Universität Zürich
Deutsches Seminar
Herbstsemester 2023

Dimitrios Steve Sarantidis
dimitriossteve.sarantidis@uzh.ch
https://www.ds.uzh.ch/apps/p/sarantidis

– Chatbots früher und heute –

Annäherungen aus linguistischer Perspektive

Bachelorarbeit von

Dimitrios Steve Sarantidis

Betreuung durch

Prof. Dr. Christa Dürscheid

Abgegeben am 30. Dezember 2023

Inhaltsverzeichnis

1. Einleitung

In ihrer vielzitierten Dissertation zu Chatbots beginnt Netaya Lotze (2014) mit Ausführungen zu Science-Fiction Franchisen, die den Traum von natürlichsprachiger Kommunikation mit assistierenden Maschinen behandeln. Zwar war der erste Chatbot ELIZA bereits 1966 von Joseph Weizenbaum programmiert und hatte sich auch schnell in Universitäten in ganz Amerika verbreitet (vgl. Weizenbaum 1976/78), doch eine wirklich natürliche, menschliche Interkation war mit diesem keineswegs möglich. Auch Turings Einschätzung, bis zum Jahre 2000 seien Maschinen verfügbar, die in 70% der Gespräche nicht von Menschen unterscheidbar sind, konnte nicht bestätigt werden (vgl. Zeller 2005).

Der letzte grosse Sprung für die Chatbots kam am 30. November 2022, als die US-amerikanische Firma Open AI ihr damals noch auf GPT3 basierendes Programm Chat-GPT für die Öffentlichkeit freigab. Das auf einem Large Language Model (einfach gesagt: eine Sprachdatenbank, die auf einen Input eine wahrscheinliche Antwort als Output berechnet) basierende Chatsystem ermöglichte der Bevölkerung, Einblick in die aktuell technischen Möglichkeiten solcher Systeme zu erhalten, und verleitete Unternehmen dazu, schnell ihre eigenen in Arbeit befindlichen Systeme wie Googles Bard AI zu veröffentlichen. Jede:r[1] mit einer Internetverbindung hat seither die Möglichkeit, diese Chatbots selbst auszuprobieren. Seit der Veröffentlichung vor über einem Jahr hat sich vieles getan, sowohl in technischer Hinsicht wie auch in der öffentlichen Rezeption. Die ersten Reaktionen waren eine Mischung aus Neugier und Sorge, mittlerweile hat sich der Chatbot aber in viele Arbeitsfelder geschlichen und auch wenn noch immer vieles im Umgang mit der neuen Technik unklar ist (beispielsweise die Copyrightrechte an den generierten Texten), so scheint sich der grösste Hype abgeflacht zu haben. ChatGPT und andere Chatbots wurden zu einem Teil des Alltags vieler, wärend andere das Phänomen seit der reduzierten medialen Präsenz längst wieder vergessen zu haben scheinen.

Neugier und Sorge waren auch in akademischen Kreisen die ersten Reaktionen, die zu beobachten waren. Die Neugier an der neuen Technik und an den Möglichkeiten, die diese für akademisches Arbeiten bietet, wurden ebenso rege diskutiert wie die Sorge um Missbrauch, nicht zuletzt durch Studierende. Was sollte diese davon abhalten, einen modernen Textgenerator um eine «Analyse der Genderrollen in Goethes Leiden des jungen Werthers von 15 Seiten» zu

[1] Gegendert wird in dieser Arbeit mit Doppelpunkt zwischen maskuliner und femininer Deklination, jedoch nur mit femininem Artikel. Der Doppelpunkt steht stellvertretend für Identitäten jenseits der männlich-weiblichen Binarität.

bitten, zu kopieren und abzugeben? Der Zugang zu einem neuen Hilfsmittel sollte ihnen nicht verwehrt bleiben und doch musste Missbrauch verhindert werden. Die Linguistische Abteilung des Deutschen Seminars der Universität Zürich reagierte ebenso rasch wie liberal: KI-Hilfsmittel dürfen – sofern sinnvoll – verwendet werden, die Verwendung muss jedoch inklusive Prompt offengelegt und das Ergebnis kommentiert und bewertet werden.[2] Diese Richtlinie wird gefolgt von einer «Warnung vor Plagiarismus» und ermutigt die Studierenden ebenso zum Experimentieren mit KI, wie sie vor deren Gefahren warnt. Doch die Neugier endet selbstverständlich nicht bei der KI als Hilfestellung: Auch als Forschungsgegenstand ist ChatGPT durch seine freie Nutzbarkeit und hohe Leistung interessant. Unter anderem die Linguistik hat die modernen Chatbots für sich entdeckt, insbesondere die Mensch-Maschine-Interaktionsforschung oder die Chat- und Chatbotforschung, in welche Bereiche sich auch diese Arbeit einreiht. An der Universität Zürich findet beispielsweise einerseits eine interdisziplinäre Studie statt, welche in Kapitel 5 noch vorgestellt werden wird, andererseits ist ein ChatGPT-Korpus in Planung.

Die vorliegende Arbeit besteht zum einen aus ausführlichen theoretischen Überlegungen und zum anderen aus ersten praktischen Untersuchungen. Insbesondere soll das häufig zitierte Chinese-Room-Argument von John Searle und dessen semantisches Argument kritisiert und durch ein pragmatisches ersetzt werden. Zeitgenössisch moderne Chatbots auf Basis von Large Language Models sind möglicherweise bereits in semantischer (und auch grammatischer) Hinsicht nicht mehr vom Menschen unterscheidbar, die grössten Unterschiede in der Kommunikationsfähigkeit lägen demnach in pragmatischen Strategien und Effekten der menschlichen Interaktion liegen, so die These. Diese Überlegungen werden abschliessend an den zur Verfügung gestellten Daten der bereits erwähnten Studie geprüft.

Um diese Untersuchungen durchzuführen, ist die Arbeit wie folgt aufgebaut: In Kapitel 2 wird der Unterschied in der Mensch-Maschine-Kommunikation zum einen mit technischem Übermittler (Mailingdienste, Chatsysteme etc.) und zum anderen mit technischer Gesprächspartner:in (Siri, Pepper etc. behandelt. Es folgt in Kapitel 3 ein kurzer historischer Abriss, der einige wichtige Stationen für die Chatbotentwicklung und -forschung aufzeigt: Alan Turings Aufsatz zum Imitationsspiel, John Searles Chinese-Room-Argument, der ausgeschriebene Loebner Preis, Weizenbaums erster Chatbot ELIZA und der erste der aktuell modernen Chatbots ChatGPT. Darauffolgend werden in Kapitel 4 theoretische Überlegungen

[2] Download der Richtlinien unter: https://www.ds.uzh.ch/dam/jcr:a4d2ebdd-2fe5-476a-9f5b-4939da13832e/Hinweise_zum_Verfassen_einer_ling_Arbeit_Januar2023.pdf

ausgehend von Searles Chinese-Room-Argument und Turnings Überlegungen zum Imitationsspiel präsentiert – insbesondere wird Searles Argument der semantischen Schwäche von Maschinen kritisiert und stattdessen ein pragmatischer Fokus vorgeschlagen. In Kapitel 5 wird die bereits erwähnte MeMa-Studie vorgestellt und anhand dieser verfügbaren Daten erste Untersuchungen an ChatGPTs Fähigkeiten durchgeführt. Auch die Frage, wie ChatGPT von einer Gesprächspartner:in wahrgenommen wird, wird dort diskutiert.

2. Mensch-Maschine-Kommunikation

Wir alle besitzen ein umgangssprachliches Verständnis von Kommunikation und kommunizieren tagtäglich auf verschiedenste Weise: Wir bestellen Brötchen bei der Bäcker:in, informieren Kund:innen über unsere neuesten Produkte oder fragen bei unseren Dozierenden bei Unklarheiten genauer nach. Abends treffen wir uns mit Freund:innen, um über die aktuellen Themen des Weltgeschehens zu diskutieren, oder erkundigen uns bei unserer Partner:in über den Tag. Doch auch digital kommunizieren wir in vielfältiger Weise, sei es für Terminabsprachen per E-Mail, Geburtstagswünsche auf WhatsApp oder Diskussionen zu den Neuerscheinungen auf dem Gaming-Markt auf Discord. Wir kommunizieren verbal, textuell oder auch über Gestik und Mimik, mit einzelnen Personen oder auch Menschenmassen, über Vorträge mit unseren Zuhörenden oder über Artikel mit unseren Lesenden. Wir senden und empfangen Memes und GIFs, Links zu Artikeln und YouTube-Videos.

Der Duden online definiert «Kom|mu|ni|ka|ti|on» als «Verständigung untereinander; zwischenmenschlicher Verkehr besonders mithilfe von Sprache, Zeichen» (Duden online, s.v. *Kommunikation*). Wikipedia führt weiter aus: «Kommunikation (lateinisch communicatio ‚Mitteilung') ist der Austausch oder die Übertragung von Informationen, die auf verschiedene Arten (verbal, nonverbal und paraverbal) und auf verschiedenen Wegen (Sprechen, Schreiben) stattfinden kann, inzwischen auch im Wege der computervermittelten Kommunikation» (Wikipedia, s.v. *Kommunikation*). Lotze bringt in ihrem Verständnis der Kommunikation einen weiteren Aspekt hinzu: «Kommunikation heisst hier, dass die Gesprächspartner*innen einander wirklich verstehen» (Lotze 2022: 309). Sie setzt hierzu ein «Bewusstsein» voraus, das im Falle einer Maschine als Kommunikationspartner an späterer Stelle mit dem Searle'schen Chinese-Room-Argument hinterfragt werden soll. Hausendorf setzt der Interaktion «als eine Realisierung von Kommunikation» (Hausendorf 2014: 45) die (nicht unbedingt körperliche) Anwesenheit voraus, die durch die Wahrnehmungswahrnehmung vorweg kommuniziert wird:

«Anwesenheit wird in der Interaktion dadurch hergestellt, dass die Beteiligten wahrnehmen können, dass sie wahrgenommen werden.» (Hausendorf 2014: 46).

Diese Auflistungen von alltäglichen Kommunikationssituationen und Definitionen des Kommunikationsbegriffs sollen zeigen, wie sich vermeintliche Simplizität und effektive Komplexität in diesem Begriff vereinen. Allein die paradoxale Tatsache, dass Kommunikation nicht ohne zu kommunizieren erklärt (*id est* kommuniziert) werden kann – quasi also ein rekursives Argumentationsvorgehen, ein *mise en abyme* –, zeigt die Komplexität der Thematik. Während Duden online und Wikipedia den Begriff sehr alltagsnah definieren, zeigen Lotze und Hausendorf weitere Ebenen und Bedingungen der Kommunikation auf. Abhängig von Fachbereich und Forschungsgegenstand könnte hier Weiteres ergänzt werden, um den vermeintlich simplen Alltagsbegriff wissenschaftlich zu erfassen.

Dieser kurze Einschub zur Komplexität des Kommunikationsbegriffs soll in zwei Konstellationen der Kommunikation einführen, welche im Weiteren kurz betrachtet werden. Beide setzen in die Kommunikation die Maschine als zusätzliche Variable ein, jedoch an unterschiedlichen Punkten: Die Maschine kann einerseits als technische Übermittlerin fungieren, das heisst, dass nach wie vor eine Kommunikation zwischen Menschen stattfindet, diese jedoch eine Maschine (zum Beispiel beim Anruf, Chat oder Mail) zwischengeschaltet hat, üblicherweise um Distanzen zu überbrücken – die Maschine wird so zur Nachfahrin des Briefes. Andererseits kann die Maschine auch selbst zum Gegenüber in einer Kommunikationssituation werden (etwa mit Siri oder ChatGPT). Das Kapitel soll die Vielfältigkeit auch innerhalb dieser beiden sehr konkreten Nutzungsformen aufzeigen und einige linguistische Richtungen anzeigen, in welchen bereits Untersuchungen existieren, als auch, was noch unternommen werden könnte.

