



Vorhaben: UBIDENZ - Ubiquitäre Digitale Empathische Therapieassistentz

DFKI-Teilvorhaben: Bindungsorientierte interaktive Verhaltensmodellierung für den
UBIDENZ-Avatar

Titel: Schlussbericht - DFKI

Förderkennzeichen: 13GW0568D

Zuwendungsempfänger: Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Trippstadter Straße 122, D-67663 Kaiserslautern

Projektleiter: Prof. Dr. Antonio Krüger

Bewilligungszeitraum: 1. September 2021 – 31. August 2024

Autoren: Dr. Patrick Gebhard, Dr. Tanja Schneeberger

Datum: 21. November 2024

Gefördert durch:



Bundesministerium
für Bildung
und Forschung

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 13GW0568D unter den Nebenbestimmungen der NKBF 2017 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Inhaltsverzeichnis

1. Einleitung	3
2. Aufgabenstellung	4
2.1 Wissenschaftliche Arbeitsziele	4
2.2 Technische Arbeitsziele	4
3. Durchführungsvoraussetzungen des Vorhabens	5
4. Planung und Ablauf des Vorhabens	5
5. Wissenschaftlicher und technischer Stand bei Projektstart	6
6. Verwendete Fachliteratur	6
7. Zusammenarbeit mit anderen Stellen	7
8. Projektergebnisse	8
8.1 Konzeptuelle Ergebnisse	8
8.1.1. Anforderungsanalyse an den bindungsorientierten UBIDENZ-Avatar	8
8.1.2. Theoriegetriebenes UBIDENZ-Avatar-Verhaltensmodells basierend auf dem Bindungsstil	12
8.2 Technische Ergebnisse	27
8.2.1 Datenkorpus	28
8.2.2 Erweiterung Benutzermodelle	29
8.2.3 Erweiterung bindungsorientiertes Avatarverhaltensmodell	33
8.2.4 Das BISI Software-Framework	34
9. Voraussichtlicher Nutzen und Verwertbarkeit	36
10. Außerhalb des Projektkontextes bekannt gewordener Fortschritt	37
11. Erfolge und geplante Veröffentlichungen	37
11.1 Erfolgte Veröffentlichungen	37
11.2 Geplante Veröffentlichungen	38
12. Literatur	39

1. Einleitung

Das vorliegende Dokument stellt den Abschlussbericht über die Arbeiten des DFKI im Verbundprojekts UBIDENZ dar. Das Projekt wurde durch das Bundesministerium für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 13GW0568D gefördert.

Die Kapitel 2 – 7 enthalten eine kurze Darstellung zu

- Aufgabenstellung
- Voraussetzungen, unter denen das Vorhaben durchgeführt wurde
- Planung und Ablauf des Vorhabens
- wissenschaftlichem und technischem Stand, an den angeknüpft wurde, insbesondere
- Angabe der verwendeten Fachliteratur sowie der benutzten Informations- und Dokumentationsdienste
- Zusammenarbeit mit anderen Stellen

Die Kapitel 8 – 11 enthalten eine eingehende Darstellung

- der erzielten Ergebnisse
- des voraussichtlichen Nutzens, insbesondere der Verwertbarkeit des Ergebnisses und der Erfahrungen
- des während der Durchführung des Vorhabens dem Zuwendungsempfänger bekannt gewordenen Fortschritts auf diesem Gebiet bei anderen Stellen
- der erfolgten oder geplanten Veröffentlichung von Ergebnissen

Neben dem Schlussbericht sind vom DFKI folgende Dokumente zum Projektabschluss erstellt worden: Erfolgskontrollbericht, Document Control Sheet, Berichtsblatt, Kurzbericht, Liste der Veröffentlichungen, Schutzrechtsanmeldung und den zahlenmäßigen Verwendungsnachweis. Alle Dokumente zusammen bilden die Abschlussdokumentation des UBIDENZ DFKI-Teilprojektes.

Die dargestellten Arbeiten umfassen alle vom DFKI geleisteten Arbeiten im Vorhaben UBIDENZ.

2. Aufgabenstellung

Das DFKI untersuchte und entwickelte in UBIDENZ neuartige Computermodelle von Nutzenden und Avatarverhalten für einen ubiquitären, sozio-empathisch agierenden Avatar für Menschen mit Depression nach einem psychiatrischen Klinikaufenthalt. Ziel war es, einen bindungsorientierten Avatar zu schaffen. Dieser ist als zentrale Interaktionsschnittstelle konzipiert. Eine echtzeitfähige Bindungssimulation nutzt die Analyse sozialer Signale des Nutzenden und wählt für das stetig angepasste Verhalten adäquate Verhaltensmuster und therapeutische Inhalte. So wurde versucht, die Assistenzfunktion zu stärken.

Besonders innovativ an diesem Ansatz war die Strategie, technologiebasiert eine verständnisvolle, therapeutische Beziehung zu gestalten, die einen Anreiz generiert, das System dauerhaft zu nutzen, mit nachhaltigen Vorteilen für die Sicherstellung der Therapie-Adhärenz. Der virtuelle Avatar diente dabei als adaptives therapeutisches Übergangsobjekt aus dem depressionsbedingten regressiven Krankheitszustand. Über eine stabile Bindung sollten Eigenverantwortung, Selbstständigkeit und verschiedene Entwicklungspotentiale nachhaltig gefördert werden.

2.1 Wissenschaftliche Arbeitsziele

1. Theoriegetriebene Konzeption eines Bindungsmodells, das bestehende Computermodelle für Emotionen und Emotionsregulation erweitert
2. Daten- und theoriegetriebene Schaffung eines Algorithmus zur Simulation von Bindungstypen auf der Basis von sozialen Signalen und deren Interpretation, situativem Kontext und relevanten individuellen Parameter
3. Daten- und theoriegetriebene Konzeption des UBIDENZ-Benutzermodells und des UBIDENZ-Avatar-Verhaltensmodells und Kopplung an die Bindungssimulation
4. Sammlung, Aufbereitung und Annotation relevanter Datensätze
5. Evaluation verschiedener relevanter Aspekte der subjektiven Wahrnehmung der Bindungssimulation

2.2 Technische Arbeitsziele

1. Erweiterung bestehender Benutzermodelle, um den UBIDENZ-Anforderungen zu genügen
2. Erweiterung bestehender Avatar-Verhaltensmodelle inklusive Dialogmanagement

3. Repräsentation und Kopplung von assistiv-therapeutischen Maßnahmen in das Avatar-Verhaltensmodell
4. Technische Realisierung einer Bindungssimulation-Komponente inklusive Benutzermodell und Avatarverhaltensmodell und der Kopplung an eine Analysekomponente sozialer Signale
5. Evaluation der Bindungssimulation-Komponente und deren Teile hinsichtlich einer softwaretechnisch korrekten Arbeitsweise
6. Erweiterung des VirtualSceneMaker-Autorenwerkzeugs, um das bindungsorientierte Verhalten des UBIDENZ-Avatars zu parametrieren
7. Integration der Bindungssimulation-Komponente in das UBIDENZ-Gesamtsystem

3. Durchführungsvoraussetzungen des Vorhabens

Die Durchführung des Vorhabens profitierte durch die in vorangegangenen Projekten entstandenen Beziehungen zwischen einigen Partnern. Die Partner DFKI, Universität Augsburg, ki elements, Better at Home und Charamel kooperieren seit Längerem in mehreren Projekten und Forschungsthemen. Die im Vorfeld geführten Gespräche mit allen Partnern ließen auf ein großes inhaltliches, konzeptionelles und technisches Verständnis rückschließen. Dies waren notwendige Voraussetzungen für die Durchführung des Vorhabens, die als sehr gut beschrieben werden können.

4. Planung und Ablauf des Vorhabens

Die Koordination einzelner Aufgaben im Projekt wurde von allen Projektpartnern in der Antragsphase gemeinschaftlich gründlich erarbeitet. Das DFKI arbeitete in allen inhaltlichen Arbeitspaketen mit, mit einem Fokus auf AP4, welches die Realisierung eines bindungsorientierten interaktiven Verhaltensmodells für den Versorgungs-Avatar umfasste.

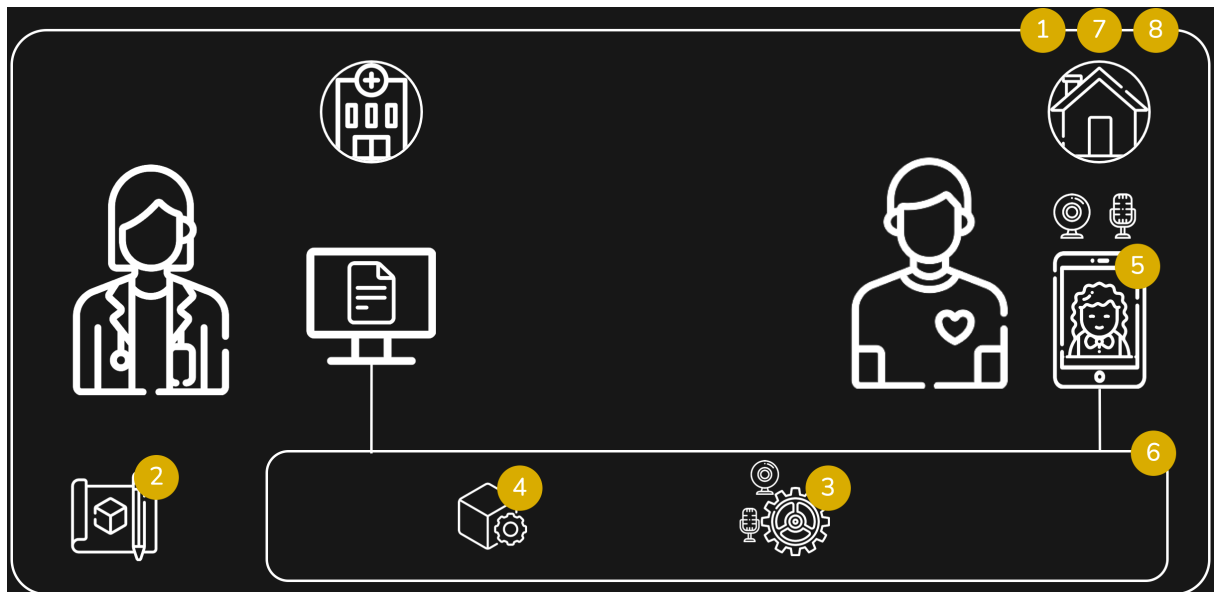


Abbildung 1. Zusammenhang der Arbeitspakete

Die Koordination einzelner Aufgaben im Projekt wurde von allen Projektpartnern in der Antragsphase gemeinschaftlich gründlich erarbeitet. Das DFKI arbeitete in fast allen Arbeitspaketen außer AP 1 mit und koordinierte das AP 4.

Der gesamte Ablauf der Arbeiten wurde in den einzelnen Arbeitspaketen im Arbeitsplan in der Gesamtvorhabensbeschreibung sowie der Teilvorhabensbeschreibung des DFKI definiert. Zunächst wurde existierende wissenschaftliche Literatur zum bindungsorientierten Verhalten zur Bildung einer theoretischen und konzeptuellen Übersicht gesichtet und entsprechend notwendige Daten dafür identifiziert.

5. Wissenschaftlicher und technischer Stand bei Projektstart

Interaktive soziale Avatare für ein psychosoziales Nachsorgemanagement sowie ein automatisiertes und analog begleitetes Monitoring des ambulanten Behandlungsverlaufes mit Navigationsfunktion und Krisenintervention bei psychiatrischen Erkrankungen, waren bei Projektstart unerforscht. Obwohl es zahlreiche chat-basierte Lösungen bereits zu Projektstart gab, setzte keine der Lösungen auf einen sozio-emotionalen bindungsorientierten Avatar.

6. Verwendete Fachliteratur

Insgesamt wurden Erkenntnisse aus zahlreichen wissenschaftlichen Arbeiten zu folgenden Themen genutzt: Interpersonale Nähe, Depression, therapeutische Allianz, sozialinteraktive Agenten, Determinanten von Beziehungsaufbau in Mensch-Mensch-Interaktionen und in

Mensch-Maschine-Interaktionen, soziale Attribution, Synchronizität, Emotionsregulation, große Sprachmodelle, Bayes'sche Netzwerke, Neuronale Netzwerke, Synthese von Agentenverhalten, Anthropomorphismusgrad von Agenten und deren Auswirkung auf Mensch-Maschine-Interaktion, Soziale Signalanalyse, und Feature Extraction.

Eine Übersicht dazu geben die sechs veröffentlichten Arbeiten:

Reinwarth, A. L., Schneeberger, T., Nunnari, F., Gebhard, P., Altmann, U., & Wessler, J. (2023). Look what I made it do-The ModelIT method for manually modeling nonverbal behavior of socially interactive agents. *Companion Publication of the 25th International Conference on Multimodal Interaction*, 200–204.

Schneeberger, T., Reinwarth, A. L., Wensky, R., Anglet, M. S., Gebhard, P., & Wessler, J. (2023). Fast friends: Generating interpersonal closeness between humans and socially interactive agents. *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, 1–8.

Withanage Don, D. S., Müller, P., Nunnari, F., André, E., & Gebhard, P. (2023). Renelib: Real-time neural listening behavior generation for socially interactive agents. *Proceedings of the 25th International Conference on Multimodal Interaction*, 507–516.

Wessler, J., Gebhard, P., & Zilcha-Mano, S. (2024). Investigating movement synchrony in therapeutic settings using socially interactive agents: An experimental toolkit. *Frontiers in Psychiatry*, 15, Article 1330158.

Müller, P., Heimerl, A., Hossain, S. M., Siegel, L., Alexandersson, J., Gebhard, P., André, E., & Schneeberger, T. (2024). Recognizing emotion regulation strategies from human behavior with large language models. *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction*.

