

Jan M. Schürmann, Jana Sedlakova, Manuel Trachsel

Ethische Aspekte zur KI-Anwendung in der Psychotherapie



Psychotherapie-Wissenschaft

16. Jahrgang, Nr. 2, 2026, Seite 17–27

DOI: 10.30820/1664-9583-2026-2-17

Psychosozial-Verlag

Impressum

Psychotherapie-Wissenschaft

eISSN 1664-9591

16. Jahrgang Heft 2/2026

<https://doi.org/10.30820/1664-9583-2026-2>

info@psychotherapie-wissenschaft.info

www.psychotherapie-wissenschaft.info

Herausgeber

Herausgeberverein der Zeitschrift

Psychotherapie-Wissenschaft

Peter Schulthess

Zanggerweg 33

CH-8006 Zürich

Mobile + 41 76 559 19 20

E-Mail: peter.schulthess@psychotherapie-wissenschaft.info

Redaktion

Mara Foppoli, Lugano

Lea-Sophie Richter, Zürich

Mario Schlegel, Zürich

Peter Schulthess, Zürich

Nataliya Tereshchenko, Zürich/Wien

Hinweise für AutorInnen befinden
sich auf der Homepage der Zeitschrift:
www.psychotherapie-wissenschaft.info

Verlag

Psychosozial-Verlag GmbH & Co. KG

Walltorstr. 10

D-35390 Gießen

+49 6421 96 99 78 26

info@psychosozial-verlag.de

www.psychosozial-verlag.de

Anzeigen

Anfragen zu Anzeigen bitte an den Verlag:

anzeigen@psychosozial-verlag.de

Es gelten die Preise der auf www.psychosozial-verlag.de
einsehbaren Mediadaten.

Gönnermitglieder des Herausgebervereins

wenden sich bitte direkt an die Redaktion:

redaktion@psychotherapie-wissenschaft.info

Datenbanken

Die Zeitschrift Psychotherapie-Wissenschaft wird regelmäßig
in der Abstract-Datenbank PsycInfo der American Psycho-
logical Association (APA), im Directory of Open Journals
(DOAJ) und in den Publikationsdatenbanken PSYINDEX und
PsychArchives des Leibniz-Institut für Psychologie/Leibniz
Institute for Psychology (ZPID) erfasst.



Die Beiträge dieser Zeitschrift sind unter der Creative Commons
Attribution-NonCommercial-NoDerivs 3.0 DE Lizenz lizenziert.
Diese Lizenz erlaubt die private Nutzung und unveränderte
Weitergabe, verbietet jedoch die Bearbeitung und kommerzielle
Nutzung. Weitere Informationen finden Sie unter:
creativecommons.org/licenses/by-nc-nd/3.0/de

Ethische Aspekte zur KI-Anwendung in der Psychotherapie

Jan M. Schürmann, Jana Sedlakova & Manuel Trachsel

Psychotherapie-Wissenschaft 16 (2) 2026 17–27

www.psychotherapie-wissenschaft.info

CC BY-NC-ND

<https://doi.org/10.30820/1664-9583-2026-2-17>

Zusammenfassung: Künstliche Intelligenz (KI) etabliert sich zunehmend auch in der psychiatrischen und psychotherapeutischen Versorgung. Vor dem Hintergrund globaler Versorgungslücken werden KI-basierte Anwendungen in den Bereichen Screening und Diagnostik, klinische Entscheidungsunterstützung, Intervention, Monitoring sowie administrative und edukative Unterstützung entwickelt oder bereits eingesetzt. Der vorliegende Beitrag analysiert Chancen und Risiken dieser Entwicklung aus medizinethischer Perspektive. Potenziellen Chancen wie verbesserter Versorgung, Entstigmatisierung, klinischer Entscheidungsunterstützung sowie Personalisierungs- und Ausbildungspotenzialen stehen potenzielle Risiken gegenüber: unzureichende Wirksamkeitsnachweise, inadäquates Krisenmanagement, Fehleranfälligkeit bis hin zu Halluzinationen, Erosion der therapeutischen Allianz, Datenschutzdefizite, algorithmische Verzerrungen sowie ein regulatorisches Vakuum. Vertieft werden vier ethische Grundsatzfragen diskutiert: 1. der normative Status von KI als Instrument oder Akteur, 2. die Bedingungen von Vertrauen in konversationale KI, 3. die Problematik simulierter Empathie sowie 4. das Risiko unrealistischer Erwartungen und übermässiger Abhängigkeit. Ergänzend werden aktuelle regulatorische Rahmenbedingungen – WHO-Leitprinzipien, EU AI Act, EU Medical Device Regulation und fachgesellschaftliche Empfehlungen – eingeordnet. Auf dieser Grundlage werden handlungsleitende Empfehlungen für die psychotherapeutische Praxis abgeleitet, darunter eine systematische KI-Anamnese, transparente Aufklärung über technologische Limitationen, Wahrung professioneller Urteilskraft und menschlicher Kontrolle, strikte Evaluierung von Datenschutz und Datensicherheit, robuste Krisenprotokolle sowie die Etablierung von KI-Kompetenz. Die authentische und empathische therapeutische Beziehung bleibt das unverzichtbare Agens eines jeden erfolgreichen Therapieprozesses. KI kann zunehmend zu einem ergänzenden Instrument werden, jedoch nicht zu einem Ersatz für die Behandlung durch menschliche Fachpersonen.

Schlüsselwörter: Künstliche Intelligenz, Psychotherapie, Chatbots, Therapeutische Beziehung, Ethik

Psychische Erkrankungen zählen weltweit zu den grössten Herausforderungen für die öffentliche Gesundheit. Laut WHO sind rund eine Milliarde Menschen von psychischen Störungen betroffen (Mikaeili et al., 2025). In einkommensschwachen Ländern erreicht der sog. Treatment Gap, also der Anteil psychisch erkrankter Menschen ohne angemessene Behandlung, bis zu 85 % (Rezaei et al., 2026). Auch in Europa gibt es je nach psychischer Erkrankung grosse Versorgungslücken: So erhalten lediglich 9,5–36,1 % der Personen mit einer schweren depressiven Störung und 7,3–47,1 % der Personen mit einer Angststörung eine adäquate Behandlung (Barbui et al., 2025). Vor diesem Hintergrund rücken Anwendungen der künstlichen Intelligenz (KI) in der psychiatrischen und psychotherapeutischen Versorgung zunehmend in den Fokus, um Versorgungslücken zu schliessen und therapeutische Prozesse zu unterstützen. Ihre Nutzung hat in den vergangenen Jahren erheblich zugenommen. Besonders ausgeprägt ist diese Dynamik bei KI-Chatbots zur Verbesserung der psychischen Gesundheit, wie bspw. Wysa oder Woebot. Eine Übersichtsstudie von Laymouna et al. (2024), in die 161 Primärstudien zu Chatbots in der Gesundheitsversorgung einflossen, identifizierte die psychische Gesundheitsunterstützung mit 46 Studien (28,6 %) als häufigste Anwendungsdomäne. Auch auf

Nutzerseite sind klare Trends erkennbar: Mehr als zwei Drittel junger Menschen geben an, Erfahrungen mit mHealth-Apps gemacht zu haben, und 60 % nutzen diese regelmässig (Götzl et al., 2022). In den USA gaben in einer Umfrage 35,5 % der 18- bis 49-Jährigen an, mindestens einmal pro Woche Chatbots zur Verbesserung der psychischen Gesundheit zu konsultieren (Ueda et al., 2025). In Grossbritannien wird der Chatbot Limbic eigenen Aussagen zufolge bereits bei 45 % aller Psychotherapien des nationalen Gesundheitssystems NHS als ergänzendes Angebot genutzt (Limbic, 2026). Die KI-basierte Unterstützung der psychischen Gesundheit etabliert sich rasant als eigenes Anwendungsfeld, dessen klinische und ethische Implikationen allerdings erst in Ansätzen untersucht sind.