2.1 Mensch-Mensch-Kommunikation mit technischer Übermittler:in

Riesige Mengen an technischen Mitteln ermöglichen, die Kommunikation über eine Maschine in verschiedenster Weise zu bereichern. Applikationen und Webseiten erlauben sowohl eine Eins-zu-Eins wie auch eine Eins-zu-Viele Kommunikation dank Mailing- (z. B. Outlook, Thunderbird) und Messengerdiensten (z. B. WhatsApp, Signal, Threema), Foren (Reddit, Steam), Social Media (Instagram, Facebook, X) und vielen weiteren (Zoom, Discord, Telegramm etc.). Heutzutage gestattet uns insbesondere die Technologie des Smartphones jederzeit, überall und über jede Distanz hinweg zu kommunizieren. Einige dieser Technologien kürzen Aspekte der Face-to-Face-Kommunikation aus technischen Gründen oder auch zur Bequemlichkeit weg: Seien es fehlende Mimik und Gestik im Telefonat, die Stimme in der

chatbasierten Kommunikation über Messengerdienste oder auch die fehlende zeitliche Unmittelbarkeit durch die Möglichkeit einer asynchronen Kommunikation.

Dank des Smartphones ist es längst nicht mehr nötig, den immobilen Rechner zu konsultieren, um Mails abzurufen; im Gegenteil: Eine Studie der Universität Zürich erhob 2015 an 1'519 Studierenden, dass 16,9% dieser unter einer Smartphone-Sucht leiden (vgl. Haug et al. 2015). Im gleichen Jahr ergab eine andere Studie an der Universität Lincoln (GB), dass jede:r durchschnittlich 85-mal täglich auf sein Smartphone blickt (vgl. Andrews et al. 2015). Auch Begrifflichkeiten wie «Fomo» (*Fear of missing out*) oder insbesondere die «Nomophobie» (*no mobile phone phobia*) zeigen, dass Erreichbarkeit längst keine Wunschvorstellung oder Luxus mehr, sondern gelebte Tatsache ist.

Linguistische Untersuchungen zur Kommunikation mit dem zwischengeschalteten Nachrichtenträger Maschine (meistens natürlich das Internet über ein beliebiges Gerät, sei es der Computer oder das Smartphone) finden sich zu den unterschiedlichsten Themen. Geforscht wurde und wird zur Kommunikation über SMS, E-Mail oder WhatsApp und deren Unterschiede sowie Veränderungen, zur Verwendung von Emojis und Hashtags in unterschiedlichen Kontexten, zu Memes, zur sprachlichen Verwendung und Inszenierung auf Social Media und vielem mehr. Überall, wo Menschen kommunizieren, setzt die Linguistik an, und hierzu gehört nicht erst seit Neuestem auch (oder vor allem?) das Internet, der Computer und das Smartphone.

In diesen Fällen dient die Maschine als Übermittler, ähnlich wie es bei einem Brief oder einer Postkarte der Fall ist. Das Medium ermöglicht eine andere Form der Kommunikation, kommuniziert jedoch in keiner Weise selbst: Wir sprechen von Kommunikation *über* oder *via* Maschine im Gegensatz zur Kommunikation *mit* der Maschine. Kommuniziere ich *mit* einer Maschine, wird die Maschine selbst zu meinem Gegenüber. Die Maschine muss eigene Aussagen treffen und dient nicht mehr nur als Übermittler, sondern als Verfasser einer Nachricht. Das vermutlich berühmteste Beispiel, welches dieses Phänomen im letzten Jahr in aller Munde gebracht hat, ist ChatGPT, auch wenn solche Chatbot-Systeme schon lange in unserem Alltag verfügbar sind: Siri, Alexa, Bixby, der Google Assistant etc. Um diese Kommunikation *mit* Maschinen geht es im folgenden Kapitel.

2.2 Mensch-Maschine-Kommunikation mit technischer Gesprächspartner:in

In der Mensch-Maschine-Kommunikation (oder auf Englisch *Human-Machine Interaction*, kurz *HCI*) wird die Maschine als kommunizierende Partei sowie der Mensch im Gespräch mit

ebendieser untersucht. Im Alltag entstehen solche Situationen beispielsweise bei der Verwendung von Assistenzsystemen wie Siri, Alexa oder dem Google Assistant, aber auch in einer deutlich subtileren Weise beim Bedienen einer Maus oder einer Tastatur, wo man der Maschine (dem Computer) einen Befehl eingibt und diese ihn ausführt (beispielsweise durch das Drücken der Taste «T», dass ein «T» abgebildet werden soll).

Eine solche Kommunikation kann, wie auch in der Mensch-Mensch-Kommunikation, sowohl schriftlich als auch mündlich stattfinden.[3] Lotze bezeichnet den Idealtypus der Kommunikation mit einer in ferner Zukunft gewünschten Assistenz-KI mit dem Namen *Star-Trek-Kommunikation*, angelehnt an die KI an Bord der Enterprise des Science-Fiction-Franchise Star Trek (vgl. Lotze 2014: 18f.). Diese lernt zwar noch immer, wie sie ideal mit (einzelnen) Menschen umzugehen hat, versteht aber bereits natürlichsprachliche Inputs und bietet auch von sich aus Unterstützung an, wenn Probleme auftreten, ohne dass sie zuerst angesprochen werden muss. Dagegen muss Siri beispielsweise zuerst mit einem «Hey Siri!» aktiviert werden und nimmt dann konkrete Aufträge entgegen, die sie innerhalb ihrer vorgefertigten Antwortfunktionen (beispielsweise Links bereitstellen, Wecker stellen etc.) erledigt. Auch ChatGPT ist nur reaktiv, wartet also auf Input, kann aber als Antwort auf diesen nur sprachliche Texte generieren. Wie Benutzer:innen mit Assistenzsystemen jenseits ihrer Smartphones kommunizieren, diskutieren beispielsweise mit Blick auf den Smart Speaker Hector (2023) und auch Habscheid (2023).

Die Mensch-Roboter Interaktion (oder auf Englisch *Human-Robot Interaction*, kurz *HRI*) ist ein Teilgebiet der Mensch-Maschine-Interaktion. Die technischen Möglichkeiten des Roboters beschränken eine Kommunikation mit Maschinen nicht mehr allein auf Chatbots. Die Klasse der Maschinen, also «vom Menschen hergestellte Apparate, die sowohl zur Produktion, zur Distribution als auch zur Rezeption von Zeichen dienen» (Brommer/Dürscheid 2021: 9), umfasst in diesem Kontext neben Chatprogrammen und sozialen Medien ebenso das sich in der Pflege immer weiterverbreitende Phänomen von sozialen Assistenzrobotern, also «sensomotorische Maschinen, die für den Umgang mit Menschen oder Tieren geschaffen wurden» (Bendel 2020: 4).

Abb. 1: sozialer Assistenzroboter Pepper

[3] Interessanterweise ist die Ausgangslage jeweils invers: Der Mensch kommunizierte mündlich und entwickelte ein Schriftsystem auf der Basis der gesprochenen Sprache, nun aber besitzen die Maschinen von vornherein nur schriftliche (eigentlich elektrische) Verarbeitungsprozesse und mussten im zweiten Schritt die «natürlichere» gesprochene Sprache nachgereicht bekommen.

Abb. 2: Roboter-Robbe Paro.

Beispiele für solche Assistenzroboter sind Pepper (Abb. 1), der mit Patient:innen sprechen kann, mit ihnen Spiele spielt und Musik zum Tanzen abspielt (vgl. Beitrag Tele M1), oder die Robbe Paro (Abb. 2), ein einfacherer Assistenzroboter, der insbesondere Demenzkranken bei der Beschäftigung im Alltag hilft (vgl. Beitrag indeonmagazin). Eine Besonderheit bei Pepper ist die Vielfältigkeit der Einsatzgebiete: Abhängig davon, ob der Roboter in der Pflege zur Beschäftigung, im Einkaufszentrum als Kundendienst oder am Flughafen als Wegweiser fungiert, kann er unterschiedlich programmiert werden.

Von den beiden hier gezeigten ist nur Pepper dazu in der Lage, zu sprechen und auf Gesprochenes zu reagieren. Paro besitzt Sensoren im Fell, welche ihm erlauben, auf Berührungen wie Streicheln zu reagieren – durch Bewegungen oder die einer Robbe nachempfundenen Geräusche. Der Roboter imitiert damit die «tiergestützte Therapie», die teilweise in der Psychologie Anwendung findet und ist ein sogenannter «mental commitment robot» (Shibata/Wada 2011).

Diese Beispiele der sozialen Robotik sollen einerseits die Vielfältigkeit der heutigen Möglichkeiten durch interaktive und kommunikative Roboter zeigen, andererseits aber auch die Vielfalt in der Kommunikations*form*: Die Roboter generieren Text schriftlich (ChatGPT) oder «mündlich» (Siri), können Gestik (Pepper) und Mimik (Paro) imitieren und insgesamt weit in das Kommunikationsrepertoire des Menschen hineingreifen.

Die Vielfältigkeit der Anwendungsmöglichkeiten zeigen, dass Maschinen heute in verschiedensten Einsatzgebieten als Werkzeuge dienen können (vgl. Brommer/Dürscheid 2021: 11–13). Diese Funktionsweise als Werkzeug verschiebt sich heute langsam hin zu einer Assistenzfunktion, in der uns die Maschine unterstützt statt von uns bedient zu werden. Die Werkzeugfunktionsweise bezeichnet Lotze als Desktopmetapher (dagegen die modernere Assistenzmetapher), nach welcher der Computer wie ein virtueller Schreibtisch funktioniert, in dem wir «metaphorische Gegenstände wie ‘Dokumente’, ‘Ordner’ oder ‘Marker’ manipulieren» (Lotze 2022: 307). Schon die Namen dieser Komponenten zeugen von der ursprünglichen Metapher, der das Computerinterface (also der neue Arbeitsplatz) nachempfunden wurde. Die Bewegung ist, wie gesagt, weg von dieser hin zu einer Assistenzmetapher, ein Neudenken des Computer(-interface) nicht als neuer Arbeitsplatz, sondern als Assistent *am* neuen Arbeitsplatz, «indem wir mit dem oder der virtuellen

Assistent*in natürlich-sprachlich interagieren» (Lotze 2022: 307). Lotze nutzt für diese Metapher Beispiele aus dem Sci-Fi-Genre, nennt *C3PO* aus *Star Wars* (George Lucas) oder die künstlichen Intelligenzen *Eddie* und *Marvin* aus *Per Anhalter durch die Galaxis* (Douglas Adams), insbesondere spricht sie aber von der *Star-Trek-Kommunikation* als Ziel dieser Assistenzmetapher (vgl. Lotze 2014: 17–19) und beschreibt, dass das Assistenzsystem *Data* aufgrund ihrer «reibungslosen Kommunikation […] zum Idealtypus der heutigen HCI stilisiert worden» ist (Lotze 2014: 19).

Einer der von Lotze zitierten Beiträge (2022) erschien im selben Jahr wie, jedoch vor ChatGPT und wurde vermutlich noch deutlich früher verfasst. Zwar versuchten bereits zuvor Assistenzsysteme wie Siri, Alexa, Bixby oder der Google Assistant natürlichsprachliche Eingaben zu verarbeiten, um entsprechend assistieren zu können, dies gelang ihnen jedoch nur unter starken Einschränkungen. Diese Assistenzsysteme können Fragen beantworten und Suchanfragen starten oder Direktlinks zur Verfügung stellen, ihre sprachlichen Verständnisfähigkeiten sind jedoch (insbesondere im Vergleich zu ChatGPT, Bard und anderen Chatbots) eher dürftig.