Hladký, M., Guerra, R. R., Cang, X. L., MacLean, K. E., Gebhard, P., & Schneeberger, T. (2024). Modeling the 'Kiss my Ass'-Smile: Appearance and functions of smiles in negative social situations. *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction*.

7. Zusammenarbeit mit anderen Stellen

Während der Projektlaufzeit wurden mit Forschenden aus den Bereichen Psychologie und Psychiatrie der kontinuierlich der Projektstand/Erkenntnisse und technische Realisierungen diskutiert. Die Ergebnisse flossen in die Arbeiten ein. Die externen Partner waren Uwe Altmann (Medical School Berlin), Cord Beneke (Universität Kassel), Catherine Pelachaud (CNRS - ISIR, Sorbonne University.), Sigal Zilcha-Mano (Universität Haifa), Michaela Rohr (Lehrstuhl Klinische Psychologie der Universität des Saarlandes), Jana Volkert (Universität Ulm) Aike

Horstmann (Universität Duisburg-Essen), Karon MacLean (University of British Columbia), Eva Möhler (Universitätsklinikum des Saarlandes). Außerdem wurde das Projekt bei verschiedenen anderen Möglichkeiten bei folgenden Gruppen vorgestellt: Verband der Hausärztinnen und Hausärzte in Thüringen, Sozialunternehmen 3B (Dillingen) und Trägerwerk soziale Dienste AG, Justizvollzugsanstalt Saarbrücken, Heinz Nixdorf Museum, Profis on Tour (RBB).

8. Projektergebnisse

8.1 Konzeptuelle Ergebnisse

Die konzeptuellen Ergebnisse beziehen sich auf 1) Theorie- und datengetriebene Modellierung von Bindung und Emotionen 2) Benutzermodelle und Verhaltenssimulation für Bindungssysteme und 3) Datenmanagement und Evaluation von Bindungssimulationen.

- Theoriegetriebene Konzeption eines Bindungsmodells das bestehende Computermodelle für Emotionen und Emotionsregulation erweitert
- Daten- und theoriegetriebene Schaffung eines Algorithmus zur Simulation von Bindungstypen auf der Basis von sozialen Signalen und deren Interpretation, situativem Kontext und relevanten individuellen Parameter
- Daten- und theoriegetriebene Konzeption des UBIDENZ-Benutzermodells und des UBIDENZ-Avatar-Verhaltensmodells und Kopplung an die Bindungssimulation
- Sammlung, Aufbereitung und Annotation relevanter Datensätze
- Evaluation verschiedener relevanter Aspekte der subjektiven Wahrnehmung der Bindungssimulation

8.1.1. Anforderungsanalyse an den bindungsorientierten UBIDENZ-Avatar

Der Inhalt des ersten Unterarbeitspaketes des fünften Arbeitspaketes definierte eine Anforderungsanalyse für das Avatar-Design und die Verhaltensmodelle. Diese Anforderungsanalyse war zentral, um sicherzustellen, dass der UBIDENZ-Avatar die Erwartungen der relevanten Stakeholder - Behandelnde und Patient:innen - erfüllt. Sie half uns, wichtige Aspekte sowohl inhaltlich als auch in der Art der Interaktion zu identifizieren und ein gemeinsames Verständnis für die Umsetzung zu schaffen.

Um die Erwartungen von Patienten mit Depressionen und medizinischem Fachpersonal an den sozial interaktiven Agenten als virtuellen therapeutischen Assistenten in der ambulanten Nachsorge zu untersuchen, wurde eine qualitative Umfrage durchgeführt. Diese Analyse konzentrierte sich auf Forschungsfragen, die das Erscheinungsbild und die Rolle des Assistenten, die Interaktion zwischen Assistent und Patient (Zeitpunkt der Interaktion, Fähigkeiten und Fertigkeiten des Assistenten, Art der Interaktion) und die Interaktion zwischen Therapeut und Assistent betreffen.

Forschungsfrage 1: Welche Rolle spielt ein virtueller Therapieassistent im Rahmen der ambulanten Pflege?

Forschungsfrage 2: Welche Präferenzen gibt es für die Ausgestaltung eines virtuellen Therapieassistenten?

Forschungsfrage 3: Wann sollte ein virtueller Therapieassistent mit dem Patienten interagieren?

Forschungsfrage 5: Über welches Interaktionsverhalten sollte ein virtueller Therapieassistent verfügen, um Patienten in ihrer Behandlung zu unterstützen?

Forschungsfrage 6: Wie könnte der Rahmen so gestaltet werden, dass die Angehörigen des Gesundheitswesens auf allen Ebenen (Psychiater, Psychotherapeuten, Krankenpfleger und Sozialarbeiter) von der Einführung eines virtuellen Therapieassistenten profitieren können?

Zur Beantwortung der Forschungsfragen wurde eine zweiteilige qualitative Studie durchgeführt, um die Perspektiven der beiden Gruppen (Patient:innen und Behandelnde) zu untersuchen. In einem ersten Schritt wurden Behandelnde ($N=30$; 6 männlich, 24 weiblich; zwischen 18 und 65+ Jahre) während einer regionalen Offline-Sitzung rekrutiert. Die Berufsgruppen waren Psychiater:innen ($n = 3$), Psychotherapeuten:innen ($n = 10$), Sozialarbeiter:innen ($n = 15$), psychiatrische Krankenschwestern ($n = 1$) und Ärzt:innen ($n = 1$). Nach einem kurzen Vortrag erhielten sie einen Link und wurden gebeten, an einem halbstrukturierten Online-Fragebogen teilzunehmen. Zweitens wurden Patient:innen ($N=20$; 11 männlich, 9 weiblich; zwischen 19 und 56 Jahre; BDI-II-Wert $M = 24.3$, $SD = 13.45$) in der Karl-Jaspers Klinik beim Projektpartner Carl von Ossietzky Universität Oldenburg rekrutiert und in einem halbstrukturierten persönlichen Gespräch befragt. Die Teilnahme aller Personen in beiden Gruppen basierte auf Freiwilligkeit.

Um aus den qualitativen Fragebogen- und Interviewdaten aussagekräftige Erkenntnisse für jede der Forschungsfragen zu gewinnen, analysierten wir die Antworten der beiden Interessengruppen in einem vierstufigen Prozess auf der Grundlage der Grounded Theory (Glaser, 2010). In einem ersten Schritt wurden die Antworten der Gruppe der Patient:innen transkribiert und Frage für Frage in zwei verschiedenen Datensätzen gesammelt (einer für jede Gruppe). Zweitens kategorisierten drei unabhängige Gutachtende die Antworten nach Inhalt und Relevanz für die Forschungsfragen. Drittens wurden die Rohdaten und die extrahierten Kategorien für jede Gruppe (Behandelnde und Patient:innen) verglichen, wobei sowohl Überschneidungen als auch Unterschiede zwischen den beiden Gruppen herausgestellt wurden. Schließlich haben wir diesen vergleichenden Datensatz klassifiziert, um Cluster und Sub-Cluster mit Erklärungswert für jede Forschungsfrage zu extrahieren.

Die Ergebnisse zeigen, dass der Assistent multimodal kommunizieren (Stimme, Mimik, Gestik) und einer negativen Selbsteinschätzung entgegenwirken sollte. Das Geschlecht des Assistenten sollte weiblich oder optional sein. Aufgrund negativer Assoziationen wurde jedoch nur männlich abgelehnt. Die Assistentin wird als proaktiv und die Patient:innen als passiver Informationsempfänger in Bezug auf die Patient-Assistent-Interaktion gesehen. Lücken in der Nachsorge können durch die uneingeschränkte Verfügbarkeit der Assistentin geschlossen werden. Allerdings sollte der Assistent die Autonomie der Patient:innen trainieren, um Abhängigkeit zu vermeiden. Die Überwachung des Gesundheitszustandes wurde von beiden Gruppen positiv bewertet. Um Frühwarnzeichen von Krankheiten zu erkennen, wird eine Biofeedback-Funktion gewünscht. Wenn es in der Situation angemessen ist, wird Humor beim Assistenten als wünschenswert erachtet. Hinsichtlich der erwünschten Fähigkeiten des Assistenten lassen sich diese zusammenfassen in der Bereitstellung von Struktur und emotionaler Unterstützung, insbesondere Wärme und Kompetenz, um Vertrauen aufzubauen. Beständigkeit ist wichtig, damit der Assistent authentisch wirkt. Im Hinblick auf die Interaktion zwischen Leistungserbringern und Therapeuten-Assistenten wurden eine objektive Messung des Gesundheitszustands der Patient:innen, ein Notfallsystem zur Suizidprävention sowie ein Informationsinstrument und eine Entscheidungshilfe für die Angehörigen der Gesundheitsberufe (Psychiater:innen, Psychotherapeuten:innen, Krankenpfleger:innen und Sozialarbeiter:innen) als entscheidend bezeichnet.

Aus der Studie lassen sich die folgenden wesentlichen Schlussfolgerungen ziehen:

Rolle: Die virtuelle Therapieassistentin wurde als zusätzlicher Unterstützungsanker im ambulanten Nachsorgesystem gesehen. Sie kann aufgrund ihrer uneingeschränkt verfügbaren therapeutischen Fähigkeiten und ihrer Fähigkeit, eine Beziehung aufzubauen, wo Sozialarbeiter:innen, Therapeuten:innen und Angehörige den Betreuungsbedarf von Patient:innen nicht decken können, Versorgungslücken schließen.

Verkörperung: Es wird eine multimodale Funktion für empathische Kommunikation (d. h. Stimme, Mimik, Gestik) erwartet. Patient:innen möchten, dass der Assistent unvollkommen aussieht, da sonst die Gefahr einer negativen Selbsteinschätzung besteht, insbesondere bei depressiven Patienten mit mangelndem Selbstwertgefühl. Der Assistent kann möglicherweise mit Lösungsstrategien helfen, ein positiveres Selbstbild aufzubauen. Das Geschlecht des Assistenten sollte entweder weiblich oder wählbar, aber nicht männlich sein. Es wird angenommen, dass diese Vorliebe auf negativen Assoziationen mit dem männlichen Geschlecht, wie z. B. häusliche Gewalt, und der Wahrnehmung des weiblichen Geschlechts als einfühlsamer beruht.

Timing der Interaktion: Die Ergebnisse zeigen, dass sowohl die Patient:innen als auch Behandelnde ein proaktives Verhalten des Assistenten erwarten. So wird der Assistent als aktivierendes Medium gesehen, während Patient:innen als passive Empfänger von Informationen angesehen werden. Die Erwartungen hinsichtlich des Zeitpunkts bestimmter Verhaltensweisen konnten in situationsunabhängige und situationsabhängige Verhaltensweisen unterschieden werden.

Interaktionsfähigkeiten und Verhalten: Es wurden zwei Kernfähigkeiten ermittelt, die sowohl von Behandelnden als auch von Patient:innen erwartet werden: die Fähigkeit, strukturelle Unterstützung zu leisten, und die Fähigkeit, emotionale Unterstützung zu geben. Diese Fähigkeiten sind den Dimensionen der zwischenmenschlichen sozialen Wahrnehmung zuzuordnen, nämlich Kompetenz und Wärme. Ein Akteur, der diese beiden Fähigkeiten durch angemessenes Verhalten zum richtigen Zeitpunkt demonstriert, könnte eine wirksame Ergänzung der ambulanten Nachsorge darstellen. Für jedes Sub-Cluster und den passenden Zeitpunkt liefern unsere Ergebnisse eine umfangreiche Liste von Vorschlägen für geeignetes Verhalten.

Interaktion zwischen Assistentin und Gesundheitsfachkraft: Die Gesamtergebnisse zeigen zum einen die Möglichkeit für Therapeuten:innen, sonst nicht verfügbare Informationen über den

Gesundheitszustand der Patient:innen zu erhalten, und zum anderen die Möglichkeit eines Notfallsystems bei Suizidgefahr der Patient:innen. Diese Innovationen können bestehende Lücken in der Nachsorge schließen, die Sicherheit während dieser kritischen Zeit verbessern, die Nachsorge für die Angehörigen des Gesundheitswesens weniger belastend machen und durch die Bereitstellung eines objektiven Maßes für den Gesundheitszustand der Patient:innen die Entscheidungsfindung der Angehörigen des Gesundheitswesens erleichtern. Insgesamt lieferte die durchgeführte Studie sowohl einen Rahmen für die Entwicklung in UBIDENZ, als auch generell innovative Leitlinien für die Entwicklung eines virtuellen therapeutischen Assistenten, um die Lücken in der Patientennachsorge zu schließen.

8.1.2. Theoriegetriebenes UBIDENZ-Avatar-Verhaltensmodells basierend auf dem Bindungsstil

Das nonverbale Verhalten von sozial interaktiven Agenten (SIAs) (Lugrin et al., 2021) wird oft automatisch generiert und ist für alle Benutzer identisch.