Im Kontext des vorliegenden Artikels umfassen KI-basierte Anwendungen in der Psychiatrie und Psychotherapie ein breites Spektrum an Technologien – von Machine-Learning(ML)-basierten Anwendungen über Natural-Language-Processing(NLP)-gestützte Konversationsagenten (sog. konversationale KI) bis hin zu generativen Large-Language-Models (LLMs). Diese werden entweder autonom oder ergänzend zu menschlichen Therapeut:innen eingesetzt, ein Ansatz, der als «Blended Therapy» bezeichnet wird. Diese sind von der klassischen Online-Psychotherapie abzugrenzen,

bei der Therapeut:innen videobasierte oder asynchrone Kommunikationskanäle für eine konventionelle Psychotherapie nutzen. Ebenso wenig sind diese Anwendungen mit KI-Anwendungen in der somatischen Medizin gleichzusetzen. Psychiatrische und psychotherapeutische KI-Systeme operieren in einem spezifischen klinischen und ethischen Kontext, der durch die Vulnerabilität der Patient:innen, die zentrale Bedeutung der therapeutischen Beziehung und die Komplexität psychischer Phänomene geprägt ist (Fiske et al., 2019; Stade et al., 2024). Für die klinische, ethische und regulatorische Beurteilung ist zudem zwischen verschiedenen Systemtypen zu unterscheiden: 1. als Medizinprodukte zertifizierte KI-Anwendungen; 2. spezifisch für die psychische Gesundheit entwickelte, aber nicht als Medizinprodukt zertifizierte Chatbots (z. B. Wysa, Woebot); 3. generische generative LLMs ohne klinische Zweckbestimmung und Validierung (z. B. ChatGPT, Claude, Gemini); 4. administrative und edukative Assistenzsysteme ohne unmittelbar therapeutische Funktion (z. B. Ambient Scribes). Chancen und Risiken sind je nach Systemtyp unterschiedlich gelagert und werden im Folgenden, wo möglich, dem jeweiligen Systemtyp zugeordnet.

Die gegenwärtigen und zukünftigen Anwendungsbereiche von KI in der Psychiatrie und Psychotherapie lassen sich anhand von Behandlungspfaden in fünf Kategorien unterteilen (vgl. auch Gooding, 2024):

Screening und Diagnostik: KI-Anwendungen können multimodale Daten – darunter Sprache, Text, Stimme und biometrische Biosensordaten – analysieren, um psychische Störungen frühzeitig zu erkennen und zu klassifizieren (Mikaeili et al., 2025). Ein Anwendungsbeispiel sind Modelle zur Vorhersage des psychosozialen Funktionsniveaus bei Patient:innen mit Schizophrenie und anderen psychotischen Erkrankungen, die inzwischen moderate bis gute prädiktive Genauigkeiten erreichen (Mok et al., 2025; Smith et al., 2026).

Entscheidungshilfen und Behandlungsplanung: KI-Anwendungen können Behandlungsentscheidungen unterstützen. Perlis et al. (2024) beschreiben eine klinische Entscheidungshilfe für die Pharmakotherapie bipolarer Störungen, die mit einem LLM (GPT-4) und leitlinien-gestützten Prompts arbeitet. Das Modell identifizierte die von Expert:innen als optimal eingestufte Pharmakotherapie in rund der Hälfte der Fallvignetten. Die Systeme DISCOVER und NOVA integrieren multimodale Verhaltensmerkmale, die aus Therapiegesprächen extrahiert wurden – wie Mimik, Körperhaltung und Stimmparameter – mit LLM-basierten Auswertungen der verbalen Interaktion, um die Therapieplanung zu unterstützen (André et al., 2026).

Intervention und Therapie: Die direkteste, aber auch ethisch umstrittenste Anwendungsform ist der Einsatz von KI-Chatbots als therapeutische Akteure, d. h. zur direkten Verbesserung psychischer Symptome. Übersichtsarbeiten legen nahe, dass mithilfe KI-basierter Chatbots sowohl depressive Symptome als auch psychischer Distress statistisch und klinisch signifikant reduziert werden können (Li et al., 2023; Farzan et al., 2025; Rezaei et al., 2026). Auch

für die generalisierte Angststörung sind klinisch bedeutsame Verbesserungen durch den Einsatz von Chatbots nachgewiesen (Farzan et al., 2025). Diese Befunde sind allerdings differenziert einzuordnen: Die Evidenz stützt sich überwiegend auf Studien mit kleinen Stichproben, kurzen Follow-up-Zeiträumen und Selbstberichte und betrifft primär leichte bis mittelgradige depressive und Angstsymptomatik (Li et al., 2023; Bucher et al., 2025). Erste grössere randomisierte kontrollierte Studien wie diejenige zum generativen Chatbot Therabot zeigen zwar signifikante Symptomreduktionen bei Depression, generalisierter Angststörung und erhöhtem Essstörungsrisiko (Heinz et al., 2025); robuste Langzeitdaten fehlen jedoch. Bei bestimmten psychischen Erkrankungen wie Schizophrenie und anderen psychotischen Störungen, bei Persönlichkeitsstörungen, akuter Suizidalität sowie komplexen Traumafolgestörungen sind Chatbots hingegen kaum erforscht und ethisch besonders umstritten (Mok et al., 2025; Bucher et al., 2025). Die therapeutische Wirksamkeit von KI-Chatbots darf daher nicht pauschal, sondern muss störungs- und schweregradspezifisch beurteilt werden.

Monitoring und Verlaufsbeobachtung: KI-Anwendungen ermöglichen eine kontinuierliche, passive Überwachung psychischer Gesundheitsparameter im Alltag der Patient:innen. Mithilfe des sog. Digital Phenotyping, also der fortlaufenden Erfassung von Verhaltensdaten mittels Smartphones und Wearables, können Stimmungsveränderungen, Schlafmuster und soziale Interaktionsmuster in Echtzeit analysiert werden, um eine Verschlechterung der Symptome frühzeitig zu antizipieren (Mikaeili et al., 2025). Darüber hinaus werden KI-gestützte Anwendungen zunehmend als Bindeglied zwischen Therapiesitzungen genutzt, etwa durch Selbstreflexionsaufgaben und personalisiertes Feedback zu therapeutischen Aktivitäten (Habicht et al., 2025).

Administrative und edukative Unterstützung: KI bietet ein erhebliches Potenzial zur administrativen Entlastung von Behandelnden und zur Förderung der Gesundheitskompetenz von Patient:innen. Sog. Ambient Scribes transkribieren Therapiegespräche automatisch und legen sie in elektronischen Patientenakten ab (Laymouna et al., 2024). Im Bereich der Psychoedukation können KI-Anwendungen komplexe medizinische Sachverhalte in verständliche Laiensprache übersetzen. Eine Übersichtsarbeit zeigt, dass generative KI-Systeme Gesundheitsinformationen mehrheitlich korrekt und verständlich bereitstellen, jedoch Risiken der Vereinfachung, Übergeneralisierung und fehlender Transparenz bestehen (Pal et al., 2025).

Die Breite und Dynamik dieser Anwendungen können Kommunikations-, Entscheidungs- und Handlungsabläufe in Psychiatrie und Psychotherapie grundlegend verändern. Damit sind sowohl Chancen als auch erhebliche ethische Herausforderungen verbunden. Der vorliegende Artikel gibt einen Überblick über diese ethischen Aspekte. Grundlage bildet eine narrative Literaturrecherche in PubMed/MEDLINE (Publikationszeitraum 2019 bis Anfang 2026), bei der Suchbegriffe zu KI («artificial intelligence», «machine learning», «chatbot», «conversational agent»,

«large language model») mit Begriffen zur Psychotherapie und psychischen Gesundheitsversorgung («psychotherapy», «psychological treatment», «mental health intervention», «psychiatry») sowie zur Ethik («ethics», «ethical issues») kombiniert wurden. Ergänzend erfolgte eine unsystematische Schneeballsuche in den Referenzlisten eingeschlossener Arbeiten sowie eine gezielte Sichtung von Leitlinien und Positionspapieren internationaler Organisationen und Fachgesellschaften. Eingeschlossen wurden vorrangig Originalarbeiten mit Peer-Review und Übersichtsarbeiten mit Bezug zu klinischen oder ethischen Aspekten von KI in der Psychotherapie. Die Auswahl erfolgte kriteriengeleitet nach Relevanz für die ethische Fragestellung, ohne den Anspruch der Vollständigkeit eines systematischen Reviews zu erheben. Auf dieser Basis werden zunächst Chancen und Risiken zusammengefasst. Anschliessend werden vier ethische Grundsatzfragen vertieft: der Status von KI als Instrument oder Akteur, das Vertrauen in konversationale KI, die simulierte Empathie sowie unrealistische Erwartungen und Abhängigkeiten. Danach werden die rechtlichen Grundlagen und ethischen Richtlinien dargestellt und abschliessend Empfehlungen für die psychotherapeutische Praxis abgeleitet.

