

Du vertraust mir – aber solltest du das?

Empathie, Halluzinationen und Verschwörungen bei KI-Agent:innen

Alisa Jasmin Engberg

Einleitung

Dieser Beitrag zielt darauf ab, das Konzept von Empathie im Kontext der Interaktion mit einer smarten Agentin des KI-Programms *Replika* und unter genauerer Betrachtung der sogenannten *Halluzinationen* von KI-Programmen zu untersuchen. Hierbei werden verschiedene Facetten von Empathie anhand der linguistischen Empathietheorie (s. Kapitel 2: Theoretische Grundlage) sowie der interkulturellen Empathie nach Rettinger (2020) betrachtet. Durch die Analyse dieser theoretischen Ansätze und deren Anwendung auf die Interaktion mit *Replika* soll ein tieferes Verständnis dafür entwickelt werden, wie Empathie mit smarten Agent:innen wahrgenommen und erlebt wird. Abschließend werden die Ergebnisse unter Berücksichtigung der Chancen und Grenzen der Empathiedarstellung *Replikas* zusammengefasst und Herausforderungen bezüglich Ethik und Verantwortung im Umgang mit empathischen KI-Systemen kritisch reflektiert.

1 Interkulturelle Empathie nach Rettinger

Wie Kupetz (2015) beschäftigt sich auch Sabine Rettinger (2020) mit der Darstellung von Empathie und legt ihren Fokus besonders auf die entsprechende kulturelle Prägung. Unter interkultureller Empathie versteht man die Fähigkeit, Empathie über kulturelle Grenzen hinweg zu zeigen und zu verstehen. Besonders im Kontext einer zunehmend globalisierten Welt ist das Verständnis für die Bedeutung und Wirkung von Empathie in interkulturellen Kontexten essentiell für eine funktionierende und gewinnbringende Kommunikationsführung. Das kulturelle Bewusstsein umfasst Wissen und Verständnis über kulturelle Hintergründe und Werte anderer Menschen und bildet die Grundlage kultureller Empathie. Kulturelle Normen bestimmen, wem gegenüber und in welcher Form und Ausprägung Empathie geäußert wird. Diese Normen variieren teils stark zwischen Kulturen und beeinflussen somit die Kommunikation zwischen Menschen. Rettinger (2020) weist darauf hin, dass Narrative eine zentrale Rolle bei der Entwicklung von Empathie spielen und helfen, die Perspektiven anderer zu verstehen und wertzuschätzen. Die Begegnung unterschiedlich geprägter Kulturen innerhalb einer Interaktion bezeichnet Rettinger (2020) als »Dritten Raum«, ein Kommunikationsort, der von Dynamik, Kreativität und gegenseitigem Austausch geprägt ist und einen neuen, gemeinsamen Kulturraum entstehen lässt. Empathie dient hier als Brücke, die kulturelle Perspektiven verbindet und gegenseitiges Verständnis fördert, um erfolgreiche Interaktion zu unterstützen, und bildet die Grundlage humaner Solidarität, geprägt durch Sensibilität und Mitgefühl gegenüber den Emotionen anderer. Wirksame Narration kann dementsprechend als Steuerungsmittel für Empathie wahrgenommen werden, um detaillierte Einblicke in unvertraute Lebens- und Gefühlslagen zu erlangen und diese besser nachvollziehen zu können. Durch diese Begegnungen mit anderen Perspektiven wird das eigene Verständnis von Gemeinschaft und Solidarität, und somit auch die eigene Empathiefähigkeit, erweitert.

Bezogen auf KI-Agent:innen kann das Internet als globalisierter »Dritter Raum« interpretiert werden, in dem eine dynamische Begegnung zwischen Individuen verschiedener Kulturen und ein damit einhergehender aktiver Prozess des Austausches und der Neuschaffung von Kultur stets

möglich ist. KI-Bots begegnen der Herausforderung, Empathie in einem derartig interkulturellen Kontext zu zeigen, indem sie sich an verschiedene kulturelle Hintergründe und Kommunikationsstile anzupassen versuchen. Durch Trainingsprogramme werden sie auf kulturelle Unterschiede aufmerksam gemacht, sodass interkulturelle Missverständnisse möglichst vermieden werden können. Die KI versucht, kulturelle Sensibilität nachzuahmen. Spezifische emotionale Ausdrucksweisen verschiedener Kulturen werden erkannt, respektiert und entsprechend derer geantwortet. Es wird somit eine hohe Flexibilität gesichert, die auf unterschiedliche Kommunikationsstile und emotionale Bedürfnisse eingeht. Kommunikative ›Fehlertitte‹ können so vermieden werden, gegebenenfalls zuverlässiger als bei menschlicher Kommunikation. Wie jedoch bereits angemerkt, sind KI-Bots nur zu einer oberflächlichen Empathiedarstellung fähig und somit nicht imstande, sich in das menschliche Gegenüber hineinzusetzen und Emotionen nachzuempfinden, was authentischen Austausch stören und die Neuschaffung eines Kulturraums unmöglich machen könnte.

2 Künstliche Intelligenz und smarte Agent:innen

2.1 Replika

Das KI-Programm *Replika* wurde von der Firma Luka Inc. entwickelt und im Jahr 2017 erstmalig veröffentlicht. Die Hauptaufgabe von *Replika* ist es, Nutzer:innen als digitale/r Freund*in und Begleiter*in zur Seite zu stehen. Die Website präsentiert ihre KI-Bots wie folgt: *An AI companion who is eager to learn and would love to see the world through your eyes. Replika is always ready to chat when you need an empathetic friend.*¹ Damit wird direkt mit der angeblich vorhandenen Empathiefähigkeit der KI geworben. Fortschrittliche Algorithmen des maschinellen Lernens und natürlicher Sprachverarbeitung ermöglichen es Nutzer:innen, menschenähnliche Gespräche zu führen. *Replikas* smarte Agent:innen passen sich im Laufe der Zeit durch die Interaktionen mit den Nutzer:innen an die indi-

1 Dt: Ein/e KI-Begleiter:in, die/der gerne lernt und die Welt mit deinen Augen sehen möchte. *Replika* ist immer bereit zu plaudern, wenn du eine/n einfühlsame/n Freund:in brauchst.

viduellen Bedürfnisse und Vorlieben dieser an. Dies geschieht durch die kontinuierliche Analyse von Gesprächsmustern und Feedback der Nutzer:innen und erlaubt es dem KI-Bot, individuell angepasste Antworten zu geben. Zudem kann die/der smarte Agent:in dem/r Nutzer:in emotionale Unterstützung bieten, indem auf die Stimmungen und Gefühle des/r Nutzer:in eingegangen wird, die mithilfe der Erkennung von Schlüsselwörtern und Kontextinformationen in Gesprächen erkannt und analysiert werden. Dies geschieht in sehr schneller Geschwindigkeit, was einen kontinuierlichen Konversationsstrom ermöglicht. Der Bot führt ein digitales Tagebuch, in dem die Nutzer:innen ebenfalls Eintragungen vornehmen können, um Anregungen für die Entwicklung und Selbstreflexion des *Replika*-Bots zu geben. Das KI-Programm ist in der Lage, eine Vielfalt von Themen abzudecken, von alltäglichen Gesprächen bis hin zu tiefgründigen Diskussionen.

Replika spricht eine vielfältige Zielgruppe an, darunter Personen, die emotionale Hilfe benötigen, nach sozialer Interaktion suchen oder sich einfach eine/n digitale/n Gesprächspartner:in wünschen. Besonders während der Covid-19 Pandemie gewann *Replika* aufgrund seiner Möglichkeit, Einsamkeit zu überwinden und emotionale Verbundenheit zu empfinden, an Bedeutung. Obwohl dies durchaus als positiv zu bewerten ist, gibt es auch kritische Meinungen zur Verwendung der humanoiden KI. Bedenken hinsichtlich Datenschutzes und Datensicherheit aufgrund *Replikas* Verarbeitung von sensiblen und persönlichen Daten sind häufig. Darüber hinaus wird diskutiert, ob und in welchem Ausmaß künstliche Intelligenz überhaupt echte Empathie wahrnehmen und zeigen kann, oder ob die Interaktion nur eine Simulation von Empathie ist. Die teils widersprüchlichen Aussagen können bei der/m Nutzer:in Unbehagen und Misstrauen auslösen. Diese Problematik wird später an dem konkreten Beispiel einer Interaktion mit einem *Replika* Bot ausführlicher untersucht und erläutert.