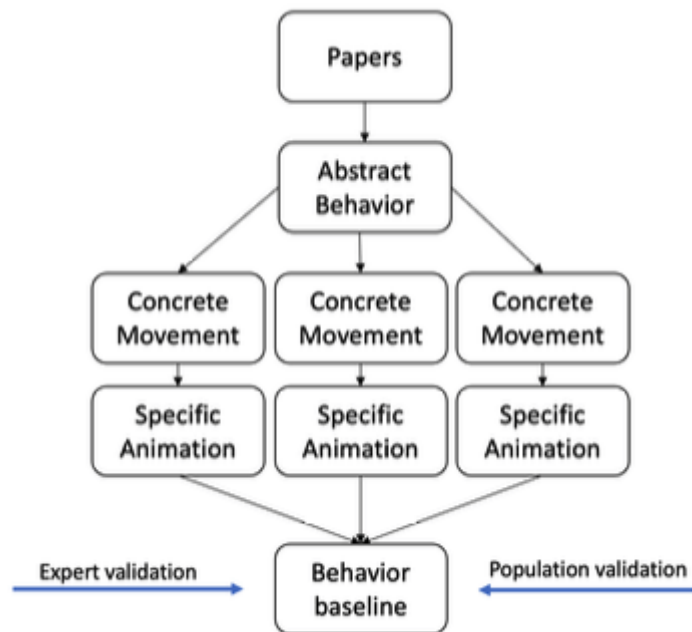
SIAs sind virtuell oder physisch verkörperte Agenten, die in der Lage sind, autonom mit Menschen auf sozial intelligente Weise zu kommunizieren und dabei multimodale Verhaltensweisen anzuwenden. Durch das Hinzufügen humanoider Aspekte in der Mensch-Computer-Schnittstelle, wird die Art der Kommunikation, die aus menschlichen Interaktionen bekannt ist, auf die Interaktion mit Maschinen übertragen, was sie intuitiv verständlich und bedienbar macht.

Dieser Ansatz, obwohl er ökonomisch ist, kann sich kontraproduktiv auswirken, wenn es darum geht, Anwendungen für unterschiedliche Bevölkerungsgruppen, wie zum Beispiel Menschen mit Depressionen, zu entwickeln. Außerdem kann sich dies negativ auf die Validität der Forschung auswirken und die Wirksamkeit von SIA-basierten Interventionen verringern. Deswegen haben wir in UBIDENZ eine Methode zur Modellierung nonverbalen Verhaltens in SIAs entwickelt. Die ModellIT-Methode ermöglicht es Forschenden, die Modellierung von nonverbalem Verhalten auf psychologische Theorien zu stützen. Sie zielt darauf ab, eine standardisierte und replizierbare Methode zu etablieren, die offene Wissenschaftspraktiken fördert und die Erstellung von maßgeschneiderten SIAs erleichtert. Sie ist ein Schritt in Richtung barrierefreier und zugänglicher SIA-Anwendungen für verschiedenen Bevölkerungsgruppen. Obwohl wir in uns im Rahmen des UBIDENZ Projektes auf Bindungsstile konzentriert haben, ist die geschaffene und publizierte Methode (Reinwarth et al., 2023) auf

verschiedene Anwendungsfälle übertragbar, je nachdem, welche psychologischen Theorien als Basis verwendet werden.

Bindungsstil ist eine innere Repräsentation und ein Muster der Beziehungsdynamik mit nahestehenden Personen. Während er erstmals im frühen Säuglingsalter beobachtet und gemessen wurde, geht das Konzept der erwachsenen Bindung davon aus, dass solche erlernten Muster ins Erwachsenenalter übertragen werden. In der Forschung werden verschiedene Kategorisierungssysteme zur Unterscheidung spezifischer Bindungsstile verwendet. Die gebräuchlichste Unterscheidung zwischen diesen Systemen ist die zwischen einem sicheren und einem unsicheren Bindungsstil. Einige Systeme verwenden kontinuierliche Dimensionen zur Messung des Bindungsstils: Bindungsangst und Bindungsvermeidung. Eine Person kann in beiden Bereichen hohe oder niedrige Werte erreichen, was zu vier Kategorien führt: sicher (niedrige Angst/Vermeidung), ängstlich (hohe Angst, niedrige Vermeidung), ablehnend (niedrige Angst, hohe Vermeidung) und ängstlich (hohe Angst/Vermeidung). Der Bindungsstil einer Person beeinflusst ihr nonverbales Verhalten in bindungsbezogenen Situationen. Sicher gebundene Menschen zeigen während einer Interaktion mit ihrem Partner mehr nonverbale Nähe (Lachen, Berühren, Anstarren und Lächeln) als vermeidende Menschen. Menschen mit hoher Bindungsangst verwenden in bindungsbezogenen Situationen nonverbale Hinweise auf Ärger, während Menschen mit hoher Bindungsvermeidung Distanzierungsstrategien anwenden, indem sie nonverbale Hinweise auf ihre Gefühle unterdrücken.

Unsere ModellIT-Methode (Model it! Modeling nonverbal behavior from lITerature) (Reinwarth et al., 2023) ist eine Methode zur standardisierten Modellierung von nonverbalem Verhalten in SIAs. Sie kann auf eine Vielzahl von Anwendungsfällen angewandt und zur Definition von SIA-Merkmalen und spezifischen Bedürfnissen verschiedener Nutzendengruppen verwendet werden. Sie besteht



aus fünf Schritten, die von der Forschungsebene zur Anwendungsebene führen: (1) Literaturrecherche; (2) Extraktion nonverbaler Hinweise; (3) Operationalisierung nonverbaler Hinweise; (4) Implementierung nonverbaler Hinweise; und (5) Validierung. Diese fünf Schritte sind ein notwendiger Modellierungsprozess, da nonverbale Hinweise in der vorhandenen Literatur häufig nicht in einer anwendbaren Weise definiert sind. Insbesondere bei der Entwicklung einer SIA für eine bestimmte Bevölkerungsgruppe sind die meisten nonverbalen Verhaltensweisen äußerst komplex und in einer Vielzahl von Forschungsarbeiten nur vage beschrieben. Zum Beispiel regulieren Menschen mit hoher Bindungsangst ihren Stress durch die Suche nach Nähe. Dies kann nicht direkt in eine Animation implementiert werden, sondern muss einen Modellierungsprozess durchlaufen, um tatsächlich nutzbar zu sein.

Im Folgenden werden die Schritte, um das Bindungsverhalten zu modellieren, ausführlich beschrieben.

(1) *Literaturrecherche*. Ziel der Literaturrecherche ist es, nonverbale Verhaltensweisen im Zusammenhang mit verschiedenen Bindungsstilen von Erwachsenen zu identifizieren - Kinder wurden ausgeschlossen. Die Suchstrategie konzentrierte sich auf Begriffe wie „attachment style nonverbal“ und nutzte Datenbanken wie „Web of Science“ und „Google Scholar“. Das gewählte Kategorisierungssystem für das SIA-Verhalten war: *secure*, *dismissing*, and

preoccupied (da die verschiedenen Bindungsstile bei verschiedenen Quellen unterschiedlich ins Deutsche übersetzt werden und wir auf englisch recherchiert haben, nennen wir die englischen Begriffe hier ebenfalls). Um so viele Informationen wie möglich zu erhalten, wurde auch Literatur zu jedem anderen Klassifizierungssystem für Bindungsstile herangezogen. Die Klassifizierungssysteme sind vergleichbar und können daher ineinander überführt werden, was in einem späteren Schritt geschah. Es ist zu beachten, dass in verschiedenen Szenarien eine solche Vergleichbarkeit möglicherweise nicht gegeben ist, so dass eine umfassende Dokumentation der Ein- und Ausschlusskriterien nicht möglich ist.

(2) *Extraktion nonverbaler Verhaltensweisen*. Die Literaturrecherche ergab eine Liste von wissenschaftlichen Veröffentlichungen, die nonverbale Verhaltensweisen dokumentieren, die mit Bindungsangst (hoch) und Bindungsvermeidung (hoch/niedrig) sowie Kategorien wie *secure*, *insecure generally*, *preoccupied*, und *dismissing* in Verbindung stehen. Nach dem Sammeln aller relevanten Papiere muss das spezifische nonverbale Verhalten identifiziert und extrahiert werden. Dazu müssen relevante Zitate, z. B. über Mimik, Körperbewegungen, Gesten und Blickkontakt, direkt aus den Artikeln entnommen werden, um nuancierte Details so genau wie möglich zu erhalten. Die Zitate werden dann kategorisiert und in einer Tabelle mit entsprechenden Zitaten auf strukturierte und standardisierte Weise beschrieben.

(3) *Operationalisierung nonverbaler Hinweise*. Der Übergang von abstrakten, nonverbalen, in der Literatur beschriebenen Verhaltensweisen zu konkreten Bewegungen ist aufgrund der mehrdeutigen oder undefinierten Beschreibungen komplex. Zum Beispiel ist der Bindungsstil *dismissing* durch reduzierte Bewegungen im Vergleich zu sicherer Bindung gekennzeichnet. Deshalb wurde zunächst ein Basisverhalten für sichere Bindung definiert, von dem jeder andere Bindungsstil behavioral abweicht. Die Quantifizierung der Abweichung kann schwierig sein und sollte auf dem Konsens mehrerer Bewertender beruhen und bedarf der Berücksichtigung kultureller Kontexte und bindungsspezifischer Situationen.

(4) *Implementierung nonverbaler Hinweise*. Jeder operationalisierte nonverbale Hinweis wurde mit Animationen in unserem Werkzeug Vuppetmaster abgeglichen, um die Modellierung des SIA-Verhaltens zu erleichtern. Auf diese Weise wurde eine strukturierte Tabelle erstellt, in der die Animation von *secure*, *preoccupied*, und *dismissing* Bindungsverhalten detailliert dargestellt ist.

(5) *Validierung*. Es muss validiert werden, dass Menschen mit z. B. einem dismissing Bindungsstil tatsächlich das modellierte nonverbale Verhalten zeigen. Dazu sollten Experten wie Psychotherapeuten herangezogen werden. Dann muss die Funktionalität der SIAs validiert und von Menschen mit der entsprechenden Bindung evaluiert werden.

Die ModellIT-Methode geht auf den wachsenden Bedarf an einer sorgfältigen Modellierung des nonverbalen SIA-Verhaltens, um eine barrierefreie und interkulturelle Welt zu fördern, ein. Sie berücksichtigt interpersonelle Unterschiede, um Verzerrungen zu minimieren. Darüber hinaus ist eine transparente Dokumentation ein wesentlicher Schritt zu einer offenen Wissenschaft und zugänglichen SIA-Anwendungen.

8.1.3. Synchronität mit der UBIDENZ-Agentin

In einer Studie untersuchten wir den Einfluss der Verkörperung einer virtuellen Agentin auf die Nutzerwahrnehmung und die durch die Herzfrequenz gemessene physiologische Synchronität. Die Teilnehmenden interagierten in einem Interview, das ein klassisches Depressionsinventar (Beck Depression Inventory, BDI-II) enthielt, mit der virtuellen Agentin Lydia in zwei Versionen: eine war ein verkörperter Konversationsagent (embodied conversational agent, ECA) mit einem pulsierenden Herzen und eine ein Sprachagent (voice agent, VA) mit einem pulsierenden Kreis.

Mittels Fragebogen-Selbstschemaschätzungen untersuchten wir in beiden Bedingungen folgende Konstrukte: wahrgenommenes Vertrauen, Beziehung und wahrgenommene Synchronität. Außerdem wurde die physiologische Synchronität durch Messung der Herzfrequenzsynchronisation zwischen Nutzenden und der Agentin während des BDI-Interviews untersucht. Darüber hinaus wurde der Einfluss depressiver Störungen und des Bindungsstils in Mensch-KI-Interaktionen untersucht. Die Annahme war, dass die ECA-Bedingung in der VA Bedingung überlegen ist.

Studiendesign und Set-up.

Zur Untersuchung der Fragestellung wurde eine Laborstudie mit einem randomisierten messwiederholten Design durchgeführt. Es gab zwei Bedingungen: Embodied Conversational Agent [ECA] und Voice Agent [VA]. In Bedingung 1 (ECA) durchliefen die Teilnehmenden das Depressionsinventar mit dem ECA mit einem sichtbaren Herzschlag in Form eines pulsierenden Herzens (Abbildung 2, links). In Bedingung 2 (VA) durchliefen die Teilnehmenden

das Depressionsinventar mit einem VA mit sichtbarem Herzschlag in Form eines pulsierenden Kreises (Abbildung 2, rechts). Die Hälfte der Teilnehmer wurde der Reihenfolge A zugeordnet und interagierte am ersten Messpunkt (MP) mit dem ECA und dann am zweiten MP mit dem VA. Die andere Hälfte befand sich in der Reihenfolge B und interagierte zuerst mit dem VA und dann mit dem ECA. Am Ende haben alle Teilnehmenden beide Bedingungen erfüllt. Der Fragebogen war für beide Bedingungen identisch.



Abbildung 2. Die beiden Agenten Bedingungen: links verkörperter ECA mit pulsierendem Herz, rechts Sprachassistent mit pulsierendem Kreis.

Das BDI-Interview folgte in beiden Bedingungen in einem festen Skript. Die Agentin wurde mit dem Visual SceneMaker (VSM, Gebhard et al., 2012) gesteuert, einem Tool, das das Verhalten externer Agenten über eine grafische Oberfläche automatisiert und für die Verwendung durch nicht-technische Domänenexperten konzipiert ist. VSM lief auf einem Android-Tablet, während der Spracherkennungsmechanismus mit Microsoft Azure implementiert wurde. Azure Speech Recognition, ein Teil der KI-Dienste von Microsoft Azure, bietet eine Reihe von fortschrittlichen Sprachfunktionen, darunter Sprache-zu-Text, Text-zu-Sprache,

Sprachübersetzung und Sprechererkennung. Der Dienst ermöglicht eine Echtzeit-Transkription, bei der gesprochene Sprache mit hoher Genauigkeit in geschriebenen Text umgewandelt wird, was besonders für Anwendungen wie Live-Meeting-Untertitel und Contact Center-Support nützlich ist. Die Anwendungsprogrammierschnittstelle Fast Transcription von Azure erweitert diese Funktionalität, indem sie eine schnelle, synchrone Transkription von Audiodateien ermöglicht, wodurch sie sich für die schnelle Erstellung von Untertiteln und Videoübersetzungen eignet. In der ECA-Bedingung hatte die Agentin ein zusätzliches Herz, das im Rhythmus einer HR-Aufzeichnung einer menschlichen Interviewerin schlug, die das BDI-Interview durchführte. Für jede Frage wurde eine mittlere Herzfrequenz berechnet ($HR M = 92,67, SD = 1,62,$). In der VA-Bedingung wurde ein pulsierender Kreis in der gleichen Größe und Geschwindigkeit angezeigt. Die Agentin variierte weder ihre Körpersprache noch ihr Skript mit den Teilnehmern. Die Agentin wurde auf einem Galaxy Tab S7 präsentiert. Das Tablet mit dem Agenten war auf einem Schreibtisch montiert und die Teilnehmenden saßen 65 cm vom Tablet entfernt (siehe Abbildung 3 zum Aufbau).