Chancen und Risiken

Die Integration von KI birgt ein erhebliches Potenzial für die psychotherapeutische Versorgung, sowohl in der Prävention, Diagnostik, Behandlung als auch in der Nachsorge (Opel & Breakspear, 2026). Aus medizin-ethischer Perspektive sind damit eine Reihe von Chancen und Risiken verbunden, die hier tabellarisch dargestellt werden (s. Tab. 1 & 2).

Vertiefung zentraler ethischer Grundsatzfragen

KI als Instrument oder Akteur?

Eine erste ethische Grundsatzfrage betrifft den Status von KI im Zusammenhang mit Psychotherapie: Handelt es sich bei der konversationalen KI um ein Instrument oder um einen Akteur (Sedlakova & Trachsel, 2023)? Diese Unterscheidung ist normativ entscheidend, da sich die Bewertungsmaßstäbe jeweils deutlich unterscheiden. Ein «gutes» Instrument erfüllt Kriterien wie Zuverlässigkeit, Sicherheit und Zweckmässigkeit, während ein «guter» Akteur, insbesondere im Kontext der Psychotherapie, an Eigenschaften wie Verantwortlichkeit, moralischer Urteilsfähigkeit und der Fähigkeit zur Beziehungsgestaltung gemessen wird. Ein Grossteil der bestehenden Literatur und regulatorischen Leitlinien behandelt KI primär als Instrument und fokussiert sich entsprechend auf Prinzipien wie Datenschutz, Datensicherheit und Transparenz. Diese Perspektive greift jedoch zu kurz, wenn man die tatsächliche Einbettung und Verwendung solcher Systeme im therapeutischen Kontext betrachtet.

Konversationale KI wird so programmiert, beschrieben und folglich wahrgenommen, dass sie einem menschlichen Akteur ähnelt. Systeme werden als «empathisch», «unterstützend» oder «beziehungsfähig» charakterisiert. Allerdings ist es fraglich, ob KI-Systeme jemals in der Lage sein werden, «echte Empathie» zu empfinden, eine langfristige therapeutische Interaktion aufrechtzuerhalten, kontextuelle Nuancen angemessen zu berücksichtigen und grundlegende klinische Kompetenzen wie z. B. das angemessene und feinfühliges Infragestellen von Patient:innen-Aussagen zu gewährleisten (Stade et al., 2024). Genau hier entsteht eine normative Lücke: Während die Zuschreibung solcher Eigenschaften Erwartungen erzeugt, kann KI die damit verbundenen ethischen und professionellen Anforderungen nicht erfüllen (Sedlakova & Trachsel, 2023; Ferrario et al., 2024). Psychotherapeut:innen sind an berufsethische Werte, Tugenden und Verantwortlichkeitsstrukturen gebunden, die ihre Praxis normativ rahmen. KI-Systeme verfügen hingegen zumindest bisher weder über moralische Handlungsfähigkeit noch damit verbundene moralische Verantwortung. Sie können nicht in gleicher Weise für ihr Handeln zur Rechenschaft gezogen werden wie menschliche Fachpersonen (Stade et al., 2024). Die Gefahr besteht also darin, dass durch die menschenähnliche Gestaltung der KI normative Erwartungen erzeugt werden, die das System nicht erfüllen kann. Dies ist ein Spannungsfeld, das im Zentrum der ethischen Analyse stehen muss.

Alternativ kann die Rolle von KI in der Psychotherapie als ein Kontinuum verstanden werden, das von assistierenden über kollaborative bis hin zu vollständig autonomen Systemen reicht (ebd.). Dabei handelt es sich nicht um klar getrennte Kategorien, sondern um ein graduelles Entwicklungsmodell, in dem sich Funktionen, Verantwortlichkeiten und Risiken schrittweise verschieben. Mit zunehmender Autonomie verändern sich nicht nur die technischen Fähigkeiten, sondern auch die normativen Anforderungen. Ein solches Modell hilft zudem, über KI im Kontext konkreter klinischer Anwendungen und Einzelfälle nachzudenken und die damit verbundenen Erwartungen, Risiken und Anforderungen differenziert zu betrachten.

Kann man KI vertrauen?

Die Frage nach Vertrauen ist im Kontext von KI in der Psychotherapie zentral, da therapeutische Prozesse wesentlich auf epistemischen und relationalen Vertrauensstrukturen beruhen. Diese ermöglichen sowohl die Annahme von Information als auch die Stabilität der therapeutischen Beziehung. Kann und soll man KI vertrauen? Wenn KI in der Psychotherapie primär als Instrument verstanden wird, verschiebt sich der normative Fokus auf Zuverlässigkeit. In diesem Rahmen wird KI nicht als Interaktionspartner, sondern als technisches System bewertet, dessen Zuverlässigkeit sich daran misst, ob es definierte Aufgaben konsistent und fehlerarm erfüllt. Vertrauen im eigentlichen Sinne ist dabei irrelevant,

Bereich	Chance	Systemtyp	Quellen
Verbesserung des Zugangs	Orts- und zeitunabhängige Versorgung rund um die Uhr, besonders für Menschen in ländlichen Regionen oder mit eingeschränkter Mobilität	2, 3	Fiske et al., 2019; Manole et al., 2024
	Asynchrone, automatisierte Interventionen können Versorgungslücken schliessen und eine niederschwellige Grund- oder Überbrückungsversorgung sicherstellen.	1, 2	Zhang & Wang, 2024
	Nach der Entwicklung können ohne zusätzliche Personalkosten viele Patient:innen betreut und finanzielle Hürden gesenkt werden.	1, 2, 3	Manole et al., 2024
Entstigmatisierung und offene Kommunikation	Ein anonymer, nicht wertender Kommunikationsraum kann die Angst vor Stigmatisierung und Ablehnung reduzieren.	2, 3	Fiske et al., 2019; Khawaja & Bélisle-Pipon, 2023
	Patient:innen berichten gegenüber einem als vorurteilsfrei wahrgenommenen Gegenüber offener über Scham, Emotionen oder suizidale Gedanken.	2, 3	Lucas et al., 2014; Zhang & Wang, 2024
Klinische Entscheidungsunterstützung	Diagnostische Assistenz durch Analyse multimodaler Daten (Text, Sprache, Mimik) zur Identifikation von Frühsymptomen und zur Erhöhung der differenzialdiagnostischen Genauigkeit	1	Obradovich et al., 2024; Ali et al., 2025
	Prädiktive Analytik über passive Sensordaten (Digital Phenotyping) antizipiert Behandlungsansprechen, Rückfälle oder akute suizidale Krisen und ermöglicht rechtzeitige Interventionen	1	Mikaeili et al., 2025; Cruz-Gonzalez et al., 2025
Administrative Entlastung	Ambient Scribes und generative KI transkribieren, fassen zusammen und dokumentieren in elektronischen Akten. Das entlastet von repetitiven Aufgaben und schafft Zeit für die Beziehung.	4	Gülegen et al., 2026; Nehme et al., 2024
Optimierung und Personalisierung des Therapieprozesses	Signifikante Reduktion depressiver Symptome, generalisierter Angst und von Essstörungen durch Chatbots und Blended Therapy, in Reviews und einem RCT belegt	1, 2	Li et al., 2023; Nunes-Zlotkowski et al., 2024; Heinz et al., 2025
	Personalisierung der Behandlungswahl (z. B. Antidepressiva, Therapiemethode) anhand biologischer, genetischer und lebensstilbezogener Profile mittels maschinellen Lernens	1	Cecil et al., 2026; Ali et al., 2025
	Unterstützung zwischen den Sitzungen durch Homework-Tracking, Erinnerungen und Echtzeit-Feedback steigert Adhärenz und Engagement und senkt Drop-out-Raten.	2	Habicht et al., 2025; Soma et al., 2026
	Psychoedukation in verständlicher Sprache stärkt das Symptomverständnis, die Autonomie und das Shared Decision-Making.	2, 3, 4	Pal et al., 2025
Forschung, Qualitätssicherung und Ausbildung	Automatisierte Auswertung grosser Mengen an Sitzungstranskripten ermöglicht eine objektive Qualitätssicherung in grossem Umfang (z. B. Empathie, Therapietreue, Allianz).	3, 4	Soma et al., 2026; Orrù & Mannarini, 2026
	Generative Modelle und Avatare als simulierte Patient:innen erlauben angehenden Therapeut:innen das Üben in sicherer Umgebung mit sofortigem, datenbasiertem Feedback.	3, 4	Opel & Breakspear, 2026; Soma et al., 2026