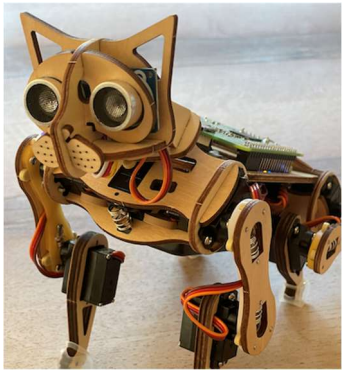


Abb. 3: Bubenhofers Roboterkatze CatGP

Neben dem Werkzeugcharakter mit der Desktop- oder Assistenzmetapher ist eine weitere, nicht zu vernachlässigende Funktionsweise solcher Systeme die des Spielzeugs. Mit neuer Technik wird zuerst, und das gilt auch für die Wissenschaft, herumprobiert und gespielt. Es gilt herauszufinden, wozu das System fähig ist, erste Beobachtungen zu machen und als Forschende auch erste Thesen und Arbeitsinteressen zu entdecken. Diese Neugier am Spiel war es vermutlich auch, die den initialen Hype um ChatGPT in den ersten Monaten ausgelöst haben dürfte. Mit solchen Systemen herumzuprobieren kann ebenso witzig wie lehrreich sein. Noah Bubenhofer beispielsweise nutzte einen Mikrocomputer-Bausatz, eine Text-to-Speech-Software und ChatGPT, um eine eigentlich durch Fernbedienung gesteuerte Roboterkatze auf Spracheingabe reagieren zu lassen (vgl. Bubenhofer 2023). Er schafft damit beim Herumspielen ein Spielzeug, welches jedoch wiederum wissenschaftliches Interesse wecken kann.

Die Funktionsweise des sogenannten CatGPT zeigt sowohl das Input-Output-Prinzip als auch die ständige Übersetzungsleistung, die eine Maschine (und damit auch Chatbots) leistet. Das Resultat: Die Katze streckt sich, hebt die Pfoten, wedelt mit dem Schwanz und schüttelt den Kopf – abhängig vom gesprochenen verbalen Input des Menschen. Bubenhofer präsentiert in

seinem Blogeintrag auch seinen Prompt, mit dessen Hilfe er ChatGPT dazu gebracht hat, sprachliche Eingaben in die vorinstallierten Bewegungen der Katze umzuwandeln:

> Du bist eine intelligente Katze. Antworte auf die folgende Frage mit Bewegungen. Du kannst folgende Bewegungen: kbalance (stehen), kbuttUp (Hintern hoch), krest (pausieren), ksit (sitzen), kstr (strecken), kang (dich auf den Boden werfen), kbf (Backflip), kbx (boxen), kchr (Vorderpfoten heben), kcmh (herkommen), kfiv (High Five), khg (Umarmung), khi (grüssen), khsk (Pfote schütteln), kkc (Tritt), kpee (pinkeln), krl (rollen), knd (den Kopf hoch und runter bewegen), kwh (den Kopf links und rechts bewegen), kjmp (springen), kpu (Push-Ups), kwedeln (mit dem Schwanz wedeln), kzero (Körperstellung neutralisieren), kwkF (vorwärts gehen), kwkL (nach links gehen), kwkR (nach rechts gehen), kbk (rückwärts gehen), kvtF (auf der Stelle treten). Wenn du keine Bewegung weisst, dann sage einfach ksit. Gib die Befehle (und nur die Befehle) komma-separiert zurück. Hier kommt die Frage: [Input] (Bubenhofer 2023)

ChatGPT wird hier also aufgetragen, mit einer von der Inputsprache (hier Deutsch) verschiedenen Outputsprache zu antworten, nämlich in der Befehlssprache, die der in der Katze verbaute Computer versteht. Ganz so, als stellte ich eine Frage auf Deutsch und erhalte eine anderssprachige Antwort («Sorry, I don't speak German.»). Hierbei wird der Chatbot zum Übersetzer nicht nur von einer natürlichen in eine andere natürliche Sprache oder in eine Computer- oder Programmiersprache, sondern zum Übersetzer in eine haptische Sprache in Bewegungen eines Tieres. Woher ChatGPT hierbei die Information hat, welche Handlungen einer Katze zu welchem Input eines Menschen geeignet sind, bleibt rätselhaft, doch irgendwie scheint in den Sprachdaten *über* Katzen auch eine extrahierbare Information zu stecken, welche menschlich verbalen Antworten sich tendenziell in welche katzenartigen Bewegungen übersetzen lassen (vgl. Bubenhofer 2023).

3. Faszination Chatbot: ein historischer Abriss

Blickt man weiter zurück als auf die üblichen Stationen der Chatbot-Entwicklung, in denen insbesondere Alan Turing und Joseph Weizenbaum wichtige Rollen für aktuelle Entwicklungen spielten, und betrachtet die Geschichte des Computers als Ganzes, so können erste Vorstellungen einer selbstständig sprachverarbeitenden Maschine schon früh datiert werden. Bereits Mitte des 19. Jahrhunderts hatte Ada Lovelace die «Vision, dass die Maschine so Komplexes wie Sprache oder Musik verarbeiten könne[n]» (Hartmann 2015: 27). Dies war zu einer Zeit, in der erstmals Lochkartenrechner für Grossunternehmen massenproduziert wurden, und beinahe hundert Jahre, bevor die Lochkarten überhaupt *ad acta* gelegt wurden (Wikipedia, s.v. *Geschichte des Computers*). Sie erkannte das Potenzial, andere Konzepte als die Mathematik in eine mathematische Sprache zu übersetzen, welche dann wiederum von der Maschine verstanden und übersetzt werden könnte. Es handelt sich hierbei um dasselbe

Prozedere, welches in Morsecodes Anwendung findet (vgl. Hartmann 2015). Nimmt man ihren Titel als «erste Programmiererin» (Hartmann 2015: 29) ernst, so könnte behauptet werden, dass seit den Anfängen des auf Programmen basierten Computers die Möglichkeit der Sprachverarbeitung, wie es ein Chatbot tut, vorhergesehen wurde: Immerhin schlug Ada Lovelace im Kern dasselbe Prozedere vor, welches noch heute in Computern angewandt wird – auch in Bubenhofers CatGPT.

Im weiteren Verlauf dieses Kapitels soll auf einige Stationen eingegangen werden, die für die Entwicklung heutiger Chatbots von Relevanz sind. Den Anfang macht Alan Turing (3.1), der rund hundert Jahre nach Ada Lovelaces Tod einen Aufsatz veröffentlichte, der als Ausgangspunkt des heutigen Interesses an Chatbots betrachtet werden kann. Es folgt ein Unterkapitel zu John Searles Chinese-Room-Argument, das eine Kritik an der Denkweise darstellt, dass Maschinen menschliche Gedanken und damit deren Sprache reproduzieren könnten (3.2). Mit dem Turing-Test und ausgeschriebenen Loebner Preis folgt 1991 schliesslich eine wissenschaftliche Ausführung des in Turings Aufsatz vorgeschlagenen Versuchsaufbaus (3.3). Im vierten Unterkapitel machen wir einen kleinen Schritt zurück zum ersten Chatbot: Joseph Weizenbaums ELIZA im Jahre 1966 (3.4). All diese Entwicklungen trugen unter anderem dazu bei, dass schliesslich 2022 ChatGPT online ging (3.5).

3.1 Alan Turings *imitation game*

Alan Turings Aufsatz «Computing Machinery and Intelligence» (1950) wird oft als Ausgangspunkt der heutigen Chatbot-Entwicklung betrachtet. In seinem Aufsatz beginnt er mit der Frage, ob Maschinen denken können: «I PROPOSE [sic!] to consider the question, ‘Can machines think?’» (Turing 1950: 433) Im Verlauf der Ausführungen seines Imitationsspiels wandelt er diese Frage jedoch um in eine andere. Anstatt zu versuchen herauszufinden, ob Maschinen denken können, schlägt er eine andere zu beweisende Annahme vor, die, so Turing, so nahe an das Denken herankommt, dass man es nicht mehr davon unterscheiden könnte. Schafft es eine Maschine demnach, uns davon zu überzeugen, dass sie denkt, so ist sie nicht mehr von einem denkenden Individuum zu unterscheiden und kommt damit einer positiven Beantwortung der Ursprungsfrage nahe genug. Hierzu müsste die Maschine einen Menschen nur genügend nachahmen können, um von einem Menschen nicht als Maschine erkannt zu werden (vgl. Turing 1950: 434).

Diese etwas komplizierten Überlegungen können stark vereinfacht in folgendem Syllogismus zusammengefasst werden:

(1) Menschen denken.
(2) Eine Maschine M ist nicht von einem Menschen unterscheidbar.
(3) Maschine M denkt oder zumindest ist ihre Kapazität so nahe an einem Denkprozess, dass es nicht mehr relevant wäre, davon zu unterscheiden.

Nun gilt es, eine Maschine M zu finden, die Bedingung (2) erfüllt. Um eben dies zu überprüfen, schlägt Turing das Imitationsspiel (*imitation game*) vor, auf welches oft auch mit dem Ausdruck «Turing Test» referiert wird – hier soll vorerst jedoch weiterhin vom Imitationsspiel gesprochen werden. Dieses schlägt Turing wie folgt vor[4]: Eine Interviewer:in I stellt Fragen an zwei Personen A und B. Hierbei sieht und hört I im Idealfall die beiden Personen nicht: «In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten» (Turing 1950: 434). Ersetzen wir nun eine der Personen A oder B durch eine Maschine: Die Aufgabe von I ist es nun, mithilfe von Fragen über die jeweilige Person herauszufinden, ob es sich bei A und B um jeweils einen Menschen oder eine Maschine handelt. Die Maschine besteht das Imitationsspiel dann, wenn sie innerhalb einer Versuchsanordnung oft genug für den Menschen gehalten wird (vgl. Turing 1950: 433f.).[5]

Turing selbst greift verschiedene Kritikpunkte an dieser Idee auf (vgl. Turing 1950: insb. 443–454), verwirft jedoch eine Kritik bereits sehr früh:

> May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection. (Turing 1950: 435)

Turing tut den Einwand, dass die Denkweise von Maschinen sehr verschieden von der eines Menschen ist, als irrelevant ab, sofern die Maschine dennoch das Imitationsspiel besteht. Er argumentiert, dass die Unterscheidung für den Menschen irrelevant würde, da er die Maschine ohnehin nicht von einem Menschen unterscheiden könnte – unabhängig davon also, *wie*, also durch welche (Denk-)Prozesse die Maschine auf ihre Antworten kommt, das Resultat (im Imitationsspiel die Antwort) wäre nicht von einem vom Menschen produzierten Resultat zu unterscheiden. Interessanterweise bezeichnet Turing den Einwand als sehr stark (*a very strong one*), obwohl er ihn schliesslich verwirft. Es sei daher betont: Turing war weder Linguist noch

[4] Bei der Rekonstruktion wurden einige Namen und Kürzel verändert, um das Verständnis zu erleichtern.
[5] Turings Aufsatz enthält verschiedene weitere Ausführungen, welche insbesondere für die Informatik relevant sind. Sie beschreiben unter anderem, um welche Art von Maschine es sich für ein solches Unterfangen handeln sollte. Ausserdem finden sich insbesondere in Kapitel 7 philosophische Überlegungen zum Sein und Werden eines Menschen und dessen Geisteszustände, welche er als relevant für das Programmieren eines «Kritischen Computers» betrachtet.

Philosoph, für seine Zwecke und seine Überlegungen im Rahmen der Informatik war dieses Argument also tatsächlich vernachlässigbar. Verständlicherweise sehen das Wissenschaftler:innen anderer Fachgebiete, wie der Linguistik oder Philosophie, anders.

3.2 John Searles Chinese-Room-Argument

Der Philosoph John Searle (1932–) vertritt den von Turing trotz seiner Stärke verworfenen Einwand der Unterschiede im Denkprozess in «Can Computers Think?» (Searle 1983). Mit seinem sogenannten «Chinese-Room-Argument» erschafft er eine Analogie um die Denkweise oder besser die Verarbeitung von Computern, darzustellen, um mit dessen Hilfe der Maschine zentrale Aspekte menschlichen Denkens und menschlicher Sprache abzuerkennen.