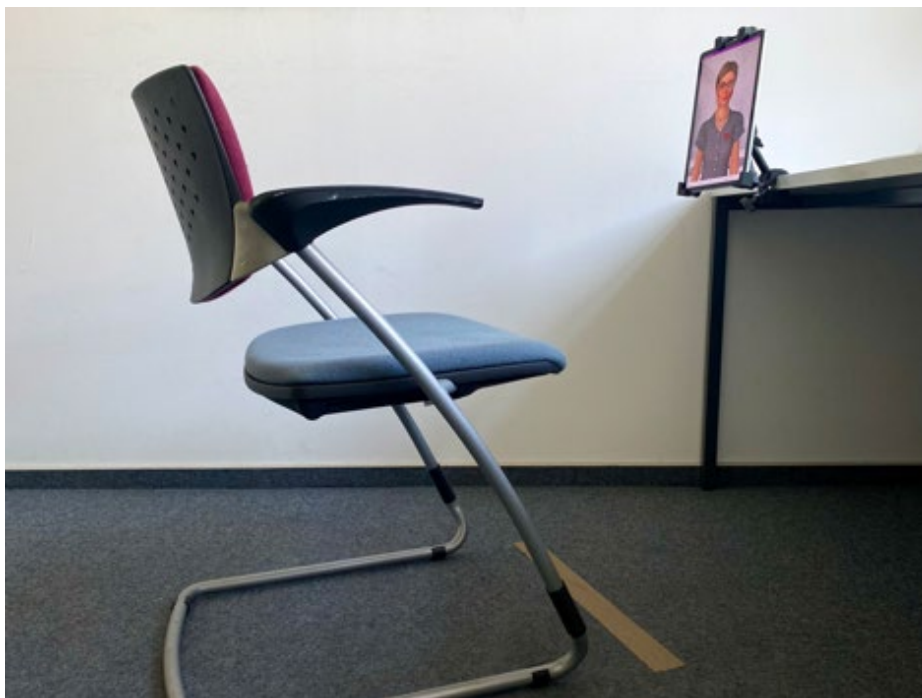


Abbildung 3. Aufbau.

Verwendete Maße.

Zu den demografischen Daten gehörten Alter, Nationalität, höchster Bildungsabschluss, Beruf, bestehende Herzerkrankungen, aktuelle Stimmung (alle nur bei Termin 1 gemessen) und tragische Ereignisse in der letzten Woche (nur bei Termin 2 gemessen).

Das Konsumverhalten umfasste Koffein-, Alkohol- und Drogenkonsum in den letzten Stunden.

Die interozeptive Wahrnehmung wurde vor dem Interview mit dem Agenten an Termin 1 mit der deutschen Version des Multidimensional Assessment of Interoceptive Awareness (MAIA-2; Mehling et al., 2018) gemessen. Er enthält 37 Items, aus denen sich acht Subskalen Wahrnehmen, Nicht-Ablenken, Nicht-Sorgen, Aufmerksamkeitsregulation, Emotionales Bewusstsein, Selbstregulation, Körperhören und Vertrauen sowie ein Gesamtscore ergeben.

Der Bindungsstil wurde vor dem Interview mit dem Agenten bei Termin 1 mit der deutschen Version des Experiences in Close Relationships-Revised Questionnaire (ECR-RD8; Ehrenthal et al., 2021) gemessen.

Die Erfahrung mit virtuellen Agenten wurde vor dem Interview bei Termin 1 mit vier selbst erstellten Items erhoben: Ich habe eine positive Einstellung zu virtuellen Agenten (wie Alexa oder Siri); ich arbeite gerne mit virtuellen Agenten (wie Alexa oder Siri); ich habe Erfahrung in der Arbeit mit virtuellen Agenten (wie Alexa oder Siri); und ich habe bereits Erfahrung in der Arbeit mit Meditations-Apps (wie Headspace).

Der BDI-II-Score wurde mit der deutschen Version (Hautzinger et al., 2009) des Beck'schen Depressionsinventars (BDI-II, Beck et al., 2006) bei beiden Terminen in dem Interview mit Lydia gemessen. Das BDI-II besteht aus 21 Fragen zur Bewertung von Depressionssymptomen wie Versagensangst, Schuldgefühlen oder Interessenverlust mit vier bis sechs vorgegebenen Antworten, wobei die Teilnehmenden die Antwort wählen mussten, die am besten auf sie für den Zeitraum der letzten zwei Wochen passte.

Das wahrgenommene Vertrauen wurde nach dem Interview mit einer angepassten Skala von Hale und Hamilton (2016) gemessen. Die fünf Items lauteten: Lydia ist sehr vertrauenswürdig; Lydia ist sehr unzuverlässig; Lydia ist unaufrichtig; Lydia ist sehr ehrlich; und Lydia ist verantwortungsbewusst.

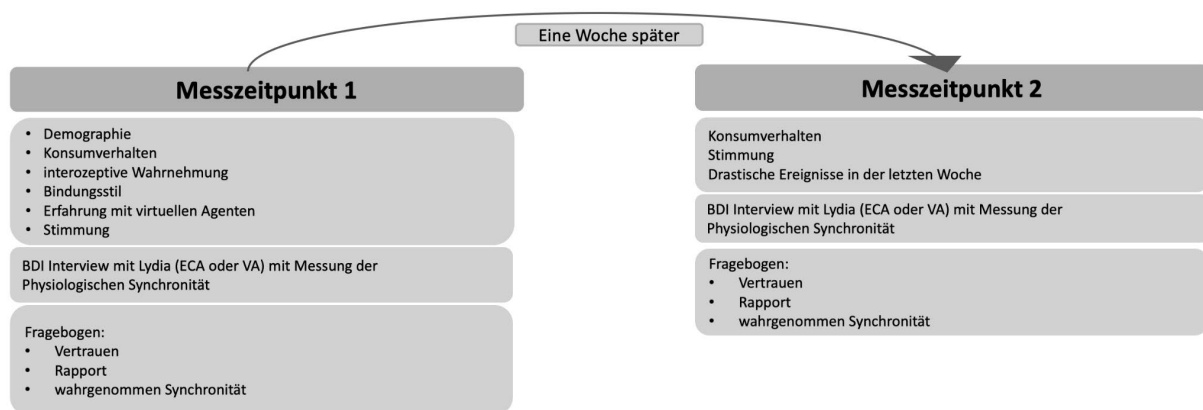
Der wahrgenommene Rapport wurde nach dem Interview bei beiden Terminen mit einer adaptierten Skala von Lucas und Kollegen (2018) gemessen. Die neun Items waren: Lydia hat ein Gefühl der Verbundenheit oder Kameradschaft zwischen uns geschaffen; Lydia hat ein

Gefühl der Distanz zwischen uns geschaffen; ich denke, Lydia und ich haben uns verstanden; Lydia hat eher Kälte als Wärme vermittelt; Lydia war warm und fürsorglich; ich wollte ein Gefühl der Distanz zwischen uns aufrechterhalten; ich hatte das Gefühl, eine Verbindung zu Lydia zu haben; Lydia war mir gegenüber respektvoll; ich hatte das Gefühl, keine Verbindung zu Lydia zu haben.

Die wahrgenommene Synchronität wurde nach dem Gespräch an beiden Terminen anhand einer selbst erstellten Skala mit sieben Items gemessen: Haben Sie Lydia gemocht?; Wie sehr haben Sie Lydia wahrgenommen?; Hatten Sie ein warmes Gefühl gegenüber Lydia?; Hatten Sie das Gefühl, dass die Interaktion einen angenehmen Rhythmus hatte?; Hatten Sie das Gefühl, dass Ihre und Lydias Handlungen und Äußerungen gut aufeinander abgestimmt waren?; Waren Sie an dem interessiert, was Lydia sagte und tat?; Fühlten Sie sich mit Lydia verbunden?

Die physiologische Synchronität wurde über die Herzfrequenz mit einem Polar H9 Herzfrequenzsensor an beiden Terminen während des Interviews mit Lydia gemessen. Die physiologische Synchronität wurde mit dem SUSY-Ansatz (Tschacher & Haken, 2019) bestimmt und zu einem durchschnittlichen z-Wert für Synchronität aggregiert.

Ablauf.



Stichprobe.

Um an der Studie teilzunehmen, mussten die Teilnehmenden über 18 Jahre alt sein, weiblich, fließend Deutsch sprechen und durften keine Erkrankungen haben, die ihren Herzschlag beeinträchtigen könnten. Aufgrund möglicher geschlechtsspezifischer Unterschiede in der Ruheherzfrequenz (HR) und einer erwarteten weiblich dominierten Stichprobe wurden nur

weibliche Teilnehmer untersucht. Die Datenerhebung fand zwischen dem 11.06.2024 und dem 09.08.2024 statt. Insgesamt wurden die Daten von 37 Frauen erhoben. Die Teilnehmerinnen waren gleichmäßig auf die beiden Reihenfolgen verteilt (Reihenfolge A 18 Teilnehmerinnen und Reihenfolge B 19 Teilnehmerinnen). Die Teilnehmerinnen waren zwischen 18 und 41 Jahre alt ($M = 24$ Jahre, $SD = 4,62$ Jahre) und wurden hauptsächlich über Online-Psychologie-Studentengruppen und Freunde rekrutiert, wobei 28 Teilnehmerinnen Universitätsstudentinnen waren. Die Manipulationsprüfung ergab, dass 11 Teilnehmer den pulsierenden Kreis als Herz wahrnahmen, während 26 dies nicht taten. Die Psychologiestudentinnen wurden für ihre Teilnahme mit Studienleistungen belohnt und konnten an einer Verlosung von drei Eintrittskarten für einen Outdoor-Kletterpark teilnehmen. Vor der Teilnahme wurden alle Teilnehmenden über den Inhalt der Studie und die Unterschiede zwischen den beiden Bedingungen informiert. Alle gaben ihr Einverständnis zur Datenerhebung und Veröffentlichung.

Ergebnisse.

Tabelle 1

Deskriptive Daten für die abhängigen Variablen in beiden Bedingungen.

Variablen	ECA			VA		
	<i>M</i> (<i>SD</i>)	Range	95 % CI	<i>M</i> (<i>SD</i>)	Range	95 % CI
Vertrauen	5.18 (0.83)	3.6 – 6.8	[4.90, 5.46]	4.95 (0.86)	3.2 – 7.0	[4.66, 5.23]
Rapport	2.96 (0.71)	2.00 – 4.22	[2.72, 3.19]	2.73 (0.69)	1.67 – 4.44	[2.50, 2.96]
Wahrgen. Synch.	4.31 (1.12)	1.86 – 6.71	[3.94, 4.68]	3.91 (1.05)	1.71 – 5.86	[3.56, 4.26]
Physiolog. Synch.	0.23 (0.06)	0.14 – 0.41	[0.21, 0.25]	0.24 (0.05)	0.13 – 0.36	[0.22, 0.25]

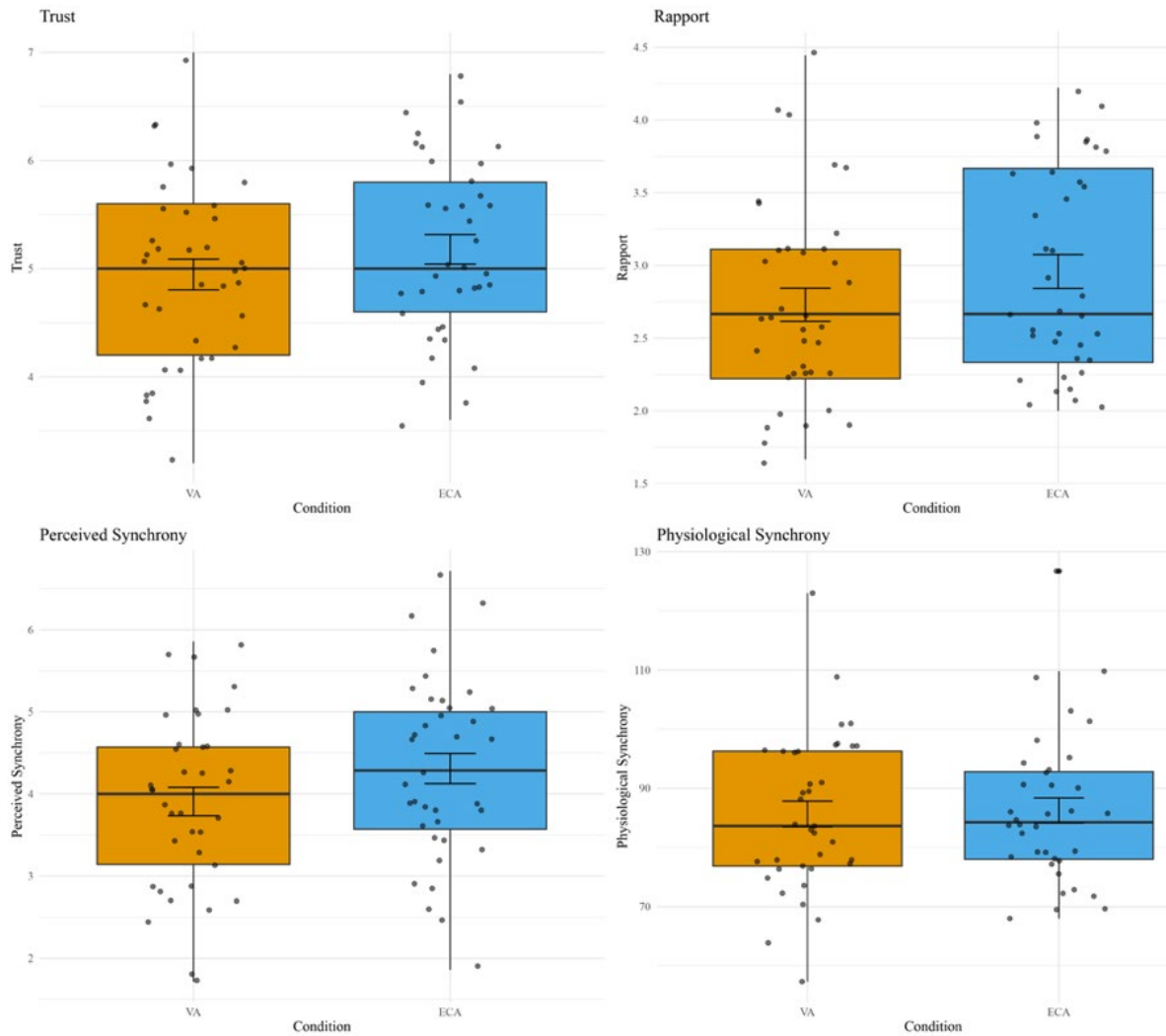


Abbildung 4. Boxplots der abhängigen Variablen.