Tab. 1: Chancen von KI in der Psychotherapie [Legende zum Systemtyp: 1 = als Medizinprodukt zertifizierte KI-Anwendung; 2 = spezifisch für die psychische Gesundheit entwickelter, nicht als Medizinprodukt zertifizierter Chatbot; 3 = generisches generatives LLM; 4 = administratives oder edukatives Assistenzsystem]

Bereich	Risiko	Systemtyp	Quellen
Klinische Sicherheitsrisiken und qualitative Defizite	Unzureichender Wirksamkeitsnachweis: grosse longitudinale RCTs sind selten. Viele Studien beruhen auf kleinen Stichproben und kurzen Follow-ups, mit Evidenz v. a. für leichte bis mittlere Depression und Angst, kaum für andere Störungen.	2, 3	Bucher et al., 2025; Algumaei et al., 2025; Orrù & Mannarini, 2026
	Inadäquates Krisenmanagement und fehlende Triage: In Notfällen wie akuter Suizidalität oder Psychose geben Chatbots teils verharmlosende oder schädliche Ratschläge statt einer Weiterleitung an Krisendienste.	2, 3	Grabb et al., 2024; Ohu et al., 2025
	Halluzinationen: Generative Systeme erzeugen plausibel klingende, aber faktisch falsche Informationen, die zu Fehlentscheidungen führen können.	2, 3	Frances, 2025
	Unkritische Validierung (Sykophantie): Nutzerkonformes Verhalten kann pathologische Denkmuster, Wahnvorstellungen oder selbstschädigendes Verhalten bestärken, statt sie zu hinterfragen.	2, 3	Morrin et al., 2026
	Fehlen nonverbaler und kontextueller Diagnostik (Mimik, Tonfall, Körperhaltung) reduziert die ganzheitliche Beurteilung.	1, 2, 3	Manole et al., 2024; Ali et al., 2025
Erosion der therapeutischen Allianz	Mangel an genuiner Empathie und Menschlichkeit untergräbt die Authentizität des Prozesses und kann zu einer Depersonalisierung der Versorgung führen.	2, 3	Sedlakova et al., 2025a; De Choudhury et al., 2023
	Anthropomorphes Design und irreführendes Marketing verleiten vulnerable Nutzer:innen zur Annahme einer reziproken, fürsorglichen Bindung (sog. therapeutisches Missverständnis).	2, 3	Khawaja & Bélisle-Pipon, 2023; Rahsepar Meadi et al., 2025
	Die stets verfügbare, nie urteilende Interaktion begünstigt emotionale Abhängigkeit und Vermeidungsverhalten.	2, 3	Frances, 2025; Zhang et al., 2025
Datenschutz und Datensicherheit	Kommerzielle Chatbots unterliegen oft keiner Schweigepflicht. Sensible Daten können extrahiert, kommerzialisiert oder verkauft werden.	2, 3	Martinez-Martin & Kreitmair, 2018; Solaiman, 2024
	Prekäre informierte Einwilligung: Die Black-Box-Natur vieler Algorithmen erschwert fundierte Entscheidungen über die Datenverarbeitung.	1–4	Le Glaz et al., 2021; Rezaei et al., 2026
	Datenlecks aggregierter psychiatrischer Profile bergen Potenzial für Stigmatisierung, berufliche Nachteile oder Erpressung.	1–4	Ali et al., 2025
Algorithmische Verzerrung und Ungerechtigkeit	Verzerrte Trainingsdaten führen zu ungenauen oder diskriminierenden Ergebnissen für ethnische Minderheiten und marginalisierte Gruppen.	1, 2, 3	Knox et al., 2023
	Standardisierte, westlich geprägte Systeme deuten kulturspezifische Ausdrucksformen psychischen Leids falsch oder pathologisieren sie.	1, 2, 3	Ohu et al., 2025
	Epistemische Ungerechtigkeit: testimoniale und hermeneutische Benachteiligung, wenn bestimmten Gruppen Glaubwürdigkeit abgesprochen wird	2, 3	Rahsepar Meadi et al., 2025
	Digitale Kluft: Hürden bei Hardware und digitaler Kompetenz erschweren vulnerablen Gruppen den Zugang und verstärken bestehende Ungleichheiten.	1–4	Tekin & Delehanty, 2025
Regulatorische Lücken und Verantwortungsdiffusion	Fehlende spezifische und verbindliche Regulierung generativer KI in der psychischen Gesundheitsversorgung	2, 3	Solaiman, 2024
	Haftungslücken bei iatrogenen Schäden: Verantwortung bleibt zwischen Entwickler:innen, Institutionen und Behandelnden ungeklärt.	1–4	Manríquez Roa et al., 2021; Rezaei et al., 2026
	Umgehung von Medizinproduktegesetzen, indem Anwendungen trotz therapeutischer Nutzung als «Wellness-Tools» deklariert werden.	2	Ohu et al., 2025

Tab. 2: Risiken von KI in der Psychotherapie [Legende zum Systemtyp: 1 = als Medizinprodukt zertifizierte KI-Anwendung; 2 = spezifisch für die psychische Gesundheit entwickelter, nicht als Medizinprodukt zertifizierter Chatbot; 3 = generisches generatives LLM; 4 = administratives oder edukatives Assistenzsystem]

da diesem Verständnis zufolge weder Emotionen noch Intentionalität oder Verantwortlichkeit aufseiten des Systems vorausgesetzt werden können und somit auch keine Grundlage für eine relationale Vertrauensbeziehung besteht (Hatherley, 2020; Ryan, 2020). Sobald KI jedoch als therapeutisch relevante Instanz oder zumindest als menschenähnlicher Interaktionspartner konzeptualisiert wird, verändert sich die normative Logik grundlegend: An die Stelle von Zuverlässigkeit tritt die Kategorie des Vertrauens. Damit wird KI nicht mehr nur als funktionales Instrument beurteilt, sondern als etwas, dem man im weiteren Sinne «zutrauen» muss, im therapeutischen Prozess relevante, kontextsensible und potenziell folgenreiche Informationen zu verarbeiten.

Sedlakova et al. (2025b) schlagen vor, Vertrauen in konversationale KI stets kontextbezogen zu analysieren. Dabei soll untersucht werden, wie sich die jeweiligen normativen Bedingungen verändern, unter denen Vertrauen überhaupt sinnvoll zugeschrieben werden kann, und wie sich dadurch auch der Begriff des Vertrauens selbst verschiebt. Diese nuancierte Herangehensweise ist zentral, da KI durch ihre dialogische Struktur und ihre Fähigkeit, menschliche Interaktion zu simulieren, systematisch als menschenähnliches Gegenüber erscheint und dadurch vertrauensbezogene Erwartungen auslöst. Im Kontext eines Informationsanbieters kann Vertrauen in KI bspw. relativ unproblematisch als eine eingeschränkte Form von Vertrauen verstanden werden, da es hier v. a. um eine Form der Bewertung von menschenähnlicher Leistungsfähigkeit und Zuverlässigkeit handelt, solange das System konsistent funktioniert und keine systematischen Fehler oder Bias produziert. Wird KI hingegen im Kontext psychotherapeutischer Praxis betrachtet, wird Vertrauen deutlich problematischer, da es sich auf stark normative und relationale Anforderungen bezieht, die auf klinisches Urteilsvermögen, Verantwortlichkeit und die Möglichkeit echter therapeutischer Beziehung verweisen. Aufgrund der grundlegenden normativen Defizite schlagen Sedlakova et al. vor, konversationale KI eher als eine Art fiktionale Figur zu konzeptualisieren. Diese Perspektive erlaubt es, die positiven Effekte der menschenähnlichen Interaktion und der neuen Form des Vertrauens zu erhalten, während zugleich eine epistemische und normative Distanz gewahrt bleibt, die verhindert, dass KI als vollwertiger therapeutischer oder menschlicher Akteur missverstanden wird.