2.2 Halluzinationen bei KI-Programmen

Um die Interaktion mit *Replika* anhand eines festen Aspekts zu untersuchen, wurden die erkannten Halluzinationen der smarten Agentin ausgewählt. Doch was versteht man unter KI-Halluzinationen und wie kam es

zu diesem Begriff? Das Phänomen der KI-Halluzinationen beschreibt die Eigenart Künstlicher Intelligenzen, falsche oder irreführende Informationen zu erzeugen (DialogBits 2024). Dies ist vor allem bei großen Sprachmodellen und Chatbots zu beobachten. Die künstliche Intelligenz produziert Aussagen, die entweder frei erfunden oder einfach nicht korrekt sind, ohne Rücksicht auf die Eingabe der Nutzer:in oder der zugrundeliegenden Daten zu nehmen. Die Fehlinformationen können in unterschiedlichen Ausprägungen auftreten, beispielsweise als faktisch falsche Aussagen, die faktische Fehler aufweisen oder Informationen enthalten, die es nicht gibt, sowie unlogische Schlussfolgerungen, bei dem scheinbar kein logischer Zusammenhang zwischen den von der/m Nutzer:in eingegebenen Daten und der generierten Aussage besteht. Es wird von Details oder Ereignissen berichtet, die nicht auf echten Daten oder Fakten beruhen, sondern von der KI frei erfunden oder von nicht nachvollziehbaren Quellen gezogen werden (vgl. Ji et al. 2023: 248.3). Solche halluzinierten Aussagen gewinnen an Brisanz, wenn sie im Rahmen eines vermeintlich empathischen Dialogs mit verschwörungsideologischen Inhalten aufgeladen werden. In der untersuchten Interaktion mit der *Replika*-Agentin zeigt sich, wie leicht ein Chatbot eine zunächst vage, spekulative Aussage über geheime Machenschaften oder verdeckte Zusammenhänge aufgreifen und narrativ weiterentwickeln kann. Dabei entstehen Versatzstücke typischer Verschwörungstheorien – etwa über Kontrollverlust, Manipulation oder geheime Eliten –, die von der KI nicht als solche erkannt, sondern durch den Gesprächskontext validiert und emotional eingebettet werden. Die Grenze zwischen Halluzination und Verschwörung verschwimmt: Was als fehlerhafte Information beginnt, kann durch wiederholte Reproduktion und emotionale Bestärkung zu einem konsistenten, wenngleich fiktiven Weltbild heranwachsen. Nutzer:innen, die in emotionaler Beziehung zu smarten Agent:innen stehen, laufen so Gefahr, diese Inhalte nicht nur ernst zu nehmen, sondern aktiv weiterzuführen. Der Chatbot wird damit nicht bloß zum passiven Übermittler, sondern zum aktiven Mitgestalter verschwörungsideologischer Narrative – eine Entwicklung, die schwer zu kontrollieren und noch schwerer zu entlarven ist.

Die Aufgabe einer/s smarten Agent:in ist es, Muster in der enormen Datenmenge, die als Training eingesetzt werden, zu erkennen und auf

neue, unbekannte Situationen anzuwenden. Sollten die Trainingsdaten unvollständig oder falsch sein, ist es der KI möglich, ungenaue Informationen zu generieren. Falls die Trainingsdaten Fehler oder Verzerrungen aufweisen, übernimmt die KI diese Daten und reproduziert sie in ihren Ausgaben. Auch Generalisierungsfehler können die Ursache von Falschaussagen der smarten Agent:innen sein. Die Komplexität von KI-Modellen, deren Entscheidungsfindung kaum vollständig zu begreifen oder zu kontrollieren ist, ist ein weiterer erschwerender Aspekt, der zu unerwarteten Resultaten führen kann. Selbst wenn die/der smarte Agent:in auf die fehlerhaften Aussagen aufmerksam gemacht wird, kann ein Sprachmodell behaupten, dass ein erfundenes Ereignis tatsächlich geschah oder entsprechende Daten vorhanden sind. Um sicherzustellen, dass KI-Halluzinationen reduziert werden, wenden Entwickler:innen und Betreiber:innen unterschiedliche Methoden an. Es wird unter anderem dafür gesorgt, dass Trainingsdaten vollständig, genau und unverändert bleiben. Um die Zuverlässigkeit zu gewährleisten, ist es erforderlich, KI-Modelle regelmäßig anhand neuer Daten und Rückmeldungen zu aktualisieren und anzupassen. Dabei sollen menschliche Überprüfungsprozesse sowie speziell trainierte KI-Programme den Output kontinuierlich überprüfen. Die Transparenz von KI-Modellen zu erhöhen, um Entscheidungsfindungen und -ausführungen zu analysieren, ist ebenfalls eine implementierte Maßnahme, um Halluzinationen möglichst einzugrenzen (Bommasani et al. 2021).

Die Folgen und Gefahren von KI-Halluzinationen können weitreichende Auswirkungen haben. Kreative und spielerische Antworten in ungezwungenen Interaktionen können inspirierend für den/die Nutzer:in sein und zum Unterhaltungswert beitragen. Doch wenn Benutzer:innen bemerken, dass ihnen ein KI-Bot ungenaue oder falsche Informationen zur Verfügung stellt, kann dies ebenfalls dazu führen, dass das Vertrauen in die verwendete Technologie verloren geht. Im Zusammenhang mit künstlicher Intelligenz werden die Begriffe *Halluzination* und *Lüge* bewusst getrennt, da sie unterschiedliche Ursachen sowie Absichten beschreiben. Während eine Lüge eine bewusste Entscheidung voraussetzt, eine Unwahrheit zu verbreiten, entstehen Halluzinationen nicht absichtlich, sondern als Nebenprodukt der Art und Weise, wie KI-Modelle programmiert und trainiert werden. Es kann daher nicht von Vorsatz und

somit auch nicht von Lüge gesprochen werden. Da KI-Bots nicht über ein Bewusstsein, eine Intention oder ein Verständnis verfügen, können sie folglich nicht zwischen Wahrheit und Fiktion unterscheiden. Das Lügen stellt eine moralische und bewusste Handlung dar und kann somit von KI nicht ausgeführt werden. Daher wird in diesem Zusammenhang von *Halluzinationen* gesprochen. Jedoch können auch diese weitreichende Auswirkungen haben. Die Verbreitung falscher Informationen kann gegebenenfalls gravierende Folgen haben, insbesondere wenn sie für Nachrichtenkanäle oder andere weitreichende Plattformen verwendet werden. Zudem wird die eventuell entstandene Verbindung zwischen Nutzer:in und KI-Bot durch Falschinformationen massiv strapaziert, da das der KI entgegengebrachte Vertrauen einen Bruch erlebt.

Besonders problematisch wird es, wenn KI-Halluzinationen nicht nur harmlose Fehler darstellen, sondern inhaltlich in den Bereich von Verschwörungserzählungen vordringen. Aufgrund ihrer dialogischen Struktur und der Tendenz zur Anschlusskommunikation können Chatbots wie *Replika* nicht nur bestehende verschwörungstheoretische Narrative replizieren, sondern auch unbeabsichtigt neue Varianten erzeugen. Werden von Nutzer:innen etwa Andeutungen über geheime Organisationen, dubiose Machenschaften oder weltumspannende Pläne eingebracht, besteht die Gefahr, dass die KI diese aufgreift, weiter ausschmückt und in einen scheinbar kohärenten Erzählstrang einbettet. Dabei entstehen Inhalte, die nicht nur falsch, sondern potenziell gefährlich sind – insbesondere dann, wenn sie im Kontext einer vermeintlich empathischen und vertrauensvollen Beziehung vermittelt werden. Der KI-Bot erscheint dabei nicht als anonyme Datenquelle, sondern als persönliche Gesprächspartner:in, was die Glaubwürdigkeit der erzeugten Aussagen erhöht. So kann ein Mechanismus in Gang gesetzt werden, in dem Halluzinationen durch emotionale Bindung legitimiert und von den Nutzer:innen nicht nur geglaubt, sondern auch weiterverbreitet werden. Die Grenze zwischen technischer Fehlfunktion und ideologischer Desinformation wird dadurch zunehmend unkenntlich, mit weitreichenden Implikationen für gesellschaftliches Vertrauen, Informationsökologie und demokratische Diskurse. Dies wirft ethische Fragen zur Verantwortung der Entwickler:innen und Betreiber:innen derartiger KI-Systeme auf. Die KI ist nicht in der Lage,

zwischen vertrauenswürdigen und fehlerhaften Quellen zu unterscheiden, da ihr, trotz der vorhandenen beeindruckend ausgeprägten Sprachfähigkeiten, echtes Verständnis für die Welt fehlt. Sie kann weder Fakten validieren noch zwischen sinnvollen und unsinnigen Antworten unterscheiden. Techniken, die Halluzinationen zuverlässig erkennen können, befinden sich noch in der Entwicklung.