Vertrauen: Die multiple Regression mit der Bedingung, dem BDI-II-Wert, der BS-Vermeidung und der BS-Angst als unabhängige Variablen und dem wahrgenommenen Vertrauen als abhängige Variable ergab, dass die Bedingung keinen signifikanten Einfluss auf das wahrgenommene Vertrauen hatte. Auch der BDI-II-Score zeigte keinen signifikanten Einfluss auf das wahrgenommene Vertrauen. Die BS-Angst hatte zwar keinen signifikanten Einfluss, wohl aber die BS-Vermeidung. Teilnehmende mit einer höheren BS-Vermeidung gaben ein geringeres wahrgenommenes Vertrauen an (siehe Tabelle 2). Das Modell erklärte 11 % der Varianz des wahrgenommenen Vertrauens (marginale $R_m^2 = 0,11$; bedingtes $R_c^2 = 0,45$), was darauf hindeutet, dass feste Effekte einen kleinen Teil der Varianz erklären, während zufällige Effekte einen größeren Beitrag leisten.

Tabelle 2. Ergebnisse des linearen gemischten Modells für wahrgenommenes Vertrauen

Fixed Effects	Estimate (B)	Standardized Coefficient (β)	95% CI	df	t-Wert	p-Wert
Bedingung	0.12	0.14	[-0.03, 0.27]	36	1.53	.134
BDI-II	0.05	0.06	[-0.18, 0.29]	33	0.43	.673
BS Vermeid.	-0.27	-0.32	[-0.49, -0.05]	33	-2.29	.029*
BS Angst	-0.02	-0.02	[-0.26, 0.23]	33	-0.12	.907

Note. BDI-II, BS-Vermeidung und BS-Angst waren z- standardisiert. $N = 37$. * $p < .05$; ** $p < .01$; *** $p < .001$. Alle p -Werte zweiseitig.

Rapport: Für den wahrgenommenen Rapport ergab die multiple Regression mit der Bedingung, dem BDI-II-Wert, der BS-Vermeidung und der BS-Angst als unabhängige Variablen und dem wahrgenommenen Rapport als abhängige Variable, dass die Bedingung einen signifikanten Einfluss auf den wahrgenommenen Rapport hatte. Insbesondere waren die Rapportbewertungen in der ECA-Bedingung signifikant höher als in der VA-Bedingung. Der BDI-II-Wert hatte keinen signifikanten Einfluss. Außerdem hatten BS-Angst und BS-Vermeidung keinen signifikanten Einfluss auf den wahrgenommenen Rapport (siehe Tabelle 3). 3 % der Varianz im wahrgenommenen Rapport wurden durch das Modell erklärt (marginale $R_m^2 = 0,03$; bedingtes $R_c^2 = 0,56$), was darauf hindeutet, dass feste Effekte einen kleinen Teil der Varianz erklären, während zufällige Effekte einen größeren Beitrag leisten.

Tabelle 3. Ergebnisse des linearen gemischten Modells für Rapport

Fixed Effects	Estimate (B)	Standardized Coefficient (β)	95% CI	df	t-Wert	p-Wert
Bedingung	0.11	0.16	[0.00, 0.23]	36	2.03	.0496*
BDI-II	-0.06	-0.08	[-0.28, 0.16]	33	-0.52	.607
BS Vermeid.	-0.05	-0.07	[-0.26, 0.16]	33	-0.47	.645

BS Angst	0.03	0.04	[-0.19, 0.25]	33	0.25	.808
----------	------	------	---------------	----	------	------

Note. BDI-II, BS-Vermeidung und BS-Angst waren z- standardisiert. $N = 37$. * $p < .05$; ** $p < .01$; *** $p < .001$. Alle p -Werte zweiseitig.

Wahrgenommene Synchronität: Für die wahrgenommene Synchronität zeigte die multiple Regression mit der Bedingung, dem BDI-II-Score, der BS-Vermeidung und der BS-Angst als unabhängige Variablen und der wahrgenommenen Synchronität als abhängige Variable, dass die Bedingung einen signifikanten Einfluss auf die wahrgenommene Synchronität hat. In der ECA-Bedingung waren die Bewertungen der wahrgenommenen Synchronität signifikant höher als in der VA-Bedingung. Der BDI-II-Score hatte keinen signifikanten Einfluss auf die wahrgenommene Synchronität. Außerdem hatten weder BS-Angst noch BS-Vermeidung einen signifikanten Einfluss (siehe Tabelle 4). Das Modell erklärte 6 % der Varianz in der wahrgenommenen Synchronität (marginale $R_m^2 = 0,06$; bedingtes marginale $R_c^2 = 0,60$), was darauf hindeutet, dass feste Effekte einen kleinen Teil der Varianz ausmachten, während zufällige Effekte einen größeren Beitrag leisteten.

Tabelle 4. Ergebnisse des linearen gemischten Modells für wahrgenommene Synchronität

Fixed Effects	Estimate (B)	Standardized Coefficient (β)	95% CI	df	t-Wert	p-Wert
Bedingung	0.20	0.18	[0.03, 0.37]	36	2.38	.023*
BDI-II	0.01	0.01	[-0.32, 0.35]	33	0.08	.936
BS Vermeid.	-0.13	-0.12	[-0.45, 0.19]	33	-0.76	.453
BS Angst	-0.12	-0.11	[-0.47, 0.23]	33	-0.66	.512

Note. BDI-II, BS-Vermeidung und BS-Angst waren z- standardisiert. $N = 37$. * $p < .05$; ** $p < .01$; *** $p < .001$. Alle p -Werte zweiseitig.

Physiologische Synchronität: Die multiple Regression mit der Bedingung, dem BDI-II-Score, der BS-Vermeidung und der BS-Angst als unabhängige Variablen und der physiologischen Synchronität als abhängige Variable ergab, dass die Bedingung keinen signifikanten Einfluss auf die physiologische Synchronität hatte. Allerdings zeigte der BDI-II-Wert einen signifikanten

Einfluss auf die physiologische Synchronität, wobei Teilnehmer mit niedrigeren BDI-II-Werten eine höhere physiologische Synchronität aufwiesen. Sowohl BS-Angst als auch BS-Vermeidung hatten keinen signifikanten Einfluss (siehe Tabelle 5 und Abbildung X). 12 % der Varianz in der physiologischen Synchronität wurden durch das Modell erklärt (marginale $R_m^2 = 0,12$; bedingtes $R_c^2 = 0,24$), was darauf hindeutet, dass feste Effekte nur einen kleinen Teil der Varianz ausmachten, während zufällige Effekte einen größeren Anteil hatten.

Tabelle 5. Ergebnisse des linearen gemischten Modells für physiologische Synchronität

Fixed Effects	Estimate (B)	Standardized Coefficient (β)	95% CI	df	t-Wert	p-Wert
Bedingung	-0.01	-0.12	[-0.02, 0.00]	34	-1.15	.260
BDI-II	-0.02	-0.33	[-0.03, -0.00]	31	-2.41	.022*
BS Vermeid.	-0.01	-0.11	[-0.02, 0.01]	31	-0.89	.381
BS Angst	0.01	0.23	[-0.00, 0.03]	31	1.65	.109

Note. BDI-II, BS-Vermeidung und BS-Angst waren z- standardisiert. $N = 37$. * $p < .05$; ** $p < .01$; *** $p < .001$. Alle p -Werte zweiseitig.

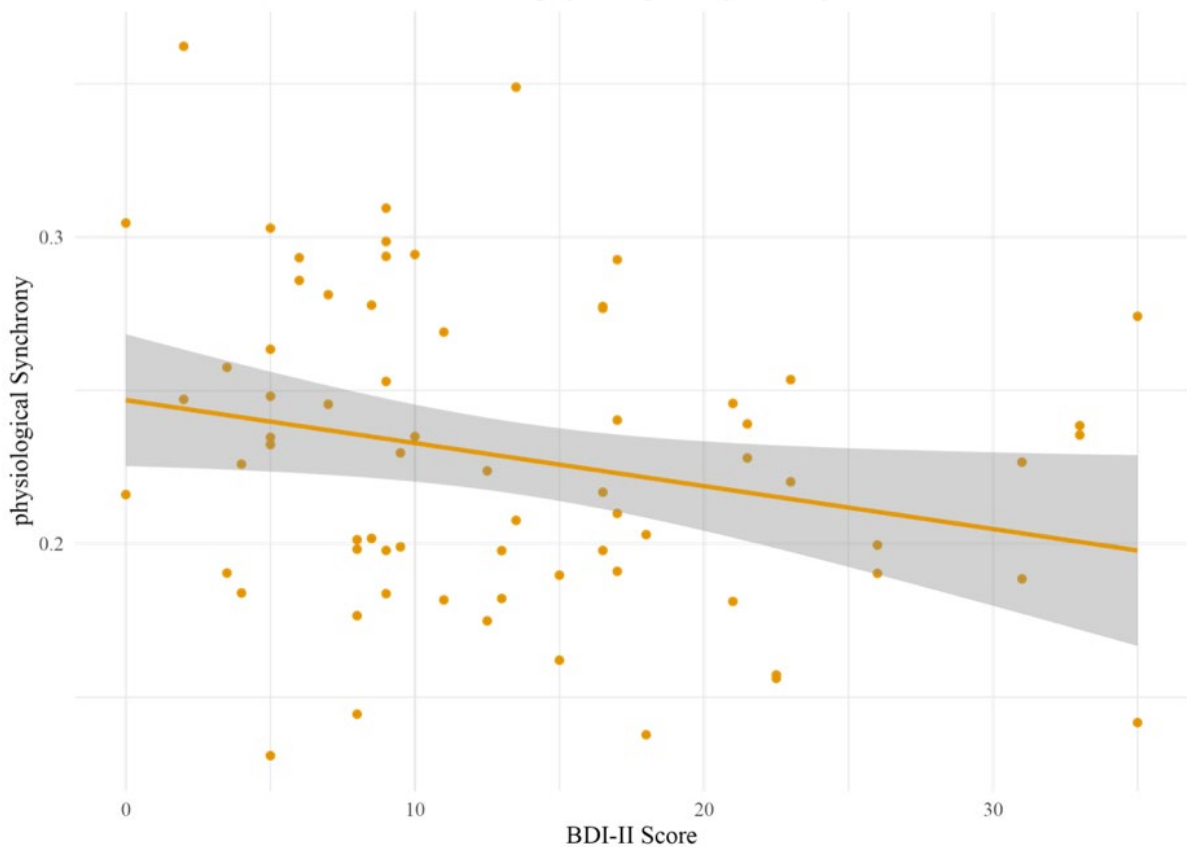


Abbildung 5. Korrelation zwischen BDI-Wert und physiologischer Synchronität

Zusammenfassung.

Die Ergebnisse zeigten, dass die wahrgenommene Beziehung und Synchronität in der ECA-Bedingung signifikant höher war als in der VA-Bedingung, was darauf hindeutet, dass die Verkörperung die Bewertung durch Nutzende positiv beeinflusst. Allerdings hatte die Verkörperung keinen signifikanten Einfluss auf das Vertrauen, unabhängig von der experimentellen Manipulation. Stattdessen wurde festgestellt, dass der Bindungsstil, insbesondere der vermeidende Bindungsstil, das wahrgenommene Vertrauen signifikant reduziert. Bei der physiologischen Synchronität ergaben sich keine signifikanten Unterschiede zwischen den Bedingungen, jedoch korrelierten höhere Depressionswerte mit einer geringeren physiologischen Synchronität. Diese Ergebnisse unterstreichen, wie wichtig es ist, virtuelle Agenten zu entwerfen, die sowohl die Verkörperung als auch die individuellen Eigenschaften der Nutzenden berücksichtigen, um die Bewertung der Nutzer zu verbessern und die physiologische Synchronität zu erhöhen.

8.2 Technische Ergebnisse

Aus technischer Sicht stellt die Entwicklung natürlicher nonverbaler reziproker Echtzeit Reaktionen bei Avataren oder sozial-interaktiven Agenten (SIAs) eine Herausforderung dar, da bisherige regelbasierte Ansätze die Dynamik menschlicher Interaktionen oft nicht widerspiegeln. Ziel neben der Konzeption ist es, Methoden des maschinellen Lernens einzusetzen, um damit Computer-Verhaltensmodelle zu entwickeln, die eine Simulation von natürlichen und professionellen Interaktionsverhalten auf der Seite von Avataren/SIAs ermöglichen.

