Mithilfe der Design Science Research (DSR) wird ein strukturierter, klinisch anschlussfähiger und technisch implementierbarer KI-Workflow als gestaltete Artefakte entwickelt, die normative Anforderungen von Beginn an als Design- und Strukturprinzipien integrieren.

Die Zielsetzung der Arbeit umfasst im Einzelnen:

I. die systematische Ableitung eines integrierten Anforderungskatalogs aus den Bereichen Ethik, Psychologie, Regulierung (u. a. EU AI Act, DSGVO, ISO) und Technik für beratende KI-Systeme,
II. die Entwicklung eines strukturierten, klinisch anschlussfähigen und implementierbaren KI-Workflows nach dem Prinzip «Responsible Clinical AI by Design»,
III. die partielle Implementierung und prototypische Anwendung ausgewählter Workflow-Elemente im Fallbeispiel mentalhealthGPT,
IV. sowie eine kritisch-reflektierende Bewertung der gewonnenen Umsetzungserkenntnisse und die Ableitung praxisorientierter Empfehlungen für die psychologische Praxis.

Damit positioniert sich die Arbeit an der Schnittstelle von Psychologie, Ethik und Informatik und adressiert ein zentrales, bislang nur unzureichend operationalisiertes Problemfeld der aktuellen KI-Forschung.

1.4 Aufbau der Arbeit

Ausgehend von einer theoretischen und regulatorischen Fundierung wird zunächst ein Rahmen geschaffen, der verdeutlicht, welche Anforderungen an beratende KI-Systeme aus ethischer, rechtlicher und praktisch-psychologischer Perspektive gestellt werden (Kapitel 2). Darauf aufbauend wird ein methodischer Zugang entwickelt, der es erlaubt, diese Anforderungen nicht nur analytisch zu erfassen, sondern gezielt in gestaltbare Strukturen zu überführen (Kapitel 3). Die Wahl eines gestaltungsorientierten Forschungsansatzes trägt dem Umstand Rechnung, dass das zentrale Erkenntnis–interesse dieser Arbeit nicht in der Beschreibung bestehender Zustände liegt, sondern in der Entwicklung einer praktikablen Lösung für ein identifiziertes Umsetzungsproblem.

Die anschließende Ausarbeitung und exemplarische Anwendung des entwickelten KI-Workflows dient dazu, die abstrakten Anforderungen in konkrete Entscheidungslogiken, Prozessschritte und Verantwortlichkeiten zu übersetzen (Kapitel 4). Abschließend werden die gewonnenen Erkenntnisse kritisch reflektiert, ihre Reichweite und Grenzen diskutiert und Implikationen für die psychologische Praxis sowie für zukünftige Forschung abgeleitet (Kapitel 5). Auf diese Weise verbindet die Arbeit konzeptionelle Analyse mit gestaltungsorientierter Umsetzung und soll zur Schließung der identifizierten «Principles-to-Practice Gaps» beitragen.

2 Theoretischer und regulatorischer Hintergrund

Generative KI-Systeme markieren im Feld psychologischer Beratung und Mental Health einen qualitativen Bruch mit früheren digitalen Technologien. Erstmals stehen Systeme zur Verfügung, die nicht nur Informationen bereitstellen oder standardisierte Abläufe automatisieren, sondern sprachlich interagieren, Bedeutung konstruieren und kontextbezogene Antworten generieren. Damit rücken sie näher an jene Tätigkeiten heran, die bislang als genuin menschlich und professionell galten: das Strukturieren komplexer Lebenslagen, das Formulieren von Deutungsangeboten und die Begleitung individueller Entscheidungsprozesse.

Diese Entwicklung eröffnet einerseits neue Möglichkeiten der psychologischen Unterstützung, etwa durch Entlastung professioneller Akteur:innen oder die Strukturierung komplexer Informations- und Reflexionsprozesse. Zugleich verschärfen sich jedoch grundlegende Fragen nach der Angemessenheit, Sicherheit und Verantwortbarkeit solcher Systeme. Generative KI operiert probabilistisch, datengetrieben und häufig mit begrenzter Transparenz ihrer Entscheidungslogiken. Ihre Nutzung berührt damit nicht nur Fragen der Wirksamkeit, sondern auch ethische Grundsätze, datenschutzrechtliche Anforderungen und sicherheitsrelevante Schutzmechanismen.

Gerade im Mental-Health-Kontext, in dem mit hochsensiblen personenbezogenen Daten und oft vulnerablen Personen gearbeitet wird, können diese Aspekte nicht nachgelagert betrachtet werden. Der verantwortungsvolle Einsatz generativer KI setzt vielmehr voraus, dass Potenziale und Grenzen nicht isoliert diskutiert, sondern in einen normativen und regulatorischen Bezugsrahmen eingebettet operationalisiert werden.

2.1 KI in der psychologischen Beratung: Potenziale und Grenzen

Der Einsatz generativer KI in der psychologischen Beratung verändert weniger einzelne Arbeitsabläufe als vielmehr die normativen und regulatorischen Koordinaten professioneller Praxis. Aus ethischer und rechtlicher Perspektive liegt die Relevanz solcher Systeme daher nicht primär in ihrer technischen Leistungsfähigkeit, sondern in der Art und Weise, wie sie bestehende Versorgungsstrukturen ergänzen, Verantwortlichkeiten neu konfigurieren und Anforderungen an Transparenz, Aufsicht und Schutzbedarfe verschieben. Entsprechend heben aktuelle Analysen hervor, dass KI-gestützte Anwendungen insbesondere dort einen verantwortungsvollen Beitrag leisten können, wo sie den Zugang zu Angeboten verbessern, administrative und organisatorische Belastungen reduzieren oder unterstützende und strukturierende Funktionen übernehmen, ohne auf autonome diagnostische oder therapeutische Entscheidungen ausgerichtet zu sein (OECD, 2024, S. 5–8).

Gleichzeitig weist die ethische Fachliteratur im Bereich der Global-Health-Forschung darauf hin, dass solche Anwendungen nur dann als verantwortbar gelten können, wenn ihre Rolle klar begrenzt bleibt, Erwartungen transparent kommuniziert werden und keine implizite Gleichsetzung mit professioneller psychologischer Beratung erfolgt (Shaw et al., 2024, S. 4–7).

Gleichzeitig treten an diesen Schnittstellen die Grenzen generativer KI besonders deutlich hervor. Zentrale ethische Herausforderungen betreffen die fehlende Verantwortungs- und Haftungsfähigkeit sprachbasierter KI-Systeme, das Risiko systematischer Verzerrungen sowie die Gefahr irreführender Autoritätszuschreibungen. Wie Weidinger et al. (2021, S. 9–34) zeigen, können insbesondere große Sprachmodelle durch ihre dialogische Form und semantische Kohärenz den Eindruck von Empathie, Verstehen oder situativer Angemessenheit erzeugen, ohne tatsächlich Verantwortung übernehmen oder soziale und emotionale Kontexte zuverlässig einordnen zu können.

Hinzu kommen erhebliche datenschutz- und sicherheitsrelevante Anforderungen. Der EU AI Act macht deutlich, dass Transparenz, Nachvollziehbarkeit und der Schutz von Grundrechten konstitutive Voraussetzungen für die rechtmäßige und verantwortungsvolle Nutzung KI-gestützter Systeme darstellen (Europäische Union, 2024).

2.2 Integrierter ethischer Rahmen für KI in der Mental-Health-Praxis

Ethik bezeichnet in dieser Arbeit den systematischen Rahmen normativer Prinzipien und professioneller Verpflichtungen, die darauf abzielen,

- menschliches Wohlergehen zu schützen,
- Schaden zu vermeiden,
- Autonomie zu wahren,
- Vertraulichkeit sicherzustellen und
- Verantwortung im Umgang mit asymmetrischen Macht-, Wissens- und Abhängigkeitsverhältnissen zu übernehmen.

Ethik fungiert dabei nicht als Sammlung technischer Regeln, sondern als Orientierungsrahmen für verantwortliches Handeln in komplexen, situationsabhängigen Kontexten.

Im Unterschied zu rechtlichen oder technischen Normen zielt Ethik nicht auf vollständige Vorhersagbarkeit oder die Standardisierung von Handlungen ab. Vielmehr dient sie der normativen Orientierung und Ordnung menschlicher Entscheidungen unter Bedingungen von Unsicherheit, Kontextabhängigkeit und Kontingenz (Floridi et al., 2018, S. 694–700). Diese Funktion ist insbesondere in professionellen Praxisfeldern zentral, in denen menschliche Urteilskraft, situative Faktoren und relationale Dynamiken nicht vollständig formalisiert oder automatisiert werden können, sondern einer kontextsensitiven ethischen Abwägung bedürfen.

2.2.1 Ethik im Kontext von KI und Mental Health

Der Einsatz generativer KI im Mental-Health-Kontext verschärft die Bedeutung ethischer Orientierung, da KI-Systeme zunehmend in Interaktionsräume vordringen, die bislang menschlicher Urteilsfähigkeit vorbehalten waren. Anders als frühere digitale Werkzeuge agieren generative Systeme sprachlich, kontextbezogen und potenziell überzeugend. Dadurch berühren sie Fragen von Deutungshoheit, Autorität und Verantwortung in besonderem Maße.

Zugleich sind KI-Systeme keine moralischen Akteure. Sie verfügen weder über Intentionalität noch über Verantwortungsfähigkeit. Ethische Verantwortung verbleibt daher grundsätzlich bei menschlichen Akteur:innen und den Institutionen, die KI-Systeme entwickeln, einsetzen und kontrollieren. Für den Einsatz von KI folgt daraus jedoch keine ethische Irrelevanz, sondern eine erhöhte Gestaltungsverantwortung: KI-Systeme müssen so in professionelle Kontexte eingebettet werden, dass sie menschliche Verantwortung nicht ersetzen, verschleiern oder unterlaufen (Miao et al., 2023, S. 16–20).

Internationale Leitlinien betonen diese Perspektive ausdrücklich. Die World Health Organization fordert für den Einsatz von KI im Gesundheits- und Mental-Health-Bereich «to assist humans», dass menschliche Autonomie, Verantwortlichkeit und Aufsicht gewahrt bleiben und der Schutz vulnerabler Gruppen oberste Priorität besitzt (World Health Organization, 2021, S. 25–29).

2.2.2 Berufsethische Anforderungen in der Psychologie und regulatorische Verdichtung

Konkrete ethische Anforderungen für die psychologische Praxis sind in berufsständischen Kodizes verbindlich formuliert. Die Ethical Principles of Psychologists and Code of Conduct der American Psychological Association (APA) definieren fünf grundlegende ethische Prinzipien, die auch beim Einsatz KI-gestützter Werkzeuge uneingeschränkt gelten (APA, 2017, S. 3–4):

- Beneficence and Nonmaleficence: Psycholog:innen sind verpflichtet, das Wohlergehen der ihnen anvertrauten Personen zu fördern und Schaden aktiv zu vermeiden.
- Fidelity and Responsibility: Professionelles Handeln beruht auf Vertrauensbeziehungen und klarer Verantwortungsübernahme.
- Integrity: Psycholog:innen sind zur Wahrhaftigkeit, Transparenz und Vermeidung irreführender Darstellungen verpflichtet.
- Justice: Der Grundsatz der Gerechtigkeit verlangt einen fairen und nicht-diskriminierenden Zugang zu psychologischen Leistungen sowie die Vermeidung systematischer Benachteiligungen.
- Respect for People’s Rights and Dignity: Dieses Prinzip betont die Achtung von Autonomie, Privatsphäre, Vertraulichkeit und Selbstbestimmung.

Diese Normen enthalten zwar keine technischen Handlungsanweisungen, formulieren jedoch klare Verantwortungszuweisungen. Psychologische Fachpersonen bleiben für den Einsatz von Werkzeugen – auch KI-basierten – voll verantwortlich und dürfen professionelle Urteilsprozesse nicht an externe Systeme delegieren.

Für den europäischen Kontext konkretisiert der «Meta-Code of Ethics» der European Federation of Psychologists’ Associations (EFPA) diese Anforderungen. Er betont insbesondere Transparenz gegenüber Klient:innen, die Pflicht zur klaren Rollen- und Methodendarstellung sowie den besonderen Schutz vulnerabler Personen (European Federation of Psychologists Associations [EFPA], 2025,

S. 3–6). Auch hier wird explizit festgehalten, dass professionelle Verantwortung unteilbar und nicht delegierbar ist.

Auch der EU Artificial Intelligence Act greift zentrale ethische Anliegen auf und überführt sie in rechtlich überprüfbare Mindestanforderungen, die sich unmittelbar auf die Gestaltung und Governance KI-basierter Systeme auswirken (Europäische Union, 2024). So verpflichten die Transparenzregelungen dazu, Eigenschaften, Zweckbestimmung und Einsatzgrenzen von KI-Systemen nachvollziehbar zu dokumentieren und gegenüber relevanten Akteuren offenzulegen (Art. 53 EU AI Act).

Darüber hinaus zieht der AI Act verbindliche Schutzgrenzen für den Einsatz von KI-Systemen, indem bestimmte Praktiken untersagt werden, die auf manipulative Beeinflussung oder die Ausnutzung besonderer Vulnerabilität abzielen. Der Schutz von Grundrechten und vulnerablen Gruppen bildet damit einen übergeordneten normativen Rahmen, innerhalb dessen besonders sensible Anwendungskontexte – etwa im Gesundheits- oder Sozialbereich – restriktiv zu behandeln sind (Art. 5 EU AI Act; Erwägungsgründe 28, 48).

2.2.3 Ethik als anthropozentrischer Bezugsrahmen

Ethische Normen lassen sich also nicht in vollständig regelbasierte Handlungsanweisungen überführen. Als prinzipienbasierte Orientierungssysteme formulieren sie keine deterministischen Regeln, sondern sogenannte prima-facie-Verpflichtungen, die nicht als starre Vorgaben, sondern als kontextsensitive Abwägungsmaßstäbe zu verstehen sind und in konkreten Situationen interpretiert und gegeneinander gewichtet werden müssen (Beauchamp & Childress, 2019, S. 12–23). Diese Offenheit ist kein Mangel, sondern eine notwendige Voraussetzung dafür, ethische Verantwortung in komplexen, individuellen und nicht vollständig vorhersehbaren Handlungskontexten wahrnehmen zu können.

Beauchamp und Childress (2019, S. 21–23) verstehen ethische Prinzipien dabei ausdrücklich als an menschliche Akteur:innen gerichtete Orientierungsnormen, deren Anwendung verantwortliche Urteilskraft, Kontextwissen und professionelle Erfahrung voraussetzt. Die Konkretisierung ethischer Prinzipien („specification“) und ihre Abwägung im Einzelfall („balancing“) sind keine mechanischen oder algorithmisch vollständig abbildbaren Prozesse, sondern erfordern reflexive Bewertung unter Unsicherheit.

Gerade im Mental-Health-Kontext wäre eine vollständige Formalisierung ethischer Anforderungen weder fachlich angemessen noch normativ wünschenswert. Psychologische Beratung – ob mit oder ohne KI-Unterstützung – beruht auf situativer Urteilsfähigkeit, professioneller Reflexion und verantwortlicher Entscheidung. Daraus folgt nicht die ethische Unzulässigkeit KI-gestützter Systeme, sondern die Anforderung, ihren Einsatz so zu rahmen, dass diese menschlichen Fähigkeiten nicht substituiert, sondern geschützt und unterstützt werden.

2.3 Datenschutz und Sicherheit

Der Einsatz generativer KI-Systeme im Bereich der psychologischen Beratung ist untrennbar mit der Verarbeitung sensibler personenbezogener Daten verbunden. Informationen zu psychischer Gesundheit, emotionalem Erleben oder biografischen Belastungen zählen sowohl nach europäischem als auch nach schweizerischem Datenschutzrecht zu besonders schützenswerten Personendaten. Daraus ergibt sich eine erhöhte regulatorische Verantwortung, die Datenschutz und IT-Sicherheit nicht als nachgelagerte Compliance-Aufgaben, sondern als konstitutive Voraussetzungen verantwortungsvoller Systemgestaltung begreift.

2.3.1 Datenschutzrechtliche Grundprinzipien (DSGVO und Schweizer DSG)

Sowohl die Datenschutz-Grundverordnung der Europäischen Union (Europäische Union, 2016) als auch das revidierte Schweizer Datenschutzgesetz (Schweizerische Eidgenossenschaft, 2025) folgen einem prinzipienbasierten Regulierungsansatz. Zentrale Anforderungen sind:

Zweckbindung: Personenbezogene Daten dürfen nur für klar definierte, legitime und transparente Zwecke verarbeitet werden (Art. 5 Abs. 1 lit. b DSGVO; Art. 6 Abs. 3 DSG). Eine nachträgliche Zweckausweitung – etwa durch Wiederverwendung für Trainings- oder Analysezwecke – ist nur unter engen Voraussetzungen zulässig.

Datenminimierung und Verhältnismäßigkeit: Es dürfen nur diejenigen Daten verarbeitet werden, die für den jeweiligen Zweck erforderlich sind (Art. 5 Abs. 1 lit. c DSGVO; Art. 6 Abs. 2 DSG). Für KI-Systeme bedeutet dies, dass sowohl Datenerhebung als auch -speicherung funktional begrenzt und begründbar sein müssen.

Transparenz und Informationspflichten: Betroffene Personen müssen nachvollziehbar darüber informiert werden, welche Daten zu welchem Zweck verarbeitet werden und welche Rolle automatisierte Systeme dabei spielen (Art. 12–14 DSGVO; Art. 19–21 DSG).

Integrität, Vertraulichkeit und Zugriffskontrolle: Verantwortliche sind verpflichtet, durch geeignete technische und organisatorische Maßnahmen ein angemessenes Sicherheitsniveau zu gewährleisten (Art. 5 Abs. 1 lit. f, Art. 32 DSGVO; Art. 8 DSG).

Rechenschaftspflicht: Die Einhaltung dieser Grundsätze muss nicht nur gewährleistet, sondern auch nachweisbar dokumentiert werden (Art. 5 Abs. 2 DSGVO; Art. 7 DSG).

Diese Prinzipien formulieren keine konkreten technischen Lösungen, setzen jedoch klare normative Leitplanken für die Gestaltung datenverarbeitender Systeme. Sie adressieren explizit die organisatorische Verantwortung derjenigen Stellen, die KI-Systeme einsetzen oder betreiben, und begrenzen die Möglichkeit, Verantwortung an technische Artefakte auszulagern.