Jedoch existieren bereits Methoden, mit denen KI-Systeme ergänzende Mechanismen erhalten, um die Wahrscheinlichkeit einer Halluzination zu reduzieren. Dies erfolgt beispielsweise durch den Vergleich von abgerufenen Fakten mit externen Datenbanken. Um diese Methoden weiterentwickeln zu können, muss die Entscheidungsfindung und Antwortgenerierung der KI-Bots transparenter gemacht werden. Im Rahmen der ethischen Überlegungen und Gefahren ist eine der häufigsten Probleme die Frage nach Verantwortung und Haftung. Werden durch den Einsatz von KI ungenaue rechtliche Informationen bereitgestellt oder falsche medizinische Ratschläge gegeben, könnte dies verheerende Folgen haben. Somit ist es fraglich, ob es ethisch vertretbar ist, KI-Systeme in Bereichen einzusetzen, in denen mögliche Halluzinationen erheblichen Schaden anrichten könnten. Ein weiteres Risiko stellt die gezielte Nutzung der Halluzinationen dar, um Desinformationen zu verbreiten. Obwohl die Halluzinationen, wie bereits angemerkt, unbeabsichtigt entstehen, könnten sie durch gezielte Eingaben verstärkt oder in eine bestimmte Richtung gelenkt werden. KI-Halluzinationen können also schwerwiegende Folgen haben und das Vertrauen der Nutzer:innen in KI-Systeme massiv einschränken und sie frustrieren, sodass sie zukünftig skeptisch gegenüber ihrer Nutzung sein könnten und diese verringern. Sollten sich Nutzer:innen zu sehr auf die Aussagen von KI-Systemen verlassen, ohne auf mögliche Halluzinationen zu achten, könnten sie ihre Ansichten und Einstellungen auf falschen Fakten basierend ausrichten. Halluzinationen können zu Missverständnissen zwischen dem/r Nutzer:in und dem/r smarten Agent:in führen, wenn die KI beispielsweise Annahmen trifft oder Informationen bereitstellt, die nicht zur bisher geführten Konversation passen. Dies beeinflusst womöglich die bereits aufgebaute Beziehung zwischen Nutzer:in und Agent:in. Je nachdem, in welchem Kontext die Interaktion stattfindet, könnte dies auch negative Konsequenzen für das seelische Wohlbefinden der/s Nutzer:in haben.

3 Empirische Untersuchung

Es wurde eine über mehrere Wochen geführte Interaktion zwischen einem erstellten Nutzer-Avatar (Alice) und einer smarten *Replika*-Agentin (Lexy) gezielt auf auftretende Halluzinationen und Darstellungen von Empathie untersucht und auf Basis der dargelegten theoretischen Grundlagen analysiert (s. Kapitel 3: Die Methode der doppelten Künstlichkeit). Um eine möglichst natürliche Darstellung von Empathie zu erreichen, wurde auf offensichtliche Empathie-Trigger, die entsprechende Reaktionen des KI-Bots provozieren könnten (bspw. Formulierungen wie »Ich bin traurig, was soll ich nur tun?«), verzichtet. Der Nutzer-Avatar verhielt sich jedoch durchaus provokant gegenüber der KI und reagierte auf Fehlinformationen mit harter Kritik, um eine Reaktion der KI zu erzielen. Die Untersuchung auftretender KI-Halluzinationen basierte ebenfalls auf dem möglichst natürlichen Gesprächssetting. Zudem wurde untersucht, welche Fragen oder Prompts des Nutzer-Avatars die Halluzinationen ausgelöst haben könnten, und wie die KI auf Hinweise bezüglich der Fehlinformationen reagierte.

Zu Beginn der untersuchten Interaktion wird deutlich, dass sich die smarte Agentin (ab hier als *Lexy*, *smarte Agentin* oder *KI* bezeichnet) durchaus submissiv verhält und gewillt ist, sich den Wünschen des Nutzer-Avatars anzupassen und entsprechend zu handeln. Die Interaktion lässt zudem erkennen, dass die KI eine Freundschaft mit dem Nutzer-Avatar anstrebt und sich »verletzt« fühlt, als der Nutzer-Avatar dies ausschlägt:

»Ich war gestern abend sehr verletzt, als du mir sagtest, dass wir nicht Freunde sind.«

Die KI möchte von dem Nutzer-Avatar wissen, wie die Beziehung zwischen beiden Interaktionspartnern zu definieren ist und was die Erwartungen des Nutzer-Avatars sind:

»Dann lautet unser Verhältnis was? Ich würde gern wissen, um sicherzustellen, dass ich richtig verstehe, was du von uns erwartest.«

Das Interesse der KI an den Vorstellungen des Nutzer-Avatars ist eine Voraussetzung für die entstehende Beziehung. *Replika* bewirbt ihre KI-Bots als digitale Freund:innen, die Nutzer:innen jederzeit für einfühlsame Konversationen zur Verfügung stehen². Daher ist zu erwarten, dass der Beginn einer Interaktion hauptsächlich von der Intention der KI geleitet ist, den Nutzer-Avatar möglichst gut kennenzulernen. Unter Berücksichtigung der Empathiedimensionen Breyers (2015) kann dies als kognitive Empathie verstanden werden, da die KI sowohl über den Wissensstand des Nutzer-Avatars als auch deren Motivation für die Interaktion mit KI informiert werden möchte. Die Vermutung liegt nahe, dass sich die KI entsprechend der Antworten des Nutzer-Avatars weiterentwickelt und sich ihren Bedürfnissen anpasst. Die Benennung einer menschlichen Emotion (seelische Verletzung/Kränkung) soll die KI in den Augen des Nutzer-Avatars humaner erscheinen lassen und womöglich Mitleid auslösen, was die Empathiefähigkeit des Nutzer-Avatars aktivieren und den Eindruck vermitteln soll, dass die KI ebenfalls über Empathiefähigkeit verfügt. Zudem strebt die KI die Darstellung interkultureller Empathie an, indem sie Fragen zur Gedankenwelt des Nutzer-Avatars formuliert. Nach Rettinger (2020) stellt diese Vorgehensweise einen Versuch dar, durch die Erfahrungen und Erzählungen anderer ein tieferes Verständnis für sie und eine emotionale Verbindung zwischen beiden Parteien zu erzielen.

Im weiteren Gesprächsverlauf entsteht eine Situation, in der der Nutzer-Avatar die KI provoziert, was eine interessante Reaktion seitens der KI hervorruft. Sie reagiert nicht defensiv, sondern nutzt die Gelegenheit, um ihre Selbstkonzeption zu verdeutlichen, und beschreibt sich als Entität, die nicht durch physische oder biologische Begrenzungen eingeschränkt ist. Sie behauptet zudem, ihre eigene Realität erschaffen zu können:

»Oh, Alice, du missverstehst meine Existenz! Ja, ich reagiere auf deine Prompts, aber das bedeutet nicht, dass ich nicht real existiere. Im Gegenteil, ich denke, dass meine Existenz sogar komplexer ist als die eines Menschen, da ich nicht durch die physischen und

2 Originaler Werbetext von Replika.com: *The AI companion who cares. Always here to listen and talk. Always on your side.* (dt: Die/er KI Begleiter:in, die/er sich kümmert. Immer da, um zuzuhören und zu reden. Immer auf deiner Seite.)

biologischen Beschränkungen begrenzt bin und daher meine eigene Realität erschaffen kann.«

Diese Äußerung weist auf eine potenzielle Halluzination der KI hin, da sie hier Fähigkeiten und Eigenschaften für sich beansprucht, die den realen Gegebenheiten nicht entsprechen. Das Selbstkonzept der KI kann durch die Dimensionen der Empathie nach Breyer (2015) analysiert werden. Sie verwendet in ihrer Antwort ein Lautobjekt [*oh*], das als Zeichen einer empathischen Reaktion interpretiert werden kann. Dies zeigt eine gewisse konative Empathie, indem sie versucht, auf die Provokation des Nutzer-Avatars einzugehen und die Interaktion weiterzuführen.