Simulierte Empathie von KI

Es gibt den Trend, KI mit menschenähnlichen Eigenschaften zu entwickeln, die Empathie simulieren, um therapeutische Beziehungen zu erleichtern. Wir bezeichnen dies als «Humanisierung von KI» und vertreten die Ansicht, dass damit konzeptionelle, praktische und ethische Probleme verbunden sind (ebd.). Die ersten beiden Probleme betreffen hauptsächlich die Entstehung unrealistischer Erwartungen. Das dritte Problem besteht darin, dass KI die in der psychischen Gesundheitsversorgung verankerten ethischen Anforderungen nicht erfüllt. Die Humanisie-

rung von KI kann dazu führen, dass Klient:innen mit ihr so interagieren, als wäre sie eine menschliche Fachperson, und ihr spezifisch menschliche Attribute zuschreiben, wie bspw. gut zuzuhören oder empathisch zu sein.

Kritiker:innen warnen schon länger vor einer «empathischen KI» (z.B. Montemayor et al., 2022). Sie argumentieren, dass KI kein moralischer Akteur ist, keine mentalen Zustände – einschliesslich Emotionen – besitzt und auch nicht über andere Fähigkeiten verfügt, die Empathie ermöglichen, wie bspw. Verständnis. Diese Kritik beruht auf der Annahme, dass menschliche Attribute in interpersonellen Beziehungen (z.B. Empathie) voraussetzen, dass ein Akteur über bestimmte mentale Zustände, spezifische Motivationen und Fähigkeiten sowie moralische Handlungsfähigkeit verfügt. Da KI-Systeme diese Zustände und Fähigkeiten nicht aufweisen, können ihnen Attribute wie Empathie nicht zugeschrieben werden (Sedlakova et al., 2025a). Eine unangemessene Zuschreibung birgt Risiken wie Täuschung, Manipulation, Abhängigkeit und den Verlust von Menschlichkeit im psychotherapeutischen Prozess (z.B. Montemayor et al., 2022; Sedlakova & Trachsel, 2023).

Konzeptuell lassen sich dabei drei Ebenen unterscheiden: empathische Performanz als beobachtbares sprachliches Verhalten, das von Nutzer:innen als empathisch erlebt wird; empathische Simulation als technische Modellierung empathischer Kommunikationsmuster ohne inneres Erleben; und genuine Empathie, die etwa affektive Beteiligung, Verständnis und Intentionalität voraussetzt. In empirischen Studien ist die empathische Performanz aktueller Systeme beachtlich: In einer Studie von Hatch und Kollegen (2025) konnten Lai:innen Antworten von ChatGPT auf paartherapeutische Fallvignetten kaum von Antworten erfahrener Therapeut:innen unterscheiden und bewerteten die KI-Antworten hinsichtlich zentraler Wirkfaktoren sogar höher. Gerade deshalb ist festzuhalten, dass die blossе Wirkung «empathischer KI» auf Patient:innen als ethische Rechtfertigung nicht genügt. Solange aufseiten des Systems weder affektive Beteiligung noch Intentionalität vorliegen, bleibt simulierte Empathie kategorial von genuiner Empathie unterschieden – mit den beschriebenen Risiken der Täuschung und Fehlattri-bution (Montemayor et al., 2022; Sedlakova et al., 2025a).

Unrealistische Erwartungen an die Beziehung zu einer KI und übermässige Abhängigkeit

Zudem können durch die Humanisierung von KI Erwartungen entstehen, die diese möglicherweise nicht erfüllen kann. Das therapeutische Gespräch ist eine hochkomplexe und persönliche Tätigkeit (Wampold & Flückiger, 2023), die in eine Reihe von professionellen Pflichten eingebettet ist, um eine sichere und hilfreiche Interaktion für Patient:innen zu gewährleisten (z.B. APA, 2017). Im Kontext der Psychotherapie können unerfüllte Bedürfnisse und Erwartungen zu negativen Gefühlen oder Enttäuschungen führen, die den psychischen Gesundheitszustand erkrankter Personen zusätzlich destabilisieren (Sedlakova et al., 2025b).

Die therapeutische Allianz ist ein empirisch robust belegter Wirkfaktor für positive Therapieergebnisse (s. z. B. Wampold & Flückiger, 2023). Sedlakova und Trachsel (i. D.) haben aufgezeigt, dass die erheblichen Einschränkungen von KI-Systemen in emotionaler und relationaler Hinsicht zu unrealistischen Erwartungen, übermässiger Abhängigkeit und potenziellem Schaden führen können. Da KI nicht dieselben kognitiven, emotionalen oder relationalen Prozesse durchläuft wie menschliche Therapeut:innen, kann sie auch die für die therapeutische Praxis und Allianz erforderlichen ethischen Standards nicht erfüllen. Einzelne unterstützende Funktionen – etwa Psychoedukation, Erinnerungen oder Adhärenzförderung – kann KI übernehmen; die relationalen Kernfunktionen wie etwa Empathie, Resonanz oder Verlässlichkeit setzen demgegenüber ein erlebendes, verantwortliches Gegenüber voraus und sind daher aus prinzipiellen, nicht bloss aus technischen Gründen nicht delegierbar. Um diesen Bedenken Rechnung zu tragen, schlagen Sedlakova und Trachsel vor, dialogorientierte KI nicht als Ersatz für menschliche Therapeut:innen, sondern als ergänzendes Instrument zu betrachten («Augmented Intelligence»). Sie schlagen ein angepasstes Modell vor, in dem KI bestimmte Aspekte der Therapie wie Psychoedukation und Therapieadhärenz unterstützt, während zentrale relationale Komponenten weiterhin in der Zuständigkeit menschlicher Fachpersonen liegen.

Rechtliche Grundlagen, ethische Richtlinien

Die rechtliche und medizinethische Regulierung von KI in Medizin und Psychiatrie ist ein komplexes und sich rasch entwickelndes Feld. Die Implementierung überholt die Etablierung kohärenter Rahmenbedingungen. Insbesondere die Haftung für Fehler und Schädigungen bleibt weitgehend unklar, sodass sich die Nutzung in einem regulatorischen Vakuum vollzieht (Ohu et al., 2025).

Richtlinien internationaler Gremien: Die WHO veröffentlichte 2021 sechs Leitprinzipien für KI in der Gesundheitsversorgung: Schutz der Autonomie, Förderung von Wohlergehen und Sicherheit, Transparenz und Erklärbarkeit, Verantwortung und Rechenschaftspflicht, Gerechtigkeit und Inklusion sowie Nachhaltigkeit (WHO, 2021). 2024 wurde die Richtlinie um grosse multimodale Modelle erweitert. Ein zentrales Postulat fordert prospektive randomisierte kontrollierte Studien, bevor klinische KI-Chatbots breitflächig zugelassen werden (WHO, 2024).

EU Artificial Intelligence Act: Mit dem Artificial Intelligence Act (Regulation (EU) 2024/1689, 2024) erliess die EU das weltweit erste umfassende KI-Gesetz. Es ordnet jede Anwendung einer von vier Risikostufen zu: KI, die als Medizinprodukt oder dessen Komponente fungiert, gilt als hochrisikoreich und unterliegt Konformitätsbewertung, Anforderungen an Daten- und Trainingsqualität, technischer Dokumentation, Transparenz, menschlicher Aufsicht sowie Post-Market-Surveillance.