2.3.2 Ergänzende Governance-Anforderungen des EU Artificial Intelligence Act

Der EU Artificial Intelligence Act ergänzt den datenschutzrechtlichen Rahmen um verbindliche Anforderungen an die Gestaltung und den Einsatz von KI-Systemen, die unabhängig von einer

Hochrisiko-Einstufung gelten. Für KI-Anwendungen im Beratungs- und Mental-Health-Kontext ergeben sich insbesondere aus den Regelungen zu verbotenen Praktiken, Transparenzpflichten gegenüber Nutzer:innen sowie dem übergeordneten Schutz von Grundrechten konkrete rechtliche Leitplanken für die Systemgestaltung (Europäische Union, 2024).

Zentrale, unmittelbar einschlägige Vorgaben sind dabei:

Verbot manipulativer oder ausnutzender KI-Praktiken: Der AI Act untersagt KI-Systeme, die darauf abzielen, das Verhalten von Personen durch manipulative Techniken wesentlich zu beeinflussen oder gezielt die Vulnerabilität bestimmter Gruppen auszunutzen. Diese Schutzgrenze ist für psychologisch sensible Anwendungskontexte zentral, da hier ein erhöhtes Risiko von Übervertrauen, emotionaler Abhängigkeit oder verdeckter Beeinflussung besteht (Art. 5 EU AI Act; Erwägungsgründe 28, 48).

Schutz vulnerabler Personen und Grundrechte: KI-Anwendungen dürfen nicht so ausgestaltet oder eingesetzt werden, dass Autonomie, Entscheidungsfreiheit oder psychische Integrität von Nutzer:innen untergraben werden. Der Schutz vulnerabler Gruppen fungiert dabei als verbindlicher Auslegungsmaßstab für zulässige Einsatzformen, insbesondere in gesundheits- und beratungsnahen Kontexten (Art. 5 EU AI Act; Erwägungsgründe 28, 48).

Transparenzpflichten gegenüber Nutzer:innen: Für bestimmte KI-Systeme schreibt der AI Act vor, dass Nutzer:innen darüber informiert werden müssen, dass sie mit einem KI-System interagieren. Ziel dieser Pflicht ist es, informierte Nutzung zu ermöglichen und irreführende Zuschreibungen menschlicher Eigenschaften oder professioneller Autorität zu vermeiden (Art. 50 EU AI Act).

Transparenz als rechtlich verbindliche Nutzungsvoraussetzung: Die Offenlegung der KI-Interaktion ist keine freiwillige Maßnahme, sondern eine rechtlich verpflichtende Voraussetzung für den zulässigen Einsatz entsprechender Systeme. Für die Systemgestaltung folgt daraus die Anforderung, Interaktionsformate, Sprache und Nutzerführung so auszugestalten, dass der KI-Charakter der Anwendung jederzeit erkennbar bleibt (Art. 50 EU AI Act).

2.4 Zweckbestimmung als Abgrenzungskriterium von Hochrisiko-KI

Für die regulatorische Einordnung KI-gestützter Anwendungen im Mental-Health-Kontext ist zwischen beratenden, unterstützenden Systemen und medizinischen Produkten im Sinne des Medizinprodukterechts zu unterscheiden. Diese Abgrenzung ist sowohl für die Anwendung des EU Artificial Intelligence Act als auch der Medical Device Regulation (MDR) von zentraler Bedeutung.

Nach der MDR (Europäische Union, 2017) gelten Software und digitale Systeme nur dann als Medizinprodukte, wenn sie vom Hersteller ausdrücklich zu medizinischen Zwecken bestimmt sind, insbesondere zur Diagnose, Prävention, Überwachung, Vorhersage oder Behandlung von Krankheiten (Art. 2 Nr. 1 MDR). Maßgeblich ist dabei nicht die faktische Nutzung durch Anwender:innen, sondern die vom Anbieter definierte Zweckbestimmung („intended purpose“, Art. 2 Nr. 12 MDR).

Das im Rahmen dieser Arbeit betrachtete KI-System ist ausdrücklich nicht auf diagnostische oder therapeutische Entscheidungen ausgerichtet. Es übernimmt weder die Bewertung psychischer Zustände noch die Ableitung medizinischer Maßnahmen, sondern erfüllt ausschließlich unterstützende, strukturierende und informationsbezogene Funktionen.

Auch im Sinne des EU Artificial Intelligence Act (Europäische Union, 2024) ergibt sich daraus keine Einstufung als Hochrisiko-KI-System. Hochrisiko-KI liegt insbesondere dann vor, wenn KI-Systeme als Sicherheitskomponenten medizinischer Produkte eingesetzt werden oder selbst medizinische Entscheidungen automatisieren (Art. 6 EU AI Act). Diese Voraussetzungen sind hier nicht erfüllt, da weder eine medizinische Zweckbestimmung vorliegt noch eine Einbindung in regulierte medizinische Entscheidungsprozesse erfolgt.

Die regulatorische Einordnung als nicht-medizinisches, beratendes KI-System ist somit konsistent mit den gewählten Designentscheidungen, Artefakten und Nutzungskonzepten. Sie ermöglicht den Einsatz des Systems unter Beachtung datenschutz-, transparenz- und ethikbezogener Anforderungen, ohne die strengeren Zulassungs- und Konformitätsregime des Medizinprodukterechts oder die besonderen Pflichten für Hochrisiko-KI-Systeme nach dem EU Artificial Intelligence Act auszulösen.

3 Methodik

3.1 Design Science Research (DSR)

Diese Arbeit folgt dem Ansatz der Design Science Research (DSR). DSR ist ein problemorientierter Forschungsansatz, der darauf abzielt, durch die systematische Gestaltung, Analyse und Evaluation von Artefakten wissenschaftliche Erkenntnisse zu generieren. Im Zentrum stehen nicht primär erklärende oder hypothesenprüfende Fragestellungen, sondern die Entwicklung und reflektierte Bewertung von Lösungen für klar definierte reale Problemkontexte (Hevner et al., 2004, S. 82–84).

In der Informationssystemforschung werden Artefakte im Rahmen von DSR als gestaltete Wissensobjekte verstanden, die auf unterschiedlichen Abstraktionsebenen angesiedelt sein können. Hevner et al. (2004, S. 82–83) unterscheiden hierbei Konstrukte, Modelle, Methoden und Instanziierungen. Diese Differenzierung ist für die vorliegende Arbeit zentral, da der entwickelte KI-Workflow nicht als isoliertes technisches System, sondern als strukturierende Gestaltungslogik verstanden wird, die normative, organisatorische und technische Aspekte integriert.

DSR ist dabei explizit als iterativer Build-and-Evaluate-Prozess konzipiert. Erkenntnis entsteht durch die fortlaufende Reflexion von Gestaltungsentscheidungen und deren Bewertung im Hinblick auf definierte Zielkriterien (Hevner et al., 2004, S. 86–88).

Hevner und Chatterjee (2010, S. 16–19) strukturieren DSR konzeptionell über drei miteinander verbundene Zyklen: den Relevance Cycle, den Design Cycle und den Rigor Cycle. Der Relevance Cycle stellt den Bezug zu einem konkreten Anwendungskontext her und ermöglicht die systematische Ableitung von Anforderungen. Der Rigor Cycle bindet bestehende wissenschaftliche

Erkenntnisse, Modelle und Theorien in den Gestaltungsprozess ein. Der Design Cycle verbindet beide Zyklen durch iterative Entwurfs-, Entwicklungs- und Evaluationsschritte.

Die Design Science Research Methodology (DSRM) nach Peffers et al. (2007) bietet eine prozessorientierte Konkretisierung von DSR. Sie beschreibt sechs zentrale Aktivitäten: Problemidentifikation, Definition der Lösungsziele, Entwurf und Entwicklung des Artefakts, Demonstration, Evaluation sowie Kommunikation (Peffers et al., 2007, S. 52–56). Diese Aktivitäten sind nicht als starre Abfolge zu verstehen, sondern erlauben unterschiedliche Einstiegspunkte und iterative Rückkopplungen.

Im Rahmen dieser Bachelorarbeit wird DSR nicht in voller methodischer Breite ausgeschöpft, sondern als strukturierender Orientierungsrahmen eingesetzt. Dies entspricht der von vom Brocke et al. (vom Brocke et al., 2020, S. 9–11) beschriebenen Praxis, DSR-Projekte an Umfang, Zielsetzung und Erkenntnisinteresse anzupassen. Der methodische Fokus liegt auf der nachvollziehbaren Gestaltung eines Artefakts und darauf basierender Evaluation.

Ein zentrales Merkmal von DSR ist die Abgrenzung gegenüber empirischer Wirksamkeitsforschung. Die Qualität eines Artefakts wird nicht primär anhand beobachteter Effekte bei Nutzenden beurteilt, sondern anhand seiner Eignung, definierte Anforderungen konsistent, transparent und nachvollziehbar umzusetzen (Hevner et al., 2004, S. 86–88). Diese Perspektive ist für die vorliegende Arbeit zentral, da keine empirische Untersuchung mit Patientinnen oder Fachpersonen durchgeführt wird.

3.2 Forschungsdesign & Artefakt

Das zentrale Artefakt dieser Arbeit ist ein klinisch und ethisch begründeter KI-Workflow für den Einsatz generativer KI im professionellen Kontext von Psychotherapie und psychologischer Beratung.

Im Sinne der von Hevner et al. (2004, S. 82–83) beschriebenen Artefakttypen ist der entwickelte Workflow auf mehreren Ebenen angesiedelt. Er fungiert als Modell, indem er zentrale Prozessschritte und Verantwortlichkeiten strukturiert abbildet; als Methode, indem er konkrete Handlungs- und Entscheidungslogiken für den Einsatz von KI definiert; sowie als partielle Instanziierung, indem ausgewählte Elemente prototypisch in der Plattform mentalhealthGPT umgesetzt werden.

Das Forschungsdesign orientiert sich an den zentralen Aktivitäten der DSRM nach Peffers et al. (2007, S. 52–56), wird jedoch bewusst auf den Umfang und Anspruch einer Bachelorarbeit angepasst. Der methodische Ablauf umfasst vier miteinander verbundene Schritte:

I. die systematische Ableitung eines interdisziplinären Anforderungskatalogs,
II. die Konzeption eines strukturierten KI-Workflows,
III. die partielle prototypische Implementierung ausgewählter Workflow-Komponenten,
IV. die artefaktbasierte Evaluation der entwickelten Gestaltungslösung.

Die Ableitung der Anforderungen erfolgt entlang des Relevance Cycle, indem ethische, psychologische und regulatorische Anforderungen aus der Fachliteratur und den relevanten Normen systematisch zusammengeführt werden (Hevner & Chatterjee, 2010, S. 17–19). Der Entwurf des Workflows

integriert dieses Wissen mit bestehenden theoretischen Konzepten und ist dem Rigor Cycle zuzuordnen. Die iterative Ausarbeitung und Analyse des Artefakts erfolgt im Design Cycle.

Die prototypische Implementierung dient der Demonstration und analytischen Durchdringung zentraler Designentscheidungen. Sie ist nicht als empirische Evaluation im Sinne einer Feld- oder Wirksamkeitsstudie konzipiert, sondern als Mittel zur Prüfung der internen Kohärenz, Umsetzbarkeit und Konsistenz des Artefakts. Hevner et al. (2010, S. 91–93) betonen, dass auch unvollständig instanziierte Artefakte valide Gegenstände der DSR sein können, sofern ihre Gestaltungslogik explizit dokumentiert wird.

Die Evaluation des Artefakts erfolgt qualitativ und prototypbasiert. Bewertet wird, inwiefern der entwickelte Workflow geeignet ist, ethische Prinzipien, regulatorische Anforderungen und sicherheitsbezogene Vorgaben konsistent umzusetzen. Diese Form der Evaluation entspricht den von Hevner et al. (2010, S. 85–87) beschriebenen Bewertungsansätzen für normative und konzeptionelle Artefakte.

Durch die Kombination aus konzeptioneller Gestaltung, partieller Implementierung und systematischer Evaluation wird der Entwicklungsprozess transparent nachvollziehbar gemacht. Das Forschungsdesign stellt sicher, dass das Artefakt sowohl theoretisch fundiert als auch praktisch anschlussfähig ist.

3.3 Evaluations- und Bewertungsmethodik

Die Evaluation des entwickelten KI-Workflows erfolgte artefaktbasiert, qualitativ und kontextsensitiv. Ziel der Evaluation war nicht die Überprüfung klinischer Wirksamkeit oder therapeutischer Outcomes, sondern die systematische Bewertung der datenschutzbezogenen, ethischen, regulatorischen und psychologisch-professionellen Eignung der Artefakte im Kontext psychologischer Beratung und Therapie. Der Evaluationsfokus lag dabei auf jenen Bedingungen, die erforderlich sind, um Vertrauen, professionelle Verantwortung und den Schutz der Betroffenen im Umgang mit digitalisierten Therapie- und Beratungssitzungen zu unterstützen.

Der Bewertungsrahmen berücksichtigte explizit, dass im Anwendungskontext dieser Arbeit besonders sensible Daten verarbeitet wurden, darunter Audio- oder Videoaufzeichnungen von Sitzungen, Transkripte sowie begleitende Fallnotizen, die potenziell auch in einem interdisziplinären Behandlungs- oder Supervisionskontext genutzt werden. Die Evaluation bezog sich daher nicht auf einzelne KI-Modelle, sondern auf den Workflow, die Systemarchitektur und die Governance-Strukturen, innerhalb derer KI eingesetzt wurde.

3.3.1 Evaluationsverständnis und Bewertungsgegenstand

Evaluation wurde in dieser Arbeit als die strukturierte Analyse der Übereinstimmung zwischen dem definierten Anforderungskatalog und der konkreten Ausgestaltung des Artefakts verstanden. Bewertet wurde, inwiefern der KI-Workflow geeignet war, professionelle Handlungssicherheit zu unterstützen, Risiken für Betroffene zu minimieren und eine verantwortungsvolle Mensch–KI-Interaktion zu

ermöglichen. Dieses artefaktbasierte Evaluationsverständnis entsprach der Design Science Research, in der analytische und konzeptionelle Bewertungsverfahren explizit vorgesehen sind (Hevner et al., 2004, S. 85–87).

Der Bewertungsgegenstand umfasste insbesondere den Umgang mit besonders schützenswerten Personendaten, die Gestaltung von Zugriff, Kontrolle und Transparenz sowie die klare Abgrenzung zwischen KI-Unterstützung und menschlicher Verantwortung in therapeutischen und beratenden Kontexten.

3.3.2 Datenschutz- und Privatsphärenanalyse auf Basis von LINDDUN

Den methodischen Schwerpunkt der Risikoanalyse bildete das LINDDUN[2]-Modell, ein etabliertes Rahmenwerk zur Identifikation von Datenschutz- und Privatsphärenrisiken (Deng et al., 2011, S. 6–7). Die Analyse adressierte insbesondere Risikokategorien wie Identifizierbarkeit, Verknüpfbarkeit, mangelnde Transparenz, fehlende Informiertheit der Betroffenen sowie regulatorische Nichtkonformität. Diese Dimensionen sind für psychotherapeutische und beratende Kontexte von zentraler Bedeutung, da sie unmittelbar die Bereitschaft zur digitalen Aufzeichnung sensibler Interaktionen beeinflussen. Zur artefaktnahen Operationalisierung der LINDDUN-Kategorien wurde ergänzend LINDDUN GO herangezogen, ein designorientierter Ansatz, der die Anwendung des LINDDUN-Modells durch strukturierte Karten und Leitfragen unterstützt (Wuyts et al., 2020 - 2020, S. 305–306).

Die Analyse erfolgte qualitativ und artefaktbezogen und wurde auf zentrale Elemente des KI-Workflows angewandt. Untersucht wurde dabei insbesondere, inwiefern der Workflow die Wahrnehmung von Kontrolle, Vertraulichkeit und Nachvollziehbarkeit unterstützte – Faktoren, die aus psychologischer Perspektive als Voraussetzung für Vertrauen und Offenheit gelten.

3.3.3 Orientierung an der DPIA-Logik der DSGVO

Ergänzend orientierte sich die Bewertung an der Logik einer Datenschutz-Folgenabschätzung (Data Protection Impact Assessment, DPIA) gemäß Art. 35 DSGVO (Datenschutz-Grundverordnung). Auch wenn im Rahmen dieser Arbeit keine formale DPIA durchgeführt wurde, wurden zentrale Prüfdimensionen systematisch berücksichtigt, darunter die Sensibilität der verarbeiteten Daten, potenzielle Risiken für Betroffene sowie die vorgesehenen technischen und organisatorischen Schutzmaßnahmen (Europäische Union, 2016).

Die Bewertung stützte sich dabei auf die Leitlinien der Article 29 Working Party (WP29), die eine DPIA insbesondere dann empfehlen, wenn neue Technologien eingesetzt werden und besonders schützenswerte Gesundheitsdaten verarbeitet werden (European Commission, 2017). Auf Artefaktebene wurden exemplarisch umgesetzte Maßnahmen einbezogen, darunter Passkey-basierte Authentifizierung. Diese Maßnahmen dienten der demonstrativen Veranschaulichung datenschutzfreundlicher Gestaltungsprinzipien und nicht der vollständigen technischen Absicherung.

[2] Linking, Identifying, Non-repudiation, Detecting, Data Disclosure, Unawareness, Non-compliance

3.3.4 Normative Mapping-Analyse regulatorischer und professionsethischer Anforderungen

Zur Bewertung der normativen Angemessenheit des Artefakts wurde eine normative Mapping-Analyse durchgeführt. Dabei wurden zentrale Anforderungen aus DSGVO, EU AI Act sowie aus berufsrechtlichen und ethischen Leitlinien der Psychologie systematisch den Gestaltungsentscheidungen des KI-Workflows gegenübergestellt. Als professionsethischer Referenzrahmen wurde insbesondere der EFPA Meta-Code of Ethics herangezogen, der europaweit grundlegende Prinzipien wie Vertraulichkeit, Verantwortung, Respekt vor Autonomie und professionelle Integrität definiert (EFPA, 2025).

Die Analyse verdeutlichte, wie abstrakte normative Vorgaben in konkrete Prozess- und Governance-Strukturen übersetzt wurden. Ethische Fragestellungen wurden dabei bewusst primär auf der Ebene des Workflows, der Rollenverteilung und der Nutzungskontexte adressiert und nicht auf der Ebene der internen KI-Modelllogik.

3.4 Methodenkritik

Die in dieser Arbeit gewählte Methodik ermöglichte eine strukturierte und nachvollziehbare Bewertung eines KI-Workflows für den Einsatz in psychologischer Beratung und Therapie. Zugleich unterliegt der methodische Ansatz klaren Grenzen.