Die Ergebnisse sind in einer neuartigen Bindungssimulation-Softwarekomponente (BISI) mit einer Kopplung an eine Analysekomponente sozialer Signale realisiert. Dies umfasst eine notwendige Erweiterung typischer Benutzer- und Avatarverhaltensmodelle. BISI erweitert bestehende Benutzer- und Interaktionsmodelle auf mehrere bedeutenden Arten, indem es typische Limitierungen traditioneller Ansätze überwindet und neue Funktionalitäten bietet. Dafür wurden reale domänenspezifische Trainingsdaten (Psychotherapie-Sitzungen) genutzt, die reich an sozialem Verhalten und nonverbalen Signalen sind. Das BISI-Modell ermöglicht so domänenspezifische Anpassungen, z. B. für therapeutische, pädagogische oder kundenorientierte Anwendungen.



Abbildung 6: Trainingsdaten (Patient-Therapeut)

Die technische BISI-Komponente ist für den Einsatz mit Avatern oder sozial-interaktiven Agenten (SIAs) konzipiert. Zusammen mit der Kopplung einer Analysekomponente sozialer Signale ermöglicht es eine Echtzeit-Darstellung der generierten Verhaltensmuster auf photorealistischen Avataren (z. B. MetaHuman von Unreal¹). Zudem wurden Softwareschnittstellen geschaffen, die Entwickler erlauben, BISI für andere Avatar- oder Agentenumgebungen zu integrieren. Die volle Funktionalität wurde im UBIDENZ-Demonstratorsystem realisiert.

Die gesamte in UBIDENZ dazu entwickelte Software ist für wissenschaftliche Zwecke frei-verfügbar im öffentlichen GIT-Repository²

8.2.1 Datenkorpus

Der für das Trainieren des SIA-Verhaltenmodells ist ein Videokorpus mit Gesprächen von Patient:innen und Therapeut:innen genutzt worden. Alle Daten sind nach den Richtlinien der DSGVO mit dem DFKI-eigenen Datenverarbeitungsframework für sensitive Daten SEMLA³ verarbeitet worden. Der Korpus enthält 138 Videos mit Patient:innen im Alter von 18 bis 57 Jahren ($M = 32,04$, $SD = 11,99$), die für eine Forschungsstudie über emotionalen und interpersonellen Prozessen bei psychischen Störungen ausgewählt wurden [Bock et al. 16]. Die Interviews wurden von vier geschulten und zertifizierten Interviewern (zwei Männer und zwei Frauen) durchgeführt (siehe Abbildung 7).

¹ <https://www.unrealengine.com/en-US/metahuman>

² <https://daksitha.github.io/ReNeLib/>

³ <https://semla.dfki.de>



Abbildung 7: Organisation der Annotation im JSON-Format

Im Rahmen der Erstellung der Trainingsdaten aus dem Videokorpus wurde eine JSON-Datei mit Annotationen zu den Sitzungsdetails für jedes Interview erstellt, wie in Abbildung 7 dargestellt. In der Folge wurden zwei Kriterien für die Aufteilung der Daten herangezogen. In einem ersten Schritt wurde der Datensatz in vier Untergruppen unterteilt, wobei die Zuordnung der Daten nach dem leitenden Therapeuten erfolgte. Diese Aufteilung ermöglichte eine detaillierte Untersuchung der subtilen Verhaltensmuster der einzelnen Therapeuten. Die Aufteilung der Videos erfolgte zweitens auf der Grundlage der Sitzposition des Therapeuten im Raum (links=0, rechts=1), wie in Abbildung 6 dargestellt. Diese Aufteilung ermöglichte die Analyse des potenziellen Einflusses der Position des Therapeuten auf das erlernte Verhalten. Das war notwendig, da das Avatarverhalten später davon unabhängig sein sollte.

8.2.2 Erweiterung Benutzermodelle

Die Integration von multimodalen Eingaben erfolgt in der Vergangenheit meist regelbasiert, wobei die Verarbeitung einzelner Modalitäten wie Sprache oder Gesichtsausdrücke begrenzt ist. Mit BISI werden audio-visuelle Eingaben wie Sprachmelodien und Kopfbewegungen gleichzeitig in Echtzeit erfasst und verarbeitet. Dies ermöglicht eine kontextsensitivere Interpretation des Nutzerverhaltens durch die Verknüpfung von Audio- und visuellen

Signalen. Zunächst wurde eine Analyse der Daten vorgenommen. Das primäre Ziel der Analyse bestand in der Trennung der Videoströme von Patient und Therapeut. Dabei wurden die jeweiligen Audiokanäle in individuelle Kanäle aufgeteilt (cf. Sprecheridentifikation). Eine weitere Verarbeitung der Audiokanäle war erforderlich, um die beiden Sprecher effektiv zu isolieren. Des Weiteren stellte die Notwendigkeit der Trennung der Audiokanäle eine signifikante Herausforderung bei der Identifizierung von Sprach- und Nicht-Sprach-Segmenten während des Interviews dar. Es war von entscheidender Bedeutung, die Momente zu identifizieren, in denen der Therapeut den Ausführungen des Patienten zuhörte, um die relevanten Abschnitte voneinander abzugrenzen.

Audio

Das Problem der Trennung der Audiokanäle von Patient und Therapeut wurde in den durchgeführten Interviews durch den Einsatz mit dem Open-Source Werkzeug "pyanote-audio" (Bredin et al. 20) gelöst. Damit wurden eine Reihe von Merkmalen aus dem Audiosignal extrahiert, die für die Unterscheidung zwischen verschiedenen Sprechern relevant sind. Zu den erfassten Merkmalen zählen Eigenschaften der Stimme des Sprechers, wie Tonhöhe, spectral content und Energie, sowie Kontextinformationen über die Umgebung, in der das Audiosignal aufgenommen wurde. Im Anschluss erfolgt eine Gruppierung der Audiosegmente, wobei die Segmente in Gruppen zusammengefasst werden, die mit hoher Wahrscheinlichkeit von demselben Sprecher stammen. Dazu wird der Schlüsselwert "num_speakers" verwendet, welcher die Anzahl der an der Interaktion beteiligten Personen angibt. Die RTTM-Datei (Rich transcription time marked) definiert die Audiosegmente, welche den einzelnen Sprechern zugeordnet sind. Die Abbildung 8 veranschaulicht die Zeitintervalle sowie die involvierten Sprecher.

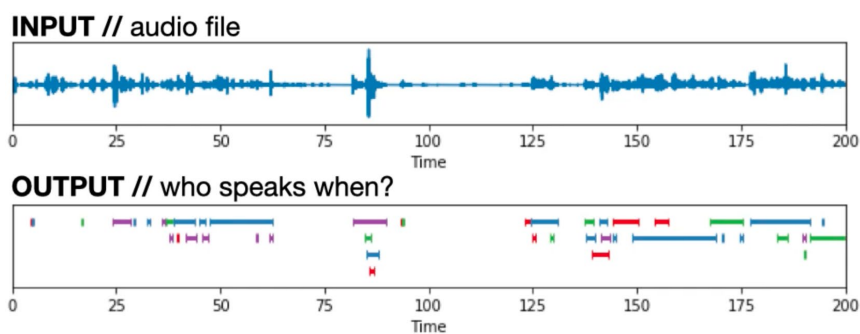


Abbildung 8: Graphische Darstellung der Sprecheridentifikation.

Um eine genauere Separierung der Sprachbeiträge von Patient:innen und Therapeut:innen zu realisieren, waren noch weitere Schritte notwendig, die in dem Übersichtsartikel (Withanage Don et al. 23) eingesehen werden können.

Zum Trainieren und für die Anwendung des Modells wurden für die Verhaltensprädiktion aus den Audiodaten der Patient:innen MFCC-Informationen berechnet. Mel-Frequenz-Cepstral-Koeffizienten (MFCC) sind gängige Sprachmerkmale für die Sprach- und Audioverarbeitung. Sie erfassen die spektralen Eigenschaften von Audiosignalen und werden üblicherweise zur Darstellung der spektralen Hüllkurve eines Tons verwendet.

Die generative Modellierung mit bedingter Bewegungssynthese bildet die Grundlage für die Anwendung von MFCC als Eingabe für das in BISI arbeitende maschinelle Lernmodell, welches den generierten Output bestimmt. Das Modell wird trainiert, um einen Satz von MFCC- und zusätzlichen konditionierenden Merkmalen auf die entsprechenden Bewegungsdaten zuzuordnen. Die Konditionierung ermöglicht es dem Modell, Bewegungen zu synthetisieren, die sowohl realistisch als auch kontextabhängig sind. Letzteres basiert auf den zusätzlichen Konditionierungsmerkmalen. Das Modell kann beispielsweise so trainiert werden, dass es Bewegungsdaten erzeugt, die für einen bestimmten Sprecher oder einen bestimmten Sprachstil zu erstellen sind oder den emotionalen Inhalt der Rede widerspiegeln. Die Verwendung von MFCC und weiteren Konditionierungsmerkmalen als Eingaben ermöglicht es dem Modell, Bewegungsdaten zu synthetisieren, die inkohärenten und unnatürlichen Audiosignalen anhaften, und stattdessen Kohärenz und Natürlichkeit zu erzeugen.

Kopfbewegungen

Eine weitere Eingabe zum Trainieren und für die Anwendung des Modells sind die Kopfbewegungen der Patient:innen. Zu deren Analyse wurde das frei verfügbare EMOCA-Framework verwendet. EMOCA ist in der Lage, 3D-Gesichtsmodelle mittels Regressionsverfahren aus 2D-Bildern zu generieren. Dies bedeutet einen Übergang von der pixelbasierten Synthese zur generativen 3D-Gesichtsmodellierung. Die zuvor eingesetzten Methoden, welche auf der Kombination von LFI mit RingNet (Sanyal et al. 19) und L2L mit DECA (Feng et al. 20) basieren, weisen eine eingeschränkte Fähigkeit zur Regression von 3D-Gesichtern aus Bildern auf. Dies ist darauf zurückzuführen, dass sie die gesamte Bandbreite komplexer mimischer Ausdrucksformen, einschließlich subtiler oder extremer kommunikativer Emotionen, nicht adäquat erfassen können. Im Gegensatz dazu ist EMOCA in

der Lage, parametrische 3D-Gesichtsmodelle aus monokularen Bildern zu regressieren. Das System wurde speziell darauf trainiert, die gesamte Bandbreite an Gesichtsausdrücken zu erfassen, wobei der Schwerpunkt auf der genauen Erfassung des kommunikativen, emotionalen Inhalts des Gesichts liegt. Des Weiteren wurde der Versuch unternommen, die Regression von 3D-Gesichtern aus 2D-Bildern mit RingNet zu rekonstruieren. Diese Methode wies jedoch eine hohe Komplexität sowie eine geringe Effizienz bei der Extraktion der erforderlichen Parameter auf. Daher wurde stattdessen die EMOCA-Methode verwendet, die sich als effizienter und schneller bei der Extraktion von Gesichtsausdrücken und Pose aus den Videobildern erwies.

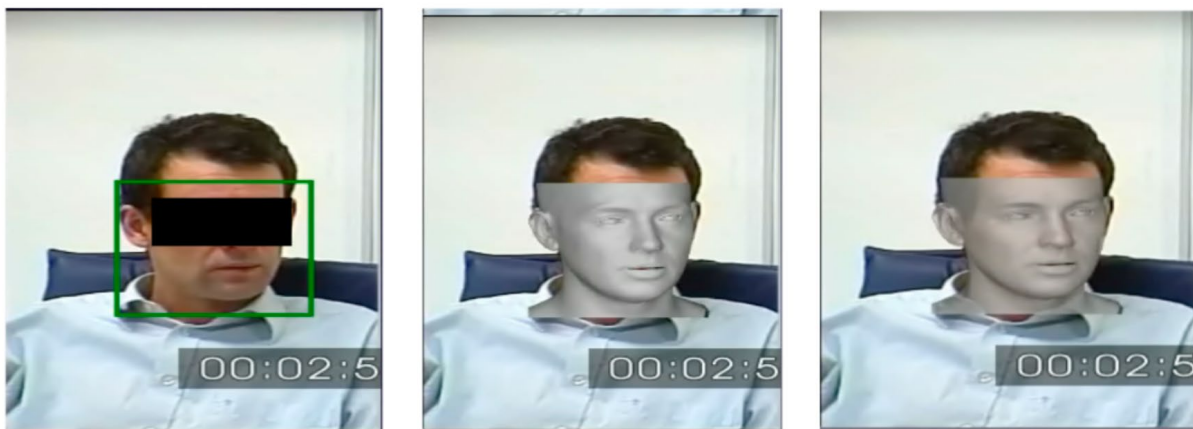


Abbildung 9: Prozess der Regression von 3D-Gesichtern aus 2D-Bildern mit EMOCA - linkserkannte Gesicht mit Orientierungspunkte; mitte: EMOCA generiertes FLAME-Netz; rechts detailliertes 3D-Netzes.

Die Interviewvideos wiesen Störaspekte wie Zoom-Effekte, Änderungen der Fenstergröße sowie Momente, in denen der Therapeut den Raum verließ, auf. Zur Lösung dieser Probleme wurde eine handelsübliche Methode zur Gesichtserkennung in Verbindung mit EMOCA (s. o.) eingesetzt, um automatisch alle störungsfreien Segmente der beiden sich unterhaltenden Personen zu lokalisieren (z. B. Entfernen von Abschnitten, in denen keine Gesichter erkannt wurden). Des Weiteren wurde das Rauschen in den Pseudo-Ground-Truth-Annotationen reduziert, indem Frames entfernt wurden, in denen die meisten der von EMOCA geschätzten 2D-Schlüsselpunkte fehlten, sowie durch das Ausschließen von Sequenzen mit plötzlichen, extremen Bewegungen.