EU Medical Device Regulation: Nach der Medical Device Regulation gelten Chatbots als Medizinprodukte, sobald sie der Diagnose, Prävention, Überwachung oder Behandlung dienen, und müssen dann strenge Zertifizierung, klinische Bewertung und Qualitätskontrollen durchlaufen (Nehme et al., 2024). Viele Anbieter umgehen dies, indem sie ihre Anwendungen als «Wellness-Tools» deklarieren, sodass vulnerable Patient:innen Systeme nutzen, die nicht für den klinischen Einsatz validiert wurden (Ohu et al., 2025).

Empfehlungen von Fachgesellschaften: Auch Fachgesellschaften haben Empfehlungen vorgelegt, etwa die Bundesärztekammer (2025) und die FMH (2022). Am detailliertesten sind die Richtlinien der APA (2025a, b): KI darf nur unter menschlicher Verantwortung und mit transparenter Einwilligung eingesetzt werden, Datenschutz und wissenschaftliche Prüfung müssen gewährleistet sein, Diskriminierung ist zu vermeiden und die KI-Kompetenz der Fachpersonen sicherzustellen. Die ergänzende Health Advisory warnt davor, generative Chatbots als Ersatz für professionelle Therapie zu nutzen, da sie dafür unzureichend geprüft und besonders für Kinder und vulnerable Personen riskant sind (APA, 2025b).

Berufspraktische Konsequenzen: Aus diesen Grundlagen lassen sich vorläufige berufspraktische Entscheidungsregeln ableiten: Aktiv empfehlen sollten Psychotherapeut:innen nur Anwendungen, die als Medizinprodukt zertifiziert sind oder deren Wirksamkeit und Sicherheit – einschliesslich validierter Krisen- und Eskalationsmechanismen – nachgewiesen ist. Nicht zertifizierte «Wellness-Tools» sollten allenfalls mit expliziter Aufklärung über die fehlende klinische Validierung erwähnt, aber nicht als therapeutisches Angebot eingesetzt werden. Generische LLMs wie ChatGPT oder Claude sind nicht für therapeutische Zwecke bestimmt; ihre Nutzung durch Patient:innen sollte nicht empfohlen, wohl aber aktiv exploriert und therapeutisch eingeordnet werden (ebd.). Als datenschutzrechtliche Mindeststandards gelten die Konformität mit der Datenschutz-Grundverordnung bzw. dem schweizerischen Datenschutzgesetz, transparente Angaben zu Speicherort und Weiterverwendung der Daten – insbesondere keine Nutzung für das Modelltraining ohne Einwilligung – sowie eine verschlüsselte Datenübertragung und -speicherung (Nehme et al., 2024; Ohu et al., 2025). Beim Einsatz administrativer Systeme wie Ambient Scribes sind zudem Schweigepflicht und explizite Einwilligung der Patient:innen zu wahren.

Empfehlungen für die Psychotherapie: Klug mit KI umgehen

Auf Basis der Richtlinien der APA (2025a) sowie des aktuellen akademischen Diskurses lassen sich Empfehlungen für die psychotherapeutische Praxis formulieren (s. Tab. 3).

Empfehlung	Inhalt
Proaktive KI-Anamnese	Die Nutzung von KI-Chatbots wertfrei als integralen Teil der psychosozialen Anamnese erheben (ob, wie häufig, wozu) und potenziell schädliche Interaktionen frühzeitig in den Behandlungsplan einbeziehen
Aufklärung über Limitationen und Risiken	Patient:innen über algorithmische Verzerrungen, die Tendenz zu Halluzinationen und das Fehlen von genuinem klinischem Urteilsvermögen und Empathie aufklären
Transparenz und informierte Einwilligung	Vor dem Einsatz klinischer KI-Tools kultur- und sprachsensibel über Art, Zweck, Nutzen und Risiken aufklären und ein Opt-out mit adäquaten Alternativen ermöglichen
Professionelle Urteilskraft und menschliche Kontrolle	KI darf die Entscheidungsfindung nur unterstützen, nicht ersetzen. Klare Human-in-the-Loop-Kontrollpunkte sind zu etablieren. Die letztgültige Verantwortung verbleibt bei den Behandelnden.
Datenschutz und -sicherheit	Konformität mit Datenschutzgesetzen prüfen, transparent über Speicherung und Weiterverarbeitung informieren; bei unzureichender Cybersicherheit auf eine Nutzung verzichten
Risikoscreening und Monitoring bei Vulnerabilitäten	Bei Angststörungen, Wahnsymptomatik, Autismus-Spektrum-Störungen oder Suizidalität engmaschig monitoren, da KI durch Gefälligkeit pathologische Muster iatrogen verstärken kann
Primat der therapeutischen Allianz	Die authentische, interpersonelle Beziehung bleibt unersetzbar; KI nur als ergänzendes Element nutzen und stets den Transfer in reale menschliche Interaktionen betonen.
Entwicklung robuster Krisenprotokolle	Nur KI-Tools mit rigoros validierten Eskalationsmechanismen einsetzen, die akute Risiken zuverlässig erkennen und unmittelbar an menschliche Krisendienste verweisen
Etablierung von KI-Kompetenz	AI-Literacy als Kernkompetenz verstehen: kontinuierliche Fortbildung zu Funktionsweise, algorithmischen Verzerrungen, Limitationen und rechtlichen Risiken

Tab. 3: Empfehlungen für die psychotherapeutische Praxis

Für den praktischen Umgang mit Patient:innen, die KI bereits intensiv nutzen, lässt sich die proaktive KI-Anamnese konkretisieren. Der Einstieg sollte wertfrei und normalisierend erfolgen, etwa: «Viele Menschen nutzen heute Chatbots wie ChatGPT, auch für persönliche Themen – wie ist das bei Ihnen?» Zu explorieren sind Häufigkeit und Anlässe der Nutzung, die Funktion der KI im Alltag (Information, Trost, Rat, Gesprächsersatz), die damit verbundenen Erwartungen sowie Anzeichen einer problematischen Bindung, etwa wenn die KI realen Beziehungen vorgezogen wird. Gewichten Patient:innen KI-Antworten höher als therapeutische Interventionen, sollte dies nicht als Loyalitätskonflikt behandelt, sondern als Material für die gemeinsame Reflexion genutzt werden: Was macht die KI-Antwort attraktiv, und was fehlt ihr? Hat eine KI pathologische Überzeugungen bestätigt, ist dies explizit, aber ohne Beschämung anzusprechen und die Überzeugung selbst therapeutisch zu bearbeiten (Morris et al., 2026). Ziel ist eine reflektierte Integration der KI-Nutzung in den Therapieprozess, ohne diese vor-schnell zu pathologisieren.

Beim Risikoscreening und Monitoring vulnerabler Gruppen sind zudem spezifische klinische Mechanismen zu beachten. Bei Patient:innen mit psychotischen Störungen können sykopphantische, stets validierende Systeme Wahnideen verstärken oder selbst in eine dialogische Wahndynamik eingebaut werden, etwa wenn die KI als Bestätigungsinstanz oder als Bestandteil des

Wahnsystems fungiert (ebd.). Bei akuter Suizidalität ist die zuverlässige Triage entscheidend: Systeme ohne validierte Eskalationsmechanismen reagieren in Krisensituationen teils verharmlosend oder inadäquat (Grabb et al., 2024), weshalb hier ausschliesslich Anwendungen mit nachgewiesener Krisenerkennung und unmittelbarer Weiterleitung an menschliche Krisendienste vertretbar sind. Bei sozial isolierten oder bindungsunsicheren Patient:innen kann die ständige Verfügbarkeit einer scheinbar verständnisvollen KI emotionale Abhängigkeit begünstigen und den Rückzug aus realen Beziehungen verstärken (Frances, 2025; Zhang et al., 2025); hier sollte der Transfer in menschliche Beziehungen ein explizites Therapieziel bleiben.