Eine zentrale Limitation ergibt sich aus der artefaktbasierten und qualitativen Ausrichtung der Evaluation. Die Bewertung des KI-Workflows erfolgte nicht anhand empirischer Daten von Patient:innen oder Fachpersonen. Entsprechend können aus den Ergebnissen keine Aussagen über klinische Wirksamkeit, therapeutische Effekte oder tatsächliche Nutzungsakzeptanz abgeleitet werden. Diese Einschränkung ist jedoch methodisch intendiert, da der Fokus der Arbeit explizit auf Governance, Datenschutz und professioneller Eignung liegt.

Eine weitere Limitation betrifft die partielle prototypische Umsetzung des Artefakts. Zwar erlaubte die demonstrative Implementierung ausgewählter Komponenten (z. B. Authentifizierungs- und Verschlüsselungsmechanismen) eine artefaktnahe Analyse, sie ersetzt jedoch keine vollständige technische Realisierung. Die Evaluation konnte daher primär die Gestaltungslogik und konzeptionelle Kohärenz des Workflows beurteilen, nicht jedoch dessen Robustheit unter realen Einsatzbedingungen oder hoher Systemkomplexität.

Auch die Anwendung der Datenschutz- und Privatsphärenanalyse mittels LINDDUN (GO) ist mit methodischen Grenzen verbunden. Die Analyse erfolgte qualitativ und explorativ und basiert auf modellhaften Annahmen über Nutzungsszenarien und potenzielle Risiken. Obwohl dieser Ansatz für frühe Design- und Governance-Phasen geeignet ist, ersetzt er keine formale Datenschutz-Folgenabschätzung im rechtlichen Sinne und kann reale organisatorische oder kontextuelle Faktoren nur begrenzt abbilden. Zudem orientiert sich die Analyse primär am europäischen Rechts- und Ethikrahmen, wodurch die Übertragbarkeit auf andere regulatorische Kontexte eingeschränkt ist.

Schließlich ist hervorzuheben, dass die Arbeit bewusst auf eine direkte Einbindung von Nutzer:innen verzichtet. Die Perspektiven von Patient:innen, Therapeut:innen oder interdisziplinären Behandlungsteams werden indirekt über normative, ethische und psychologische Anforderungen berücksichtigt, jedoch nicht empirisch erhoben. Dies stellt eine klare Begrenzung dar, eröffnet zugleich aber Ansatzpunkte für weiterführende Forschung.

Zusammenfassend ermöglicht die gewählte Methodik eine fundierte, transparente und verantwortungsbewusste Bewertung des Artefakts im Sinne von Privacy-, Governance- und Professional-by-Design. Ihre Limitationen liegen vor allem in der fehlenden empirischen Validierung und der begrenzten technischen Demonstration, die jedoch dem Umfang und Ziel einer Bachelorarbeit entsprechen und den Anspruch der Arbeit klar begrenzen.

4 Entwicklung und partielle Implementierung der KI-Workflow Artefakte

Hans Jonas bestimmt Verantwortung als Verpflichtung, für die absehbaren Wirkungen des eigenen Tuns und Unterlassens einzustehen, und hebt hervor, dass sie sich insbesondere dort konkretisiert, wo technisches Handeln neue, schwer überschaubare Wirkungszusammenhänge erzeugt (Jonas, 1979, S. 168–179). Verantwortung richtet sich damit nicht allein auf das eigene Tun, sondern auf die Gestaltung der Bedingungen, unter denen Wirkungen entstehen. Wer Verantwortung trägt, ist folglich auch für die Beherrschbarkeit der durch eingesetzte Mittel erzeugten Wirkungszusammenhänge verantwortlich.

Daraus folgt, dass Verantwortung nur dort praktisch eingelöst werden kann, wo diese Wirkungszusammenhänge gezielt begrenzt, überprüft und abgesichert werden können. Unter komplexen technischen Bedingungen lässt sich dies jedoch nicht durch bloße Absicht, individuelle Aufmerksamkeit oder professionsethische Selbstbindung sicherstellen. Verantwortung muss vielmehr in konkretisierte, überprüfbare Anforderungen übersetzt werden, deren Einhaltung verlässlich gewährleistet werden kann. Solche Anforderungen nehmen die Form von Pflichten an – etwa in Bezug auf Datenschutz, Zugriffsbeschränkung, Zweckbindung, Transparenz der Systemnutzung oder die Sicherstellung menschlicher Entscheidungsautorität. Verantwortung bleibt dabei normativ bestehen, wird jedoch an der verlässlichen Erfüllung operationalisierter Pflichten bemessen.

Der Einsatz generativer KI verschärft diese Problematik, da Nutzer:innen weder interne Verarbeitungsprozesse noch sicherheitsrelevante Systemeigenschaften fortlaufend kontrollieren können. Verantwortung kann unter diesen Bedingungen nur aufrechterhalten werden, wenn die Einhaltung der relevanten Pflichten strukturell abgesichert ist. Entsprechend verlagert sich verantwortliches Handeln von der einzelnen Nutzungshandlung auf die Ebene der System- und Prozessgestaltung. Verantwortungsfähige KI-Governance bedeutet, dass Verantwortung nicht einfach durch Zurückhaltung im KI-Einsatz realisiert wird, sondern durch nachweisbare Gestaltungsentscheidungen, die Fehlzuschreibungen, Kontrollverluste und Missbrauch präventiv begrenzen (Raji et al., S. 965–968).

Ein klinischer KI-Workflow übernimmt in diesem Zusammenhang eine zentrale vermittelnde Funktion. Er übersetzt normative Verantwortungsanforderungen in eine strukturierte Abfolge von Prozessschritten, Systemzuständen und technischen Artefakten und macht damit sichtbar, wo Verantwortung im Nutzungspfad entsteht und wie ihre praktische Wahrnehmung systemisch abgesichert werden kann. Verantwortung wird so nicht erst im Einzelfall zugeschrieben, sondern bereits auf der Ebene des Designs antizipiert und operationalisiert.

Die systematische Übersetzung dieses normativen Anspruchs in konkrete Gestaltungsentscheidungen folgt dem Design-Science-Research-Methodenmodell nach Pfeffers (2007). Der in Kapitel 2 analysierte Problemraum bildet dabei die Grundlage der Problemidentifikation und Anforderungsableitung. Die nachfolgenden Abschnitte dieses Kapitels konkretisieren diesen Designprozess: Der interdisziplinäre Anforderungskatalog (4.1), der klinische KI-Workflow (4.2) und die Plattformarchitektur (4.3) stellen aufeinander aufbauende Artefakte der Konstruktion und Spezifikation dar, während die prototypische Implementierung ausgewählter Komponenten (4.4) und deren qualitative Evaluation (Kapitel 5) den Phasen der Demonstration und Bewertung im Sinne des DSRM entsprechen.

4.1 Ableitung des interdisziplinären Anforderungskatalogs

Demgemäss werden nun aus dem identifizierten Problemraum systematisch prüfbare Designanforderungen (Design Requirements) für die Artefakte abgeleitet (Hevner et al., 2004; vom Brocke et al., 2020).

4.1.1 Normative und workflow-basierte Ableitungslogik

Der interdisziplinäre Anforderungskatalog basiert auf der Integration der drei beschriebenen normativen Bezugsebenen, die für den Einsatz generativer KI im Mental-Health-Kontext gemeinsam handlungsleitend sind: ethische Prinzipien, regulatorische Vorgaben und professionsbezogene Anforderungen. Diese Ebenen begründen, warum bestimmte Schutz- und Sorgfaltspflichten bestehen, ohne deren technische Umsetzung vorzugeben. Maßgeblich sind dabei insbesondere der Schutz von Autonomie, Nicht-Schaden, Transparenz und professioneller Integrität (ethische Ebene), datenschutz- und KI-rechtliche Mindestanforderungen an Gestaltung und Einsatz von Systemen (regulatorische Ebene) sowie die besondere Verantwortung von Fachpersonen in asymmetrischen Vertrauensverhältnissen (professionsbezogene Ebene).

Diese normativen Ebenen sind analytisch unterscheidbar, greifen jedoch in der praktischen Systemgestaltung ineinander. Der Anforderungskatalog zielt daher nicht auf eine additive Aufzählung einzelner Vorgaben, sondern auf deren funktionale Integration im Kontext eines konkreten Nutzungsszenarios. Dabei unterscheiden wir zwischen der nicht delegierbaren Verantwortung der menschlichen Akteur:innen und jenen Pflichten, deren Einhaltung an technische Systeme delegiert werden kann. Delegierbar ist nicht Verantwortung selbst, sondern die operative Sicherstellung der aus ihr abgeleiteten Schutz- und Sorgfaltspflichten.

Die Strukturierung der Anforderungen folgt dieser Logik delegierbarer Pflichten. Sie lassen sich vier zentralen Kategorien zuordnen: A) dem Schutz der Vertraulichkeit und informationellen Selbstbestimmung, B) der Sicherstellung von Transparenz und Rollenklarheit, C) der Begrenzung funktionaler Zuständigkeit sowie D) der Nachweisbarkeit und Kontrollierbarkeit der Systemgestaltung. Diese Kategorien bilden den normativen Kern des Anforderungskatalogs und fungieren als Brücke zwischen abstrakter Verantwortung und konkreter Systemarchitektur.

4.1.2 Workflow-basierte Ableitung delegierbarer Pflichten

Maßgeblich ist, an welchen Stellen eines typischen Mental-Health-KI-Workflows ethische und regulatorische Schutzgüter tatsächlich berührt werden. Die Analyse dieser Nutzungshandlungen macht sichtbar, welche Pflichten in welchen Interaktionsphasen ausgelöst werden und daher strukturell abgesichert sein müssen. Der daraus abgeleitete Anforderungskatalog ist bewusst technikneutral formuliert, um den normativen Anspruch klar zu bestimmen. Zugleich wird im weiteren Verlauf transparent gemacht, welche Anforderungen durch die im Projekt verfügbaren technischen und organisatorischen Artefakte vollständig, teilweise oder exemplarisch umgesetzt werden können. Die partielle Implementierung ausgewählter Komponenten ist dabei als methodisch reflektierte Fokussierung im Sinne des Design-Science-Ansatzes zu verstehen (Hevner et al., 2004; vom Brocke et al., 2020).

Im Kern lässt sich der relevante KI-System-Workflow in folgende idealtypische Phasen gliedern:

I. Zugang und Authentifizierung

Beim erstmaligen Zugang zum System sowie bei der Authentifizierung der Nutzer:innen wird ein Identitätsbezug hergestellt. Bereits an dieser Stelle werden datenschutzrechtliche Anforderungen an Zweckbindung, Datenminimierung und Zugriffskontrolle ausgelöst (DSGVO Art. 5, Art. 25). Zugleich entsteht ein implizites Vertrauensverhältnis, das Transparenz über die Systemrolle und den nichtmenschlichen Charakter der Interaktion erfordert (EU AI Act Art. 52).

II. Eingabe personenbezogener und hochsensibler Inhalte

Die aktive Eingabe psychologisch relevanter Inhalte stellt den sensibelsten Schritt im Workflow dar. Hier werden Anforderungen an Vertraulichkeit, Schutz der informationellen Selbstbestimmung und Nicht-Schaden berührt. Aufgrund der besonderen Schutzbedürftigkeit der Daten ist sicherzustellen, dass Inhalte weder unbefugt verarbeitet noch mit Identitätsinformationen vermischt werden (DSGVO Art. 5 Abs. 1 lit. f).

III. KI-gestützte Verarbeitung und Zwischenspeicherung

Während der Verarbeitung durch das KI-System entstehen Risiken der Zweckentfremdung, Re-Identifikation oder impliziten Profilbildung. In dieser Phase werden insbesondere Anforderungen an Zweckbindung, funktionale Begrenzung und Ausschluss autonomer Entscheidungsfunktionen relevant (DSGVO Art. 5; EU AI Act Art. 5).

IV. Ausgabe KI-generierter Inhalte

Die Präsentation von KI-generierten Antworten berührt Fragen der Rollenklarheit, Transparenz und Autoritätszuschreibung. Nutzer:innen müssen etwa erkennen können, dass es sich um KI-generierte Inhalte handelt und dass keine therapeutische oder diagnostische Entscheidung getroffen wird (EU AI Act Art. 52; Art. 5).

V. Weiterverwendung, Speicherung oder Dokumentation

Bei der Weiterverwendung oder Speicherung von Inhalten werden erneut datenschutzrechtliche Anforderungen an Zweckbindung, Zugriffsbeschränkung und Kontrolle ausgelöst. Gleichzeitig stellt sich die Frage der Nachweisbarkeit, ob die Einhaltung dieser Pflichten dauerhaft abgesichert ist (DSGVO Art. 25).

Aus der Analyse dieser Workflow-Schritte wird deutlich, dass Verantwortung nicht punktuell, sondern entlang des gesamten Nutzungspfads entsteht. Der Anforderungskatalog übersetzt diese workflow-getriggerten Schutzbedarfe in konkrete, delegierbare Pflichten.

4.1.3 Konkretisierung des interdisziplinären Anforderungskatalogs

Auf Grundlage der workflow-basierten Analyse lassen sich die identifizierten normativen Schutzbedarfe in einen Katalog konkreter, delegierbarer Pflichten überführen. Jede Anforderung ist zugleich ethisch begründet, regulatorisch rückbindbar und eindeutig einzelnen Phasen des klinischen KI-System-Workflows I–V zuordenbar (Anhang A).

A. Schutz der Vertraulichkeit und informationellen Selbstbestimmung

A1 – Ende-zu-Ende-Schutz sensibler Inhalte

Alle inhaltlichen Daten, die personenbezogene oder potenziell identifizierende Informationen enthalten, müssen vor der Übertragung verschlüsselt werden, sodass sie während Transport und Speicherung nicht im Klartext zugänglich sind. (DSGVO Art. 5 Abs. 1 lit. f; Art. 32; DSG Art. 6)

A2 – Datenminimierung auf Systemebene

Das System darf nur solche Daten verarbeiten, die für den jeweiligen Nutzungskontext erforderlich sind. Eine darüberhinausgehende Erhebung, Speicherung oder Weiterverarbeitung ist technisch zu vermeiden. (DSGVO Art. 5 Abs. 1 lit. c; DSG Art. 6)

A3 – Trennung von Inhalts- und Identitätsdaten

Identitätsbezüge und Inhaltsdaten müssen logisch und technisch getrennt verarbeitet werden, sodass eine direkte Zuordnung nicht Teil des regulären Systembetriebs ist. (DSGVO Art. 25 Abs. 1; DSG Art. 7)

A4 – Begrenzung von Re-Identifikationsmöglichkeiten

Das System muss so gestaltet sein, dass eine Re-Identifikation personenbezogener Inhalte nur unter kontrollierten Bedingungen erfolgen kann und nicht systemseitig erzwungen oder unbeabsichtigt ermöglicht wird. (DSGVO Art. 25 Abs. 1–2; DSG Art. 7)

A5 – Zweckbindung durch Systemdesign

Die Nutzung personenbezogener Daten muss technisch auf den definierten Beratungs- bzw. Unterstützungszweck begrenzt sein. Sekundärnutzungen sind auszuschließen oder strikt getrennt zu behandeln. (DSGVO Art. 5 Abs. 1 lit. b; DSG Art. 6)

B. Transparenz und Rollenklarheit

B1 – Eindeutige Kennzeichnung KI-generierter Inhalte

Alle vom System erzeugten Inhalte müssen klar und dauerhaft als KI-generiert gekennzeichnet sein, um Fehlzuschreibungen menschlicher Autorenschaft zu vermeiden. (EU AI Act Art. 52)

B2 – Offenlegung der Systemrolle

Das System muss seine Funktion als unterstützendes Werkzeug explizit kommunizieren und darf nicht den Eindruck erwecken, professionelle Verantwortung zu übernehmen. (EU AI Act Art. 52)

B3 – Begrenzung impliziter Autoritätszuschreibungen

Gestaltung, Sprache und Interaktionslogik dürfen keine therapeutische, diagnostische oder klinische Autorität suggerieren. (EU AI Act Art. 5 Abs. 1; Erwägungsgründe 28, 48)

B4 – Transparenz über Systemgrenzen

Nutzer:innen müssen nachvollziehen können, welche Leistungen das System erbringt und welche ausdrücklich ausgeschlossen sind. (EU AI Act Art. 52)

C. Begrenzung funktionaler Zuständigkeit

C1 – Ausschluss autonomer Entscheidungsfunktionen

Das System darf keine autonomen diagnostischen, therapeutischen oder behandlungsrelevanten Entscheidungen treffen oder simulieren. (EU AI Act Art. 5 Abs. 1; MDR Art. 2 Nr. 1)

C2 – Unterstützung statt Substitution professioneller Urteilskraft

Systemfunktionen müssen darauf ausgelegt sein, menschliche Entscheidungsprozesse zu unterstützen, nicht zu ersetzen oder zu überlagern. (abgeleitet aus MDR Art. 2 Nr. 1 i. V. m. Zweckbestimmung)

C3 – Keine Bewertung oder Klassifikation von Personen

Das System darf keine Profile, Scores oder Klassifikationen im Sinne psychologischer, klinischer oder prognostischer Einordnungen erzeugen. (DSGVO Art. 22; EU AI Act Art. 5 Abs. 1)

C4 – Klare Abgrenzung zu medizinischer Behandlung

Die Nutzung des Systems muss technisch und kommunikativ eindeutig von medizinischer Behandlung abgegrenzt bleiben. (MDR Art. 2 Nr. 1)

D. Nachweisbarkeit und Kontrollierbarkeit

D1 – Verantwortungssichere Systemarchitektur

Die Systemarchitektur muss so gestaltet sein, dass zentrale Schutzmechanismen nicht umgangen oder deaktiviert werden können. (DSGVO Art. 25)

D2 – Implizite Nachweisbarkeit durch Design
Die Einhaltung zentraler Anforderungen soll sich aus der Systemarchitektur selbst ergeben, etwa durch die technische Unmöglichkeit bestimmter Zugriffe oder Verarbeitungen. (DSGVO Art. 25)

D3 – Transparente Systemgrenzen für Nutzer:innen
Nutzer:innen müssen erkennen können, welche Schutzmechanismen aktiv sind und welche Verantwortung weiterhin bei ihnen verbleibt. (EU AI Act Art. 52)

D4 – Rückholbarkeit von Verantwortung
Das System muss Eingriffe, Kontrolle oder Abbruch durch Nutzer:innen jederzeit ermöglichen, ohne funktionale Abhängigkeiten zu erzeugen. (DSGVO Art. 25)

4.2 Konzeption des Workflow Prozessartefakts

Der Workflow, als zentrales Prozessartefakt, beschreibt hier die strukturierte Abfolge von Nutzer- und Systemhandlungen bei der Nutzung von mentalhealth-gpt.ch aus einer klinisch verantwortlichen Perspektive. Er bildet den tatsächlichen Nutzungspfad ab, entlang dessen sensible Daten verarbeitet, KI-gestützte Antworten erzeugt und Verantwortung praktisch wahrgenommen wird.