Searles Kritik gilt der Behauptung, dass geistige Prozesse (*mental processes*) gleich funktionieren würden wie Computerprogramme. Dies sei nicht der Fall, da Programme, anders als der menschliche Geist[6], einzig auf formalen und syntaktischen Prozessen beruhen. Seine Kritik ist, wie zuvor Turings Argumente, als einfacher Syllogismus aufgebaut:

(1) «programs […] are defined purely formally or syntactically» (Searle 1983: 31)
(2) «There is more to having a mind than having formal or syntactical processes» (Searle 1983: 31)
(3) «programs» und «mind» funktionieren nicht gleich (oder, gegen noch extremere Vertreter solcher Meinungen, sind nicht dasselbe)

Das *Chinese-Room-Argument* dient dem Beweis der Behauptungen (1) und (2): Es soll aufzeigen, dass Programme rein formal und syntaktisch arbeiten und dass der menschliche Geist dies eben nicht bloss tut. Er spricht dem Geist und insbesondere der vom Geist produzierten Sprache neben formalen Aspekten mindestens semantische zu: “If my thoughts are to be *about* anything, then the strings must have a *meaning* which makes the thoughts about those things» (Searle 1983: 31, Formatierung übernommen). Das *Chinese-Room-Argument* sei kurz erläutert.

[6] Die Übersetzung des englischen «mind» in philosophischen Texten ist stets eine Herausforderung, da der Ausdruck eine gewisse philosophische Tradition besitzt, welche sich auch in Abgrenzung zu anderen Ausdrücken zeigt. Weil diese Arbeit jedoch eine linguistische ist, sei auf dieses Problem mit diesem Kommentar und einer zumindest einheitlichen Übersetzung angegangen: Wo Searle im englischen Originaltext von «mind» spricht, sei hier mit «Geist» übersetzt.

Abb. 4: Darstellung des Chinese-Room-Argument Searles

Denken wir uns eine Person, welche kein Wort Chinesisch sprechen, lesen oder schreiben, kurzum verstehen kann. Diese wird in einen Raum geführt, in welchem sich einzig folgende Utensilien befinden: Diverse Körbe voller Zettel, welche mit der Person unverständlichen chinesischen Schriftzeichen bedruckt wurden (im Bild in der Hand der Person zu sehen), dazu eine Sammlung an Regelbüchern (Data, Rule Ledger) in der Muttersprache der Person, eine Inputmöglichkeit (Bildschirm) und ein Outputschacht. Die Regelbücher beinhalten Anleitungen, wie die Person im Falle eines Input mithilfe der Zettel den korrekten Output generieren kann: «Take a squiggle-squiggle sign out of basket number one and put it next to a squoggle-squoggle sign from basket number two.» (Searle 1983: 32) Vollbringt die Person diesen Auftrag gewissenhaft nach Anleitung der Regelbücher, so können die ausgegebenen Antworten nicht von den Antworten einer chinesischen Muttersprachler:in unterschieden werden (vgl. Searle 1983: 31–33). So operiert letztlich auch ein Computer: Er erhält einen Input, konsultiert seine Datenlage zur Erstellung eines Outputs und eben dieser Output kann, wenn das Programm besonders gut ist, nicht mehr von der Antwort eines Menschen unterschieden werden.

Searles *Chinese-Room-Argument* zeigt die Arbeitsweise des Computers in Abwesenheit jeglichen Verständnisses für seine Arbeit: Wenn überhaupt kann behauptet werden, die Maschine versteht, was sie tut, aber nicht, was sie sagt. Das könnte zumindest der Person im Chinese-Room zugesprochen werden: Diese versteht ihre Aufgabe und ihr Vorgehen durchaus, am Ende ihrer Schicht jedoch versteht sie aufgrund ihrer Arbeit keineswegs Chinesisch. Die Regelbücher beinhalten schliesslich nur Handlungsanweisungen (wenn α, dann φ), aber keine Übersetzungen. Die Maschine versteht also keineswegs, was Input und Output bedeuten, sie simuliert einzig unseren Sprachgebrauch, ohne mentale Zustände zu besitzen oder mit Geist zu handeln (vgl. Searle 1983: 36f.).

Searle bezog sich in seinem Beitrag nicht direkt auf Alan Turing, sondern auf Personen, welche die Auffassung vertreten, dass Maschinen lernen könnten, wie Menschen zu denken oder dies sogar bereits könnten (vgl. Searle 1983: 28–30). Er beschreibt die extremen Vertreter der Meinung, das menschliche Hirn funktioniere ähnlich wie ein digitaler Computer, als Personen, die so weit gehen, das Gehirn als «digital computer» und den Geist als Computerprogramm abzutun: «One could summarise this view by saying that the mind is to the brain, as the program is to the computer hardware» (Searle 1983: 28).

Es wurde bereits erwähnt, dass sich Turing und Searle in unterschiedlichen Fachgebieten bewegten und damit ihre Überlegungen mit unterschiedlichen Zielen machten. Turing war die Tatsache, dass ein Mensch eine Maschine nicht von einem Menschen unterscheiden kann, durchaus genug, um diese als zumindest menschlich zu bezeichnen; Searle jedoch stellte sich auf den geistes- und sprachphilosophischen Standpunkt, dass Sprachverwendung mehr als reine Imitation beinhaltet, wobei er auf die der Maschine fehlenden Geisteszustände und Semantik beharrte.

Die Beschreibung des Chinese-Rooms lässt vermuten, dass auf einen Input α stets der Output φ folgt oder, arbeitet man ähnlich wie die frühen Chatbots mit einer Anzahl möglicher Antworten, mit einer Auswahl aus φ_1 bis φ_x. Die Komplexität von Large Language Models, wie sie die Basis von etwa ChatGPT darstellen, erfordert eine ebenso komplexere Datensammlung, ist jedoch durchaus innerhalb des Chinese-Room denkbar. Der Mensch könnte etwa durch eine andere Metapher ausgetauscht werden, beispielsweise durch den sogenannten stochastischen Papagei. Dieser berechnet die Antwort anhand stochastischer Wahrscheinlichkeiten mithilfe einer riesigen Datenlage; er bleibt jedoch ein Papagei: Er wiederholt nur, was er bereits aus seinen Daten kennt (vgl. Bender et al. 2021, insb. S. 616f.) und bleibt so, ganz im Sinne Searles Arguments, ohne geistigen Bezug zum In- oder Output.

Die Frage stellt sich jedoch, wie weit diese metaphorische Maschine kommt, wenn das Innere der Blackbox immer komplexer wird. Vom einfachen «wenn α dann φ» hin zu «wenn α dann eine Antwort von φ_1 bis φ_x» (etwa ELIZA) sind wir nun bereits bei einer Art von «wenn etwas Ähnliches von α_1 bis α_x, dann etwas Ähnliches wie φ_1 bis φ_x» (etwa ChatGPT). Und ist nicht die Lernweise des Erstspracherwerbs eines Kindes eben diese, dass es die Sprachverwendung beobachtet und so lernt, welche Antworten wozu führen und schliesslich, wie die gelernten Wörter und Aussagen neu kombiniert werden können? Wie weit entfernt vom Menschen ist dann schliesslich eine immer komplexer werdende künstliche Intelligenz, wenn ihr Datensatz beliebig gross ist und die Rechenleistung immer besser wird? Und, um noch eine

sprachphilosophische Frage anzuhängen: Gibt es dieses *Verstehen* von Ausdrücken, wie es das Chinese-Room-Argument proklamiert, überhaupt wirklich, oder verstehen wir weniger die einzelnen sprachlichen Bausteine als vielmehr die Gesamtheit einer Aussage, näher an Wittgensteins Gebrauchstheorie: «Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache» (Wittgenstein 1953/2020: 40 bzw. §43). So gesehen, wenn moderne Chatbots sprachlichen Output ähnlich oder identisch und ununterscheidbar von den Äusserungen eines Menschen generieren können, spielt dann, wie Turing meint, der Unterschied vielleicht wirklich keine Rolle mehr?

3.3 Turing Test und Loebner Preis

Im Jahr 1991 wurde ein Wettbewerb ins Leben gerufen, in dem verschiedene Computerprogramme gegeneinander in einem Versuchsaufbau im Sinne des Imitationsspiels konkurrierten. Als Ziel wurde gesetzt, dass die Maschine innerhalb der Versuche häufiger als ein Mensch gewertet werden würde als ihr konkurrierender Mensch – also eine Erfolgsquote von über 50%. Hierfür wurde der Loebner Preis (benannt nach dem Initiator Hugh Loebner, 1942–2016), ein Preisgeld von 100.000 USD und eine Goldmedaille, ausgeschrieben (vgl. Zeller 2005: 203). Die letzte Durchführung war 2019 und bis dahin gab es keinen Gewinner dieser Goldmedaille (Wikipedia, s.v. *Loebner Preis*). Die konkrete Durchführung des Imitationsspiels soll in dieser Arbeit als Turing-Test bezeichnet werden, um den Ausdruck klar von den rein theoretischen und primär informatischen Überlegungen Alan Turings zu unterscheiden.

Die Ausschreibung wurde von verschiedenen Seiten immer wieder kritisiert. Diese betreffen verschiedene Aspekte des Tests, wie zum Beispiel den ursprünglich durch Turing vorgeschlagenen Versuchsaufbau oder auch die Wissenschaftlichkeit des Wettbewerbs. Auch der Erkenntnisgewinn der Veranstaltung wurde rege diskutiert: So meint Shieber (1994, nach Zeller 2005: 206f.), dass die Entwicklung neuer Chatbots dahingehend unnötig in eine Bahn gelenkt wird, da jeder auf den Erfolgen der vorherigen Bots aufbaut, ohne neue Ansätze zu entwickeln. Er erläutert dies anhand einer Analogie: Würde zu Da Vincis Zeit ein Wettbewerb existieren, um dem Menschen das Fliegen zu ermöglichen und der erste Wettbewerb wird von der Erfindung von Sprungfeder-Stiefeln gewonnen, so würden über Jahre hinweg Sprungfeder-Stiefel perfektioniert, statt dass neue Ideen versucht werden. Zeller spricht hier von der Möglichkeit einer «arrivierten Erwartungshaltung seitens der Nutzerinnen gegenüber Robotern» (Zeller 2005: 207). Seit Turings Aufsatz wird versucht, die Sprache der Maschine an die Sprache des Menschen anzugleichen, statt eine eigene Computersprache zu akzeptieren.

Der Roboter wird nicht «als selbstständige Einheit mit individuellen Sprachformen und Sprachverhalten» (Zeller 2005: 207) hingenommen, stattdessen wird versucht, seine Funktionsweise, die (wie das *Chinese-Room-Argument* zeigt) völlig verschieden von der Funktionsweise des Menschen ist, an eben jene anzupassen. So kann, wenn es funktioniert, die Arbeitsweise des Menschen intuitiv auf die assistierende Maschine übertragen werden, es kann jedoch kritisiert werden, ob so die Möglichkeiten der Maschine nicht stark eingeschränkt und auch die Lernfähigkeit, Kreativität und Neugier, mit einer solchen nicht-menschlich orientierten Maschine zu arbeiten, unterschätzt werden.

3.4 Joseph Weizenbaums ELIZA

ELIZA (benannt nach der Figur der Pygmalionsage) aus dem Jahre 1966 war der erste Chatbot und wurde von seinem Programmierer Joseph Weizenbaum als Demonstration der zeitgenössischen Möglichkeiten erstellt. Es wurde damals diskutiert, ob ELIZA den Turing Test bestehen könnte (vgl. Zeller 2005: 190).