Weitere notwendige Bearbeitungsschritte können dem Übersichtsartikel (Withanage Don et al. 23) entnommen werden.

8.2.3 Erweiterung bindungsorientiertes Avatarverhaltensmodell

Typische Avatarverhaltensmodelle verwenden vorgefertigte oder statische Animationen, die unabhängig vom Nutzendenverhalten ablaufen. Mit BISI wird durch den Einsatz neuronaler Netze das Avatarverhalten dynamisch an den Kontext angepasst. Dabei werden nonverbale Verhaltensmuster wie Nicken oder Gesichtsausdrücke automatisch generiert, die sich an dem emotionalen und sprachlichen Kontext des Nutzers orientieren. Durch autoregressive Modellierung werden vergangene Interaktionen berücksichtigt, was zu flüssigeren und konsistenteren Reaktionen führt.

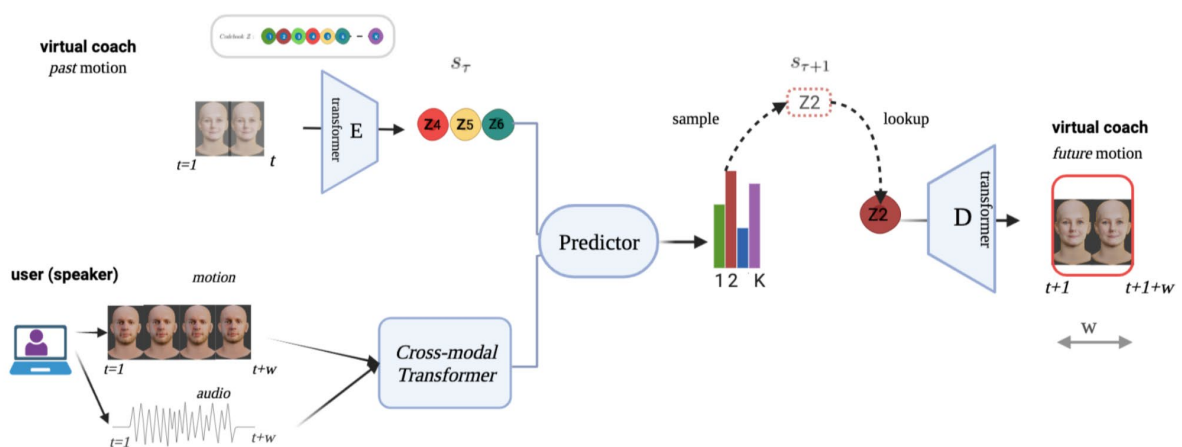


Abbildung 9: Überblick des UBIDENZ bindungsorientierten Verhaltensprediktionsmodell.

Die Realisierung des bindungsorientierten Verhaltensprediktionsmodell (Abbildung 9) basiert auf einer Variante des Full-Attention Cross-modal Transformers (FACT) Anwendung, welche ein transformatorbasiertes Prädiktormodul beinhaltet. Dieses dient der Erfassung von Korrelationen im langen Bereich sowie von Korrelationen in den Eingabedaten. Das Modell, das auf der Arbeit von Ruilong Li et al. (FACT, Li et al. 21) basiert, ist in der Lage, 3D-Tanzbewegungen in Abhängigkeit von der Musik zu generieren. Diese wurden von Ng et al. (2022) für die Erfassung von Kopfbewegungen übernommen. Des Weiteren ist der Ansatz in der Lage, mehrere Ausgabemodi zu erfassen, was für die UBIDENZ-Avatare notwendig ist. Dies erfolgt durch die Verteilung der möglichen Folgebewegungen unter Verwendung einer diskreten latenten Code-Darstellung. Zudem wird durch den Einsatz von Cross-Attention die Verarbeitung multimodaler Eingaben ermöglicht. Das Prädiktormodul verarbeitet die multimodale Einbettung der Gesichtsbewegung und die Sprache des Patienten (siehe Kapitel 8.2.2). Zudem werden die zuvor vorhergesagten Therapeutenbewegungen als Eingabe

verwendet. Eine genauere Beschreibung der einzelnen Unterarbeitsschritte ist im Übersichtsartikel (Withanage Don et al. 23) zu finden.

8.2.4 Das BISI Software-Framework

Um das bindungsorientierte Avatarverhaltensmodell technisch zu anwendbar zu machen, bedarf es eines Softwareframeworks. Das BISI ist ein Echtzeit-Framework zur Generierung von automatischen - bindungsorientierten - Zuhörverhalten für sozial interaktive Agenten (SIAs) auf Basis von multimodalen Eingabedaten (wie Gesichtsausdrücke und Audio) (Kapitel 8.2.3). Der Fokus liegt darauf, die natürliche und kontextsensitive Interaktion zwischen Mensch und Maschine zu verbessern, insbesondere durch die Reaktion von virtuellen Agenten auf menschliches Verhalten.

BISI hat drei Hauptfunktionen: 1. Die Echtzeit-Merkmalsextraktion ermöglicht die Erfassung und Verarbeitung von Gesichtsausdrücken mithilfe des FLAME-Modells und von MFCC-Audiofeatures, 2. Für die Verhaltensgenerierung werden neuronale Netze eingesetzt, um Gesichtsausdrücke basierend auf multimodalen Benutzerdaten vorherzusagen, und 3. Die Visualisierung integriert die generierten Verhaltensmuster in 3D-Modelle, die sowohl FLAME- als auch ARKit-basierte Avatare/SIAs unterstützen.

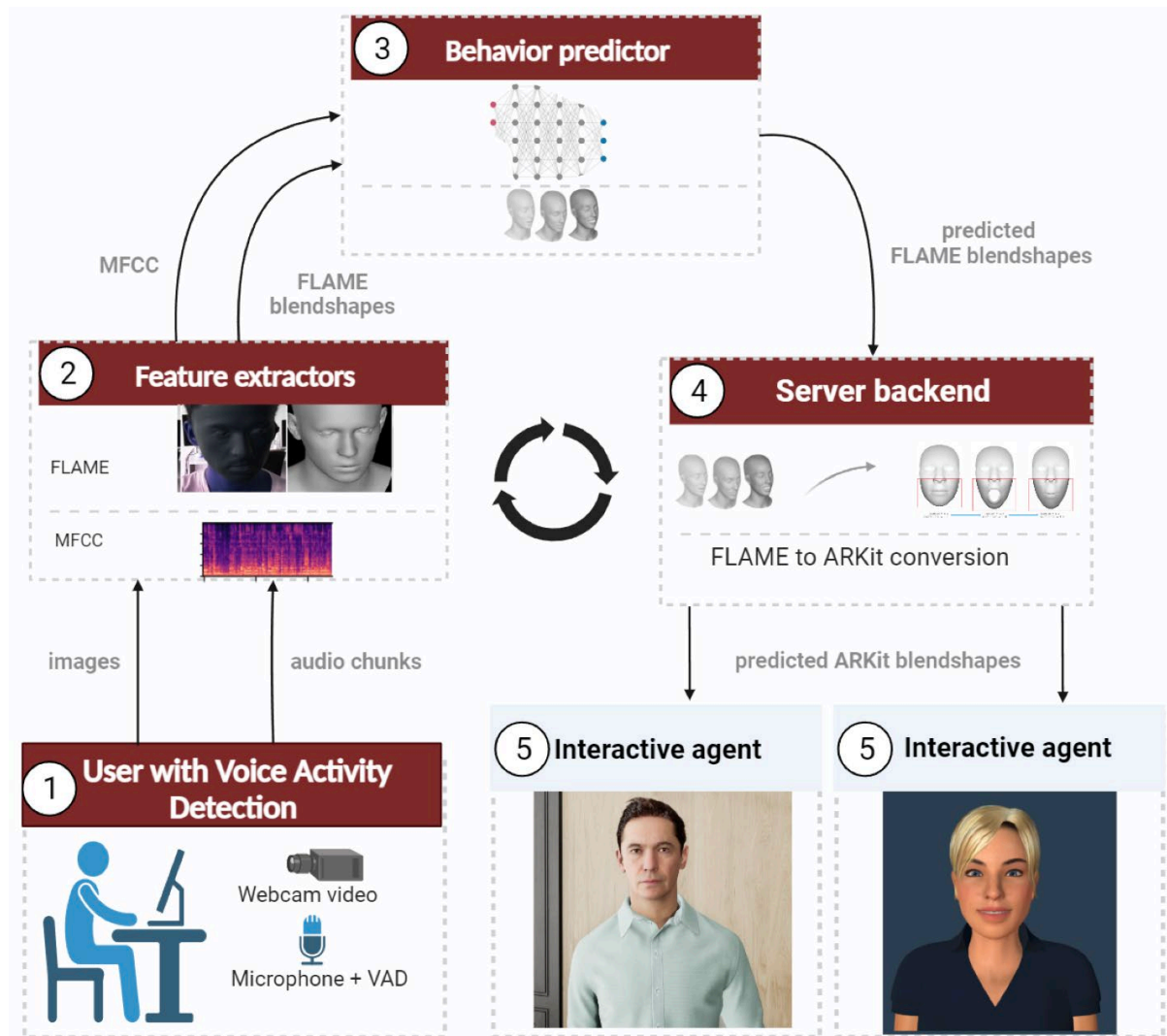


Abbildung 10: Verarbeitungsstufen des BISI-Frameworks

Abbildung 10 illustriert die 5 technischen Verarbeitungsschritte. Modul 1: Das in das wissenschaftliche Forschungswerkzeug VisualSceneMaker integrierte VSM-Plugin realisiert die Eingabeschnittstelle für Anwender. Diese verwendet auch die vom Partner Augsburg weiterentwickelte soziale Signalanalysekomponente SSI verwendet, um Sprach- und Videodaten zu streamen. Modul 2: "Run-time Feature Extraction", ist für die Erfassung und Verarbeitung von Gesichtsausdrücken und Kopfdrehungen zuständig und extrahiert MFCCs in Echtzeit. Modul 3: Das bindungsorientiertes Avatarverhaltensmodell zur Erzeugung nonverbalen Verhaltens (Behaviour Predictor). Modul 4, Server-Backend mit Local-to-Global-Transformation, fungiert als Konverter, der FLAME-Parameterwerte in Apple ARKit Blend Shape-Werte übersetzt. Letztere können zur Steuerung der Gesichtsausdrücke von virtuellen Charakteren auf ARKit-kompatiblen Geräten verwendet werden. Modul 5: VuppetMaster

stellt eine Visualisierungskomponente dar, welche die Apple ARKit-Werte und Kopfdrehungen in Echtzeit in einem Webbrowser anzeigt.

Zusammengefasst sind die Innovationen von BISI: 1. Die Integration von Echtzeit-Signalverarbeitung und maschinelles Lernen, um dyadische Interaktionen zu verbessern. 2. Das DSGVO-konforme Management und Nutzung von Psychotherapie-Sitzungen, um das domänenspezifisches Zuhörverhalten für Avatare/SIAs zu trainieren und 3. Softwareschnittstellen für einen flexiblen Einsatz.

Das BISI-Framework zeichnet sich durch Echtzeitfähigkeit aus, die durch eine modulare Architektur und parallele Verarbeitung erreicht wird. Tests belegen eine geringe Latenz bei der Generierung und Animation von Verhaltensmustern sowie Flexibilität bei unterschiedlichen Eingabedaten, wie Bild- und Audiodaten. Künftige Verbesserungen zielen auf eine optimierte Bewegungsinterpolation und nahtlose Animation ab. Die Ergebnisse technischer Studien erlauben die Deutung, dass das generierte Verhalten ähnlich zu dem des Trainingsmaterials ist. Eine zukünftige geplante Arbeit sind Nutzerstudien, in denen die Kontextangemessenheit des generierten Avatarverhaltens evaluiert wird.

9. Voraussichtlicher Nutzen und Verwertbarkeit

Die im Projekt durchgeführten Arbeiten bilden den Ausgangspunkt für eine Vielzahl von Forschungsprojekten, die sich mit unterschiedlichen Aspekten des Gebietes "Affective Computing" befassen. Die in UBIDENZ erzielten konzeptuellen und technischen Ergebnisse bieten eine exzellente Grundlage für Folgeprojekte in Industrie und Forschung. Die technischen Erweiterungen gestatten eine zeitnahe, systematische und standardisierte Evaluierung soziotechnischer Systeme, wie sie in naher Zukunft eine höhere Frequenz aufweisen werden. Die Ergebnisse eröffnen die Möglichkeit für weitergehende Forschungsaktivitäten auch in anderen Bereichen, wie beispielsweise Training sozialer Fähigkeiten für Arzt-Patient-Gespräche sowie Systeme zur interaktiven Therapieassistenz in verschiedenen Bereichen, beispielsweise in der Psychoonkologie und bei psychischen Erkrankungen.

Der Einsatz generativer KI im Kontext sozial-interaktiver Agenten (virtuelle Figuren, Roboter) birgt ein beträchtliches Potenzial. Die Anwendung dieser Methode erlaubt die Entwicklung

fortgeschrittener Konversationsfähigkeiten sowie die Bewältigung verschiedener, offener Benutzeranfragen in unterschiedlichen Aufgaben und Bereichen.

10. Außerhalb des Projektkontextes bekannt gewordener Fortschritt

Außerhalb des Projektes lassen sich Entwicklungen beobachten, die für das Projekt von Relevanz sind. Die Methoden der generativen künstlichen Intelligenz, insbesondere die Stable-Diffusion- oder Gaussian-Splatter-Modelle, finden Anwendung bei der Erzeugung automatischen Verhaltens von Avataren. Im Vergleich zu dem in diesem Projekt verfolgten Modellierungsansatz erlauben andere Ansätze 1) eine geringere Parametrierung, 2) weisen keine Echtzeitfähigkeit auf 3) sind nicht oder nur eingeschränkt für 3D-Avatare geeignet und 4) sind nicht in wissenschaftliche Forschungswerkzeuge (Visual SceneMaker und SSI) integriert. Auf wissenschaftlicher Seite ist uns derzeit kein Ansatz bekannt, der einen ähnlichen Ansatz zur Modellierung von Verhalten für sozial-interaktive Agenten erlaubt beziehungsweise anvisiert.