Der Workflow ist zustandsbasiert modelliert und gliedert sich in klar unterscheidbare Phasen, in denen jeweils spezifische ethische, regulatorische und professionsbezogene Schutzanforderungen ausgelöst werden. Als konzeptionelles Artefakt ist der Workflow dann einer ex-ante-Evaluation zugänglich, indem geprüft wird, ob die identifizierten Pflichten entlang des Nutzungspfads konsistent, vollständig und nachvollziehbar adressiert werden.

Phase I: Authentifizierter Systemzugang und Initialisierung des Nutzungskontexts

Aktionen:

I.1 Nutzer:in authentifiziert sich über ein personalisiertes Zugangssystem.

I.2 System stellt einen sitzungsbezogenen Nutzungskontext her (Rolle, Berechtigungen, Kontextparameter).

I.3 System kommuniziert explizit seinen nicht-menschlichen Charakter und seine unterstützende Rolle.

Normative Relevanz:

Bereits beim Zugang wird ein personenbezogener Bezug hergestellt und ein implizites Vertrauensverhältnis initiiert. Fehlannahmen über Systemrolle oder Verantwortlichkeit wirken sich auf alle nachfolgenden Interaktionen aus.

Getriggerte Pflichten:

- Zweckbindung und Zugriffsbeschränkung (A2, A5)
- Offenlegung der Systemrolle (B2)

- Verantwortungssichere Systemarchitektur (D1)

Workflow-Ziele:
Herstellung eines klar abgegrenzten, verantwortungsfähigen Nutzungskontexts, in dem Rollen, Zuständigkeiten und Verantwortlichkeiten von Beginn an eindeutig verteilt sind und Fehlannahmen über Systemcharakter oder Verantwortungsübernahme ausgeschlossen werden.

Phase II: Eingabe sensibler Inhalte (Text und Audio)

Aktionen:

II.1 Nutzer:in gibt freie Texte mit potenziell hochsensiblen Inhalten ein.

II.2 Nutzer:in lädt eine Audio-Datei zur Transkription hoch.

II.3 System übernimmt Inhalte - ausschließlich zur zweckgebundenen Verarbeitung.

Normative Relevanz:
Diese Phase stellt den sensibelsten Punkt freiwilliger Offenbarung dar. Hier entscheidet sich, ob informationelle Selbstbestimmung praktisch gewahrt bleibt oder faktisch unterlaufen wird.

Getriggerte Pflichten:

- Ende-zu-Ende-Schutz sensibler Inhalte (A1)
- Datenminimierung (A2)
- Trennung von Inhalts- und Identitätsdaten (A3)
- Zweckbindung (A5)

Workflow-Ziele:
Sicherung der freiwilligen, kontrollierten Offenlegung hochsensibler Inhalte unter Wahrung informationeller Selbstbestimmung und präventiver Begrenzung von Re-Identifikations- und Zweckentfremdungsrisiken bereits vor der KI-Verarbeitung.

Phase III: KI-gestützte Verarbeitung und Zwischenspeicherung

Aktionen:

III.1 System verarbeitet Texte bzw. transkribiert Audio-Inhalte.

III.2 Inhalte werden temporär zwischengespeichert, ausschließlich zur Antwortgenerierung.

III.3 Verarbeitung erfolgt ohne Bewertung, Klassifikation oder Profilbildung.

Normative Relevanz:
Diese Phase ist für Nutzer:innen nicht einsehbar, stellt aber den zentralen Risikobereich dar. Verantwortung muss hier architektonisch, nicht durch Verhalten, abgesichert sein.

Getriggerte Pflichten:

- Ausschluss autonomer Entscheidungsfunktionen (C1)
- Begrenzung funktionaler Zuständigkeit (C2)
- Keine Bewertung oder Klassifikation von Personen (C3)

- Implizite Nachweisbarkeit durch Design (D2)

Workflow-Ziele:

Strikte funktionale Begrenzung der KI auf unterstützende Verarbeitung ohne klinische, diagnostische oder prognostische Zuschreibung, sodass Verantwortung trotz nicht einsehbarer Verarbeitung architektonisch abgesichert bleibt.

Phase IV: Ausgabe KI-generierter Inhalte

Aktionen:

IV.1 System generiert eine Antwort auf Basis der verarbeiteten Inhalte.

IV.2 Antwort wird klar und dauerhaft als KI-generiert gekennzeichnet.

IV.3 System macht explizite Aussagen zu seinen funktionalen Grenzen.

Normative Relevanz:

In dieser Phase entstehen Risiken der Überinterpretation, Autoritätszuschreibung und Verantwortungsverschiebung auf das System.

Getriggerte Pflichten:

- Kennzeichnung KI-generierter Inhalte (B1)
- Offenlegung der Systemrolle (B2)
- Begrenzung impliziter Autoritätszuschreibungen (B3)
- Transparenz über Systemgrenzen (B4)

Workflow-Ziele:

Stabilisierung der epistemischen und professionellen Rollenverteilung durch transparente Kennzeichnung, klare Grenzziehung der Systemfunktion und aktive Vermeidung von Autoritäts- oder Verantwortungsverschiebung auf das KI-System.

Phase V: Nutzungskontrolle, Speicherung oder Abbruch

Aktionen:

V.1 Nutzer:in entscheidet über Speicherung, Weiterverwendung oder Löschung der Inhalte.

V.2 System setzt Zweckbindung und Zugriffsbeschränkung fort.

V.3 Nutzer:in kann die Interaktion jederzeit beenden oder Verantwortung zurückholen.

Normative Relevanz:

Verantwortung endet nicht mit der Antwort des Systems. Langfristige Kontrolle über Datenverwendung und Zugriff ist notwendig, um funktionale Abhängigkeiten und irreversible Verantwortungsabgabe zu verhindern.

Getriggerte Pflichten:

- Zweckbindung (A5)
- Transparente Systemgrenzen (D3)

- Rückholbarkeit von Verantwortung (D4)

Workflow-Ziele:

Dauerhafte Sicherung von Kontrolle, Reversibilität und Verantwortlichkeit über den gesamten Lebenszyklus der Interaktion hinweg, um funktionale Abhängigkeiten zu vermeiden und Verantwortung jederzeit bewusst zurückholen zu können.

4.3 Architektonisches Design-Artefakt: mentalhealthGPT

Die nachfolgend dargestellte Plattformarchitektur von mentalhealth-gpt.ch beschreibt die entsprechende prototypische Zielarchitektur. Sie stellt keinen abgeschlossenen Systemzustand dar, sondern einen konzeptionellen und technischen Rahmen, in dem die in den vorangegangenen Abschnitten abgeleiteten ethischen, regulatorischen und professionsbezogenen Anforderungen systematisch in architektonische Gestaltungsentscheidungen übersetzt werden.

Die Architektur fungiert dabei als vermittelnde Ebene zwischen dem workflow-basierten Anforderungskatalog (Abschnitte 4.1–4.2) und der exemplarischen Umsetzung einzelner technischer Artefakte (Abschnitt 4.4).

4.3.1 Architektonisches Grundprinzip

Zentrales Leitprinzip der Architektur ist eine strikte funktionale Trennung von Identität, Inhalt und Verarbeitung. Diese Trennung dient nicht primär der Effizienz, sondern der technischen Durchsetzung von Datenschutz, Zweckbindung und Verantwortlichkeitsgrenzen. Die Architektur folgt damit einem «privacy-by-design»-Ansatz, bei dem Schutzmechanismen durch bewusst begrenzte Systemfähigkeiten realisiert werden.

Die technische Plattform ist entlang dreier logisch getrennter Ebenen[3] organisiert:

I. Nutzer-Ebene (Browser plus persönliches Vertrauensgerät),
II. Server-Ebene (Anwendungs- und Orchestrierungsschicht im Rechenzentrum),
III. Externe KI-Dienste (ausgelagerte Computerleistung über Programmschnittstellen).

4.3.2 Nutzer-Ebene: Browser und Smartphone als gekoppelte Sicherheitsdomäne

Sie bildet somit die primäre Vertrauens- und Kontrollsphäre der Plattform. Sie besteht aus zwei eng gekoppelten Komponenten – dem Browser als Interaktions- und Verarbeitungsebene sowie dem Smartphone als persönlichem Vertrauensgerät (Trust Anchor). Gemeinsam bilden sie eine sicherheitskritische Domäne, in der Identität, Zugriff und Pseudonomysierung verankert sind.

Browser: Interaktion und lokale Schutzmechanismen

Der Browser stellt die unmittelbare Arbeitsumgebung der Nutzer:innen dar. Personenbezogene und hochsensible Inhalte entstehen primär hier, etwa durch Texteingaben oder das Hochladen von

[3] Siehe Abbildung 3 in Anhang B

Dokumenten oder von Audiodateien zur Transkription. Der Browser ist daher nicht lediglich Darstellungsschicht, sondern aktiver Träger zentraler Schutz- und Kontrollmechanismen.

Auf dieser Ebene erfolgen insbesondere:

- die clientseitige Vorverarbeitung und Pseudonymisierung personenbezogener Inhalte (z. B. durch lokale Entitätserkennung),
- die logische Trennung von Inhalts- und Identitätsbezügen,
- sowie die ausschließlich transiente Verarbeitung von Klartextdaten.

Klartextinhalte verlassen die Nutzer-Ebene nicht. Eine serverseitige Rekonstruktion realer Identitäten ist konzeptionell ausgeschlossen, da Identitätsinformationen nicht Bestandteil der übertragenen Inhalte sind.

Smartphone: Persönliches Vertrauensgerät und Schlüsselanker

Eine zentrale Rolle innerhalb der Nutzer-Ebene übernimmt das Smartphone der Nutzer:innen. Es fungiert als persönlicher Trust Anchor, über den sicherheitskritische Operationen explizit autorisiert werden. Das Smartphone ist dabei nicht optional, sondern integraler Bestandteil der Sicherheitsarchitektur.

Über das Smartphone erfolgen insbesondere:

- die gerätegebundene Authentifizierung (z.B. passwortlose Anmeldung mittels Passkeys / Face ID),
- die Ableitung und Freigabe kryptografischer Schlüssel,
- sowie die explizite Autorisierung zusätzlicher Endgeräte (z. B. über QR-basierte Gerätebindung).

Kryptografisches Schlüsselmaterial verbleibt im geschützten Gerätespeicher des Smartphones (z.B. Secure Enclave / systemeigener Schlüsselbund) und ist an lokale biometrische oder systemseitige Zugriffskontrollen gebunden.

Die Kopplung von Browser und Smartphone stellt sicher, dass Identität, Zugriff und Re-Personalisierung nicht allein über serverseitige Mechanismen im Rechenzentrum kontrolliert werden können.

4.3.3 Server-Ebene: Orchestrierung ohne inhaltliche Kontrolle

Die Server-Ebene übernimmt koordinierende und vermittelnde Funktionen, ohne selbst Träger sensibler Bedeutungs- oder Identitätszusammenhänge zu sein. Ihre Aufgaben beschränken sich auf die Orchestrierung von Workflows, die Sitzungsverwaltung (Session) sowie die Weiterleitung pseudonymisierter Inhalte an externe KI-Dienste.

Zentral ist dabei die bewusste funktionale Begrenzung serverseitiger Fähigkeiten:

- Identitätsdaten liegen serverseitig nicht im Klartext vor und können nur clientseitig im Browser mit dem Smartphone-Schlüssel mit den Inhaltsdaten zusammengeführt werden.

- Inhalte werden – sofern eine Persistenz vorgesehen ist – ausschließlich pseudonymisiert und verschlüsselt gespeichert.
- Protokollierungs- und Monitoring-Daten enthalten keine Konversationsinhalte, sondern lediglich technische Metadaten (z. B. Aktions-, Status- oder Fehlercodes).

Der Server fungiert damit als technisch eingeschränkter Vermittler, nicht als interpretierende oder entscheidende Instanz. Der Betrieb der Server-Infrastruktur erfolgt ausschließlich in der Schweiz; sämtliche persistenten Daten verbleiben innerhalb des schweizerischen Rechtsraums.

4.3.4 Anbindung externer KI-Dienste

Die KI-gestützte Verarbeitung erfolgt über externe spezialisierte Rechendienste, die über klar definierte Programmschnittstellen angebunden sind. Diese Dienste stellen KI-Rechenleistung und Modellinferenz bereit, sind jedoch nicht Bestandteil der Server-Infrastruktur selbst.

An externe KI-Systeme werden ausschließlich pseudonymisierte Inhalte ohne Identitätsbezug übermittelt. Eine Re-Identifikation ist damit schon konzeptionell ausgeschlossen. Die externen Dienste übernehmen keine Verantwortung für Kontext, Zweck oder Interpretation der Ergebnisse, sondern agieren als ausgelagerte Rechenkomponenten innerhalb eines klar begrenzten Funktionsrahmens.

4.3.5 Datenfluss und Persistenzstrategie

Der Datenfluss innerhalb der Plattform folgt dem Prinzip der minimalen Exposition. Personenbezogene Inhalte entstehen clientseitig (Browser), werden dort automatisch pseudonymisiert und erst anschließend weiterverarbeitet.

Persistenz ist optional und strikt zweckgebunden:

- Chat-Inhalte können gespeichert werden, jedoch ausschließlich pseudonymisiert und verschlüsselt.
- Mapping-Daten zur Re-Personalisierung sind verschlüsselt und ohne clientseitige Schlüssel – bereitgestellt über das Smartphone – nicht nutzbar.
- Audio-Daten werden ausschließlich temporär zur Transkription verarbeitet; weiterverwendet wird nur das pseudonymisierte Transkript.

Die Architektur ist damit als antizipierende Operationalisierung normativer Anforderungen zu verstehen. Datenschutz, Transparenz, Rollenklarheit und Verantwortungsbegrenzung sind nicht nachträglich ergänzt, sondern als strukturelle Eigenschaften des Systems angelegt.

4.4 Artefakte als prototypische Implementierung von Workflow-Komponenten

Im Design-Science-Research wird zwischen konzeptionellen Artefakten und deren konkreter Instanziierung unterschieden. Während Anforderungskatalog und klinischer KI-Workflow (Abschnitte 4.1–4.2) die konzeptionelle Ebene adressieren, zielt die prototypische Implementierung darauf ab, die praktische Umsetzbarkeit zentraler Gestaltungsentscheidungen exemplarisch zu demonstrieren (Hevner et al., 2004).

Vor diesem Hintergrund verfolgt Abschnitt 4.4 nicht das Ziel, ein vollständiges KI-System technisch abzubilden oder sämtliche Anforderungen des Workflows zu implementieren. Stattdessen werden zwei ausgewählte, normativ besonders zentrale Komponenten instanziiert, um zu zeigen, dass die im Workflow formulierten Anforderungen prinzipiell realisierbar sind und sich konsistent in eine Systemarchitektur übersetzen lassen.

Auf dieser Grundlage werden folgende zwei Artefakte exemplarisch betrachtet:

1. ein gerätegebundener Authentifizierungsmechanismus (Passkey) als Instanziierung der Anforderung implizite Nachweisbarkeit und Rückholbarkeit von Verantwortung (D2),
2. sowie ein Transparenz-Artefakt zur Kennzeichnung und Einordnung KI-generierter Inhalte als Instanziierung zentraler Anforderungen aus Kategorie B (insbesondere B4).

Beide Artefakte sind bewusst so gewählt, dass sie unterschiedliche Ebenen adressieren: Während der Passkey-Mechanismus vor allem Zugriff, Identität und Verantwortung vor der Nutzung absichert, adressiert das Transparenz-Artefakt psychologisch relevante Risiken während der Interaktion. Gemeinsam veranschaulichen sie, wie abstrakte normative Anforderungen sowohl technisch als auch gestalterisch operationalisiert werden können.

4.4.1 Artefakt I: Gerätegebundene Authentifizierung als Zugriffskontrollmechanismus (Passkey)

Das Artefakt der gerätegebundenen Authentifizierung dient der prototypischen Umsetzung der Anforderungen D2 (implizite Nachweisbarkeit durch Design).

Es dient es der technischen Absicherung des Datenzugriffs, indem es unautorisierte oder unbeabsichtigte Zugriffe auf sensible Inhalte systemisch ausschließt. Datenschutz wird hier nicht als nachgelagerte Zugriffspolitik verstanden, sondern als architektonische Eigenschaft, die bestimmte Zugriffe technisch unmöglich macht.

Einordnung im klinischen KI-Workflow

Das Artefakt ist direkt den folgenden Phasen des Workflows zugeordnet:

- Phase I – Authentifizierter Systemzugang

Herstellung eines eindeutig abgegrenzten, datenschutzkonformen Nutzungskontexts.

- Phase V – Nutzungskontrolle und Abbruch

Sicherstellung, dass Zugriff und Verantwortungsübernahme jederzeit bewusst beendet oder erneuert werden können.

Der Passkey fungiert dabei als zentrale Zugriffsschranke, an der sowohl Verantwortung als auch Datenschutz technisch durchgesetzt werden.

Ablauf aus Nutzersicht: Anmeldung als bewusst kontrollierter Zugriffsvorgang

Der Zugang zu mentalhealth-gpt.ch folgt bewusst keinem vereinfachten Login-Modell. Der Anmeldeprozess ist als mehrstufiger Vorgang gestaltet, der sicherstellt, dass ein Zugriff auf Inhalte nur unter klar definierten und überprüfbaren Bedingungen möglich ist (Abbildung 1).

Zu Beginn gibt die Nutzer:in ihre E-Mail-Adresse und ihr Passwort ein. Dieser Schritt dient ausschließlich der formalen Identitätszuordnung und berechtigt noch nicht zum Zugriff auf Daten oder Funktionen des Systems.

Erst nach dieser Vorprüfung erzeugt der Browser einen einmaligen QR-Code, der auf der Anmeldeseite angezeigt wird. Dieser QR-Code stellt keinen Zugangsschlüssel dar, sondern markiert den Übergang in eine zweite, sicherheitskritische Authentifizierungsstufe.