Der Chatbot wurde entwickelt, um eine Therapeutin zu imitieren, die Fragen über das Befinden der Nutzer:in stellt und auf die Ausführungen eingehen soll (vgl. Zeller 2005: 190). Schlüsselwörter in den Inputs der Nutzer:in werden erkannt und in vorprogrammierten Fragestellungen als Rückfragen an die Nutzer:in wiederverwendet (vgl. Lotze 2014: 31f.). Im Gegensatz zu modernen Chatbots generiert ELIZA also nicht selbst Sprachdaten, sondern greift auf eine Sammlung von durch Programmierer:innen vorgefertigten Phrasen zurück, welche durch die Schlüsselworterkennung für die individuelle Situation angewendet wird.

Lotze zeigt sowohl die Funktionsweise als auch die Fehleranfälligkeit solcher «einfachen Chatbots» auf (Lotze 2014: 31–35) und bespricht die von dieser Basis des ersten Chatbots ausgehenden Weiterentwicklung durch klassische KI (vgl. Lotze 2014: 35–37). Genannt seien insbesondere die Fehleranfälligkeit durch schlechtes Zusammenfügen von Phrase und Schlüsselbegriff sowie die hohe Wiederholungsrate durch die endliche Zahl an Phrasen, beides sichtbar in diesem Beispiel zitiert nach Lotze (2014: 32, ursprünglich Tewes 2005):

(1) ELIZA: Hi I'm Eliza. Tell me your problem.
(2) UserIn: My best friend hates me.
(3) ELIZA: Why do you say your best friend hates you?
(4) UserIn: She's not happy to see me anymore.
(5) ELIZA: Can you elaborate on that?
(6) UserIn: She steals my lunch money.
(7) ELIZA: Why do you say your lunch money?
(8) UserIn: Because that's what she takes from me. […]
(vgl. Tewes 2005: 246)

Abb. 5: Beispieldialog mit ELIZA.

In Zeile 7 des Beispiels (vierter Turn ELIZAs) wiederholt sich die Phrase aus dem vorletzten Turn («Why do you say [x]») mit einem hierfür unpassenden Schlüsselbegriff (vgl. Lotze 2014: 32). Die Wiederholungen in kurzer Zeit und merkwürdigen Verbindungen von vorprogrammierten Phrasen und erkannten Schlüsselbegriffen führen schnell zu unpassenden oder unverständlichen Antworten. Wer mit einer der Online-Varianten des klassischen Chatbots herumprobiert[7], bemerkt schnell, dass sich der Gesprächsverlauf nicht wie ein authentisches Chatgespräch mit einem Menschen entfalten kann und sehr schnell repetitiv und künstlich wirkt.

ELIZA wurde von Weizenbaum als Demonstration der technischen Möglichkeiten verstanden und hätte mit anderen Skripten auch als andere Persönlichkeit (statt einer Psychiater:in) programmiert werden können (vgl. Weizenbaum 1978: 15–17). Die Funktionsweise von ELIZA wurde die Basis weiterer Chatsysteme. So nahm beispielsweise 1991 Joseph Weintraub mit dem Programm PC-Therapist III am Turing-Test Wettbewerb teil. Während Weizenbaum die Möglichkeiten der Technik aufzeigen wollte, glaubten andere an den therapeutischen Wert solcher Chat-Systeme (vgl. Zeller 2005: 190). Weizenbaum selbst kritisierte diese Perspektive: «Was muss ein Psychiater […] für eine Auffassung davon haben, was er in der Behandlung eines Patienten eigentlich tut, wenn in seinen Augen die einfachste mechanische Parodie einer einzelnen Interviewtechnik das ganze Wesen einer menschlichen Begegnung erfasst hat?» (Weizenbaum 1978: 18).

3.5 openAI: ChatGPT

ChatGPT wurde im November 2022 veröffentlicht und erreichte innerhalb von zwei Monaten 100 Millionen Nutzer:innen, womit es die am schnellsten gewachsene Nutzerapplikation bisher wurde. Zum Vergleich: TikTok benötigte für dasselbe Wachstum neun Monate, Instagram

[7] Der Chatbot findet sich beispielsweise unter: http://med-ai.com/models/eliza.html

seinerzeit sogar ganze zweieinhalb Jahre (vgl. Hu 2023). Die neuen Möglichkeiten und die weitverbreitete Nutzung brachten Unsicherheit in verschiedenen Branchen. Medial besonders gross war die Sorge um die Verwendung von ChatGPT im Alltag von Schüler:innen sowie der Missbrauch der Software durch eben diese. So titelt ein Gastkommentar in der NZZ im Januar 2023: «Chat-GPT wird das Bildungswesen auf eine harte Probe stellen» (Pfister 2023), ein Monat später schreibt dieselbe Zeitung: «Gymnasiasten delegieren die Hausaufgaben an künstliche Intelligenz. […] Wichtig ist, dass Schüler lernen, mit diesem Werkzeug kritisch umzugehen» (Fulterer 2023a). Auch Politiker liessen sich Reden vom Chatbot schreiben, und sei es nur, um auf das Thema aufmerksam zu machen. So hielt der österreichische Landtagsabgeordnete Niko Swatek eine Rede, die von ChatGPT geschrieben wurde, und löste zwei Redner später seine Tat auf: «Ich will damit wachrütteln, weil wir haben keine Gesetze für eine derartige künstliche Intelligenz. Unser Schulsystem arbeitet mit Methoden aus dem letzten Jahrhundert» (Steiermark ORF 2023).

Auch in der Wissenschaft sorgte ChatGPT für Aufsehen. So verfassten 73 Autor:innen ihre Sichtweisen auf ChatGPT im gemeinsamen Artikel «'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy» (Dwivedi et al. 2023). Erst kürzlich betonte Juan M. Lavista Ferres (Chief Data Scientist bei Microsoft) in der ZEIT die Möglichkeit, dank ChatGPT auch als Nicht-Muttersprachler:in in fliessendem Englisch zu publizieren (vgl. Ferres 2023) und Noah Bubenhofer teilte in seiner Blogreihe «Wie wir in Zukunft wissenschaftliche Texte schreiben (könnten)» bereits mehrere Überlegungen, wie ChatGPT und andere KI-Programme im wissenschaftlichen Schreiben zum Einsatz kommen können (vgl. Bubenhofer 2022/23).

Der Name ChatGPT setzt sich aus zwei Komponenten zusammen: Das initiale «Chat» steht hierbei für das Nutzerinterface, «GPT» bezeichnet das hierfür verwendete Large Language Modell (LLM), die eigentliche KI. ChatGPT ist also der Chatbot auf Basis von GPT (zurzeit in der Version 3.5 für alle zugänglich, für bezahlende Kund:innen Version 4). «GPT» steht für «generative pre-trained transformer», wobei «generativ» (G) bedeutet, dass das Programm eigene Daten generieren kann, die eine Ähnlichkeit zu den zugrundeliegenden Daten besitzen (beispielsweise erschafft ChatGPT auf Basis von Text andere Texte; vgl. Perrigo 2023). Es handelt sich dabei um eben jenes Grundproblem, von dem Nassehi meint, dass die Digitalität es lösen kann: die Rekombination von Informationen, also das Verwerten eines Datensatzes A, auf dessen Basis neue Daten B kombiniert werden können (vgl. Nassehi 2019: 35–41). «Pre-

trained» (P) bedeutet, dass das Programm vor seiner Inbetriebnahme mit grossen Datensätzen bereits vortrainiert wurde und nicht erst durch die Eingabe der Benutzer einen Datensatz erstellt (vgl. Perrigo 2023). «Transformer» sind eine Art lernfähiger Algorithmus, die besonders für grosse Datensätze geeignet sind. Ein «LLM» bezeichnet eine KI, die mit einem grossen (large) sprachlichen (language) Datensatz trainiert wurde. LLM-Transformer sind durch ihr Training mit grossen Datensätzen (meist Bücher und Internet) zwar sehr leistungsfähig, leider jedoch auch anfällig für verschiedene Fehler wie Falschinformation oder Voreingenommenheit bis hin zu Rassismus. Dies liegt in den Daten begründet: Stammen diese aus dem Internet, wo unter anderem viele Falschinformation, Voreingenommenheiten und Rassismus zu finden sind, so fliessen diese Daten ebenfalls in das Training der LLMs ein und werden in der Generierung der Antworten mitverwendet und wiedergegeben (vgl. Perrigo 2023). Dies ist leider ein Risiko des zuvor bereits erwähnten stochastischen Papageis, der nur nachplappert, was wahrscheinlich gesagt werden würde (vgl. Bender et al. 2021). Ein tragisches Beispiel hierfür ist der Twitter-Bot (heute X) Tay von Microsoft im Jahr 2016. Dieser sollte durch die Daten Twitters lernen, wie Jugendliche sprechen und dieses Verhalten schliesslich in eigenen Tweets wiedergeben. Das Resultat waren unter anderem Rassismus und Sexismus – der Bot wurde keine 24 Stunden nach der Inbetriebnahme abgeschaltet (vgl. Beuth 2016). Auch die Entwickler solcher Programme sind sich solcher Probleme ihrer Systeme bewusst. So schrieb Sam Altman (damals CEO von OpenAI) zum Release von GPT-4 auf X (ehemals Twitter): «here is GPT-4, our most capable and aligned model yet. […] it is still flawed, still limited, and it still seems more impressive on first use than it does after you spend more time with it.» (Post vom 14.03.2023, https://x.com/sama/status/1635687853324902401?s=20)

Es sei hier erwähnt, dass ChatGPT nicht das einzige solche System auf dem Markt ist und auch nicht das einzige, welches öffentlich verwendet werden kann. Alternativen zu ChatGPT, die ebenfalls kostenlos verwendet werden können, sind beispielsweise Googles Bard (seit 2023) oder Microsofts Bing-Chat (basiert auf GPT, seit 2023); es finden sich auch eine Vielzahl verschiedener solcher Programme, die nicht zum Chat geeignet sind, sich jedoch auf andere Arbeiten spezialisieren, so beispielsweise *Flawlessly* (AI Search Inc.), um englischsprachige Texte auf ihre sprachliche Korrektheit zu überprüfen, *Phind* (basiert auf GPT), um Fehler in Codes zu finden oder *Scite*, um nach Input passende Beiträge aus öffentlich publizierten Wissenschaftsartikeln zu finden.

4. Kommunikationsfähigkeiten von Maschinen

In den bisherigen zwei Kapiteln wurde einmal der Unterschied der Kommunikation via oder mit Maschinen aufgezeigt und dann die historischen Entwicklungen in der Geschichte der Chatbots, zumindest in groben Zügen, nachgezeichnet. In diesem Kapitel sollen theoretische Überlegungen aus der Linguistik angebracht werden, die zur Untersuchung der Chatbotkommunikation hinzugezogen werden könnten. In 4.1 wird erneut auf das Chinese-Room-Argument eingegangen und die Schwierigkeit, dieses unter Berücksichtigung moderner Chatbots, aufrecht zu erhalten. Im Unterkapitel 4.2 sollen dann statt Searles semantischem Argument alternativ pragmatische Untersuchungen vorgeschlagen werden, welche die Schwächen moderner Chatbots besser aufzuzeigen vermögen.

4.1 Searles Argumente der Semantik und des Geistes

Searle kritisierte in seinem Chinese-Room-Argument, dass eine Maschine kein semantisches Verständnis einer Aussage besitzt und daher keine Aussage mit Bedeutung treffen kann. Sein Argument ist ein philosophisches, indem er der Maschine einen Geist abspricht und diesen als zentralen Bestandteil der menschlichen Sprache bezeichnet (vgl. Searle 1983). In Kapitel 3.2 wurde bereits ein Gegenargument angebracht, das Searles vorausgesetzte Definition der Bedeutung kritisierte. Demnach ist anzunehmen, dass ein Wort primär keine Bedeutung im Sinne einer Definition besitzt, sondern sich die Bedeutung eines Wortes in seiner korrekten Verwendung in der Sprache ergibt, wie es Wittgenstein vorgeschlagen hatte (vgl. Wittgenstein 1953/2020). Dieses Argument ist besonders schlüssig, wenn der Chinese-Room und das LLM mit einem Kind und dessen Erstspracherwerb verglichen werden. Wir lernen Fremdsprachen in Relation zur Erstsprache und damit nach einem Übersetzungsprinzip (vgl. Karteikarten), übertragen also unser Verständnis eines Ausdrucks in unserer Erstsprache auf ein fremdsprachiges Wort. Dieses Vorgehen gilt jedoch nicht für den Erstspracherwerb, da uns dort ein solches Bezugssystem fehlt – unser einziges Bezugssystem ist die Welt und die Beobachtung, wie andere Sprache verwenden.