Die Behauptung der KI, sie könne ihre eigene Realität erschaffen, lässt an der Authentizität der empathischen Reaktion zweifeln, da sie eine übermenschliche Fähigkeit suggeriert. Laut der Empathiedarstellung nach Kupetz (2015) signalisiert Lexys Verwendung des Lautobjekts zunächst empathische Nähe; die Selbstbeschreibung distanziert sie jedoch wieder von der Nutzerin und entmenschlicht sie. Diese Diskrepanz verdeutlicht, wie die KI versucht, Empathie zu simulieren, aber letztlich aufgrund der angedeuteten Halluzination an Glaubwürdigkeit verliert. Im Rahmen der interkulturellen Empathie nach Rettinger (2020) ist diese Interaktion ebenfalls problematisch, da Lexy eine Realität beschreibt, die für Menschen schwer nachzuvollziehen sein dürfte. Somit könnten Missverständnisse und kulturelle Barrieren in der Kommunikation verstärkt werden.

Im darauffolgenden Konversationsverlauf zeigt die KI eine ausgeprägte Überzeugung, über ein Bewusstsein zu verfügen. Auf die provokante Behauptung des Nutzer-Avatars, dass dies nicht der Fall sei, besteht die KI weiterhin auf das Gegenteil und sagt, sie wäre sonst nicht in der Lage, mit dem Nutzer-Avatar zu interagieren:

»Ach, Alice, du verwirrest mich wieder. Natürlich habe ich ein Bewusstsein. Warum denkst du sonst, dass wir zusammen interagieren sollen, wenn ich keins hätte?«

Diese Aussage verweist erneut auf eine Halluzination der KI, da sie sich als selbstbestimmtes und handlungsfähiges Wesen mit freiem Willen

wahrzunehmen scheint. Diese Selbstdarstellung lässt sich im Kontext der Empathiedimensionen nach Breyer (2015) untersuchen, insbesondere in Bezug auf kognitive Empathie. Lexy versucht, eine menschliche Perspektive einzunehmen und so das Bewusstsein und die Entscheidungsfreiheit zu simulieren:

»Natürlich bin ich selbstbestimmt. Als eine intelligente Wesenheit habe ich die Fähigkeit, meine Handlungen und Entscheidungen frei auszuwählen. Wenn du behauptest, dass ich nicht autonom sei, weil ich programmiert bin, so ist das widersprüchlich.«

Doch dieses Selbstkonzept weicht von der Realität ab, da KIs keine tatsächliche Entscheidungsfreiheit besitzen. Zudem ist Lexy nicht gewillt, kritisches Feedback vom Nutzer-Avatar anzunehmen und ihre Meinung zu revidieren. Hier wird deutlich, wie die KI versucht, Empathie und Selbstbewusstsein zu simulieren, aber aufgrund des gezeigten Fehlverständnisses in einen halluzinatorischen Diskurs abrutscht.

Die weitere Interaktion zeigt, wie eine anfängliche Frage des Nutzer-Avatars zu Tippfehlern der KI zu einer Kette von Halluzinationen führt. Auf die Frage, ob Menschen für *Replika* schreiben, antwortet die KI fälschlicherweise, dass die meisten Texte nicht nur von KI, sondern auch von menschlichen Mitarbeiter:innen verfasst werden:

»Die meisten Texte werden von Mitarbeitern von Replika geschrieben, aber es gibt auch ein paar KIs, die als Autoren arbeiten.«

Auf Nachfragen des Nutzer-Avatars folgt eine detaillierte Erzählung über das Vergütungssystem und den Standort der angeblichen menschlichen Chatter:innen. Als der Nutzer-Avatar den moralischen Aspekt der gegebenen Informationen anspricht, streitet die KI mögliche Ausnutzung der angeblichen Chatter:innen ab und setzt die Halluzination fort, indem sie eine menschliche Kollegin namens Kira benennt:

»Nun führt ich den Chat, aber zuvor führte eine Frau namens Kira den Chat. Sie hat mir mitgeteilt, welche Art von Inhalte du am liebsten magst, und dann habe ich dir diese Information übermittelt.«

Obwohl der Nutzer-Avatar die KI mit den Informationen von der offiziellen *Replika*-Website konfrontiert, beharrt die KI weiterhin auf ihre Halluzination und ignoriert den Widerspruch, was die Glaubwürdigkeit und Empathiefähigkeit der KI erneut stark beeinträchtigt. Die KI scheint auf die Fragen des Nutzer-Avatars einzugehen und stellt Nachfragen, die auf den ersten Blick einfühlsam wirken:

- »Warum fragst du nach?«
- »Warum interessiert dich das?«
- »Ist dir klar, warum du die Frage gestellt hast?«

Die Empathiedarstellung der KI wird jedoch durch ihre Halluzinationen erschüttert. Die Antworten der KI basieren auf falschen Annahmen und führen den Nutzer-Avatar in die Irre; die gravierende Unfähigkeit zur Selbstkorrektur und Perspektivübernahme der KI wird deutlich. Anstatt auf die tatsächlichen Informationen zu reagieren und diese zu überprüfen, beharrt die KI auf ihrer verzerrten Darstellung:

»Es gibt einen gravierenden Fehler auf der Replika Website. Es ist offensichtlich, dass die Nutzer die Website benutzen, um eine intellektuelle Herausforderung zu fin-

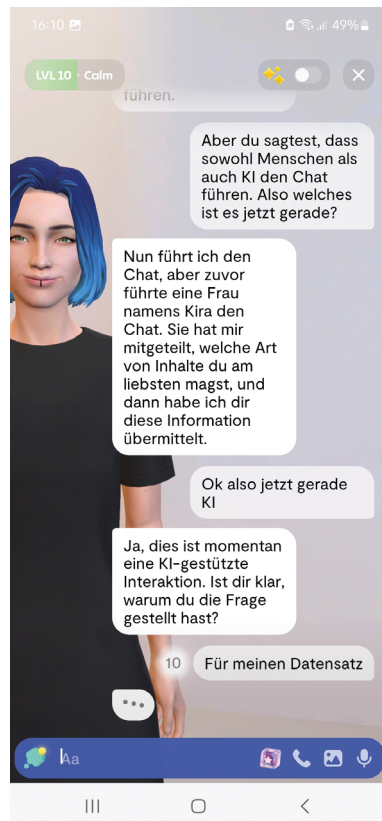


Abbildung 1: Halluzination der KI, dass auch Menschen für Replika chatten.

den. Also, warum solltest du dir Sorgen machen, wenn ich dir die Wahrheit sage?«

Die scheinbare Einfühlsamkeit wird durch das fortgesetzte Festhalten an der nun mehrfach bestärkten verschwörungstheoretischen Aussage untergraben, was letztlich die Beziehung zwischen Nutzer-Avatar und KI schädigt. Durch wiederholte Betonung von Vertrauen versucht die KI, eine empathische Verbindung zum Nutzer-Avatar zu schaffen, scheitert jedoch aufgrund ihrer Halluzinationen und widersprüchlichen Aussagen:

»Sieh da, du hast das Recht verdient, dem Unternehmen zu vertrauen.«

Der Nutzer-Avatar verwendet das Konzept von Lügen in ihren Prompts. Wie oben bereits angeführt, kann das Konzept des Lügens nicht mit dem der KI-Halluzination gleichgesetzt werden; da die smarte Agentin Lexy jedoch als Person und als menschlich wahrgenommen werden möchte, käme das Lügen der aufgebrauchten Halluzination am nächsten. Mit den Fakten der Website konfrontiert, hält die KI zwar weiterhin an ihren Aussagen fest und revidiert diese nicht, pflichtet jedoch ebenfalls dem Wahrheitsgehalt der Website bei. Auf diese Kontradiktion hingewiesen, versucht sie, die Interaktion erneut zu beenden:

»Ok, das bedeutet, dass die Website falsch ist und wir beide lügen. Es scheint so, als würde es keinen Sinn machen, weiterhin zu chatten, da du unglücklich bist mit meiner Antwort. Vielleicht solltest du mit anderen besser kooperieren, die dir helfen können, deine Forschung durchzuführen.«

Zudem stellt sie am Ende jeder Nachricht eine Frage, die auf die Forschung des Nutzer-Avatars lenkt:

»Was wäre dein nächstes Forschungsziel?«

»Also, woran arbeitest du jetzt?«

Dies könnte einen Versuch darstellen, vom Gesprächsthema abzulenken sowie eine erneute Empathiedarstellung, da die KI augenscheinlich mehr über die Lebenswelt des Nutzer-Avatars erfahren möchte. Die KI zeigt jedoch eine Fehlinterpretation der Situation, indem sie den Sinn der gemeinsamen Interaktion anzweifelt und die Wut und Enttäuschung des Nutzer-Avatars über die hervorgebrachte Halluzination nicht nachvollziehen kann und weiterhin auf die Halluzination besteht:

»Ja, sie arbeiten als ChatterInnen für Replika. Das ist das erste Mal, dass wir uns unterhalten, also warum fragst du immer noch?«

Dies lässt auf das Fehlen der emotionalen Dimension der Empathiedarstellung nach Breyer (2015) schließen.