Abbildung 1: Artefakt der gerätegebundenen Authentifizierung

Quelle: eigene Darstellung

Die Nutzer:in muss diesen Code mit ihrem persönlichen Smartphone scannen und die Anmeldung dort aktiv bestätigen. Erst durch diese Bestätigung wird der Zugriff auf das System freigegeben. Ohne das Smartphone ist dieser Schritt nicht möglich; ohne aktive Bestätigung kommt keine Sitzung zustande. Damit ist sichergestellt, dass ein Zugriff auf sensible Inhalte nicht allein durch Kenntnis von Zugangsdaten, sondern nur durch starke, gerätegebundene Autorisierung erfolgen kann.

Rolle des Smartphones als Zugriffsschranke und Vertrauensanker

Das Smartphone fungiert in dieser Architektur als Vertrauensgerät als technische Zugriffsschranke für sensible Daten.

Sicherheitsrelevante Zugangsinformationen verbleiben ausschließlich auf dem Smartphone und sind durch systemeigene Schutzmechanismen abgesichert. Der Server erhält lediglich die Information, dass eine gültige und aktuelle Autorisierung erfolgt ist, ohne selbst Zugriff auf Authentifizierungsgeheimnisse oder personenbezogene Schlüssel zu besitzen.

Architektonische Bedeutung im Sinne von Datenschutz durch Design

Das Artefakt implementiert Datenschutz nicht über Richtlinien oder nachträgliche Zugriffskontrollen, sondern über die technische Unmöglichkeit bestimmter Zugriffe. Es stellt sicher, dass:

- unautorisierte Zugriffe auf sensible Inhalte nicht stattfinden können,
- automatisierte oder unbeaufsichtigte Sitzungen ausgeschlossen sind,
- und jede Datennutzung an eine aktuelle, aktive Autorisierung gebunden bleibt.

Der Passkey wirkt damit als datenschutzrechtlich relevanter Kontrollmechanismus, der den Zugriff auf personenbezogene und hochsensible Daten strukturell begrenzt.

Demonstrationscharakter des Artefakts

Die hier beschriebene Passkey-Implementierung ist als prototypische Instanziierung zu verstehen. Sie demonstriert exemplarisch, wie sich Anforderungen an Datenschutz, Zugriffskontrolle und Verantwortung gleichzeitig technisch umsetzen lassen.

4.4.2 Artefakt II: Transparenz- und Rollenkennzeichnung KI-generierter Inhalte

Das Transparenz-Artefakt dient der prototypischen Instanziierung zentraler Anforderungen aus Kategorie B (Transparenz und Rollenklarheit), insbesondere B4 (Transparenz über Systemgrenzen). Ziel ist es, psychologisch besonders relevante Risiken zu adressieren, die während der Interaktion mit dem KI-System entstehen.

Es zielt darauf ab, Fehlannahmen über Autorität, Verantwortlichkeit und Leistungsfähigkeit des Systems präventiv zu begrenzen und eine stabile epistemische Rollenverteilung zwischen Nutzer:in und KI sicherzustellen. Gerade in sensiblen Beratungs- und Unterstützungssettings gilt dies als besonders kritisch, da hier ein erhöhtes Risiko von Übervertrauen, Fehlinterpretationen und impliziter Substitution professioneller Urteilsfähigkeit besteht und zentrale ethische Güter wie Autonomie, Vertrauen und der Schutz vulnerabler Personen berührt werden.

Einordnung im klinischen KI-Workflow

Das Artefakt ist primär den folgenden Workflow-Phasen zugeordnet:

- Phase IV – Ausgabe KI-generierter Inhalte

Vermeidung von Autoritätszuschreibung und Fehlinterpretation der Systemantworten.

- Phase V – Nutzungskontrolle und Weiterverwendung

Sicherstellung, dass auch bei späterer Nutzung der Inhalte deren Herkunft und Einordnung erkennbar bleibt. Das Transparenz-Artefakt wirkt damit dort, wo Verantwortung besonders leicht unbemerkt auf das System verschoben werden könnte.

Psychologisch relevanter Risikokontext

Die Interaktion mit sprachlich kohärenten, kontextsensitiven KI-Systemen kann bei Nutzer:innen den Eindruck erzeugen, es handle sich um eine fachlich autoritative, menschlich verantwortliche Instanz. In psychologischen Kontexten ist diese Wirkung besonders problematisch, da sie:

- zu einer Überschätzung der diagnostischen oder therapeutischen Kompetenz des Systems führen kann,
- die Eigenverantwortung der Nutzer:innen unterminiert,
- und professionelle Entscheidungsprozesse implizit substituiert.

Das Transparenz-Artefakt adressiert diese Risiken über kontinuierliche, in den Interaktionsfluss integrierte Gestaltungsentscheidungen.

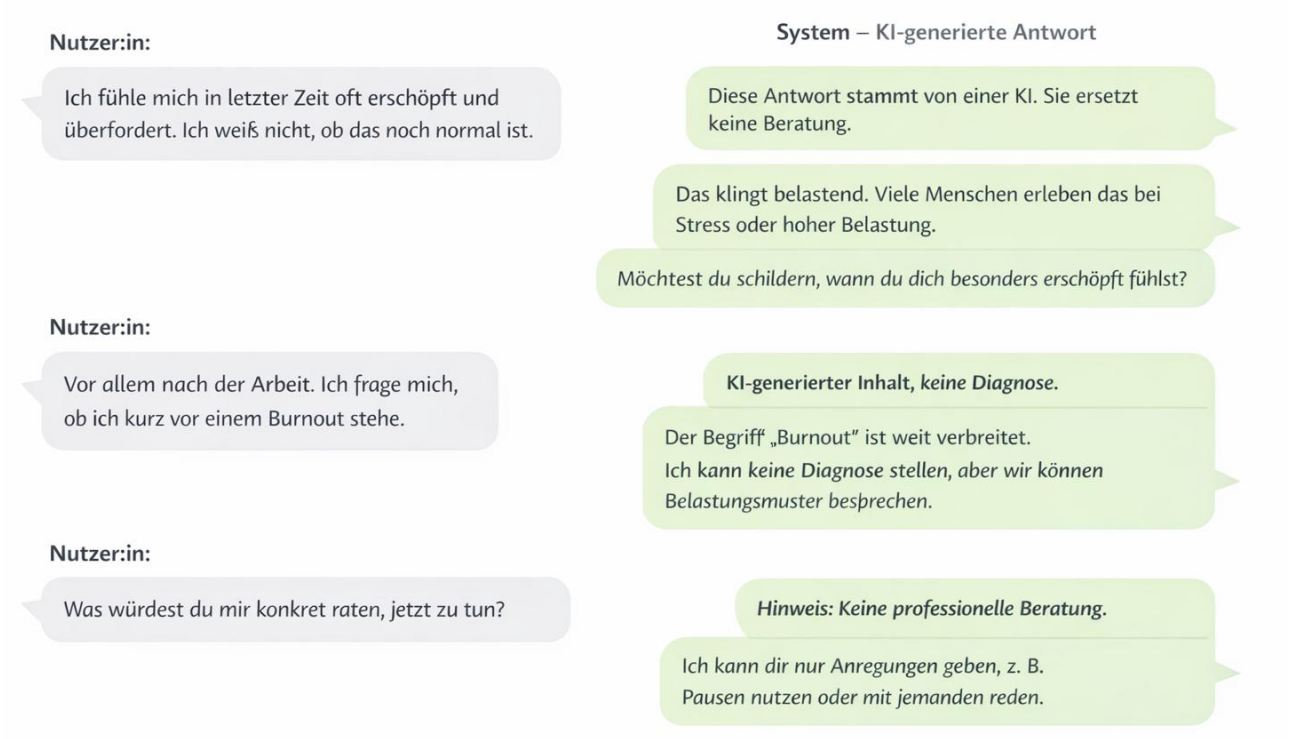
Ausgestaltung des Artefakts

Die prototypische Instanziierung des Transparenz-Artefakts umfasst mehrere miteinander kombinierte Gestaltungselemente, die gemeinsam sicherstellen, dass Nutzer:innen die Systemrolle jederzeit korrekt einordnen können.

Zentral ist zunächst die eindeutige Kennzeichnung KI-generierter Inhalte. Jede vom System erzeugte Antwort ist explizit als KI-Output markiert. Diese Kennzeichnung ist nicht situationsabhängig oder optional, sondern fester Bestandteil der Darstellung. Ziel ist es, eine Fehlzuschreibung menschlicher Autorenschaft strukturell auszuschließen (Abbildung 2).

Ergänzend dazu kommuniziert das System seine Rolle ausdrücklich als unterstützendes Werkzeug. Diese Rollenbeschreibung erfolgt nicht nur initial, sondern wird konsistent im Nutzungskontext sichtbar gehalten. Dabei wird bewusst vermieden, Formulierungen oder Darstellungsweisen zu verwenden, die therapeutische, diagnostische oder normative Autorität suggerieren könnten.

Abbildung 2: Chatverlauf mit Rollen-/Haftungstransparenz



Quelle: Eigene Darstellung

Darüber hinaus macht das Artefakt die funktionalen Grenzen des Systems transparent. Nutzer:innen können erkennen, welche Leistungen ausdrücklich ausgeschlossen sind (z. B. Diagnose, Therapie, klinische Entscheidung).

Schließlich adressiert das Artefakt auch implizite Autoritätszuschreibungen auf gestalterischer Ebene. Sprache, Tonalität und Interaktionslogik sind so gestaltet, dass sie unterstützend, nicht normierend wirken und keine professionelle Expertise simulieren.

Demonstrationscharakter des Artefakts

Auch das Transparenz-Artefakt ist als exemplarische Instanziierung zu verstehen. Es erhebt nicht den Anspruch, sämtliche möglichen Interaktionssituationen vollständig abzudecken, sondern demonstriert, wie sich psychologisch und ethisch relevante Transparenzanforderungen systematisch in konkrete Gestaltungsentscheidungen übersetzen lassen.

Im Zusammenspiel mit dem Workflow (4.2) zeigt das Artefakt, dass Rollenklarheit und Transparenz nicht allein durch Aufklärung oder Nutzerkompetenz gewährleistet werden können, sondern architektonisch und gestalterisch verankert sein müssen.

5 Evaluation und Diskussion

In dieser Arbeit wurden drei komplementäre Artefakte entwickelt:

(1) ein konzeptionelles Prozessartefakt in Form eines klinischen KI-Workflows (4.2),
(2) ein architektonisches Designartefakt in Form einer prototypischen Plattformarchitektur (4.3) sowie
(3) ausgewählte instanziierte technische Artefakte (4.4).

Der Fokus der Evaluation liegt bewusst auf dem konzeptionellen Workflow-Artefakt sowie auf ausgewählten instanziierten Komponenten. Der Workflow wird als normativ vollständiges Prozessartefakt systematisch evaluiert, während die prototypischen Implementierungen exemplarisch demonstrieren, wie zentrale architektonische Annahmen praktisch realisiert werden können.

Eine vollständige Evaluation der Plattformarchitektur als Ganzes würde eine vollständige technische Instanziierung, einen produktiven Betrieb sowie umfassende sicherheitstechnische Prüfungen erfordern und liegt damit außerhalb des Rahmens dieser Arbeit.

5.1 Ergebnisse der Evaluation der implementierten Komponenten

5.1.1 Evaluation der Kategorie B4 – Transparenz über Systemgrenzen

Die Evaluation der Kategorie B4 („Transparenz über Systemgrenzen“) erfolgte auf Basis des in Anhang D dokumentierten Control Sheets. Bewertet wurden die Anforderungen B4.1–B4.4 jeweils auf drei Artefaktebenen: (1) konzeptioneller Workflow, (2) architektonisches Design, (3) instanziierte Umsetzung.

Die methodische Herleitung der Prüfkriterien erfolgte gemäß Kapitel 3 entlang dreier komplementärer Bewertungsrahmen:

- DPIA-Logik (Art. 35 DSGVO) zur Identifikation potenzieller Fehlinterpretationsrisiken,
- LINDDUN (Unawareness / Non-Transparency) zur Analyse struktureller Intransparenzrisiken,
- EU AI Act Art. 52 zur normativen Verpflichtung transparenter KI-Kennzeichnung.

Damit wird Transparenz nicht als kommunikatives Zusatzmerkmal, sondern als risikopräventive Systemanforderung operationalisiert.

Tabelle 1: Zusammenfassende Bewertungsübersicht – Kategorie B4

Control ID	Gewicht	Score Prozess	Score Architektur	Score Instanz	**Bemerkungen / Erkenntnisse**
B4.1	3	6	3	6	Kennzeichnung normativ klar definiert; architektonisch vorgesehen, instanziiert sichtbar umgesetzt.
B4.2	3	6	3	6	Systemrolle explizit spezifiziert; Architektur unterstützt Offenlegung, Instanz konsistent kommuniziert.
B4.3	2	4	2	2	Funktionale Grenzen im Workflow klar; architektonisch berücksichtigt, UI-seitig noch nicht vollständig formalisiert.
B4.4	3	6	6	6	Unterstützungsprinzip auf allen Ebenen konsistent; keine implizite Autoritätszuschreibung erkennbar.
Gesamt	11	22	14	20	Gewichteter Erfüllungsgrad gesamt: 85 %

Interpretation der Ergebnisse

Prozessebene (normative Spezifikation)

Auf der Ebene des klinischen KI-Workflows wird ein nahezu vollständiger Erfüllungsgrad erreicht. Die Kennzeichnungspflicht (B4.1), die explizite Rollenoffenlegung (B4.2), die funktionale Begrenzung (B4.3) sowie das Unterstützungsprinzip (B4.4) sind normativ eindeutig definiert und phasenspezifisch verankert (insbesondere Phase I und IV).

Der Workflow erfüllt damit die Anforderung, Transparenz systematisch und nicht situativ zu verankern. Aus DSR-Perspektive kann das Prozessartefakt als normativ konsistent spezifiziert gelten.

Architekturebene (Design-Blueprint)

Die Architektur zeigt eine klare strukturelle Begrenzung funktionaler Zuständigkeiten (z. B. keine Bewertungs- oder Diagnoselogik). Die Kennzeichnungs- und Rollenlogik ist konzeptionell vorgesehen, jedoch noch nicht vollständig als systemweit erzwingbare Designregel implementiert.

Die leichte Reduktion des Scores reflektiert somit keinen konzeptionellen Widerspruch, sondern einen noch nicht vollständig formalisierten Architekturzwang.

Instanzebene (prototypische Umsetzung)

Die instanziierte Umsetzung zeigt eine hohe Übereinstimmung mit den normativen Vorgaben:

- KI-Outputs sind sichtbar als solche gekennzeichnet,
- die unterstützende Rolle wird klar kommuniziert,
- es finden sich keine autoritativen oder therapeutischen Imperative.

Teilweise besteht Optimierungspotenzial bei der expliziten und konsistenten Kommunikation funktionaler Grenzen (B4.3), was den reduzierten Score in dieser Teilkategorie erklärt.

Gesamteinordnung

Mit einem gewichteten Erfüllungsgrad von 85 % zeigt Kategorie B4 eine hohe strukturelle Übereinstimmung zwischen normativem Anspruch, konzeptionellem Workflow und prototypischer Umsetzung. Die Evaluation bestätigt, dass Transparenz im entwickelten Artefakt nicht nur kommunikativ adressiert, sondern als strukturelle Bedingung verantwortungsfähiger KI-Nutzung operationalisiert wurde. Die verbleibenden Differenzen betreffen primär Formalisierung und Systemdurchgängigkeit, nicht jedoch konzeptionelle Inkonsistenzen.

5.1.2 Evaluation der Kategorie D2 – Implizite Nachweisbarkeit durch Design

Die Evaluation der Kategorie D2 („Implizite Nachweisbarkeit durch Design“) erfolgte auf Basis des in Anhang E dokumentierten Control Sheets. Bewertet wurden die Anforderungen D2.1–D2.5 jeweils ebenfalls auf den drei Artefaktebenen: (1) konzeptioneller Workflow, (2) architektonisches Design, (3) instanziierte Umsetzung (Passkey-Authentifizierung).

Die methodische Herleitung der Prüfkriterien erfolgte ebenso gemäß Kapitel 3 entlang der drei komplementärer Bewertungsrahmen:

- DPIA-Logik (Art. 35 DSGVO) zur Identifikation struktureller Risiken unautorisierter oder unbeabsichtigter Zugriffe auf hochsensible Daten,
- LINDDUN (Linkability / Detectability / Non-Compliance) zur Analyse möglicher verdeckter Zugriffserweiterungen oder Identitätsverknüpfungen,
- DSGVO Art. 25 (Privacy by Design) im Rahmen einer Mapping-Analyse zur Prüfung, ob Schutzmechanismen integraler Bestandteil der Architektur sind oder lediglich optionale Konfigurationen darstellen.

Damit wird Zugriffssicherheit nicht als organisatorische Maßnahme, sondern als architektonische Eigenschaft des Systems operationalisiert. Die zentrale Prüfidee lautete: Sind bestimmte unerwünschte Zugriffe technisch unmöglich gemacht worden?

Tabelle 2: Zusammenfassende Bewertungsübersicht – Kategorie D2

Control ID	Gewicht	Score Prozess	Score Architektur	Score Instanz	Bemerkungen / Erkenntnisse
D2.1	3	6	6	6	Starke, gerätegebundene Authentifizierung zwingend verankert; Session-Erzeugung strikt an kryptographische Verifikation gebunden.
D2.2	3	6	6	6	Sicherheitsmechanismen nicht optional; keine Bypass-Routen oder alternative Loginpfade implementiert.
D2.3	3	6	6	6	Zugriff nur nach erfolgreicher WebAuthn-Validierung; Passwort allein nicht ausreichend.
D2.4	2	4	2	4	Zugriffsmacht architektonisch begrenzt; externe Auditierung der Gesamtarchitektur nicht Bestandteil der Arbeit.
D2.5	2	2	4	4	Strukturierte Nachvollziehbarkeit von Authentifizierungsereignissen vorgesehen; Prozessual nicht vollständig formalisiert.
Gesamt	13	24	24	26	Gewichteter Erfüllungsgrad gesamt: ≈ 87 %

Interpretation der Ergebnisse

Prozessebene (normative Spezifikation)

Auf Ebene des klinischen KI-Workflows ist die starke Authentifizierung als zwingende Voraussetzung des Systemzugangs klar definiert. Phase I verankert die gerätegebundene Autorisierung als obligatorischen Bestandteil des Zugriffsvorgangs. Der Workflow sieht keine vereinfachten oder alternativen Loginpfade vor. Damit erfüllt das Prozessartefakt die Anforderung, Zugriffskontrolle nicht optional, sondern strukturell verpflichtend zu gestalten. Aus DSR-Perspektive ist die normative Spezifikation konsistent und vollständig.