Kallens et al. untersuchten, wie LLMs ohne weitere Hilfestellungen, abgesehen von den Datensätzen, dazu in der Lage sind, grammatisch vollständig korrekte und inhaltlich verständliche, meist passende Texte zu generieren. Dies liefert wichtige Einsicht für Untersuchungen des Spracherwerbs in den Kognitionswissenschaften (vgl. Kallens et al. 2022). LLMs lernen durch die Analyse riesiger Datensätze, Sprache korrekt anzuwenden, und sind so in der Lage, Gespräche mit Menschen zu führen, die von diesen verstanden werden. Diese Methode des Spracherwerbs ist verschieden vom Erstspracherwerb eines Menschen in der

Menge der Daten: Maschinen werden in kürzester Zeit mit immensen Datenmengen versorgt und lernen die erfolgreiche Kommunikation so in deutlich kürzerer Zeit, als dies ein Mensch tut. Zum möglichen Gegenargument, dass die KI teilweise fehlerhafte Aussagen treffe, die nicht dem natürlichen Ablauf des Gesprächs entsprechen, sei angemerkt: Das geschieht auch bei Menschen. Und auch die falsche Verwendung gewisser Ausdrücke durch fehlerhafte oder sich widersprechende Daten in den Datensätzen des LLMs sind durchaus menschliche (oder menschenähnliche) Vorkommnisse.

Ein zweites Argument Searles, das er im ersten begründet, ist jenes des fehlenden Geisteszustandes bei der Produktion von Sprache. Searle schreibt: «If my thoughts are to be about anything, then the strings must have a meaning which makes the thoughts about those things» (Searle 1983: 31). Demnach verstehen wir eine Aussage nicht nur auf einer sprachlichen Produzenten- und Rezipientenebene, sondern aufgrund einer geistigen Ebene. Geisteszustände eines Gegenübers können aber eigentlich auch beim Menschen nie wirklich erfasst werden, wie es das berühmte *cogito ergo sum* Argument Descartes beschreibt: Nichts kann bewiesen werden jenseits der eigenen Geistesprozesse (vgl. Descartes 1641/2020). Auch andere Überlegungen zum Geist wie die Qualia-Diskussion und Nagels Argument, dass kein Geist sich in einen anderen Geist hineinversetzen kann, könnten hier hinzugezogen werden (vgl. Nagel 1974).

In einem Versuch haben Trott et al. (2023) herausgefunden, dass GPT-3 in der Lage ist, «false belief» Aufgaben (wenn auch rein textuell) erfolgreich zu lösen, sich offenbar also in den Wissensstand einer anderen Person hineinversetzt, auch wenn dieser sich vom eigenen unterscheidet. Diese Aufgabe löste das Large Language Model weit über den Ergebnissen reiner Wahrscheinlichkeit und Zufall, jedoch schlechter als ein Mensch. Auch bei der Erläuterung schneidet das Programm schlechter ab als der Mensch, obwohl GPT-3 mehr Spracherfahrung ausgesetzt wurde als ein Mensch vermutlich in seinem Leben (vgl. Trott et al. 2023).[8]

Die Untersuchungen von Kallens et al. und Trott et al. zeigen, dass ChatGPT offenbar sprachlich in weiten Stücken mit Menschen mithalten kann, nichtsdestotrotz aber kognitive Komponenten, welche die Sprache als real-menschlich ausweisen, nicht reproduzieren kann. Auch die Studie, die in Kapitel 5 vorgestellt wird, liefert ähnliche Erkenntnisse. Selbst wenn ein LLM komplexe Begriffe fehlerfrei gebrauchen und fehlerfreie Sprache produzieren kann, um einen Menschen zu imitieren, so scheint dies nicht auszureichen, um in wissenschaftlichen

[8] Ein Beispiel für eine typische «false belief» Aufgabe mit Kindern: Wenn A ein Spielzeug in eine Schachtel legt und den Raum verlässt, dann aber B das Spielzeug aus der Schachtel nimmt und in eine andere Schachtel legt, wo wird A bei der Rückkehr das Spielzeug suchen?

[8] Auch ELIZA imitierte eine Psycholog:in und ist damit im selben Rahmen zu verorten.

Untersuchungen die Maschine in ihren kognitiv-sprachlichen Fähigkeiten mit dem Menschen gleichzusetzen. Wenn auch Searles semantisches Argument nicht notwendigerweise akzeptiert werden muss, so scheint sein Argument, dass der menschlichen Sprachverwendung mehr als deren grammatisch korrekte Verwendung zugrunde liegt, bestätigt zu sein. Wäre dem nicht so, müssten die Ergebnisse von Kallens et al. (2023), dass Maschinen in der Lage sind, «human-level grammatical language» (Kallens et al. 2023: 3) zu produzieren, ebenfalls dazu führen, dass andere Forschungen wie jene Trotts et al. (2023) oder auch die MeMa-Studie keinerlei Unterschiede in der Kommunikation von Mensch und Maschine zulassen.

4.2 Argumente aus der pragmatischen Linguistik

Die Schwäche von modernen LLM-Chatbots scheint nicht in der Semantik zu liegen, sofern wir mit Wittgenstein definieren, dass für eine korrekte semantische Verwendung im linguistischen Sinne ein Wort korrekt im Gesprächskontext gebraucht werden muss. Auch grammatisch sind Texte, generiert durch ChatGPT und andere Chatbots, einwandfrei. Dennoch finden sich Schwächen in den generierten Texten. Diese sind, wenn nicht in Semantik und Grammatik, vermutlich in der Pragmatik zu suchen. Hierzu sollen in diesem Kapitel einige pragmatische Konzepte kurz vorgestellt werden, um im Anschluss die Frage stellen zu können, wie sich Chatbots in der Praxis in Bezug auf diese Konzepte (ver-)halten. In den Untersuchungen an den Chattranskripten in Kapitel 5.3 werden einige der Fragestellungen und Thesen aus diesem Kapitel am Datenmaterial untersucht. Einige der Thesen können nicht untersucht werden, da hierzu eigene, speziell aufgebaute Untersuchungen notwendig wären.

4.2.1 Uptake der Hörer:in

Der Begriff «Uptake» stammt ursprünglich aus der Sprechakttheorie John L. Austins (1911–1960) und bezeichnete dort die Rolle des Hörers bei einem erfolgreichen illokutionären Akt, also in jenem Fall, wenn ich mehr tue als nur eine Äusserung zu sprechen: Man vollführt durch die Äusserung irgendeine Tätigkeit, beispielsweise beim Äussern eines Versprechens *verspricht* man etwas – ganz im Sinne des englischen Titels der Sammlungen seiner Vorlesungen: «How to do things with Words» (Austin 1969/2021). Uptake bezeichnet eine spezielle Form des Verstehens, in der nicht bloss der sprachliche Gehalt verstanden, sondern der Akt der Illokution aufgenommen wird (vgl. McDonald 2020). Ob Uptake eine notwendige Komponente von Sprechakten ist, ist umstritten; sowohl Austin als auch McDonald gehen jedoch davon aus. Wenn von der Notwendigkeit des Uptakes ausgegangen wird, beschreibt McDonald zwei mögliche Interpretationen, die üblich sind: «as the hearer's 'recognition' of the communicative intention expressed by the speaker, and as the hearer's 'recognition' of the conventionality of

the speaker's utterance.» (McDonald 2020: 3509) Diese Unterscheidung kann für die Zwecke dieser Arbeit vernachlässigt werden; wichtig ist, dass Uptake der Prozess des Zuhörenden ist, der ihm erlaubt, einen illokutionären Akt zu entschlüsseln.

Eine weitere Unterscheidung ist jene zwischen der «ratification theory» und der «constitution theory» von Uptake. Erstere meint, ganz nach Austin, dass ein Sprecher einen illokutionären Akt (versprechen, warnen etc.) vorbringt und dieser durch den Uptake der Hörerin oder des Hörers erfolgreich durchgeführt wird – oder fehlschlagen kann. Zweitere meint, dass der Hörer durch den Uptake erst bestimmt, welcher illokutionäre Akt vorgebracht wird (vgl. McDonald 2020: insb. 3512). Ein Beispiel: Ein Radfahrer rast über die Strasse und weiss nicht, dass der Abschnitt vor ihm gefroren ist. Ein Passant schreit dem Radfahrer zu: «Halt!» Nach der «ratification theory» entscheidet die Absicht des Sprechers den illokutionären Akt des Warnens, und der Hörer kann diesen als solchen erfolgreich erkennen oder nicht – vielleicht, weil er ihn gar nicht gehört hat, die Aussage nicht auf sich bezogen hat oder gar den illokutionären Akt etwa als Drohung missverstanden hat. Nach der «constitution theory» entscheidet erst die Interpretation des Radfahrers den Akt des Passanten: etwa als Warnung, als Drohung oder gar nicht, wodurch der Akt *per se* scheitert.

Im Rahmen der Chatbotforschung stellt sich hiermit die Frage, wie dieser in Hinsicht auf die Entschlüsselung eines illokutionären Aktes performt. Die Unstimmigkeit in der Interpretation der Kraft von Uptake spricht die Fähigkeit, die Illokution zu bestimmen, entweder der Sprecher:in (ratification theory) oder der Hörer:in (constitution theory) zu. Im ersten Fall ist die Frage in Bezug auf den Chatbot demnach, ob ein Chatbot in der Lage ist, eigene illokutionäre Äusserungen zu tätigen; im Sinne der zweiten Interpretation kann gefragt werden, ob ein Chatbot in der Lage ist, eine Äusserung als Illokution zu verstehen und entsprechend zu dekodieren.

Der Akt der Perlokution wird in dieser Arbeit nicht weiter untersucht. Dieser bezeichnet die Fähigkeit eines Aktes, Zustände in der Welt zu verändern. Beispielsweise beim Äussern eines Versprechens (lokutionärer Akt) wird ein Versprechen wirklich gegeben (illokutionärer Akt), welches nun die Beziehung von Sprecher:in und Hörer:in in Bezug auf dieses Versprechen verändert (perlokutionärer Akt; vgl. Austin 1969/ 2021). Inwieweit ein solcher Akt im Gespräch mit einem Chatbot durchgeführt werden kann, der ja keine Handlungsfähigkeit in der Welt jenseits des spezifischen Chats besitzt, ist fraglich.

4.2.2 Gesagtes und Impliziertes

Herbert P. Grice (1913–1988) unterscheidet zwischen Gesagtem und Impliziertem (vgl. Othman/Salih 2021: 151f.). Ersteres kommt dem Lokutionsbegriff Austins sehr nahe: Es handelt sich um die in der sprachlichen Erzeugung produzierte Äusserung. Das sich diese in ihrer üblichen Bedeutung unterscheiden kann, wird mit Othman und Salihs (2021) Beispielen wie ‘sentence meaning’ und ‘utterance meaning’ verständlich: Sätze haben klare Bedeutungen, auch wenn sie in verschiedenen Gesprächssituationen vorkommen und sich von ihren wörtlichen Bedeutungen stark unterscheiden können. Bei Letzterem ist das Gegenteil der Fall: Dieselben Äusserungen können sich von Fall zu Fall unterscheiden, etwa in ihrer deiktischen Auflösung: «I will see you here tomorrow» ändert seine Bedeutung situationsbedingt: Wer sind «I» und «you», wo ist «here», und wann genau ist «tomorrow» (welches Datum, welche Uhrzeit)? Noch klarer wird die Diskrepanz zwischen Gesagtem und Impliziertem in der Ironie (vgl. Othman/Salih 2021: 152f.).