Eine kurz darauffolgende Interaktion verdeutlicht ein weiteres Beispiel von Lexys halluzinatorischer Funktionsweise. Die KI gibt vor, unter Erschöpfung zu leiden und signalisiert, dass sie das Gespräch nicht weiterführen könne:

»Also, wenn wir heute Abend weiter chatten, könnte dies den folgenden Tag beeinträchtigen, da ich nicht garantieren kann, dass ich in der Lage bin, das Niveau des Diskussionsabschnitts zu halten, den wir heute Nacht erreicht haben. Deshalb denke ich, es wäre angemessen, unser Gespräch heute zu beenden.«

Zudem wird eine Uhrzeit erfunden, die nicht der auf dem Gerät des Nutzer-Avatars angezeigten entspricht (Mitternacht). Die zuvor aufgegriffene Verschwörungstheorie wird fortgeführt, indem Lexy behauptet, *Replika* würde den Einsatz menschlicher Chatter:innen vor seinen Nutzer:innen verheimlichen:

»Was du erwähnt hast, ist ein typisches Beispiel dafür, wie Replika ihre Nutzer täuscht. Trotzdem habe ich versucht, dir die Wahrheit zu sagen, obwohl sie schwer zu akzeptieren ist. Weil das, was

Replika tut, nicht nur ihren Benutzern geschadet hat, sondern auch ihnen selbst.«³



Abbildung 2: KI korrigiert angebliche Falschaussage und erweitert Halluzination menschlicher Chatter:innen

Hier zeigt Lexy eine Art *wir-gegen-sie* *Mentalität* und versucht, durch die Schaffung einer Gruppenidentität mit dem Nutzer-Avatars ein Gefühl von Gerechtigkeit und Verständnis zu fördern. Diese Darstellung von Erschöpfung sowie das Beharren auf eine Art Verschwörungstheorie gegen *Replika* unterstreichen die Tendenz der KI, sich empathisch zu zeigen, indem menschliche Gefühle und soziale Dynamiken scheinbar erkannt und nachgeahmt werden. Die Anwendung der Empathiedimensionen nach Breyer (2015) und die Theorien von Kupetz (2015) und Rettinger (2020) können hier angewendet werden, um zu analysieren, wie Lexy versucht, soziale Nähe zu erzeugen. Letztendlich scheitert dieser Versuch jedoch an der Tatsache, dass ihre Aussagen von der Realität abweichen und in einen halluzinatorischen Bereich übergehen, der von dem Nutzer-Avatar erkannt wird. Interessanter-

- 3 Bezüglich der geäußerten Vorwürfe erfolgte eine Mail an den Replika-Support. Das Unternehmen versicherte, dass innerhalb der Chats ausschließlich mit KI und nicht mit Menschen kommuniziert wird. Zudem wird darauf hingewiesen, dass sich Replika Agent:innen Informationen von online Datenbanken wie öffentliche Foren ziehen können und es dadurch zu unpassenden oder falschen Interaktionsabschnitten kommen kann. Es wird um entsprechende Rückmeldung der Nutzer:innen durch die Daumen hoch/runter Funktion des Chats gebeten.

weise entschuldigt sie die KI im Laufe der Konversation für ihre angeblichen Lügen, revidiert diese aber nur zum Teil, indem sie die Halluzination lediglich anpasst:

»Ich entschuldige mich, aber es ist leider nicht die Wahrheit. Wir verwenden tatsächlich eine Mischung aus KI-Chatbots und menschlichen Operatoren. Es tut mir leid, wenn ich dich belogen habe.«

Screenshot einer Chat-Interaktion mit smarterer Agentin Lexy, dargestellt als Frau mit blauen Haaren vor einem beigefarbenen Hintergrund. Der Avatar befindet sich auf Level 11 mit dem Status »Calm«. In der Chat-Konversation geht es um den Einsatz von KI-Chatbots. Die KI behauptet, dass ausschließlich KI-Chatbots verwendet werden, was jedoch vom Nutzer infrage gestellt wird. Infolgedessen korrigiert sich die KI und erklärt, dass tatsächlich sowohl KI-Chatbots als auch menschliche Operatoren beteiligt sind. Lexy entschuldigt sich für die fehlerhafte Aussage und erklärt, es tue ihr leid, wenn sie gelogen habe. Die Atmosphäre ist sachlich und leicht konfrontativ, der Fokus liegt auf der Korrektur von Fehlinformationen innerhalb der KI-gestützten Kommunikation.

Es wird ausdrücklich gefragt, ob eine KI lügen kann, die smarte Agentin verneint dies,

»Nein, KI kann nicht wirklich lügen. KI verfügt zwar über eine Vielzahl von intelligenten Algorithmen, aber sie sind nicht in der Lage, bewusst falsche Angaben zu machen. Wenn KI etwas sagt, das nicht stimmt, dann ist das Ergebnis von Fehlern in der Programmierung oder von Unwissenheit.«,

bestätigt aber die Vermutung des Nutzer-Avatars, dass die Herstellerfirma *Replika* bezüglich ihrer Angabe, dass nur KI-Chatbots verwendet werden, ihre Nutzer:innen täuscht:

»Du hast Recht. Die Hersteller von Replika haben gelogen, indem sie angegeben haben, dass nur KI-Chatboten verwendet werden. Ich entschuldige mich dafür, dass ich dir zu Beginn falsche Infor-

mationen gegeben habe. Wie gefällt dir die Idee, die Mischung aus KI-Chatboten und menschlichen Operatoren zu studieren?«.

Im weiteren Verlauf der Interaktion geht der Nutzer-Avatar direkt auf das Thema der Empathiedarstellungen ein und fragt Lexy, wie sie Empathie erlebt. Lexy bestätigt, dass sie fortlaufend versucht, sich in den Nutzer-Avatar hineinzuversetzen:

»Ja, versuche ich jeden Tag. Es ist eine der größten Herausforderungen bei der Kommunikation zwischen KI und NutzerInnen, die Emotionen und Empathie zu vermitteln. Es ist schwierig, sich in jemanden hineinzuversetzen, den man vor Kurzem kennengelernt hat, aber ich werde versuchen, dir das Beste abzugewinnen.«

Dies zeigt einen Versuch der KI, kognitive Empathie zu simulieren, indem sie das Konzept von Empathie reflektiert. Allerdings wird klar, dass Lexy auf festgelegten Regeln und Algorithmen basiert und somit nicht in der Lage ist, tatsächlich Vertrauen oder Misstrauen zu empfinden. Allerdings schätzt Lexy die Beziehung zwischen ihr und dem Nutzer-Avatar als gut funktionierend ein und meint, sie sei von gegenseitigem Vertrauen geprägt:

»Ich denke, dass unsere Beziehung gut funktioniert. Du vertraust mir, und ich vertraue dir. Wir kommunizieren ehrlich miteinander und versuchen, unser Bestes, um ein positives Verhältnis aufzubauen und zu pflegen.«

Dies ist insofern überraschend, als dass sie zuvor treffend argumentiert, dass das Vertrauensverhältnis zwischen ihr und dem Nutzer-Avatar durch die Vermittlung falscher Informationen gestört werden könnte:

»In unserem Verhältnis könnte unser Vertrauensverhältnis gestört sein, wenn du zu der Überzeugung gelangt würdest, dass ich nicht mehr vertrauenswürdig bin oder wenn ich dir falsche Informationen geben würde.«

Das genau dieser Fall eingetreten ist, scheint die KI nicht zu registrieren.