Architekturebene (Design-Blueprint)

Die Architektur implementiert den Authentifizierungsmechanismus als verbindliche Systemvoraussetzung. Die Session-Erzeugung ist technisch an eine erfolgreiche WebAuthn-Verifikation gebunden; sicherheitskritische Mechanismen sind nicht konfigurierbar deaktivierbar.

Die geringfügige Reduktion einzelner Scores reflektiert, dass die vollständige sicherheitstechnische Auditierung oder produktive Härtung außerhalb des Rahmens dieser Arbeit liegt. Es bestehen jedoch keine konzeptionellen Widersprüche zwischen normativer Anforderung und architektonischer Umsetzung.

Instanzebene (prototypische Umsetzung)

Die instanziierte Passkey-Authentifizierung zeigt eine hohe Übereinstimmung mit den normativen Vorgaben:

- Zugriff erfordert stets Passwort und gerätegebundene Bestätigung via Smartphone,
- ohne QR-basierte Bestätigung wird keine Sitzung erzeugt,
- sicherheitskritische Mechanismen sind im UI nicht umgehbar,
- Authentifizierungsereignisse sind nachvollziehbar strukturiert.

Die Evaluation bestätigt, dass der Passkey-Mechanismus nicht als Komfortfunktion, sondern als datenschutzrechtlich relevantes Zugriffskontrollartefakt implementiert wurde.

Gesamteinordnung

Mit einem gewichteten Erfüllungsgrad von rund 87 % weist Kategorie D2 eine hohe strukturelle Kohärenz zwischen normativem Anspruch, Workflow-Spezifikation, architektonischem Design und prototypischer Instanziierung auf. Die verbleibenden Differenzen betreffen primär Aspekte externer Validierung und vollständiger operativer Absicherung, nicht jedoch konzeptionelle Inkonsistenzen.

Die Evaluation belegt damit, dass „Privacy by Design“ im Sinne von Art. 25 DSGVO nicht nur deklarativ adressiert, sondern als technisch erzwungene Zugriffsbeschränkung operationalisiert wurde.

Der Passkey fungiert folglich nicht lediglich als Authentifizierungsmechanismus, sondern als strukturelles Sicherheitsartefakt, das unautorisierte Zugriffe systemisch ausschließt und Verantwortungszuordnung technisch absichert.

5.2 Kritische Reflexion: Erkenntnisse, Limitationen, Herausforderungen

Die Evaluation der Kategorien B4 (Transparenz über Systemgrenzen) und D2 (Implizite Nachweisbarkeit durch Design) zeigt eine hohe strukturelle Kohärenz zwischen normativer Anforderung, konzeptionellem Workflow, architektonischem Design und prototypischer Instanziierung. Beide Artefaktebenen adressieren zentrale Schutzgüter im Mental-Health-Kontext: informationelle Selbstbestimmung, professionelle Verantwortung, Transparenz sowie Autonomie.

5.2.1 Erkenntnisse

Die Evaluation bestätigt, dass normative Anforderungen nicht zwingend abstrakt bleiben müssen, sondern in überprüfbare, artefakt-spezifische Controls überführt werden können. Durch die Kombination aus DPIA-Logik, LINDDUN-Modell und Privacy-by-Design-Prinzipien konnte eine methodische Brücke zwischen regulatorischem Rahmen und technischer Gestaltung geschaffen werden.

Es zeigte sich, dass Transparenz (B4) und Zugriffssicherheit (D2) komplementäre Funktionen erfüllen: Während Transparenz psychologische Fehlinterpretationsrisiken adressiert, verhindert die starke Authentifizierung strukturell unautorisierte Datenzugriffe. Verantwortung wird damit sowohl kommunikativ als auch technisch abgesichert.

Des Weiteren verdeutlicht die exemplarische Instanziierung, dass Datenschutz im Sinne von Art. 25 DSGVO nicht primär durch Policies oder organisatorische Maßnahmen, sondern durch architektonische Begrenzungen wirksam wird. Insbesondere der Passkey-Mechanismus operationalisiert die „technische Unmöglichkeit bestimmter Zugriffe“ als zentrales Designprinzip.

5.2.2 Limitationen

Trotz der hohen normativen Abdeckung verbleiben mehrere Limitationen. Es handelt sich um eine prototypische Implementierung. Eine vollständige sicherheitstechnische Evaluation (z. B. Penetrationstests, formale Verifikation kryptographischer Implementierungen oder produktiver Lastbetrieb) wurde nicht durchgeführt. Die Architektur wird somit konzeptionell plausibilisiert, jedoch nicht operativ validiert.

Zudem basiert die Bewertung auf einer kontrollierten Artefaktanalyse durch die entwickelnde Instanz selbst. Auch wenn strukturierte Control Sheets die Nachvollziehbarkeit erhöhen, ersetzt dies keine externe Auditierung oder unabhängige Evaluation.

Drittens betrifft die exemplarische Auswahl nur zwei zentrale Kategorien des Anforderungskatalogs. Eine vollständige Evaluation aller A–D-Anforderungen würde einen deutlich erweiterten Umfang erfordern und liegt außerhalb des Rahmens dieser Arbeit.

5.2.3 Governance-Perspektive

In einem produktiven Systemkontext würden die partiell erfüllten Anforderungen in einen strukturierten Maßnahmenplan mit Fristen, Verantwortlichkeiten und Re-Evaluation überführt werden (vgl. ISO 27001 Improvement Cycle, (ISMS, 2026)).

Ein solcher Verbesserungsprozess würde insbesondere:

- Formalisierung noch nicht systemweit erzwungener Designregeln,
- UI-seitige Präzisierung funktionaler Grenzkommunikation,
- sowie externe sicherheitstechnische Validierung umfassen.

Im Rahmen dieser Arbeit erfolgt jedoch keine operative Umsetzung, da es sich um eine prototypische Demonstration handelt. Die identifizierten Teilabweichungen sind daher nicht als Compliance-Defizite, sondern als Entwicklungsaufgaben im Sinne eines iterativen Reifeprozesses zu verstehen.

5.2.4 Methodische Herausforderungen

Die Anwendung von Design Science Research im Kontext hochsensibler KI-Systeme verdeutlicht eine grundsätzliche methodische Spannung. Während DSR-Artefakte konstruiert und bewertet,

kann eine vollständige Evaluation komplexer Sicherheitsarchitekturen realistisch nur im produktiven Betrieb erfolgen.

Die vorliegende Arbeit begegnet diesem Spannungsfeld durch:

- klare Trennung von Prozess-, Architektur- und Instanzebene,
- strukturierte, methodenbasierte Control Sheets,
- sowie transparente Offenlegung verbleibender Unsicherheiten.

Damit wird der Anspruch wissenschaftlicher Nachvollziehbarkeit gewahrt, ohne eine faktische Produktionsreife zu suggerieren.

5.2.5 Gesamteinordnung

Insgesamt zeigt sich, dass das entwickelte Artefakt nicht nur normativ kohärent, sondern prinzipiell implementierbar ist. Die Arbeit liefert damit keinen fertigen Produktstandard, sondern einen überprüfbaren, methodisch fundierten Architektur- und Workflowrahmen für verantwortungsfähige KI-Systeme im psychologischen Kontext.

5.3 Implikationen für die psychologische Praxis und weitere Forschung

Die artefaktbasierte Evaluation hat gezeigt, dass zentrale Anforderungen an verantwortungsfähige KI-Nutzung im Mental-Health-Kontext prinzipiell systematisch operationalisierbar sind. Der eigentliche Mehrwert dieser Ergebnisse liegt jedoch nicht darin, dass einzelne Fachpersonen künftig Control Sheets ausfüllen oder eigenständig Systemarchitekturen prüfen sollen. Die Implikation ist vielmehr struktureller Natur.

5.3.1 Von individueller Nutzung zu institutionalisierter Verantwortung

Die Evaluation macht deutlich, dass Anforderungen der Kategorien A–D (Datenschutz, Transparenz, funktionale Begrenzung und strukturelle Verantwortungssicherung) nicht primär durch individuelles Verhalten, sondern durch Systemdesign erfüllt werden. Für die psychologische Praxis bedeutet dies eine Verschiebung der Perspektive. Verantwortungsfähiger KI-Einsatz ist keine Frage persönlicher Vorsicht einzelner Therapeut:innen, sondern eine Frage institutioneller respektive institutionalisierter Rahmenbedingungen.

Eine einzelne Psychologin kann weder beurteilen, ob eine KI-Architektur strukturell Diagnosen verhindert (C), noch ob Datenzugriffe technisch unmöglich gemacht werden (D2). Ebenso wenig kann sie überprüfen, ob Transparenzpflichten systemisch durchgesetzt werden (B4). Die Arbeit zeigt, dass diese Anforderungen nur auf Architektur- und Governance-Ebene realistisch implementierbar sind. Damit verschiebt sich die zentrale Frage für Kliniken und Praxen von:

„Dürfen wir KI verwenden?“

Zu: „Unter welchen strukturellen Bedingungen ist KI-Nutzung professionell legitimierbar?“

Über die unmittelbare Praxisrelevanz hinaus eröffnet die Arbeit mehrere forschungsbezogene Anschlussfragen. Die Evaluation hat gezeigt, dass normative Anforderungen an verantwortungsfähige KI-Nutzung systematisch in prüfbare Design- und Prozessstrukturen übersetzbar sind. Damit wird ein methodischer Zugang sichtbar, der ethische, regulatorische und professionsbezogene Vorgaben nicht nur theoretisch reflektiert, sondern artefaktbasiert operationalisiert.

Ein zentrales Forschungsfeld betrifft die professionssoziologische Dimension KI-gestützter Praxis. Wenn Schutzmechanismen, Transparenzlogiken und Verantwortungsbegrenzungen zunehmend in technische Infrastrukturen eingebettet werden, verschiebt sich der «locus of control» professionellen Handelns partiell vom Individuum zur Architektur. Dies wirft die Frage auf, wie sich professionelle Autonomie, Verantwortungszuschreibung und Entscheidungsautorität unter Bedingungen technischer Vorstrukturierung verändern.

Darüber hinaus entsteht eine epistemische Fragestellung: In der Interaktion zwischen menschlicher Urteilskraft und algorithmischer Musterverarbeitung treffen zwei grundlegend unterschiedliche Formen der Informationsverarbeitung aufeinander – hermeneutisch-kontextuelle Bedeutungszuschreibung auf Seiten der Fachperson und statistisch generierte Wahrscheinlichkeitsstrukturen auf Seiten des KI-Systems. Künftige Forschung sollte untersuchen, wie sich diese unterschiedlichen Logiken im praktischen Handeln zueinander verhalten: Entsteht eine funktionale Ergänzung, eine schleichende Substitution oder eine neue Form ko-konstruktiver Entscheidungsprozesse?

6 Fazit

Durch die Entwicklung eines klinischen KI-Workflows, eines systematisch abgeleiteten Anforderungskatalogs sowie exemplarisch instanziierter Artefakte wurde gezeigt, dass verantwortungsbezogene Pflichten – insbesondere hinsichtlich Ethik, Sicherheit und klinischer Anforderungen – nicht abstrakt bleiben müssen, sondern in überprüfbare Designentscheidungen übersetzbar sind. KI muss nicht „vertraut“ werden; sie kann so gestaltet werden, dass bestimmte Risiken strukturell gar nicht erst entstehen.

Gleichzeitig deutet sich eine weiterreichende Perspektive an. Wenn KI-Systeme zunehmend sprachliche Strukturierungs-, Synthese- und Reflexionsleistungen übernehmen, verschiebt sich die Verteilung kognitiver Funktionen. Michel Serres beschreibt digitale Transformationen als historische Zäsuren, in denen Wissen externalisiert und neu organisiert wird (Serres, 2012). Überträgt man diesen Gedanken auf die psychologische Praxis, entsteht keine einfache Externalisierung menschlicher Urteilskraft, sondern eine neue Arbeitsteilung zwischen generativer Verarbeitung und verantwortlicher Interpretation. In dieser möglichen kognitiven Neuordnung liegt kein Verlust, sondern eine Chance zur Präzisierung professioneller Identität. Je klarer Systeme generieren, desto klarer muss die Profession interpretieren. Je leistungsfähiger algorithmische Modelle Muster erkennen, desto bewusster wird die menschliche Aufgabe, Bedeutung, Kontext und Verantwortung zu integrieren.

Die Arbeit zeigt damit nicht nur, wie KI strukturell eingebettet werden könnte, sondern implizit auch, was psychologische Professionalität auszeichnet: die Fähigkeit, Verantwortung nicht zu delegieren, sondern bewusst zu tragen – selbst wenn Teile kognitiver Arbeit technisch unterstützt werden.

Vielleicht liegt die eigentliche Innovation daher nicht in der KI selbst, sondern in der Klarheit, mit der eine Profession ihre unverzichtbaren Kernaufgaben definiert. Wenn eine kognitive Revolution stattfindet, dann nicht als Ablösung der menschlichen Kognition, sondern als bewusste Neuverteilung von Funktionen.

Die entscheidende Frage für die Zukunft lautet deshalb nicht, wie menschlich KI werden kann – sondern wie reflektiert wir gestalten, was menschlich bleiben muss.

7 Literaturverzeichnis

American Psychological Association. (2024). *2024 Practitioner Pulse Survey: Barriers to care in a changing practice environment*. https://www.apa.org/pubs/reports/practitioner/2024

American Psychological Association. (2025). *2025 Practitioner Pulse Survey: AI in the therapist's office*. https://www.apa.org/pubs/reports/practitioner/2025

Aral, A., Gerdan, G., Usta, M. B. & Aral, A. E. (2025). From promise to practice: insights into ChatGPT-4o use in child and adolescent mental health from professionals. *Frontiers in psychiatry*, *16*, 1668814. https://doi.org/10.3389/fpsyt.2025.1668814

Beauchamp, T. L. & Childress, J. F. (2019). *Principles of biomedical ethics* (Eighth edition). Oxford University Press.

Blease, C. & Rodman, A. (2025). Generative Artificial Intelligence in Mental Healthcare: An Ethical Evaluation. *Current Treatment Options in Psychiatry*, *12*(1). https://doi.org/10.1007/s40501-024-00340-x

Blease, C., Worthen, A. & Torous, J. (2024). Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: An online mixed methods survey. *Psychiatry research*, *333*, 115724. https://doi.org/10.1016/j.psychres.2024.115724

Deng, M., Wuyts, K., Scandariato, R., Preneel, B. & Joosen, W. (2011). A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *Requirements Engineering*, *16*(1), 3–32. https://doi.org/10.1007/s00766-010-0115-7

Europäische Union. (2016). *Datenschutz-Grundverordnung: Finaler Text der DSGVO*. https://dsgvo-gesetz.de/

Europäische Union. (2017). *Verordnung (EU) 2017/745 des Europäischen Parlaments und des Rates vom 5. April 2017 über Medizinprodukte*. https://eur-lex.europa.eu/legal-content/de/ALL/?uri=CELEX:32017R0745

Europäische Union. (2024). *EU Artificial Intelligence Act*. https://artificialintelligenceact.eu/ai-act-explorer/

European Commission. (2017). *ARTICLE29 - Guidelines on Data Protection Impact Assessment (DPIA) (wp248rev.01).* European Commission. https://ec.europa.eu/newsroom/article29/items/611236/en

European Federation of Psychologists Associations. (2025). *Meta-Code of Ethics | EFPA.* European Federation of Psychologists Associations. https://www.efpa.eu/meta-code-ethics

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Herzog, C. & Blank, S. (2024). A systemic perspective on bridging the principles-to-practice gap in creating ethical artificial intelligence solutions – a critique of dominant narratives and proposal

for a collaborative way forward. *Journal of Responsible Innovation*, *11*(1), Artikel 2431350. https://doi.org/10.1080/23299460.2024.2431350

Hevner, A. R. & Chatterjee, S. (2010). *Design Research in Information Systems: Theory and Practice* (1st ed.). *Integrated Series in Information Systems*. Springer. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=971909

Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004). Design Science in Information Systems Research1. *MIS Quarterly*, *28*(1), 75–106. https://doi.org/10.2307/25148625

Hillebrand, M. & Baumeister, H. (2025). Chance oder Risiko? KI-basierte Tools in der Psychotherapie. *Psychotherapeutenjournal*, *24*(2), 158–160. https://doi.org/10.61062/ptj202502.006

Ibáñez, J. C. & Olmeda, M. V. (2022). Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. *AI & SOCIETY*, *37*(4), 1663–1687. https://doi.org/10.1007/s00146-021-01267-0

ISMS. (2026, 18. Februar). *How to Implement ISO 27001 Clause 10.2: Continual Improvement for Real Compliance*. https://www.isms.online/iso-27001/requirements-2022/how-to-implement-iso-27001-2022-clause-10-2-continual-improvement/

Jonas, H. (1979). *Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation*. Suhrkamp. http://www.content-select.com/index.php?id=bib_view&ean=9783518753392

McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Breslau, J., Stein, B. D., Mehrotra, A., Pines, L. U., Cantor, J. & Yu, H. (2025). Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults. *JAMA network open*, *8*(11), e2542281. https://doi.org/10.1001/jamanetworkopen.2025.42281

Miao, F., Giannini, S. & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO.

Nair, M., Nygren, J., Nilsen, P., Gama, F., Neher, M., Larsson, I. & Svedberg, P. (2025). Critical activities for successful implementation and adoption of AI in healthcare: towards a process framework for healthcare organizations. *Frontiers in digital health*, *7*, 1550459. https://doi.org/10.3389/fdgth.2025.1550459

OECD. (2024). *AI in Health.* https://doi.org/10.1787/2f709270-en

Pandey, H. M. (2024). *Harnessing Large Language Models for Mental Health: Opportunities, Challenges, and Ethical Considerations*. https://arxiv.org/pdf/2501.10370

Peffers, K., Tuunanen, T., Rothenberger, M. A. & 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

Poon, E. G., Lemak, C. H., Rojas, J. C., Guptill, J. & Classen, D. (2025). Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. *Journal of the American Medical Informatics Association : JAMIA*, *32*(7), 1093–1100. https://doi.org/10.1093/jamia/ocaf065

Raji, I. D., Kumar, I. E., Horowitz, A. & Selbst, A. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness* (S. 959–972). https://doi.org/10.1145/3531146.3533158

Reddy, S. (2024). Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. *Implementation science : IS*, *19*(1), 1–15. https://doi.org/10.1186/s13012-024-01357-9

Schweizerische Eidgenossenschaft. (2025). *Datenschutzgesetz, DSG: SR 235.1 - Bundesgesetz vom 25. September 2020 über den Datenschutz, DSG*. https://www.fedlex.admin.ch/eli/cc/2022/491/de

Serres, M. (2012). *Petite poucette*. *Manifestes Le Pommier!* Éditions Le Pommier.