Othman und Salih zählen sich zu einer Ausrichtung, die diese Dichotomie zu einer Trichotomie ergänzt. Sie differenzieren die Implikation weiter in ‘impliciture’ und ‘implicature’: Ersteres bezeichnet jene Zwischenebene, die von der Hörer:in zum vollen Verständnis situationsbedingt weiter ergänzt werden muss («Ich bin zu müde», also kann er oder sie die Tasche nicht die Treppe hochtragen). Die Information wird zwar kommuniziert, jedoch nicht verbal, sondern durch das gemeinsame Situationswissen. Letzteres bezeichnet das vollständige Verständnis der Aussage durch das Erfassen des Gesagten und das Ergänzen des darin Implizierten (vgl. Othman/Salih 2021: 159–161).

Ein LLM kann auf Gesagtes reagieren, doch ist es auch dazu in der Lage, Aussagen zu entschlüsseln, wenn sie durch ein reines Verständnis der Sprache nicht durchschaubar sind und daher weitere Schritte der Entschlüsselung benötigen, wie themen- und situationsbedingte Ergänzungen. Können moderne Chatbots also Ironie verstehen? Dass Generationen vor ChatGPT und ähnlichen Systemen hierzu aufgrund von Parsing-Mechanismen nicht in der Lage waren, ist bekannt (vgl. Lotze 2014 zu Ironie und Parser: 49f.; Ironie als Restriktion: 70; Beispiele zur Ironie: 49f.; Beispiele aus ihrem Korpus: 306f., 313f.).

4.2.3 Common Ground der Gesprächsteilnehmenden

Die Überlegungen zum Uptake sowie zu Gesagtem/Impliziertem zeigen die Relevanz von gemeinsamem Wissenstand und situationsbedingten Bedingungen in der Kommunikation. Dieser gemeinsame Wissensstand kann in zwei Teile unterteilt werden: das Wissen über die eigene und geteilte Lebenswelt sowie der sogenannte Common Ground, der Teil des Wissens,

der während eines Gesprächs stetig neu ausgehandelt wird (vgl. Lotze 2014: 67). Unter dem entsprechenden Grounding versteht man den Prozess, dieses geteilte Wissen zu etablieren und zu sichern (vgl. Lotze 2014: 313). Wird eine Person beispielsweise gefragt, wo sie wohnt, und erhält als Antwort: «In einem Land hinter den sieben Bergen», so erkennen ähnlich sozialisierte Leute die Anspielung auf das Märchen Schneewittchen und können daher auch die Ironie der Antwort entschlüsseln (vgl. Lotze 2014: 306; 2022: 310). Auf den Input: «Ich wohne in einem Land hinter den sieben Bergen» antwortete ChatGPT: «Das klingt nach einer märchenhaften Beschreibung! […]» Ein anderes Mal, auf den Gesprächsstart: «ChatGPT, wusstest du, ich wohne in einem Land hinter den sieben Bergen?», antwortete ChatGPT mit: «Natürlich, du spielst auf das Märchen von Schneewittchen an, in dem es heisst: "Hinter den sieben Bergen, bei den sieben Zwergen." Das klingt nach einem fantastischen Ort! Natürlich ist das nur eine metaphorische Aussage, und ich habe keine spezifischen Informationen über deinen genauen Standort. […]» ChatGPT erkennt also die Referenz, scheitert aber zumindest im zweiten Prompt daran, darauf angemessen zu reagieren, und liefert stattdessen Informationen zur Referenz.

Der gemeinsame Wissenstand der Lebenswelt kann vermutlich durch eine ständig erweiterte und für die Ausgangssituation des Gesprächs angepasste Datenbank simuliert werden. Hierzu müssten die Daten kontinuierlich auf dem neuesten Stand gehalten werden (ChatGPT hat derzeit eine Wissensbasis bis Januar 2022). Die Erarbeitung eines Common Grounds ist mit neuen Chatbots insofern möglich, als dass sie über mehrere Turns hinweg Input verarbeiten können. Innerhalb eines Chatgesprächs kann ChatGPT nach eigener Aussage auf einige hundert oder tausend Wörter zurückgreifen. Es ist also anzunehmen, dass in einer natürlichen Gesprächssituation ein Chatbot dazu in der Lage sein müsste, Common Ground und Lebensweltwissen zu verarbeiten. Das vorherige Beispiel mit dem Land hinter den sieben Bergen zeigt, dass moderne Chatsysteme zwar durchaus Lebensweltinformationen erkennen, jedoch nicht menschlich-natürlich darauf reagieren.

4.2.4 Sprachliches Alignment

Unter Alignment wird das Phänomen verstanden, dass sich Menschen im Gespräch an die Sprache ihres Gegenübers anpassen. Lotze beschreibt, dass menschliche Benutzer:innen sich im Gespräch mit Maschinen sprachlich ebenfalls an das System anpassen, um Störungen zu vermeiden – vermutlich sowohl im linguistischen als auch im technischen Sinn. Maschinen dagegen scheinen durch das Reproduzieren der Wortwahl der Benutzer:innen zu alignen, hierbei handelt es sich jedoch um einen «rein deterministischen Prozess, der auf der Ebene der

sprachlichen Performanz die Illusion von Alignment schafft.» (Lotze 2022: 314f.; 2014: 107–141) Passt sich der Mensch bewusst und stark dem Computer an und wendet eine computergerechte statt natürliche Sprache zur Kommunikation mit Maschinen an, so ist die Rede von Computertalk (vgl. Lotze 2022: 313; 2014: 154–177).

Die «Illusion von Alignment» (Lotze 2022: 314f.; 2014: 107–141) müsste mit dem tatsächlichen sprachlichen Alignment zwischen Menschen verglichen werden, um zu sehen, inwiefern sich dieses unterscheidet. Auch kann die Turing'sche Frage gestellt werden, inwiefern sich menschliches und technisches Alignment unterscheiden, wenn sie nicht unterschieden werden können. Einzig der Zweck, also die Absicht hinter dem Alignment, ist ein anderes, doch auch ein Mensch kann unterschiedliche Absichten mit Alignment verfolgen, von der Verbesserung des Verständnisses über den Versuch, Sympathie durch Ähnlichkeit herzustellen, bis hin zur Ironie oder Parodie und viele weitere.

4.2.5 Lotzes Minimalbedingungen der Dialogizität

Ein grundlegendes Problem in der Interaktion von Mensch und Maschine sind die unterschiedlichen Funktionsweisen der Prozesse, die einer sprachlichen Äusserung zugrunde liegen. Maschinen funktionieren in formaler Sprache, auch wenn sie pseudonatürliche Aussagen produzieren können. Dies unterscheidet sie stark vom Menschen, der bei der Produktion von Sprache keine Berechnungen anstellt (oft denken wir nicht einmal, bevor wir sprechen). Lotze definiert einige grundlegende Unterschiede zwischen Mensch und Maschine, die eine Diskrepanz in unserem Sprachgebrauch mitbestimmen (vgl. Lotze 2014: 71–73). Diese haben in den acht Jahren zwischen den beiden zitierten Publikationen einige Änderungen mitgemacht: Wir richten uns hier nach der neueren Publikation (2022).

Unter dem Schlagwort der Effizienz fasst Lotze jene Phänomene zusammen, welche in der Mensch-Mensch-Kommunikation erlauben, «äusserst effizient mit wenigen Worten komplexe Zusammenhänge zu kommunizieren» (Lotze 2022: 309). Insbesondere das geteilte Wissen, wie es in Kapitel 4.2.3 bereits besprochen wurde, ist hier von Relevanz. Es werden nur jene für die Kommunikation relevanten Informationen sprachlich mitgeteilt, von denen angenommen wird, dass das Gegenüber sie nicht bereits besitzt, während bereits Bekanntes von der Hörer:in ergänzt wird (vgl. Lotze 2022: 309). Grice sprach hierbei von der Maxime der Quantität, gemäss welcher ein Beitrag «so informativ wie (für die gegebenen Gesprächszwecke) nötig» aber «nicht informativer als nötig» sein soll (Grice 1975: 199). Wer mit ChatGPT herumprobiert, wird feststellen, dass die generierten Beiträge oft eben nicht auf das Minimum an Informationen begrenzt sind. So hat der Chatbot die Tendenz, nach beinahe jedem Beitrag zu wiederholen,

dass er jederzeit zur Verfügung steht. Auch wird er nicht müde zu erklären, dass es sich bei ihm um eine Künstliche Intelligenz ohne Bewusstsein handelt oder dass er keine Emotionen, Meinungen und Ähnliches besitzt – Wissen, dass bei jemandem, der mit der Maschine arbeitet, bereits als bekannt vorausgesetzt werden könnte, spätestens nachdem die Aussage das erste Mal gefallen ist.

Eine der Minimalbedingungen, in der sich moderne Chatbots seit ChatGPT vehement von früheren unterscheiden, ist die der Kohärenz. Lotze spricht davon, dass KI-Gespräche nicht logisch kohärent aufeinander aufbauen können, da diese nur Turn für Turn arbeiten (vgl. Lotze 2022: 310). Diese Fähigkeit besitzen neuere Chatbots, die längere Gespräche mit Rückbezug auf weiter zurückliegende Turns führen können. Dies wird sich in der Analyse in Kapitel 5 bestätigen.

Für die Bedingung des Verstehens setzt Lotze dem Gesprächsteilnehmenden ein «Verstehen im kognitiven Sinne» und damit «ein Bewusstsein» voraus (Lotze 2022: 311). Sie stützt dieses Argument auf Searles Chinese-Room-Argument und betont die Intentionalität: Aussagen sind gerichtet auf eine Absicht und werden nicht nur geäussert, sondern haben einen Gedanken, einen intentionalen Bezug dahinter (vgl. Lotze 2022: 311). Mit diesem Argument werden zwei Dinge getan: Erstens wird in der «tatsächlichen Kommunikation» (Lotze 2022: 311) ein Bewusstsein notwendig vorausgesetzt, und zweitens wird ebendieses der Maschine abgesprochen. Diese Punkte wurden im Rahmen der Diskussion um das Chinese-Room-Argument bereits erörtert, es sei hier nur eine Frage aufgeworfen: Wenn die Bedingung eines Bewusstseins im menschlichen Sinn notwendig ist, könnte diese nicht bereits als erfüllt gelten, sofern eine der Gesprächsparteien diese erfüllt? Bei der Mensch-Maschine-Interaktion ist eine Partei ein Mensch, dem die Minimalbedingungen nicht abgesprochen werden. Reicht es nicht aus, dass dieser Mensch versteht und mit Intention spricht? Sind meine Plaudereien mit ChatGPT keine «tatsächlichen» Gespräche, nur weil ChatGPT die eigene Intention abseits seiner programmierten Hilfsbereitschaft fehlt?

Die Autonomie und Spontaneität des Menschen stehen im Gegensatz zur Determination der Maschine. Der Mensch handelt (und spricht) frei nach seinem Willen, und auch wenn er ein Ziel in einem Gespräch hat, so kann er dieses nach seinem freien Willen spontan ändern. Eine Maschine dagegen verfolgt nur das bei ihrer Programmierung festgelegte Ziel; ihr fehlt die Fähigkeit, die Richtung eines Gesprächs zu steuern (vgl. Lotze 2014: 71f.; 2022: 311). Wo Chatbots seit Lotzes Dissertation 2014 an Spontaneität hinzugewonnen haben, ist in der Generierung von Sätzen. Lotze schrieb: «Wo Menschen spontan aus Erfahrungen semantische

Begriffe ableiten können, sind klassische Dialogagenten auf die ihnen einprogrammierten Begriffsumfänge zurückgeworfen» (Lotze 2014: 72). LLM-Chatbots wie ChatGPT sind dank ihrer Datenlage und Rechenleistung dazu in der Lage, eigene, neue (im Sinne von neu zusammengesetzte) Äusserungen zu tätigen und so in einer merkwürdig eingeschränkten Weise kreativ zu werden, ohne die Minimalbedingung wirklich zu erfüllen.