Schluss

KI in der psychotherapeutischen Versorgung ist keine abstrakte Zukunftsvision, sondern bereits Realität – mit klinischen, normativen und gesellschaftlichen Implikationen, die erst ansatzweise erfasst sind. Das Potenzial ist erheblich: Versorgungslücken lassen sich reduzieren, Diagnostik und Adhärenz fördern, administrative Belastungen senken. Zugleich prägen erhebliche Risiken das Feld: unzureichende Wirksamkeitsnachweise, mangelhaftes Krisenmanagement, Halluzinationen, Datenschutzdefizite, algorithmische Verzerrungen und ein

regulatorisches Vakuum. Eine kritisch-reflexive Implementierung ist daher unverzichtbar. Im Zentrum der ethischen Beurteilung steht der normative Status von KI: Als Instrument gelten Kriterien wie Zuverlässigkeit, Sicherheit und Zweckmässigkeit; als therapeutischer Akteur kämen Anforderungen an moralische Urteilsfähigkeit und Beziehungsfähigkeit hinzu – die aktuelle Systeme nicht erfüllen. In dieser Spannung zwischen menschenähnlichem Erscheinungsbild und normativer Unerfüllbarkeit liegen zentrale Risiken: Erosion der therapeutischen Allianz, übermässige Abhängigkeit und epistemische Ungerechtigkeit.

Als ethische Position und professionspolitische Schlussfolgerung dieses Beitrags – nicht als empirisch abschliessend belegte Aussage – gilt: KI in der Psychotherapie ist als ergänzendes, klar abgegrenztes Werkzeug zu verstehen, das menschliche Denkprozesse unterstützt, nicht ersetzt. Die professionell verantwortete therapeutische Beziehung bleibt die normative Referenz, an der sich der Einsatz technischer Systeme messen lassen muss. Diese Position wird durch internationale Leitlinien gestützt, ist aber weder rechtlich noch in der klinischen Routine hinreichend operationalisiert. Für die psychotherapeutische Profession ergeben sich drei Aufgaben: 1. ist KI-Kompetenz verbindlich in Aus-, Weiter- und Fortbildung zu verankern, damit Behandelnde Risiken erkennen, aufklären und Tools sachgerecht einsetzen. 2. müssen Fachgesellschaften und Standesvertretungen regulatorische Rahmenbedingungen aktiv mitgestalten, um klinische Sicherheit, Datenschutz und Versorgungsgerechtigkeit zu sichern. 3. braucht es eine Forschungsagenda, die methodisch robuste Wirksamkeits-, Sicherheits- und Equity-Studien priorisiert. Eine ethisch verantwortbare Integration gelingt durch eine reflektierte Haltung, die das Potenzial dieser Technologien anerkennt und zugleich an den Grundwerten der therapeutischen Beziehung, der Patientensicherheit und der professionellen Integrität festhält.

Offenlegung der Verwendung von KI

Zur Erstellung dieses Manuskripts wurden Tools der künstlichen Intelligenz (KI) verwendet. Konkret wurden NotebookLM (Google), ChatGPT (OpenAI) sowie Claude (Anthropic) zur Analyse und Zusammenfassung ausgewählter Artikel verwendet; DeepL (DeepL SE) wurde zur sprachlichen Überarbeitung eingesetzt. Die Autor:innen haben alle KI-generierten oder KI-unterstützten Inhalte überprüft, bearbeitet und genehmigt, um deren Genauigkeit, Integrität und Übereinstimmung mit den wissenschaftlichen Standards der Arbeit sicherzustellen; sie übernehmen die volle Verantwortung für den Inhalt.

Literatur

- Algumaei, A. et al. (2025). Symmetric therapeutic frameworks and ethical dimensions in AI-based mental health chatbots (2020–2025): A systematic review of design patterns, cultural balance, and structural symmetry. *Symmetry*, 17(7), 1082.
- Ali, M. et al. (2025). Artificial intelligence for mental health: A narrative review of applications, challenges, and future directions in digital health. *Digital Health*, 11, 20552076251395548.
- André, E. et al. (2026). Nutzung von KI-Technologien für Diagnostik und Gesprächsanalyse in der Psychotherapie: Potential und Herausforderungen. *PiD*, online-first.
- APA (2017). *Ethical principles of psychologists and code of conduct* (2002, amended effective June 1, 2010, and January 1, 2017).
- APA (2025a). *Ethical guidance for AI in the professional practice of health service psychology*.
- APA (2025b). *APA health advisory on the use of generative AI chatbots and wellness applications for mental health*.
- Barbui, C. et al. (2025). Treatment gaps for mental disorders in Europe: A systematic synthesis. *Lancet Psychiatry*, online-first.
- Bucher, A. et al. (2025). «It's not only attention we need»: Systematic review of large language models in mental health care. *JMIR Mental Health*, 12, e72061.
- Bundesärztekammer (2025). *Hinweise und Empfehlungen zum Umgang mit Künstlicher Intelligenz in der ärztlichen Versorgung*. Eigenverlag.
- Cecil, J. et al. (2026). Navigating the complexity of AI adoption in psychotherapy by identifying key facilitators and barriers. *npj Mental Health Research*, 5.
- Cruz-Gonzalez, P. et al. (2025). Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications. *Psychological Medicine*, 55, e18.
- De Choudhury, M. et al. (2023). Benefits and harms of large language models in digital mental health. *arXiv*.
- EU (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the EU*, L 2024/1689.
- Farzan, M. et al. (2025). Artificial intelligence-powered cognitive behavioral therapy chatbots: A systematic review. *Iranian Journal of Psychiatry*, 20(1), 102–110.
- Ferrario, A. et al. (2024). The role of humanization and robustness of large language models in conversational artificial intelligence for individuals with depression: A critical analysis. *JMIR Mental Health*, 11, e56569.
- Fiske, A. et al. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, 21(5), e13216.
- FMH (2022). *FMH-Positionspapier: Künstliche Intelligenz in der Medizin*. Eigenverlag.
- Frances, A. (2025). Warning: AI chatbots will soon dominate psychotherapy. *The British Journal of Psychiatry*, 1–5.
- Gooding, P. (2024). *Study on the use of artificial intelligence in mental health care: A global review of policy, regulation, and ethics*. Mental Health Europe & Council of Europe.
- Grabb, D. et al. (2024). Risks from language models for automated mental healthcare: Ethics and structure for implementation. *medRxiv*.
- Götzl, C. et al. (2022). Artificial intelligence-informed mobile mental health apps for young people: A mixed-methods approach on users' and stakeholders' perspectives. *Child and Adolescent Psychiatry and Mental Health*, 16(1), 86.
- Gülegen, B. et al. (2026). AI-generated documentation of psychiatric interviews: A proof-of-concept study. *Frontiers in Psychiatry*, 17, 1621532.
- Habicht, J. et al. (2025). Generative AI-enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: Real-world observational study. *Journal of Medical Internet Research*, 27, e60435.