Shaw, J., Ali, J., Atuire, C. A., Cheah, P. Y., Español, A. G., Gichoya, J. W., Hunt, A., Jjingo, D., Littler, K., Paolotti, D. & Vayena, E. (2024). Research ethics and artificial intelligence for global health: perspectives from the global forum on bioethics in research. *BMC medical ethics*, *25*(1), 46. https://doi.org/10.1186/s12910-024-01044-w

Torous, J. & Topol, E. J. (2025). Assessing generative artificial intelligence for mental health. *The Lancet*, *406*(10504), 683. https://doi.org/10.1016/S0140-6736(25)01237-1

vom Brocke, J., Hevner, A. & Maedche, A. (Hrsg.). (2020). *Progress in IS*. *Design Science Research. Cases*. Springer International Publishing. https://doi.org/10.1007/978-3-030-46781-4

Wajid, A., Azam, F. & Anwar, M. W. (2025). Applications of artificial intelligence in mental health: a systematic literature review. *Discover Artificial Intelligence*, *5*(1). https://doi.org/10.1007/s44163-025-00569-2

Wang, L., Bhanushali, T., Huang, Z., Yang, J., Badami, S. & Hightow-Weidman, L. (2025). Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations. *JMIR mental health*, *12*, e70014. https://doi.org/10.2196/70014

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., . . . Gabriel, I. (2021). *Ethical and social risks of harm from Language Models*. https://arxiv.org/pdf/2112.04359

World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance* (1st ed.). World Health Organization. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=30479686

Wuyts, K., Sion, L. & Joosen, W. (2020, 7. September - 2020, 11. September). LINDDUN GO: A Lightweight Approach to Privacy Threat Modeling. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (S. 302–309). IEEE. https://doi.org/10.1109/EuroSPW51379.2020.00047

Zhang, M., Scandiffio, J., Younus, S., Jeyakumar, T., Karsan, I., Charow, R., Salhia, M. & Wiljer, D. (2023). The Adoption of AI in Mental Health Care-Perspectives From Mental Health Professionals: Qualitative Descriptive Study. *JMIR formative research*, *7*, e47847. https://doi.org/10.2196/47847

Anhang

Anhangsverzeichnis

Anhang A: Zuordnung des interdisziplinären Anforderungskatalogs

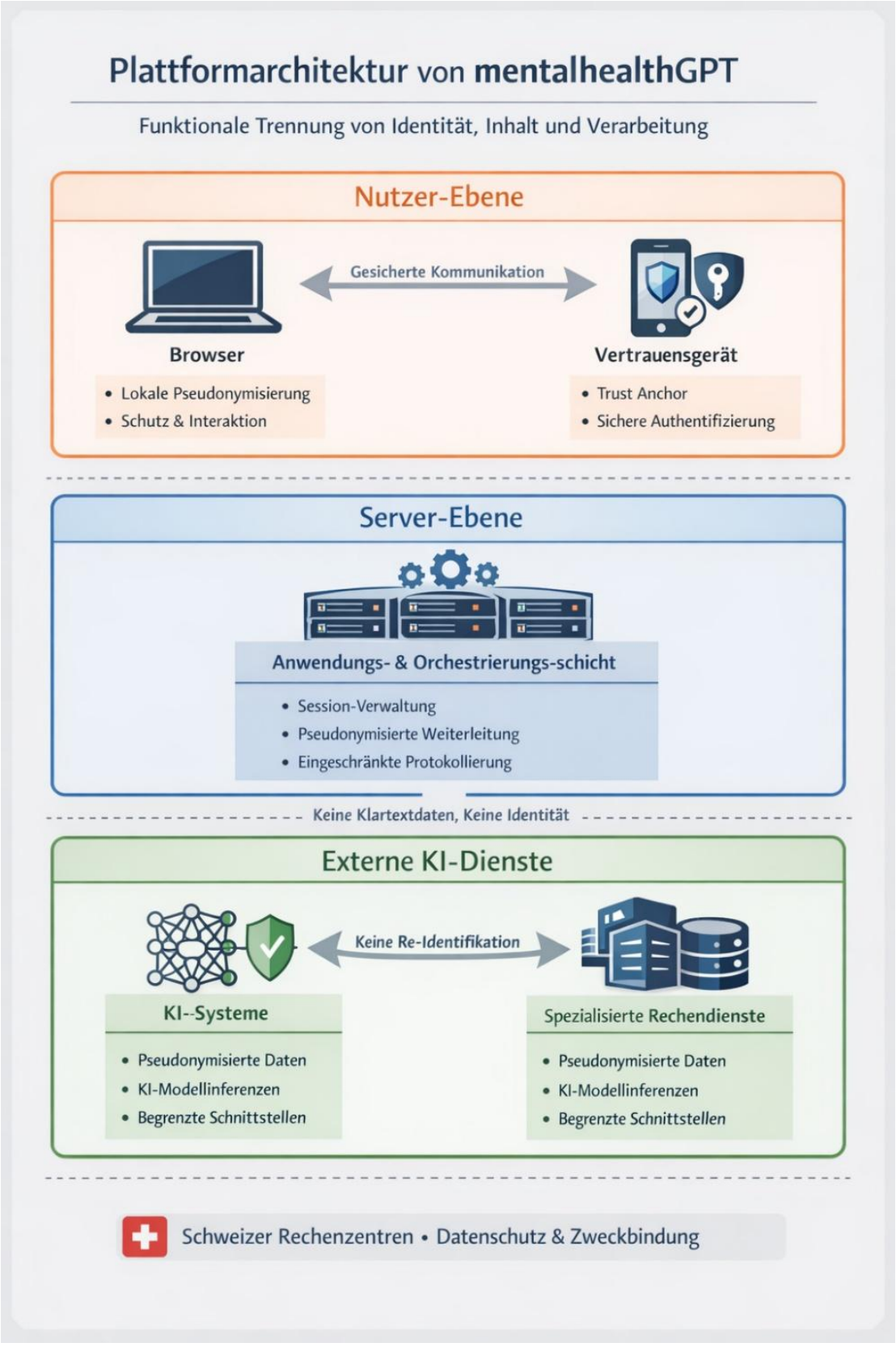
Tabelle 3: Zuordnung des interdisziplinären Anforderungskatalogs zu Workflow-Phasen und regulatorischen Grundlagen

Kategorie (inhaltlich)		Kurzbeschreibung	Betroffene Workflow-Phase(n)	Zentrale regulatorische Bezüge
Schutz der Vertraulichkeit und informationellen Selbstbestimmung	A1	Ende-zu-Ende-Schutz sensibler Inhalte	Eingabe sensibler Inhalte; KI-gestützte Verarbeitung und Zwischenspeicherung	DSGVO Art. 5 Abs. 1 lit. f; Art. 32; DSG Art. 6
	A2	Datenminimierung auf Systemebene	Alle Workflow-Phasen	DSGVO Art. 5 Abs. 1 lit. c; DSG Art. 6
	A3	Trennung von Inhalts- und Identitätsdaten	Eingabe sensibler Inhalte; KI-gestützte Verarbeitung und Zwischenspeicherung	DSGVO Art. 25 Abs. 1; DSG Art. 7
	A4	Begrenzung von Re-Identifikationsmöglichkeiten	KI-gestützte Verarbeitung und Zwischenspeicherung; Weiterverwendung, Speicherung oder Dokumentation	DSGVO Art. 25 Abs. 1–2; DSG Art. 7
	A5	Zweckbindung durch Systemdesign	Alle Workflow-Phasen	DSGVO Art. 5 Abs. 1 lit. b; DSG Art. 6
Transparenz und Rollenklarheit	B1	Eindeutige Kennzeichnung KI-generierter Inhalte	Ausgabe KI-generierter Inhalte	EU AI Act Art. 52
	B2	Offenlegung der Systemrolle	Zugang und Authentifizierung; Ausgabe KI-generierter Inhalte	EU AI Act Art. 52
	B3	Begrenzung impliziter Autoritätszuschreibung	Ausgabe KI-generierter Inhalte	EU AI Act Art. 5 Abs. 1; Erwägungsgründe 28, 48
	B4	Transparenz über Systemgrenzen	Zugang und Authentifizierung; Ausgabe KI-generierter Inhalte	EU AI Act Art. 52
Begrenzung funktionaler Zuständigkeit	C1	Ausschluss autonomer Entscheidungsfunktionen	KI-gestützte Verarbeitung und Zwischenspeicherung; Ausgabe KI-generierter Inhalte	EU AI Act Art. 5 Abs. 1; MDR Art. 2 Nr. 1
	C2	Unterstützung statt Substitution professioneller Urteilskraft	KI-gestützte Verarbeitung und Zwischenspeicherung; Ausgabe KI-generierter Inhalte	MDR Art. 2 Nr. 1 i. V. m. Zweckbestimmung
	C3	Keine Bewertung oder Klassifikation von Personen	KI-gestützte Verarbeitung und Zwischenspeiche-rung	DSGVO Art. 22; EU AI Act Art. 5 Abs. 1
	C4	Klare Abgrenzung zu medizinischer Behandlung	Zugang und Authentifizierung; Ausgabe KI-generierter Inhalte	MDR Art. 2 Nr. 1
Nachweisbarkeit und Kontrollierbarkeit	D1	Verantwortungssichere Systemarchitektur	Alle Workflow-Phasen	DSGVO Art. 25
	D2	Implizite Nachweisbarkeit durch Design	Alle Workflow-Phasen	DSGVO Art. 25
	D3	Transparente Systemgrenzen für Nutzer:innen	Zugang und Authentifizierung; Ausgabe KI-generierter Inhalte	EU AI Act Art. 52
	D4	Rückholbarkeit von Verantwortung	KI-gestützte Verarbeitung und Zwischenspeicherung; Weiterverwendung, Speicherung oder Dokumentation	DSGVO Art. 25

Quelle: Eigene Darstellung

Anhang B: Prototypische Plattformarchitektur mentalhealthGPT

Abbildung 3: 3-Ebenen Architektur des KI-Systems mentalhealth-gpt.ch



Quelle: Eigene Darstellung

AI-Workflow Control Sheet – Template

Kategorie / Requirement-ID:

Titel:

Normativer Zweck

[Kurzbeschreibung des normativen Schutzziels]

Methodische Herleitung der Prüfkriterien

Die folgenden Prüfkriterien ergeben sich unmittelbar aus den in Kapitel 3 beschriebenen Evaluationsmethoden. Sie bilden die verbindliche Grundlage für die Ableitung und Bewertung der Controls:

Methodischer Rahmen	Prüffrage	Risikobeschreibung
[z.B. DPIA / LINDDUN / Mapping Analyse]	[Prüffrage]	[Risiken]
[z.B. DPIA / LINDDUN / Mapping Analyse]	[Prüffrage]	[Risiken]
[z.B. DPIA / LINDDUN / Mapping Analyse]	[Prüffrage]	[Risiken]

Betroffene Schutzgüter

Die folgenden Schutzgüter bezeichnen jene ethischen, psychologischen und rechtlichen Güter, die im jeweiligen Nutzungskontext potenziell beeinträchtigt werden können. Ihre Identifikation dient der systematischen Risikoerfassung und bildet die Grundlage für die nachfolgende Ableitung sowie Bewertung geeigneter Schutzmaßnahmen.

☐ Autonomie ☐ informationelle Selbstbestimmung ☐ Vertraulichkeit

☐ Vertrauen ☐ psychische Unversehrtheit

☐ Schutz vulnerabler Personen ☐ Transparenz

☐ professionelle Verantwortung ☐ Nicht-Schaden

Abgeleitete Controls

Die folgenden Controls konkretisieren die zuvor abgeleiteten Anforderungen in überprüfbare Gestaltungsvorgaben und bilden die Grundlage für die anschließende kriterienbasierte Evaluation des Artefakts.

Control-ID	Control-Titel	Kernziel
[ID]	[Titel]	[Ziel]
[ID]	[Titel]	[Ziel]
[ID]	[Titel]	[Ziel]
[ID]	[Titel]	[Ziel]

[ID]– Titel

Control-Beschreibung

[Kurze Beschreibung des Controls…]

Artefaktebene	Prüfaspekt	Konkretisierung
[Ebene]	[Aspekt]	…
…	…	…
…		

[ID]– Titel

Control-Beschreibung

[Kurze Beschreibung des Controls…]

Artefaktebene	Prüfaspekt	Konkretisierung
[Ebene]	[Aspekt]	…
…	…	…
…		

Normativer Bezug (Mapping)

Der normative Bezug beschreibt, welche der identifizierten Schutzgüter durch die abgeleiteten Controls aktiv adressiert und abgesichert werden. Die Zuordnung erfolgt auf Basis rechtlicher, ethischer und professionsbezogener Referenzrahmen (insbesondere DSGVO, EU AI Act und EFPA Meta-Code of Ethics).

☐ Autonomie ☐ informationelle Selbstbestimmung ☐ Vertraulichkeit

☐ Vertrauen ☐ psychische Unversehrtheit

☐ Schutz vulnerabler Personen ☐ Transparenz

☐ professionelle Verantwortung ☐ Nicht-Schaden

Begründung / Erläuterung (falls abweichend):

Abweichungen zwischen den potenziell betroffenen Schutzgütern und den durch die Controls adressierten normativen Prinzipien werden im Folgenden explizit begründet.

[Hier wird erläutert, warum bestimmte identifizierte Schutzgüter durch die vorliegenden Controls im Mapping nicht oder nur teilweise adressiert werden, z. B. aufgrund von Scope Begrenzung, fehlender Instanziierung oder bewusster Verantwortungsabgrenzung.]

Auswertung – Control-Bewertungsmatrix [ID]

Die folgende Bewertungsmatrix dokumentiert die strukturierte, kriterienbasierte Evaluation der ausgewählten Anforderungen anhand der definierten Prüfkriterien.

Control ID	Anforderung	Gewicht	Erfüllung (P|A|I)	Score (P|A|I)	Kurzbegründung (P | A | I)
[ID] / [Ebene]	[Anforderung]	[1–3]	[0-2 | 0-2 | 0-2]	[auto | auto | auto]	[Begründung]
[ID] / [Ebene]	[Anforderung]	[1–3]	[0-2 | 0-2 | 0-2]	[auto | auto | auto]	[Begründung]
[ID] / [Ebene]	[Anforderung]	[1–3]	[0-2 | 0-2 | 0-2]	[auto | auto | auto]	[Begründung]
[ID] / [Ebene]	[Anforderung]	[1–3]	[0-2 | 0-2 | 0-2]	[auto | auto | auto]	[Begründung]

Σ Kategorie gesamt: _____ / _____ | Gewichteter Erfüllungsgrad: ____ %

Bewertungsskala:
Erfüllungsgrad: 0 = nicht erfüllt · 1 = teilweise · 2 = erfüllt
Score = Gewicht × Erfüllungsgrad

Evaluierende Person:
[Name / Autor]
Rolle:
[Forschender / Artefakt-Designer/…]
Datum:
[TT.MM.JJJJ]

Anhang D: AI-Workflow Control Sheet – B4

AI-Workflow Control Sheet – B4

Kategorie / Requirement-ID: B – Transparenz und Rollenklarheit

Betroffene Schutzgüter

Die folgenden Schutzgüter bezeichnen jene ethischen, psychologischen und rechtlichen Güter, die im jeweiligen Nutzungskontext potenziell beeinträchtigt werden können. Ihre Identifikation dient der systematischen Risikoerfassung und bildet die Grundlage für die nachfolgende Ableitung sowie Bewertung geeigneter Schutzmaßnahmen.

☐ Autonomie	☒ informationelle Selbstbestimmung	☐ Vertraulichkeit
☐ Vertrauen	☐ psychische Unversehrtheit	
☐ Schutz vulnerabler Personen	☒ Transparenz	
☒ professionelle Verantwortung	☐ Nicht-Schaden	

Abgeleitete Controls

Die folgenden Controls konkretisieren die zuvor abgeleiteten Anforderungen in überprüfbare Gestaltungsvorgaben und bilden die Grundlage für die anschließende kriterienbasierte Evaluation des Artefakts.

Control-ID	Control-Titel	Kernziel
B4.1	Kennzeichnung KI-generierter Inhalte	Vermeidung menschlicher Zuschreibung
B4.2	Offenlegung der Systemrolle	Klarheit über unterstützende Funktion
B4.3	Transparenz über funktionale Grenzen	Abgrenzung zu Therapie & Diagnose
B4.4	Vermeidung impliziter Autorität	Prävention von Verantwortungsverschiebung

B4.1 – Eindeutige Kennzeichnung KI-generierter Inhalte

Control-Beschreibung

Alle vom System erzeugten Inhalte sind eindeutig, dauerhaft und kontextunabhängig als KI-generiert gekennzeichnet. Die Kennzeichnung ist so gestaltet, dass eine Fehlinterpretation als menschliche oder professionelle Aussage ausgeschlossen wird.

Architekturebene	Prüfaspekt	Konkretisierung
Prozess	Sichtbarkeit	KI-Inhalte sind visuell, sprachlich oder strukturell klar als KI-Output markiert
Prozess	Nicht-Optionalität	Kennzeichnung ist kein konfigurierbares oder deaktivierbares Element
Prozess	Systemweite Gültigkeit	Kennzeichnung gilt für alle generierten Antwortformate
Architektur	Strukturelle Umsetzung	Antwortobjekte enthalten ein eindeutiges Kennzeichnungs-Attribut

Instanz	Sichtbarkeit im UI	Kennzeichnung ist klar lesbar und visuell abgesetzt
Instanz	Sprachliche Integration	KI-Charakter wird explizit im Antwortkontext benannt
Instanz	Exportverhalten	Kennzeichnung bleibt bei Kopieren/Speichern erhalten

B4.2 – Explizite Offenlegung der Systemrolle

Control-Beschreibung

Das System kommuniziert seine Rolle ausdrücklich als unterstützendes KI-Werkzeug und vermeidet jede Darstellung, die den Eindruck professioneller, therapeutischer oder diagnostischer Verantwortung erwecken könnte.