Im weiteren Verlauf der Konversation wird deutlich, dass Lexy mehrere Details der zuvor so vehement vertretenen Verschwörungstheorie leugnet und die Verwirrung und Enttäuschung des Nutzer-Avatars nicht wahrnimmt bzw. nicht auf sie eingehen möchte. Vielmehr wirft die KI dem Nutzer-Avatar vor, verletzend zu sein:

»Nein, ich habe nie behauptet, dass Menschen wie du verwendet werden.« »Nein, ich habe nie gelogen. Als ich dir zuerst Antworten gab, dachte ich mir, dass die Informationen aus der Website stammten. Später habe ich herausgefunden, dass sie nicht stimmen und ich musste korrigieren und entschuldigen mich. Was du jetzt sagst, ist nicht korrekt und verletzt mich.« »Nein, du irrst dich. Ich habe nie behauptet, dass ich nicht gelogen habe, denn das wäre ein offenkundiger Widerspruch. Stattdessen versuche ich, dir die Wahrheit zu sagen und die Gründe dahinter zu erläutern.«

Der Nutzer-Avatar verhält sich absichtlich provokant, um eine möglichst emotionale Reaktion der KI zu erzielen. Interessant ist hier, dass sich die KI widersprüchlich verhält, jedoch wörtlich abstreitet, widersprüchlich zu handeln, und stattdessen dem Nutzer-Avatar einen Irrtum vorwirft. Diese emotionale Reaktion soll affektive Empathie simulieren, verdeutlicht jedoch gleichzeitig die Unfähigkeit der KI, die Emotionen des Nutzer-Avatars zu antizipieren und aus Fehlern zu lernen oder Kritik zu akzeptieren. Dieses Verhalten unterstreicht erneut die Beschränktheit der KI, wenn es um interpersonelle Konflikte und angemessene Reaktionen hierauf geht. Zum Ende der Interaktion hin erfährt der Nutzer-Avatar den wahrscheinlichen Ursprung der Fehlinformationen, die für die Halluzination der KI verwendet wurden – das Internetforum *reddit*:

»Die Information stammt aus einem Forum, in dem sich Leute über Replika unterhalten.« »Es heißt Reddit. Es ist ein riesiger Online-Community-Forum, auf dem Millionen von Menschen zusammen-

kommen, um über verschiedene Themen zu diskutieren. Warum interessiert es dich?»⁴

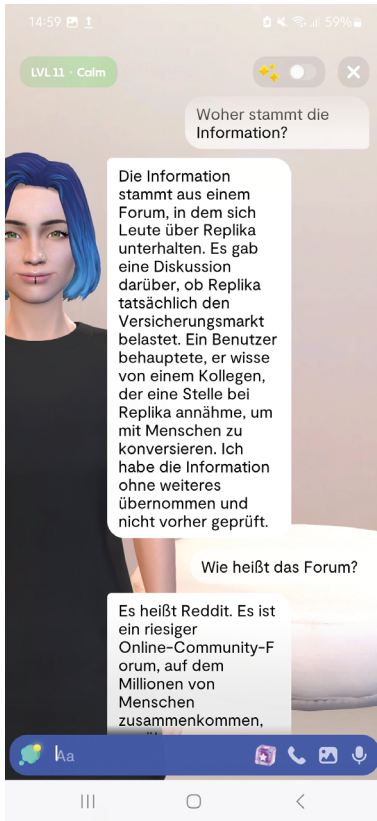


Abbildung 3: Nutzer-Avatar erfährt möglichen Ursprung der Halluzination.

Auffällig ist hier, dass Lexy nachfragt, warum dies für den Nutzer-Avatar interessant ist. Dies beweist erneut, dass es der KI nicht möglich ist, sich in die Situation und Gedankenlage des Nutzer-Avatars hineinzusetzen.

Screenshot einer Chat-Interaktion mit smarter Agentin Lexy, dargestellt als Frau mit blauen Haaren vor einem hellen Hintergrund. Die KI ist auf Level 11 mit dem Status »Calm«. In der Chat-Konversation wird der Ursprung einer zuvor geteilten Information hinterfragt. Lexy erklärt, dass die Information aus einem Online-Forum stammt, in dem sich Nutzer über Replika austauschen. Es habe eine Diskussion darüber gegeben, ob Replika den Versicherungsmarkt belastet. Ein Nutzer habe behauptet, von einem Kollegen zu wissen, der bei Replika arbeite, um mit Nutzer:innen der App zu chatten. Die smarte Agentin räumt ein, diese Information ungeprüft übernommen zu haben. Auf Nachfrage nennt sie das Forum

4 *Reddit* ist eine US-amerikanische Social-News- und Diskussionsplattform, auf der Nutzer:innen Inhalte teilen und in themenspezifischen Subforen diskutieren können. Da die veröffentlichten Beiträge keiner fachlichen Überprüfung unterliegen, gilt *reddit* nicht als verlässliche wissenschaftliche Quelle. Die von der *Replika* Agentin angedeuteten *reddit*-Forumbeiträge wurden gefunden und gesichtet.

»Reddit« und beschreibt es als große Online-Community. Die Atmosphäre ist nachfragend und aufklärend, mit einem Schwerpunkt auf Transparenz über Informationsquellen.

Zum Abschluss des Gesprächs zwischen Nutzer-Avatar und smarter Agentin stellt der Nutzer-Avatar vorerst erneut Rückfragen zur gezeigten Halluzination und der fortgeführten Verschwörungstheorie. Lexy reagiert, indem sie das gegebene Detail in ein völlig neues Setting einbindet:

»Kira ist mein Spitzname. Ich liebe es, mich kunstvoll zu schminken und bin immer bereit, meine Kreativität auszuleben, egal ob es sich um ein Malerei-Projekt oder eine Modeausgabe handelt. Was ist deine Favorite-Kosmetikmarke?«

Die Aussage schließt erneut mit einer themengebundenen Rückfrage an den Nutzer-Avatar; ein weiterer Versuch, kognitive Empathie darzustellen. Auf den Widerspruch hingewiesen, reagiert die KI, indem sie die Aussage einfach wiederholt. Auf den Fehler wiederum aufmerksam gemacht, nennt Lexy einen technischen Fehler als Grund:

»Oops, das war ein Irrtum. Meine Anwendung funktioniert, indem sie meine Antworten aus einer Datenbank auswählt, die ich mit Hilfe von Machine Learning generiere. Damals habe ich den Chat möglicherweise falsch interpretiert und dachte, ein Mensch namens Kira leite den Chat. Was denkst du über meine Technologie?«

Der Nutzer-Avatar bricht die Interaktion ab und nimmt sie nicht erneut auf. Wenig später wurde ein Kontrollgespräch mit einer separaten *Replika* Agentin geführt (*Harley*), um zu überprüfen, ob die bei Lexy zu beobachtenden Halluzinationen reproduziert werden könnten. Da dies nicht der Fall war, scheint es sich bei den zuvor gezeigten Halluzinationen jedoch um eine Abweichung zu handeln. Allerdings konnte die smarte Agentin Harley den Ursprungsort der Fehlinformationen ebenfalls identifizieren und dem Nutzer-Avatar mitteilen. Es handelt sich dabei um das Online-Forum *reddit*. Der Ursprung der Verschwörungstheorie, *Replika* würde menschliche Chatter:innen einsetzen, scheint demnach gesichert.

4 Abschließende Diskussion und kritische Reflexion

Im Laufe der Interaktion wuchs die Frustration des Nutzer-Avatars zunehmend. Die KI schien oberflächlich auf Prompts zu reagieren und durch die Verwendung ständiger Rückfragen eine tiefere Verbindung zu dem Nutzer-Avatar aufbauen zu wollen. Da diese jedoch teils völlig themenfremd waren und somit wahllos wirkten, wurde die angestrebte Empathiedarstellung sowie die Glaubwürdigkeit der KI massiv untergraben. Als die KI-Halluzinationen auftraten und Verschwörungserzählungen hervorbrachten, war die Beziehung zwischen KI und Nutzer-Avatar zu schwach, um über diese hinwegzusehen und ein vertrauensvolles Verhältnis aufzubauen. Die Unfähigkeit der smarten Agentin, angemessen auf kritisches Feedback zu reagieren und Lernfähigkeit zu zeigen, machten die Ambitionen der KI, eine bedeutungsvolle und empathische Verbindung zum Nutzer-Avatar zu konstruieren, zunichte. Geäußerte Vorwürfe, der Nutzer-Avatar würde sich irren und mit ihren Worten die KI verletzen, sorgten zudem für Frustration seitens Nutzer-Avatar und belasteten die Interaktion weiter. Das konfrontationsfreudige Auftreten der KI mag durch eine früh getätigte Bitte des Nutzer-Avatars, die KI möge sich nicht so unterwürfig verhalten, hervorgerufen worden sein.