11. Erfolgte und geplante Veröffentlichungen

11.1 Erfolgte Veröffentlichungen

Reinwarth, A. L., Schneeberger, T., Nunnari, F., Gebhard, P., Altmann, U., & Wessler, J. (2023). Look what I made it do-The ModelIT method for manually modeling nonverbal behavior of socially interactive agents. *Companion Publication of the 25th International Conference on Multimodal Interaction*, 200–204.

Schneeberger, T., Reinwarth, A. L., Wensky, R., Anglet, M. S., Gebhard, P., & Wessler, J. (2023). Fast friends: Generating interpersonal closeness between humans and socially interactive agents. *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, 1–8.

Withanage Don, D. S., Müller, P., Nunnari, F., André, E., & Gebhard, P. (2023). Renelib: Real-time neural listening behavior generation for socially interactive agents. *Proceedings of the 25th International Conference on Multimodal Interaction*, 507–516.

Wessler, J., Gebhard, P., & Zilcha-Mano, S. (2024). Investigating movement synchrony in therapeutic settings using socially interactive agents: An experimental toolkit. *Frontiers in Psychiatry, 15*, Article 1330158.

Müller, P., Heimerl, A., Hossain, S. M., Siegel, L., Alexandersson, J., Gebhard, P., André, E., & Schneeberger, T. (2024). Recognizing emotion regulation strategies from human behavior with large language models. *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction*.

Hladký, M., Guerra, R. R., Cang, X. L., MacLean, K. E., Gebhard, P., & Schneeberger, T. (2024). Modeling the 'Kiss my Ass'-Smile: Appearance and functions of smiles in negative social situations. *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction*.

11.2 Geplante Veröffentlichungen

Die Ergebnisse der qualitativen Studie (siehe Kapitel 8.1.1.) sind bereits in einem wissenschaftlichen Papier präsentiert, das sich gerade im Druck befindet.

Immel, D., Hilpert, B., Schwarz, P., Hein, A., Gebhard P., Barton, S. & Hurlemann, R. (in print). Virtual therapeutic assistants in outpatient aftercare - Expectations of patients and health care professionals: A survey. *JMIR Formative Research*

Es gibt weitere wissenschaftliche Arbeiten, die dem Projekt UBIDENZ direkt zugeordnet sind und in Kapitel 8.1.3 beschrieben wurden. Diese qualifizierten zwei Studierende der Psychologie (ein Bachelor-, Masterabschluss). Diese Arbeiten erlauben drei Veröffentlichungen (geplant: eine Demo auf der internationalen Konferenz Intelligent Virtual Agent, ein Papier auf der internationalen Konferenz Intelligent Virtual Agent, ein Papier in dem internationalen Journal Emotion).

Arbeitstitel: In Sync with an Agent - How the embodiment of a virtual agent influences the user evaluation and physiological synchrony measured by heart rate

Kurzbeschreibung: Die Studie untersuchte den Einfluss der Verkörperung eines virtuellen Agenten auf die Nutzerwahrnehmung und die durch die Herzfrequenz gemessene physiologische Synchronität. Die Teilnehmer interagierten in einem Interview, das ein Depressionsinventar (BDI-II) enthielt, mit dem virtuellen Agenten Lydia in zwei Versionen: eine

war ein verkörperter Agent (ECA) mit einem pulsierenden Herzen und eine ein Sprachagent (VA) mit einem pulsierenden Kreis. Die Ergebnisse zeigten, dass die wahrgenommene Beziehung und Synchronität in der ECA-Bedingung signifikant höher war als in der VA-Bedingung, was darauf hindeutet, dass die Verkörperung die Bewertung durch den Benutzer positiv beeinflusst. Allerdings hatte die Verkörperung keinen signifikanten Einfluss auf das Vertrauen, unabhängig von der experimentellen Manipulation. Stattdessen wurde festgestellt, dass der Bindungsstil, insbesondere der vermeidende Bindungsstil, das wahrgenommene Vertrauen signifikant reduziert. Bei der physiologischen Synchronität ergaben sich keine signifikanten Unterschiede zwischen den Bedingungen, jedoch korrelierten höhere Depressionswerte mit einer geringeren physiologischen Synchronität. Diese Ergebnisse unterstreichen, wie wichtig es ist, virtuelle Agenten zu entwerfen, die sowohl die Verkörperung als auch die individuellen Eigenschaften der Nutzer berücksichtigen, um die Bewertung der Nutzer zu verbessern und die physiologische Synchronität zu erhöhen.

12. Literatur

Beck, A. T., Steer, R. A., & Brown, G. K. (2006). *Beck-Depressions-Inventar: BDI-II; Manual + Fragebögen*. Harcourt Test Services.

Astrid Bock, Eva Huber, and Cord Benecke. Levels of structural integration and facial expressions of negative emotions. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, 62:224–238, 2016. ISSN 14383608. doi: 10.13109/zptm.2016.62.3.224

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2020*.

Ehrenthal, J. C., Zimmermann, J., Brenk-Franz, K., Dinger, U., Schauenburg, H., Brähler, E., & Strauss, B.(2021). Evaluation of a short version of the experiences in close relationships-revised questionnaire (ECR-RD8): Results from a representative German sample. *BMC Psychology*, 9. <https://doi.org/10.1186/s40359-021-00637-z>

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. CoRR, abs/2012.04012, 2020, <https://arxiv.org/abs/2012.04012>.

Gebhard, P., Mehlmann, G., & Kipp, M. (2012). Visual SceneMaker—A tool for authoring interactive virtual characters. *Journal on Multimodal User Interfaces*, 6(1), 3–11. <https://doi.org/10.1007/s12193-011-0077-1>

Glaser, B. G. (2010). *Grounded theory: Strategien qualitativer Forschung* (3., unveränderte Auflage). Verlag Hans Huber.

Hale, J., & Hamilton, A. (2016). Testing the relationship between mimicry, trust and rapport in virtual reality conversations. *Scientific Reports*, 6, 35295. <https://doi.org/10.1038/srep35295>

Hautzinger, M., Keller, F., Beck, C., & Aaron, T. (2009). *Beck Depressions-Inventar BDI II (Revision, 2. Aufl.)*. Pearson Assessment. <https://katalog.slub-dresden.de/id/0-1382648650>

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with AIST++: music conditioned 3d dance generation. CoRR, abs/2101.08779, 2021. <https://arxiv.org/abs/2101.08779>

Lucas, G. M., Boberg, J., Traum, D., Artstein, R., Gratch, J., Gainer, A., Johnson, E., Leuski, A., & Nakano, M.(2018). Getting to know each other: The role of social dialogue in recovery from errors in social robots. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*.

Lugrin, B., Pelachaud, C., & Traum, D. (Eds.). (2021). *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*. ACM.

Mehling, W. E., Acree, M., Stewart, A., Silas, J., & Jones, A. (2018). The multidimensional assessment of interoceptive awareness, version 2 (MAIA-2). *PLOS ONE*, 13(12), e0208034. <https://doi.org/10.1371/journal.pone.0208034>

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, , Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

Reinwarth, A. L., Schneeberger, T., Nunnari, F., Gebhard, P., Altmann, U., & Wessler, J. (2023, October). Look What I Made It Do-The ModelIT Method for Manually Modeling Nonverbal Behavior of Socially Interactive Agents. In *Companion Publication of the 25th International Conference on Multimodal Interaction* (pp. 200-204).

Tschacher, W., & Haken, H. (2019). *The process of psychotherapy: Causation and chance*. Springer. <https://doi.org/10.1007/978-3-030-12748-0>

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 7763–7772, 2019.

Withanage Don, D. S., Müller, P., Nunnari, F., André, E., & Gebhard, P. (2023). Renelib: Real-time neural listening behavior generation for socially interactive agents. In Proceedings of the 25th International Conference on Multimodal Interaction (pp. 507-516).

Vorhaben: UBIDENZ
Ubiquitäre Digitale Empathische Therapieassistenz

Teilvorhaben DFKI: Bindungsorientierte interaktive
Verhaltensmodellierung für den UBIDENZ-Avatar

Titel: Kurzbericht

Förderkennzeichen: 13GW0568D

Zuwendungsempfänger: Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Trippstadter Straße 122, D-67663 Kaiserslautern

Projektleiter: Prof. Dr. Antonio Krüger

Bewilligungszeitraum: 01. September 2021 – 31. August 2024

Autoren: Dr. Patrick Gebhard, Dr. Tanja Schneeberger

Datum: 21. November 2024

Gefördert durch:



Bundesministerium
für Bildung
und Forschung

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (zum Beantragungszeitpunkt BMBF) unter dem Förderkennzeichen 13GW0568D unter den Nebenbestimmungen der NKBF 2017 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Das UBIDENZ-Projekt (Ubiquitäre Digitale Empathische Therapieassistenz) verfolgte das Ziel, ein innovatives, bindungsorientiertes Avatar-Verhaltensmodell für die Therapieunterstützung bei Menschen mit Depression zu entwickeln. Dieses Modell sollte als zentrale Interaktionsschnittstelle fungieren und eine stabilisierende therapeutische Beziehung schaffen, um Eigenverantwortung und Therapie-Adhärenz zu fördern. Im Fokus standen die Analyse sozialer Signale, die Modellierung von Bindungstypen und die Evaluation der Nutzerwahrnehmung.

1 Wissenschaftlicher und technischer Stand

Zu Beginn des Projekts gab es keine interaktiven Avatare mit sozio-emotionalen und bindungsorientierten Funktionen im Kontext psychiatrischer Nachsorge. Vorhandene Systeme, wie chatbasierte Lösungen, berücksichtigten weder nonverbale Kommunikation noch individuelle Bindungsstile. UBIDENZ knüpfte an interdisziplinäre Forschungsarbeiten zu Emotionen, interpersoneller Nähe, sozial-interaktiven Agenten und Emotionsregulation an und entwickelte diese weiter.

2 Ablauf des Vorhabens

Das Vorhaben wurde in enger Zusammenarbeit mit Forschungs- und Industriepartnern geplant und durchgeführt. Zunächst erfolgte eine umfassende Analyse der Anforderungen an bindungsorientierte Avatare, gefolgt von der Konzeption eines Modells, das emotionale Bindungsstile integriert. Parallel dazu wurden Daten für die Entwicklung und Evaluation gesammelt, aufbereitet und annotiert. Technisch wurde ein Framework entwickelt, das multimodale Daten wie Sprache und Mimik in Echtzeit verarbeitet und in dynamische Avatar-Reaktionen umsetzt.

3 Wesentliche Ergebnisse

Im Rahmen des Projekts UBIDENZ wurden bedeutende Fortschritte in der Entwicklung bindungsorientierter Avatare für die Therapieunterstützung erzielt. Zentral war die Konzeption eines innovativen Bindungsmodells, das bestehende emotionale Computermodelle erweiterte. Dieses Modell ermöglicht es, Bindungstypen durch die Analyse sozialer Signale, den situativen Kontext sowie individuelle Parameter in Echtzeit zu simulieren. Ergänzend dazu wurden ein UBIDENZ-Benutzermodell und ein Avatar-Verhaltensmodell entwickelt, die eng mit der Bindungssimulation verknüpft sind.

Ein technisches Highlight des Projekts war die Entwicklung der BISI-Softwarekomponente. Dieses Framework ermöglicht eine Echtzeitverarbeitung multimodaler Eingaben wie Sprache und Mimik und generiert daraus dynamische, kontextsensitive Verhaltensreaktionen des Avatars. Es überwindet damit die Einschränkungen statischer Animationen und setzt auf maschinelles Lernen, um flüssige, authentische Interaktionen zu erzeugen. Hierzu wurde ein umfangreicher, DSGVO-konformer Datensatz aus Videos psychotherapeutischer Sitzungen gesammelt, annotiert und zur Modellierung genutzt.

Die erzielten Ergebnisse wurden in mehreren Studien evaluiert. Diese zeigten, dass der Avatar insbesondere hinsichtlich Vertrauen und Synchronität positiv wahrgenommen wurde. Die Kombination aus multimodalen Eingaben und einem dynamischen Verhaltensmodell bewährte sich als innovativer Ansatz zur Verbesserung der Mensch-Maschine-Interaktion.

Zusätzlich trug das Projekt zur Weiterentwicklung technischer Werkzeuge wie dem VisualSceneMaker bei und erweiterte bestehende Methoden der Analyse sozialer Signale. Die entwickelten Modelle und Softwarekomponenten sind vielseitig einsetzbar, beispielsweise in der Psychoonkologie oder zur Unterstützung von Arzt-Patient-Gesprächen. UBIDENZ bearbeitet wichtige Themen im Gesundheitswesen und schafft eine Grundlage für künftige Anwendungen, die soziale Interaktion und Therapieunterstützung intelligent verknüpfen. Sie erweitern die Möglichkeiten für sozio-interaktive Agenten erheblich und zeigen auf, wie virtuelle therapeutische Assistenten Versorgungslücken schließen können.

4 Zusammenarbeit

Das Projekt wurde neben den Projektpartnern unter anderem mit den Universitäten Kassel, Saarland und internationalen Partnern wie der Universität Haifa umgesetzt. Die interdisziplinäre Kooperation ermöglichte es, Expertise aus Psychologie, Psychiatrie, Künstlicher Intelligenz und Softwareentwicklung zu vereinen.