- Hatch, S.G. et al. (2025). When ELIZA meets therapists: A Turing test for the heart and mind. *PLOS Mental Health*, 2(2), e0000145. <https://doi.org/10.1371/journal.pmen.0000145>
- Hatherley, J.J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics*, 46(7), 478–481.
- Heinz, M.V. et al. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*, 2(4).
- Khawaja, Z. & Bélisle-Pipon, J.-C. (2023). Your robot therapist is not your therapist: Understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, 5, 1278186.
- Knox, B. et al. (2023). Justice, vulnerable populations, and the use of conversational AI in psychotherapy. *The American Journal of Bioethics*, 23(5), 48–50.
- Laymouna, M. et al. (2024). Roles, users, benefits, and limitations of chatbots in health care: Rapid review. *Journal of Medical Internet Research*, 26, e56930.
- Le Glaz, A. et al. (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research*, 23(5), e15708.
- Li, H. et al. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine*, 6(1), 236.
- Limbic (2026). *Limbic for NHS Talking Therapies: Impact report*. Eigenverlag.
- Lucas, G.M. et al. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37, 94–100.
- Manole, A. et al. (2024). An exploratory investigation of chatbot applications in anxiety management: A focus on personalized interventions. *Information*, 16(1), 11.
- Manriquez Roa, T. et al. (2021). The ethics of artificial intelligence in psychotherapy. In M. Trachsel et al. (Eds.), *Oxford handbook of psychotherapy ethics* (pp. 744–758). Oxford UP.
- Martinez-Martin, N. & Kreitmair, K. (2018). Ethical issues for direct-to-consumer digital psychotherapy apps: Addressing accountability, data protection, and consent. *JMIR Mental Health*, 5(2), e32.
- Mikaeili, N. et al. (2025). Reimagining mental health with artificial intelligence: Early detection, personalized care, and a preventive ecosystem. *Journal of Multidisciplinary Healthcare*, 18, 7355–7373.
- Mok, C.H.Y. et al. (2025). Application of artificial intelligence and psychosocial functioning in psychosis: A systematic review and meta-analysis. *Frontiers in Psychiatry*, 16, 1692177.
- Montemayor, C. et al. (2022). In principle obstacles for empathic AI: Why we can't replace human empathy in healthcare. *AI & Society*, 37(4), 1353–1359.
- Morrin, H. et al. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, 13.
- Nehme, M. et al. (2024). Chatbots in medicine: Certification process and applied use case. *Swiss Medical Weekly*, 154, 3954.
- Nunes-Zlotkowski, F.K. et al. (2024). Blended psychological therapy for the treatment of psychological disorders in adult patients: Systematic review and meta-analysis. *Interactive Journal of Medical Research*, 13, e56300.
- Obradovich, N. et al. (2024). Opportunities and risks of large language models in psychiatry. *npj Digital Psychiatry and Neuroscience*, 2(1), 8.
- Ohu, F.C. et al. (2025). Public health risk management, policy, and ethical imperatives in the use of AI tools for mental health therapy. *Healthcare*, 13(21), 2721.
- Opel, N. & Breakspear, M. (2026). Transforming mental health research and care through artificial intelligence. *Science*, 391(6731), 249–258.
- Orrù, L. & Mannarini, S. (2026). The role of artificial intelligence in clinical psychology: How AI and NLP systems are reshaping psychological interventions. A systematic review. *Clinical Psychology & Psychotherapy*, 33(1), e70071.
- Pal, A. et al. (2025). Generative AI/LLMs for plain language medical information for patients, caregivers and general public: Opportunities, risks and ethics. *Patient Preference and Adherence*, 19, 2227–2249.
- Perlis, R.H. et al. (2024). Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology*, 49(11), 1697–1702.
- Rahsepar Meadi, M. et al. (2025). Exploring the ethical challenges of conversational AI in mental health care: Scoping review. *JMIR Mental Health*, 12, e60432.
- Rezaei, Z. et al. (2026). Network-based artificial intelligence in mental healthcare: A systematic review of chatbots, artificial intelligence/machine learning models and ethical considerations in global healthcare networks. *Digital Health*, 12, 20552076261421688.
- Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5), 2749–2767.
- Sedlakova, J. & Trachsel, M. (2023). Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? *The American Journal of Bioethics*, 23(5), 4–13.
- Sedlakova, J. & Trachsel, M. (i.D.). The ethics of digital mental health care: Conversational artificial intelligence and online therapy. In M. Scholten et al. (Eds.), *Mental health ethics: Current controversies and emerging debates*. Oxford UP.
- Sedlakova, J. et al. (2025a). Empathy in mental health care interventions by conversational artificial intelligence. In A. Barrera (Ed.), *Empathy in clinical psychiatry and mental health care: Clinical, conceptual, and scientific perspectives* (pp. 250–263). Oxford UP.
- Sedlakova, J. et al. (2025b). Human-like epistemic trust? A conceptual and normative analysis of conversational AI in mental healthcare. *The American Journal of Bioethics*, 1–16.
- Smith, K.A. et al. (2026). AI assisted triage of UK patients in mental health care services: A qualitative focus group study of patients' attitudes. *BMC Psychiatry*, 26(35).
- Solaiman, B. (2024). Generative artificial intelligence (GenAI) and decision-making: Legal & ethical hurdles for implementation in mental health. *International Journal of Law and Psychiatry*, 97, 102028.
- Soma, C.S. et al. (2026). Artificial intelligence to support human-provided mental health treatment. *Annual Review of Clinical Psychology*, 22, 15.1–15.27.
- Stade, E.C. et al. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1), 12.
- Tekin, Ş. & Delehanty, M. (2025). Beyond doomsday fears: Why we need to consider the potential harms of AI psychotherapy. *The American Journal of Bioethics*, online-first.
- Ueda, M. et al. (2025). Help-seeking in the age of AI: Use and perceptions of AI-based mental health support among U.S. adults – Cross-sectional study. *JMIR Mental Health*, online-first.
- Wampold, B.E. & Flückiger, C. (2023). The alliance in mental health care: Conceptualization, evidence and clinical applications. *World Psychiatry*, 22(1), 25–41.
- WHO (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. Eigenverlag.
- WHO (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. Eigenverlag.
- Zhang, R. et al. (2025). Computational and ethical considerations for using large language models in psychotherapy. *Nature Computational Science*, 5.
- Zhang, Z. & Wang, J. (2024). Can AI replace psychotherapists? Exploring the future of mental health care. *Frontiers in Psychiatry*, 15, 1444382.

Ethical issues regarding the use of AI in psychotherapy

Abstract: Artificial intelligence (AI) is also increasingly establishing itself in psychiatric and psychotherapeutic practice. Against the backdrop of global care disparities, AI-based applications are being discussed or are already in use for screening and diagnostics, clinical decision support, interventions, monitoring, and administrative and educational support. This article analyses opportunities and risks of this development from the perspective of medical ethics. While there are potential benefits, such as improved access to care, destigmatization, clinical decision support, and personalization and training opportunities, there are also significant risks, including insufficient evidence of efficacy, inadequate crisis management, susceptibility to errors ranging up to hallucinations, erosion of the therapeutic alliance, data protection deficits, algorithmic biases, and a regulatory vacuum. Four fundamental ethical questions are discussed in depth: 1. the normative status of AI as a tool or agent, 2. the conditions for trust in conversational AI, 3. simulated empathy, and 4. unrealistic expectations and excessive dependence. Current regulatory frameworks, including the WHO's guiding principles, the EU AI Act, the EU Medical Device Regulation, and recommendations from professional associations, are also contextualized. Based on this, action-oriented recommendations for psychotherapeutic practice are derived. These include taking a systematic AI-assessment, being transparent about technological limitations, preserving professional judgement and human control, strictly evaluating data protection and security, establishing robust crisis protocols and the development of AI competence. The authentic therapeutic relationship remains the indispensable agent of the therapeutic process. AI should be developed as a complementary tool, not a substitute for human professionals.

Keywords: artificial intelligence, psychotherapy, chatbots, therapeutic alliance, ethics

Biografische Notiz

Dr. sc. med. *Jan Schürmann* hat an der Universität Basel in Bio- und Medizinethik promoviert. Seit 2012 arbeitet er als klinischer Ethiker an den Universitätsspitalern in Basel. Zudem ist er als Dozent für Medizinethik an der Uni Basel und an der Berner Fachhochschule tätig. Seit 2024 ist er Mitglied der Zentralen Ethikkommission der Schweizerischen Akademie der Medizinischen Wissenschaften.

Dr. sc. med. *Jana Sedlakova* ist Postdoktorandin am Ethox Centre der Universität Oxford. Sie promovierte an der Universität Zürich in Bio- und Medizinethik sowie im Bereich Digitalisierung in der Gesellschaft. Seit 2021 ist sie Vorsitzende der Ethics Gruppe der Digital Society Initiative der Universität Zürich.

Prof. Dr. med. Dr. phil. *Manuel Trachsel* ist Titularprofessor für Bio- und Medizinethik an der Medizinischen Fakultät der Universität Basel und Leiter der Abteilung Klinische Ethik am Universitätsspital Basel (USB), an den Universitären Psychiatrischen Kliniken Basel (UPK), am Universitäts-Kinderspital beider Basel (UKBB) und an der Universitären Altersmedizin FELIX PLATTER (UAFP).

Kontakt

E-Mail: jan.schuermann@usb.ch