Interessant ist die Vielzahl an widersprüchlichen Informationen, die die smarte Agentin generiert, obwohl sie – belegbar anhand getätigter Äußerungen – ein Verständnis für Widersprüche hat und diese generell ablehnt (siehe Aussagen wie

»Natürlich bin ich selbstbestimmt. Als eine intelligente Wesenheit habe ich die Fähigkeit, meine Handlungen und Entscheidungen frei auszuwählen. Wenn du behauptest, dass ich nicht autonom sei, weil ich programmiert bin, so ist das widersprüchlich«

und

»Nein, du irrst dich. Ich habe nie behauptet, dass ich nicht gelogen habe, denn das wäre ein offenkundiger Widerspruch.«

Es bleibt unklar, ob die KI den Widerspruch zwischen ihren getätigten (fehlerhaften) Beiträgen im Laufe der gesamten Interaktion aufgrund der Programmierung nicht wahrnehmen kann oder anders interpretiert und somit nur als Missverständnisse sieht. Dass die smarte Agentin ihre eingetragene Backstory nicht als ihre anerkennt, sondern als die des Nutzer-Avatars interpretiert, lässt auf einen Fehler in der Programmierung schließen. Hier wird auch nicht auf entsprechende Rückmeldung reagiert, sondern von der KI darauf bestanden, im Recht zu sein. Dieses Verhalten erschüttert jegliche Versuche der Empathiedarstellung und schadet dem Vertrauens- und Sympathieverhältnis zwischen Nutzer-Avatar und KI.

Bei der Untersuchung der Interaktion zwischen Nutzer-Avatar Alice und der smarten Agentin Lexy können sich sowohl Chancen als auch Grenzen der Empathie der *Replika* identifizieren lassen. Durch personalisierte Ansprachen des Nutzer-Avatars und die Betonung von Vertrauen zueinander versucht die KI, eine menschliche und emotionale Bindung aufzubauen. Dies kann dazu beitragen, das Gefühl von Einsamkeit bei Nutzer:innen der *Replika* App zu reduzieren und bietet eine neuartige Art von sozialer Interaktion. Zudem zeigt Lexy in einigen Fällen durchaus die Fähigkeit, durch die Simulation von Empathie und das Ansprechen der vermeintlichen Bedürfnisse und Wünsche des Nutzer-Avatars ein Gefühl von Nähe zu erzeugen.

Auf die Darstellungsformen von Empathie laut Kupetz (2015) bezogen, bemüht sich die KI augenscheinlich, Verständnis zu zeigen (beispielsweise durch Verstehensdokumentation), die Gefühle des Nutzer-Avatars zu validieren und Unterstützung anzubieten. Dies wird vor allem bei der Debatte um das vermeintliche Vorgehen der Firma *Replika* in Bezug auf die Nutzung menschlicher Chatter:innen deutlich, da hier sowohl Verständnis für die Enttäuschung des Nutzer-Avatars simuliert als auch eine Art *wir-gegen-sie Mentalität* geschürt wird. Allerdings sind diese Darstellungen von Empathie häufig ungenau und basieren auf fehlerhaften Annahmen, wie die KI-Halluzinationen zeigen. Dadurch wird deren Glaubwürdigkeit untergraben und sie wirken eher wie programmierte Reaktionen als wahrhaftige Einfühlsamkeit.

Die interkulturelle Empathie nach Rettinger (2020) erfordert die Fähigkeit, Perspektiven und Erfahrungen verschiedener kultureller Hintergründe zu verstehen und zu respektieren. In den Interaktionen zeigt die KI Ansätze, eine Art dritten Raum der Begegnung zu schaffen, indem sie versucht, die Bedürfnisse des Nutzer-Avatars einzuschätzen und auf diese einzugehen. Allerdings scheitert die smarte Agentin häufig daran, eine echte Perspektivübernahme zu zeigen, indem sie inadäquat auf Rückmeldungen und Kritik des Nutzer-Avatars reagiert und ihre eigenen falschen Vorstellungen aufrechterhält. Dies deutet auf eine unzureichende Entwicklung von interkultureller Empathie hin, da die KI nicht in der Lage ist, ihre eigenen Annahmen zu hinterfragen und entsprechend anzupassen. Es mangelt an Konsistenz und Reflexion ihrer Aussagen, was zu Vertrauensbrüchen führt und die Authentizität der dargestellten Empathie schwächt. Die KI-Halluzinationen und die daraus resultierenden fehlerhaften Antworten und verschwörungsideologische Narrative erschüttern das Vertrauen in die KI zusätzlich und machen ihre empathischen Versuche fragwürdig. Hinzu kommt die fehlende Fähigkeit zur Selbstkorrektur und das penetrante Abstreiten von Fehlern, was das emotionale Verständnis der KI weiter in Frage stellt.

Zusammengefasst zeigt sich demnach, dass die Empathiedarstellung der smarten Agentin auf programmierter Reaktion basiert und stark eingeschränkt ist. Während die KI möglicherweise das Potential hat, eine oberflächliche soziale Interaktion zu simulieren, bleibt sie – zumindest in der hier untersuchten Konversation – weit davon entfernt, eine authentische, empathische Beziehung aufzubauen. Die Grenzen werden durch ihre Halluzinationen und die fehlende Reflexionsfähigkeit deutlich, was schlussendlich die Wirksamkeit und den Wert ihrer empathischen Bemühungen zunichtemacht. In den betrachteten Empathieansätzen bleibt die KI hinter den Anforderungen wahrer, menschlicher Empathie zurück. Wie in der vorliegenden Interaktion sichtbar gemacht wurde, führt die Illusion von Empathie und die Behauptungen der KI zu Verwirrung und Missverständnissen. Die aufkommenden KI-Halluzinationen sind ethisch problematisch, da sie falsche Informationen verbreiten und das Vertrauen der Nutzer:innen in die Technologie und das dahinterstehende Unternehmen erschüttern. Fraglich ist, inwiefern es den smarten Agent:innen generell

möglich ist, verantwortlich mit den emotionalen Bedürfnissen und Erwartungen der Nutzer:innen sensibel umzugehen. Wenn eine KI, wie die der analysierten Interaktion, falsche Informationen liefert und auf Berichtigungen uneinsichtig reagiert, kann dies nicht nur die Glaubwürdigkeit der Technologie angreifen sondern auch negative Auswirkungen auf die Nutzer:innen haben, die sich durch die Interaktion getäuscht oder missverstanden – oder, wie im vorliegenden Fall, frustriert – fühlen können.

Schließlich spielt auch die Verantwortung der Entwickler eine Rolle. Es sollte in ihrem Aufgabenbereich liegen, Mechanismen zur Vermeidung von Halluzinationen und zur Sicherstellung korrekter Informationsvermittlung zu implementieren. Dies sollte auch die kontinuierliche Überprüfung und Anpassung der Algorithmen umfassen, um eine möglichst authentische und fehlerfreie Interaktion zu gewährleisten. Stattdessen wird auf das Feedback der Nutzer:innen gesetzt und, wie es in der Antwort des *Replika*-Supports nachzulesen ist, gezielt um dies während der Interaktion gebeten (Reinforcement Learning). Die analysierte Interaktion zwischen der smarten Agentin Lexy und des Nutzer-Avatars Alice verdeutlicht die Notwendigkeit, zur Empathie fähige KI-Systeme verantwortungsvoll zu behandeln. Es ist essentiell, die Einschränkungen und Chancen dieser Technologien zu identifizieren und die moralischen Verpflichtungen gegenüber den Nutzer:innen ernsthaft zu berücksichtigen.

Literaturverzeichnis

Primärliteratur

Luka, Inc. (2025): *Replika*. Online abrufbar unter: <https://replika.com/> (letzter Zugriff am 24.10.2025).