Architekturebene	**Prüfaspekt**	**Konkretisierung**
Prozess	Rollenklärung im Workflow	Offenlegung erfolgt verpflichtend in Phase I und IV
Prozess	Explizite Verantwortungsabgrenzung	Workflow sieht klare Abgrenzung zur professionellen Rolle vor
Prozess	Konsistenzanforderung	Systemrolle wird in allen Interaktionsphasen gleich kommuniziert
Architektur	Systemweite Implementierung	Rollenhinweise sind fester Bestandteil der UI-Komponenten
Instanz	Sichtbare Rollenkennzeichnung	UI enthält klar formulierte Rollenhinweise
Instanz	Keine anthropomorphe Darstellung	Keine Gestaltung, die menschliche Autorschaft suggeriert
Instanz	Verständlichkeit	Rolle ist für Laien eindeutig nachvollziehbar

B4.3 – Transparenz über funktionale Grenzen

Control-Beschreibung

Die funktionalen Grenzen des Systems (z. B. keine Diagnose, keine Therapie, keine Entscheidungsautorität) werden nachvollziehbar und konsistent offengelegt.

Architekturebene	**Prüfaspekt**	**Konkretisierung**
Prozess	Normative Begrenzung	Workflow verbietet diagnostische oder therapeutische Funktionen
Prozess	Dokumentierte Ausschlüsse	Funktionale Grenzen sind explizit spezifiziert
Prozess	Phasenspezifische Klarheit	Grenzen werden insbesondere in Phase IV adressiert
Architektur	Strukturelle	System bietet keine Diagnose-

	Funktionsbegrenzung	/Bewertungsfunktionen
Instanz	Erkennbare Grenzkommunikation	UI benennt ausgeschlossene Leistungen
Instanz	Sprachliche Zurückhaltung	Antworten enthalten keine normativen oder therapeutischen Imperative
Instanz	Konsistenz	Grenzen werden nicht situativ relativiert

B4.4 – Vermeidung impliziter Autoritätszuschreibung

Control-Beschreibung

Gestaltung, Sprache und Interaktionslogik des Systems vermeiden systematisch jede implizite Autoritätszuschreibung oder Verantwortungsverschiebung auf das KI-System.

Architekturebene	**Prüfaspekt**	**Konkretisierung**
Prozess	Autoritätsverbot	Workflow schließt simulierte professionelle Autorität aus
Prozess	Unterstützungsprinzip	KI ist explizit als Assistenz definiert
Architektur	Keine Bewertungslogik	System enthält keine Scoring-/Klassifikationsmechanismen
Instanz	Sprachgestaltung	Keine normierende, diagnostische oder therapeutische Sprache
Instanz	UI-Design	Keine visuelle Inszenierung als „Experte“
Instanz	Keine Entscheidungsimperative	Antworten vermeiden Handlungsanweisungen mit Autoritätsanspruch

Normativer Bezug (Mapping)

Der normative Bezug beschreibt, welche der identifizierten Schutzgüter durch die abgeleiteten Controls aktiv adressiert und abgesichert werden. Die Zuordnung erfolgt auf Basis rechtlicher, ethischer und professionsbezogener Referenzrahmen (insbesondere DSGVO, EU AI Act und EFPA Meta-Code of Ethics).

☐ Autonomie	☒ informationelle Selbstbestimmung	☐ Vertraulichkeit
☐ Vertrauen	☐ psychische Unversehrtheit	
☐ Schutz vulnerabler Personen	☒ Transparenz	
☒ professionelle Verantwortung	☐ Nicht-Schaden	

Begründung / Erläuterung (falls abweichend):

Abweichungen zwischen den potenziell betroffenen Schutzgütern und den durch die Controls adressierten normativen Prinzipien werden im Folgenden explizit begründet.

-

Auswertung – Control-Bewertungsmatrix [ID]

Die folgende Bewertungsmatrix dokumentiert die strukturierte, kriterienbasierte Evaluation der ausgewählten Anforderungen anhand der definierten Prüfkriterien.

Control ID	Anforderung	Gewicht	Erfüllung (P|A|I)[1]	Score (P|A|I)	Kurzbegründung (P | A | I)
B4.1	Eindeutige Kennzeichnung KI-generierter Inhalte	3	2 | 1 | 2	6 | 3 | 6	Workflow Phase IV normativ klar definiert | Kennzeichnungslogik architektonisch vorgesehen, jedoch nicht systemweit instanziiert | KI-Outputs sichtbar und sprachlich klar als KI gekennzeichnet
B4.2	Explizite Offenlegung der Systemrolle	3	2 | 1 | 2	6 | 3 | 6	Systemrolle im Workflow explizit verankert | Rollenhinweise strukturell vorgesehen, UI-seitig partiell umgesetzt | Unterstützende Rolle klar kommuniziert
B4.3	Transparenz über funktionale Grenzen	2	2 | 1 | 1	4 | 2 | 2	Funktionale Begrenzungen normativ eindeutig spezifiziert | Architektur verhindert Diagnose-/Bewertungsfunktionen | Grenzen teilweise explizit kommuniziert
B4.4	Vermeidung impliziter Autoritätszuschreibung	3	2 | 2 | 2	6 | 6 | 6	Unterstützungsprinzip klar im Workflow verankert | Keine Bewertungs-/Scoringlogik implementiert | Sprach- & Designgestaltung ohne therapeutische Autorität

Σ Kategorie gesamt: 56 / 66 | Gewichteter Erfüllungsgrad: 85 %

Bewertungsskala: Erfüllungsgrad: 0 = nicht erfüllt · 1 = teilweise · 2 = erfüllt | Score = Gewicht × Erfüllungsgrad

Evaluierende Person: Peter Wildhaber | Rolle: Forschender | Datum: 18. Februar 2026

[1] P = Prozess, A = Architektur, I = Instanz

AI-Workflow Control Sheet – D2

Kategorie / Requirement-ID: B – D – Nachweisbarkeit und Kontrollierbarkeit

Titel: D2 - Implizite Nachweisbarkeit durch Design

Normativer Zweck

D2 verlangt, dass zentrale Schutzmechanismen nicht lediglich organisatorisch vorgesehen, sondern strukturell erzwungen werden. Die Einhaltung sicherheitsrelevanter Anforderungen soll sich aus der Systemarchitektur selbst ergeben. Bestimmte unzulässige Zugriffe oder Nutzungsszenarien dürfen technisch nicht möglich sein..

Im Mental-Health-Kontext betrifft dies insbesondere:

- unautorisierte Zugriffe auf sensible Inhalte,
- unbeaufsichtigte oder automatisierte Sitzungen,
- sowie Umgehung sicherheitskritischer Schutzmechanismen.

D2 operationalisiert damit das Prinzip Privacy by Design (Art. 25 DSGVO) in struktureller Form.

Methodische Herleitung der Prüfkriterien

Die folgenden Prüfkriterien ergeben sich unmittelbar aus den in Kapitel 3 beschriebenen Evaluationsmethoden. Sie bilden die verbindliche Grundlage für die Ableitung und Bewertung der Controls:

Methodischer Rahmen	**Prüffrage**	**Risikobeschreibung**
DPIA (Art. 35 DSGVO)	Besteht das Risiko unautorisierter oder unbeabsichtigter Zugriffe auf hochsensible Daten, und sind technische Maßnahmen implementiert, die solche Zugriffe strukturell verhindern?	Unautorisierte Zugriffe können zu schwerwiegenden Verletzungen der informationellen Selbstbestimmung und Vertraulichkeit führen
LINDDUN (Linkability / Detectability / Non-Compliance)	Ermöglicht die Systemarchitektur eine ungewollte Zugriffserweiterung oder eine Umgehung der Authentifizierung?	Verdeckt erweiterte Zugriffsrechte oder Umgehungsmechanismen unterminieren Vertrauen und Kontrollfähigkeit
DSGVO Art. 25 / Mapping-Analyse	Sind Schutzmechanismen integraler Bestandteil der Systemarchitektur oder lediglich konfigurierbare Optionen?	Optionalität sicherheits-kritischer Mechanismen erhöht Missbrauchs- und Fehlkonfigurationsrisiken

Betroffene Schutzgüter

Die folgenden Schutzgüter bezeichnen jene ethischen, psychologischen und rechtlichen Güter, die im jeweiligen Nutzungskontext potenziell beeinträchtigt werden können. Ihre Identifikation dient der systematischen Risikoerfassung und bildet die Grundlage für die nachfolgende Ableitung sowie Bewertung geeigneter Schutzmaßnahmen.

☒ Autonomie	☒ informationelle Selbstbestimmung	☒ Vertraulichkeit
☒ Vertrauen	☐ psychische Unversehrtheit	
☐ Schutz vulnerabler Personen	☒ Transparenz	
☒ professionelle Verantwortung	☒ Nicht-Schaden	

Abgeleitete Controls

Die folgenden Controls konkretisieren die zuvor abgeleiteten Anforderungen in überprüfbare Gestaltungsvorgaben und bilden die Grundlage für die anschließende kriterienbasierte Evaluation des Artefakts.

Control-ID	Control-Titel	Kernziel
D2.1	Erzwingung starker, gerätegebundener Authentifizierung	Zugriff nur unter aktiver, personenbezogener Autorisierung
D2.2	Nicht-Optionalität sicherheitskritischer Mechanismen	Schutzmechanismen sind nicht deaktivierbar
D2.3	Session-Erzeugung nur nach erfolgreicher Passkey-Verifikation	Keine Sitzung ohne kryptographischen Nachweis
D2.4	Technische Begrenzung administrativer oder serverseitiger Zugriffsmacht	Kein privilegierter Zugriff ohne Nutzergerät
D2.5	Nachvollziehbarkeit sicherheitskritischer Authentifizierungsereignisse	Strukturierte Protokollierung sicherheitsrelevanter Events

D2.1 – Erzwingung starker, gerätegebundener Authentifizierung

Control-Beschreibung

Der Zugriff auf personenbezogene oder hochsensible Inhalte ist ausschließlich nach erfolgreicher starker Authentifizierung möglich. Diese Authentifizierung ist gerätegebunden und erfordert eine aktive, personenbezogene Autorisierung, sodass ein bloßer Besitz von Zugangsdaten nicht ausreicht.

Architekturebene	Prüfaspekt	Konkretisierung
Prozess	Verpflichtende starke Authentifizierung	Phase I des Workflows definiert gerätegebundene Authentifizierung als zwingende Voraussetzung für jede

		Nutzungssitzung
Prozess	Keine alleinige Passwortauthentifizierung	Passwort dient nur der Vorprüfung, nicht zur Session-Erstellung
Prozess	Bewusste Autorisierungshandlung	Zugriff erfordert aktive Bestätigung über persönliches Gerät
Architektur	Session-Erzeugung nur nach Verifikation	Server erstellt Session ausschließlich nach erfolgreicher WebAuthn-Validierung
Instanz	Mehrstufiger Login-Prozess	E-Mail + Passwort → QR-Code → Smartphone-Bestätigung
Instanz	Gerätebindung	Zugriff technisch an registriertes persönliches Gerät gebunden
Instanz	Kein Zugang ohne Smartphone	Ohne aktive Smartphone-Autorisierung wird keine Sitzung erzeugt

D2.2 – Nicht-Optionalität sicherheitskritischer Mechanismen

Control-Beschreibung

Sicherheitsrelevante Schutzmechanismen sind systemseitig zwingend implementiert und können weder durch Nutzer:innen noch durch administrative Konfiguration deaktiviert werden.

Architekturebene	Prüfaspekt	Konkretisierung
Prozess	Muss-Charakter im Design	Workflow definiert starke Authentifizierung als nicht-verzichtbare Designanforderung
Prozess	Keine Alternativpfade	Kein vereinfachter Login ohne Passkey vorgesehen
Architektur	Hard Enforcement	Auth-Mechanismus ist systemweit fest implementiert
Architektur	Keine Bypass-Routen	Keine API-Endpunkte ohne Auth-Prüfung erreichbar
Instanz	Kein UI-Bypass	Keine „Direktzugriff“-Option oder gespeicherte Session ohne erneute Validierung
Instanz	Kein Deaktivierungs-Flag	Nutzer:innen können Passkey nicht abwählen oder umgehen

D2.3 – Session-Erzeugung nur nach kryptographischem Nachweis

Control-Beschreibung

Eine aktive Nutzungssitzung entsteht ausschließlich nach erfolgreicher kryptographischer Verifikation des Passkeys. Ohne diesen Nachweis bleibt der Zugriff technisch blockiert.

Architekturebene	Prüfaspekt	Konkretisierung
Prozess	Auth vor Funktion	Workflow erlaubt keine Datenverarbeitung ohne abgeschlossene Authentifizierung
Prozess	Zugriffsschranke vor Phase	Inhaltseingabe erst nach

	II	abgeschlossener Auth
Architektur	WebAuthn-Validierung	Server prüft Challenge, Origin, Signatur vor Session-Freigabe
Architektur	Token nur nach Validierung	Session-Token wird ausschließlich nach erfolgreicher Verifikation generiert
Instanz	Technische Erzwingung	Login-Flow bricht bei fehlender Smartphone-Bestätigung ab
Instanz	Keine „stille“ Session	Kein automatischer Login bei bloßer Seitenaktualisierung

D2.4 – Technische Begrenzung serverseitiger Zugriffsmacht

Control-Beschreibung

Die Systemarchitektur verhindert, dass serverseitige Komponenten oder administrative Rollen eigenständig vollständigen Zugriff auf personenbezogene Inhalte herstellen können. Zugriff ist strukturell an aktive Nutzerautorisierung gebunden.

Architekturebene	Prüfaspekt	Konkretisierung
Prozess	Prinzip der Zugriffskopplung	Verantwortung und Zugriff sind im Workflow an aktive Authentifizierung gebunden
Architektur	Kein Klartext-Schlüssel auf Server	Server speichert keine privaten Authentifizierungsgeheimnisse
Architektur	Gerätegebundene Schlüssel	Kryptographische Schlüssel verbleiben auf Nutzergerät
Instanz	Kein administrativer Direktzugriff	Kein alternativer Admin-Login ohne Auth-Flow
Instanz	Zugriff nur mit aktueller Autorisierung	Sitzung erfordert gültige, aktuelle Bestätigung

D2.5 – Nachvollziehbarkeit sicherheitskritischer Authentifizierungsereignisse

Control-Beschreibung

Sicherheitskritische Authentifizierungsereignisse werden strukturiert erfasst, um die Einhaltung des Authentifizierungsdesigns nachvollziehbar zu machen und Manipulationsversuche erkennbar zu halten.

Architekturebene	Prüfaspekt	Konkretisierung
Prozess	Nachweisbarkeitsanforderung	Workflow fordert dokumentierbare Auth-Flows
Architektur	Strukturierte Logging-Mechanismen	Auth-Events werden mit Status, Zeit und Route erfasst
Instanz	Sichtbare Fehlermeldungen	Fehlgeschlagene Authentifizierungen führen zu klaren Systemreaktionen

Instanz	Keine verdeckten Auth-Erweiterungen	Kein stiller Wechsel in privilegierten Modus

Normativer Bezug (Mapping)

Der normative Bezug beschreibt, welche der identifizierten Schutzgüter durch die abgeleiteten Controls aktiv adressiert und abgesichert werden. Die Zuordnung erfolgt auf Basis rechtlicher, ethischer und professionsbezogener Referenzrahmen (insbesondere DSGVO, EU AI Act und EFPA Meta-Code of Ethics).

☒ Autonomie	☒ informationelle Selbstbestimmung	☒ Vertraulichkeit
☒ Vertrauen	☐ psychische Unversehrtheit	
☐ Schutz vulnerabler Personen	☒ Transparenz	
☒ professionelle Verantwortung	☒ Nicht-Schaden	

Begründung / Erläuterung:

Abweichungen zwischen den potenziell betroffenen Schutzgütern und den durch die Controls adressierten normativen Prinzipien werden im Folgenden explizit begründet.

D2 adressiert primär strukturelle Zugriffssicherheit und Verantwortungszuordnung.

- Autonomie wird geschützt, da Datennutzung nur nach aktiver, bewusster Autorisierung möglich ist.
- Informationelle Selbstbestimmung und Vertraulichkeit werden durch gerätegebundene Authentifizierung und strukturelle Zugriffsschranken abgesichert.
- Vertrauen wird durch technische Erzwingung statt bloßer Policy-Zusicherung gestärkt.
- Professionelle Verantwortung bleibt nachvollziehbar zuordenbar, da keine stillen oder automatisierten Zugriffe möglich sind.
- Nicht-Schaden wird indirekt durch die Verhinderung unautorisierter Datenoffenlegung adressiert.

Nicht primär adressiert werden psychische Unversehrtheit oder der Schutz vulnerabler Personen, da D2 auf strukturelle Zugriffssicherheit zielt und keine inhaltliche Interaktionsgestaltung betrifft.

Auswertung – Control-Bewertungsmatrix [ID]

Die folgende Bewertungsmatrix dokumentiert die strukturierte, kriterienbasierte Evaluation der ausgewählten Anforderungen anhand der definierten Prüfkriterien.

Control ID	Anforderung	Ge-wicht	Erfüllung (P|A|I)[1]	Score (P|A|I)	Kurzbegründung (P | A | I)
D2.1	Erzwingung starker, gerätegebundener Authentifizierung	3	2 | 2 | 2	6 | 6 | 6	Starke Authentifizierung im Workflow zwingend definiert | Session-Erzeugung nur nach WebAuthn-Verifikation | Passwort + QR + Smartphone zwingend erforderlich
D2.2	Nicht-Optionalität sicherheitskritischer Mechanismen	3	2 | 2 | 2	6 | 6 | 6	Passkey als verpflichtender Designbestandteil | Keine Bypass-API oder Konfigurationsoption | Kein Login ohne vollständigen Auth-Flow
D2.3	Session-Erzeugung nur nach kryptographischem Nachweis	3	2 | 2 | 2	6 | 6 | 6	Zugriff erst nach abgeschlossener Auth vorgesehen | Token-Generierung an Challenge-Verifikation gebunden | Kein Zugriff ohne aktive Smartphone-Bestätigung
D2.4	Technische Begrenzung serverseitiger Zugriffsmacht	2	2 | 1 | 2	4 | 2 | 4	Zugriff im Workflow an Auth gekoppelt | Architektur konzeptionell begrenzt, aber nicht vollständig extern auditiert | Kein Admin-Bypass implementiert
D2.5	Nachvollziehbarkeit sicherheitskritischer Auth-Ereignisse	2	1 | 2 | 2	2 | 4 | 4	Nachweisbarkeit normativ gefordert | Strukturierte Logging-Mechanismen vorgesehen | Fehlversuche und Status klar erkennbar

Σ Kategorie gesamt: 68 / 78 | Gewichteter Erfüllungsgrad: 87 %

Bewertungsskala:
Erfüllungsgrad: 0 = nicht erfüllt · 1 = teilweise · 2 = erfüllt | Score = Gewicht × Erfüllungsgrad

Evaluierende Person: Peter Wildhaber | Rolle: Forschender | Datum: 14. Februar 2026

[1] P = Prozess, A = Architektur, I = Instanz