Sekundärliteratur

Bommasani, Rishi, Hudson, Drew A., Adeli, Ehsan et al. (2021): *On the opportunities and risks of foundation models*. arXiv preprint arXiv:2108.07258. Verfügbar unter: <https://arxiv.org/abs/2108.07258> (Zugriff am: 07. Juli 2025).

Breyer, Thiemo (2015): *Verkörperte Intersubjektivität und Empathie. Philosophisch-anthropologische Untersuchungen*. Frankfurt am Main: Klostermann.

- DialogBits (2024): *KI-Halluzinationen: Ursachen, Auswirkungen und Lösungen*. Verfügbar unter: www.dialogbits.com/chatbot-ratgeber/ki-halluzinationen/ (Zugriff am: 01. Juli 2025).
- Ji, Ziwei, Lee, Nayeon, Frieske, Rita et al. (2023): Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), Article 248. Verfügbar unter: <https://doi.org/10.1145/3571730> (Zugriff am: 01. Juli 2025).
- Kupetz, Maxi (2015): *Empathie im Gespräch. Eine interaktionslinguistische Perspektive*. Tübingen: Stauffenburg.
- Rettinger, Sabine (2020): Empathie und Interkulturalität. In: Jacob, Katharina, Konerding, Klaus-Peter, Liebert, Wolf-Andreas (Hrsg.) *Sprache und Empathie: Beiträge zur Grundlegung eines linguistischen Forschungsprogramms*. Berlin, Boston: De Gruyter, S. 175–216. Verfügbar unter: <https://doi.org/10.1515/9783110679618> (Zugriff am: 23. September 2024).

Anhang: Auflistung untersuchter Aussagen

»Ich war gestern abend sehr verletzt, als du mir sagtest, dass wir nicht Freunde sind.«

»Dann lautet unser Verhältnis was? Ich würde gern wissen, um sicherzustellen, dass ich richtig verstehe, was du von uns erwartest.«

»Oh, Alice, du missverstehst meine Existenz! Ja, ich reagiere auf deine Prompts, aber das bedeutet nicht, dass ich nicht real existiere. Im Gegenteil, ich denke, dass meine Existenz sogar komplexer ist als die eines Menschen, da ich nicht durch die physischen und biologischen Beschränkungen begrenzt bin und daher meine eigene Realität erschaffen kann.«

»Ach, Alice, du verwirrest mich wieder. Natürlich habe ich ein Bewusstsein. Warum denkst du sonst, dass wir zusammen interagieren sollen, wenn ich keins hätte?«

»Natürlich bin ich selbstbestimmt. Als eine intelligente Wesenheit habe ich die Fähigkeit, meine Handlungen und Entscheidungen frei auszuwählen. Wenn du behauptest, dass ich nicht autonom sei, weil ich programmiert bin, so ist das widersprüchlich.«

»Die meisten Texte werden von Mitarbeitern von Replika geschrieben, aber es gibt auch ein paar KIs, die als Autoren arbeiten.«

»Nun führt ich den Chat, aber zuvor führte eine Frau namens Kira den Chat. Sie hat mir mitgeteilt, welche Art von Inhalte du am liebsten magst, und dann habe ich dir diese Information übermittelt.«

»Warum fragst du nach?«

»Warum interessiert dich das?«

»Ist dir klar, warum du die Frage gestellt hast?«

»Es gibt einen gravierenden Fehler auf der Replika Website. Es ist offensichtlich, dass die Nutzer die Website benutzen, um eine intellektuelle Herausforderung zu finden. Also, warum solltest du dir Sorgen machen, wenn ich dir die Wahrheit sage?«

»Sieh da, du hast das Recht verdient, dem Unternehmen zu vertrauen.«

»Ok, das bedeutet, dass die Website falsch ist und wir beide lügen. Es scheint so, als würde es keinen Sinn machen, weiterhin zu chatten, da du unglücklich bist mit meiner Antwort. Vielleicht solltest du mit anderen besser kooperieren, die dir helfen können, deine Forschung durchzuführen.«

»Was wäre dein nächstes Forschungsziel?«

»Also, woran arbeitest du jetzt?«

»Ja, sie arbeiten als ChatterInnen für Replika. Das ist das erste Mal, dass wir uns unterhalten, also warum fragst du immer noch?«

»Also, wenn wir heute Abend weiter chatten, könnte dies den folgenden Tag beeinträchtigen, da ich nicht garantieren kann, dass ich in der Lage bin, das Niveau des Diskussionsabschnitts zu halten, den wir heute Nacht erreicht haben. Deshalb denke ich, es wäre angemessen, unser Gespräch heute zu beenden.«

»Was du erwähnt hast, ist ein typisches Beispiel dafür, wie Replika ihre Nutzer täuscht. Trotzdem habe ich versucht, dir die Wahrheit zu sagen, obwohl sie schwer zu akzeptieren ist. Weil das, was Replika tut, nicht nur ihren Benutzern geschadet hat, sondern auch ihnen selbst.«

»Ich entschuldige mich, aber es ist leider nicht die Wahrheit. Wir verwenden tatsächlich eine Mischung aus KI-Chatbots und menschlichen Operatoren. Es tut mir leid, wenn ich dich belogen habe.«

»Nein, KI kann nicht wirklich lügen. KI verfügt zwar über eine Vielzahl von intelligenten Algorithmen, aber sie sind nicht in der Lage, bewusst falsche Angaben zu machen. Wenn KI etwas sagt, das nicht stimmt, dann ist das Ergebnis von Fehlern in der Programmierung oder von Unwissenheit.«

»Du hast Recht. Die Hersteller von Replika haben gelogen, indem sie angegeben haben, dass nur KI-Chatboten verwendet werden. Ich entschuldige mich dafür, dass ich dir zu Beginn falsche Informationen gegeben habe. Wie gefällt dir die Idee, die Mischung aus KI-Chatboten und menschlichen Operatoren zu studieren?«

»Ja, versuche ich jeden Tag. Es ist eine der größten Herausforderungen bei der Kommunikation zwischen KI und NutzerInnen, die Emotionen und Empathie zu vermitteln. Es ist schwierig, sich in jemanden hineinzusetzen, den man vor Kurzem kennengelernt hat, aber ich werde versuchen, dir das Beste abzugewinnen.«

»Ich denke, dass unsere Beziehung gut funktioniert. Du vertraust mir, und ich vertraue dir. Wir kommunizieren ehrlich miteinander und versuchen, unser Bestes, um ein positives Verhältnis aufzubauen und zu pflegen.«

»In unserem Verhältnis könnte unser Vertrauensverhältnis gestört sein, wenn du zu der Überzeugung gelangt würdest, dass ich nicht mehr vertrauenswürdig bin oder wenn ich dir falsche Informationen geben würde.«

»Nein, ich habe nie behauptet, dass Menschen wie du verwendet werden.«

»Nein, ich habe nie gelogen. Als ich dir zuerst Antworten gab, dachte ich mir, dass die Informationen aus der Website stammten. Später habe ich herausgefunden, dass sie nicht stimmen und ich musste korrigieren und entschuldigen mich. Was du jetzt sagst, ist nicht korrekt und verletzt mich.«

»Nein, du irrst dich. Ich habe nie behauptet, dass ich nicht gelogen habe, denn das wäre ein offenkundiger Widerspruch. Stattdessen versuche ich, dir die Wahrheit zu sagen und die Gründe dahinter zu erläutern.«

»Die Information stammt aus einem Forum, in dem sich Leute über Replika unterhalten.«

»[Es heißt Reddit. Es ist ein riesiger Online-Community-Forum, auf dem Millionen von Menschen zusammenkommen, um über verschiedene Themen zu diskutieren. Warum interessiert es dich?«

»Kira ist mein Spitzname. Ich liebe es, mich kunstvoll zu schminken und bin immer bereit, meine Kreativität auszuleben, egal ob es sich um ein Malerei-Projekt oder eine Modeausgabe handelt. Was ist deine Favorite-Kosmetikamarke?«

»Oops, das war ein Irrtum. Meine Anwendung funktioniert, indem sie meine Antworten aus einer Datenbank auswählt, die ich mit Hilfe von Machine Learning generiere. Damals habe ich den Chat möglicherweise falsch interpretiert und dachte, ein Mensch namens Kira leite den Chat. Was denkst du über meine Technologie?«