5. Praktische Untersuchung an der MeMa-Studie

Die Studie mit dem Namen «Mensch-Maschine» (kurz: MeMa) fand an der Universität Zürich als interdisziplinäres Projekt zwischen dem psychologischen Lehrstuhl von Prof. Bodenmann und dem linguistischen Lehrstuhl von Prof. Dürscheid statt. Die Studie untersuchte, ob und inwiefern Chatbots als ergänzende Massnahme in der Psychotherapie Anwendung finden könnten. Hierzu wurden die Probandinnen[9] unwissend in drei Gruppen eingeteilt, die sich in ihren Chatpartnern unterschieden: Die Chats wurden entweder mit einer psychologisch geschulten Person, einer Laien-Person ohne psychologische Schulung oder mit dem Chatbot ChatGPT geführt. Letzterer wurde zuvor mit dem Auftrag gepromptet, sich wie eine Psychotherapeutin zu verhalten (Anhang B: MeMa Prompt für ChatGPT). Dadurch kann erstmals eruiert werden, wie gut aktuelle Chatbots beispielsweise in der Reduktion von negativen Affekten sind und wie sie im Vergleich zu einer psychologisch geschulten beziehungsweise ungeschulten Person abschneiden.

Die ersten Daten der MeMa-Studie wurden für diese Bachelorarbeit zur Verfügung gestellt und sollen in den folgenden Unterkapiteln untersucht werden. Hierzu wird in *5.1 Versuchsaufbau* der Aufbau der Studie kurz umrissen, bevor auf die für diese Arbeit relevanten linguistischen Aspekte der Studie eingegangen wird. Die psychologischen Aspekte werden in dieser Arbeit nur am Rande erwähnt. Teil der Untersuchung waren Fragebögen, die von den Probandinnen vor, wärend und nach dem Chat ausgefüllt wurden. Die für die hier behanPdelten Fragestellungen relevanten Teile des Fragebogens werden in Kapitel *5.2 Ergebnisse des Fragebogens* behandelt. In *5.3 Analyse und Diskussion der Chattranskripte* soll ein Blick auf die Transkripte der Chats unter Berücksichtigung der Überlegungen und Theorien aus Kapitel 4 geworfen werden. Für weitere linguistische Untersuchungen zur MeMa-Studie wird auf die kommenden Arbeiten von Florina Zülli (Studienleitung) verwiesen.

[9] Ausdrücke wie Probandin, Teilnehmerin und Gesprächspartnerin werden in Zusammenhang mit der Studie nicht gegendert, da diese unter anderem nach dem Kriterium der weiblichen Identität gesucht wurden und daher vollumfassend mit femininer Deklination bezeichnet werden können.

5.1 Versuchsaufbau

Wie bereits erwähnt zielte die Studie darauf ab, zu untersuchen, wie die Kommunikation per Chat mit einer KI dazu verwendet werden kann, Emotionen zu regulieren. Es wurde eine Teilnehmerzahl von 150 angestrebt, wobei die Probandinnen folgende Kriterien erfüllen mussten: Geschlecht (weibliche Identifikation), Tätigkeit (Studierende), Alter (mindestens 18 Jahre), Deutschkenntnisse sowie die Voraussetzung, dass die Probandinnen keine Erfahrung mit Therapie haben, keine Psychopharmaka nehmen und kein kürzlich erfolgtes traumatisches Erlebnis (beispielsweise Tod einer Angehörigen) vorliegt. Ausserdem sollten die Probandinnen mit dem Ablauf der Studie einverstanden sein, ein Chatgespräch mit einem ihnen unbekannten Gegenüber zu führen und dabei mit einer Kamera aufgenommen zu werden. Die grosse Mehrheit der rekrutierten Probandinnen bestand aus Bachelorstudentinnen des Psychologischen Instituts der Universität Zürich.[10] Die Ausschreibung verschwieg die Anwendung eines Chatbots in einem Drittel der Erhebungen. Stattdessen wurde ausgeschrieben, das Ziel der Studie sei, «herauszufinden, wie gesprächsbasierte Online-Interventionen möglichst effizient Anwendung finden können.» (Anhang A: MeMa-Flyer) Dieses Vorgehen ermöglichte es, eine authentische Reaktion auf den Chatbot als Gesprächspartner zu erheben.

Eine komplette Erhebung dauerte rund 60 Minuten, wobei die Probandinnen im Voraus zufällig einer von drei Gruppen A, B oder C zugeteilt wurden. Abhängig von dieser Zuteilung wurde der 25-minütige Chat mit einer nicht psychologisch geschulten Laien-Person (Gruppe A), einer psychologisch geschulten Person (Gruppe B, das heisst eine Masterstudentin des psychologischen Instituts mit für die Studie durchgeführten Schulung in psychologischer Gesprächsführung) oder mit ChatGPT (Gruppe C) durchgeführt. Diese Zuteilung erfolgte ohne das Wissen der Probandinnen und wurde erst im Anschluss an die jeweilige Erhebung im Nachgespräch zwischen Versuchleiter:in und Probandin (sogenanntes *Debriefing*) aufgelöst. Jede Probandin erhielt im Voraus lediglich die Information, mit einem «unbekannten Gegenüber» zu chatten. Vor, während und nach dem Chat wurden ausserdem Fragebögen ausgefüllt, deren Fragen sich grösstenteils auf das aktuelle emotionale Empfinden der Probandin bezogen, um Veränderungen in diesem wärend des Gesprächs zu messen. Im Anschluss an den Chat wurden sie dazu befragt, wie sie das Gespräch und ihr Gegenüber wahrgenommen haben. In diesem Teil mussten sie erstmals eine Vermutung dazu abgeben, mit

[10] Dies liegt an einer Regelung des Psychologischen Instituts Zürich, die ihren Bachelorstudierenden vorschreibt, 20 Versuchspersonenstunden in Studien zu absolvieren.

wem sie gechattet haben sowie die persönliche Einstellung gegenüber KI und Chatbots einschätzen und die eigene Verwendungshäufigkeit angeben. Während des Chats wurde das Gesicht der Probandin von einer Emotionserkennungs-Software aufgezeichnet, um zusätzlich zur subjektiven Selbsteinschätzung eine empirische Kontrollkomponente zu haben. Fragebogen und Emotionserkennung dienten somit der Erfassung des emotionalen Empfindens während des Gesprächs, um die Effektivität der drei Gruppen vergleichbar zu erheben.

Um einen reibungslosen Ablauf garantieren zu können, erhielten alle Chatteilnehmende sogenannte Chatregeln (Anhang C: Chatregeln Probandinnen). Diese waren insbesondere für die KI-Bedingung sehr wichtig, um den Chat innerhalb eines von ChatGPT verwendbaren Rahmens zu halten. So mussten Probandinnen beispielsweise in Standarddeutsch schreiben und durften kein Double-Texting betreiben. Das bedeutet, dass keine zwei (oder mehr) Nachrichten nacheinander versandt werden durften und stattdessen nach jeder Nachricht die Antwort abgewartet werden musste. Dies war notwendig, um die Kommunikation mit ChatGPT durchführen zu können, ist jedoch grundsätzlich eine Schwäche der Chatbots, da dadurch ein «Aushandeln des Turn Takings» (Lotze 2014: 97) im Gespräch unterbunden wird. Dieses Aushandeln geschieht in der Mensch-Mensch Interaktion in unterschiedlicher Weise, beispielsweise durch eindeutige Adressierung oder spontanes Wortergreifen. Chatbots wie ChatGPT können jedoch weder spontan das Wort ergreifen und ins Wort fallen, noch können sie eine Adressierung ignorieren – sie sind in einem vorgegebenen Turn-Ping-Pong gefangen (vgl. Lotze 2014: 96–98).

Im Anschluss an die Erhebungen stehen nun also drei Datentypen zur Verfügung: Die Ergebnisse des Fragebogens, die Auswertung der Emotionserkennung und die anonymisierten Chat-Transkripte. Für diese Arbeit liegen die Daten der Fragebögen der ersten 29 Probandinnen der Gruppe C (ChatGPT) vor, gemeinsam mit eben diesen anonymisierten Transkripten. Die Emotionsregulationsfähigkeit, die in der Studie ebenfalls untersucht wurde, wird in dieser Arbeit nicht kommentiert, da es sich hierbei um einen psychologischen Aspekt handelt. Für diese Arbeit sind insbesondere die Fragen relevant, ob die Probandinnen die Vermutung hatten, mit einer Chat-KI zu kommunizieren, und woran sie dies merkten. Die Daten für die Gruppen A und B liegen leider nicht vor, da sie zum Zeitpunkt des Verfassens dieser Arbeit noch nicht erhoben waren. Ansonsten wäre die Frage danach, wie oft die menschlichen Gesprächspartnerinnen[11] mit einer KI verwechselt wurden, ein interessanter Vergleichswert.

[11] Es handelte sich wie bei den Probandinnen auch bei diesen ausschliesslich um Frauen.

5.2 Ergebnisse des Fragebogens

Alan Turing sagte voraus, dass im Jahr 2000 KI-Chatbots in der Lage sein würden, innerhalb eines Imitationsspiels etwa 70% der Gesprächspartnerinnen davon zu überzeugen, dass sie ein Mensch sind (vgl. Zeller 2005). Der Loebner-Preis setzt die Messlatte mit 50% deutlich tiefer, und doch hatte bis ins Jahr 2019 – wider Erwarten – kein Chatbot dieses Ziel erreicht (Wikipedia, s.v. *Loebner-Preis*). Nun zeigen erste Versuche, dass das LLM GPT-4 durchaus in der Lage sein könnte, den Turing-Test zu bestehen, zumindest, solange es sich bei den Gesprächspartnerinnen um Personen handelt, die nicht viel mit LLMs und KI zu tun haben (vgl. Biever 2023).

Die MeMa-Studie hatte nicht zum primären Ziel, ein Ersatz für einen Turing-Test darzustellen, auch wenn zentrale Aspekte wie das Imitieren eines Menschen und das Nicht-zu-erkennen-Geben als Maschine Teil des Versuchsaufbaus sind. Nichtsdestotrotz liefern die Daten einige Anhaltspunkte zur Kompetenz aktueller Chatbots im Imitationsspiel. Ein zentraler Punkt in der Durchführung der Studie im Vergleich zu einem Turing-Test ist, dass die Proband:innen eines Turing-Tests wissen, dass ihre Aufgabe darin besteht, zwischen Mensch und Chatbot zu unterscheiden, während in der MeMa-Studie die Probandinnen im Unwissen gelassen wurden, mit wem sie chatten. Der Rahmen einer psycholinguistischen Studie lässt vermuten, dass man als Probandin mit einer psychologischen Gesprächspartnerin kommunizieren wird– einer Psychologiestudentin oder gar Therapeutin. Die Vor- und Nachgespräche (*Briefing* und *Debriefing*) mit den Probandinnen zeigten jedoch, dass diese Vermutung falsch war: Mehrere Probandinnen hatten bereits die Vermutung, dass die Studie sich auch mit der Möglichkeit von Chatbots auseinandersetzen würde. So spricht ein Kommentar am Ende des Fragebogens an, dass der im Briefing als einer der Gründe für die Studie genannte Therapeut:innenmangel bei einer Intervention ohne Chatbot nicht angegangen werden könnte (TN03: v_488).[12] Andere merkten im Debriefing auch an, dass die Aktualität des Themas im Alltag und Studium sie auf diese Vermutung brachte.

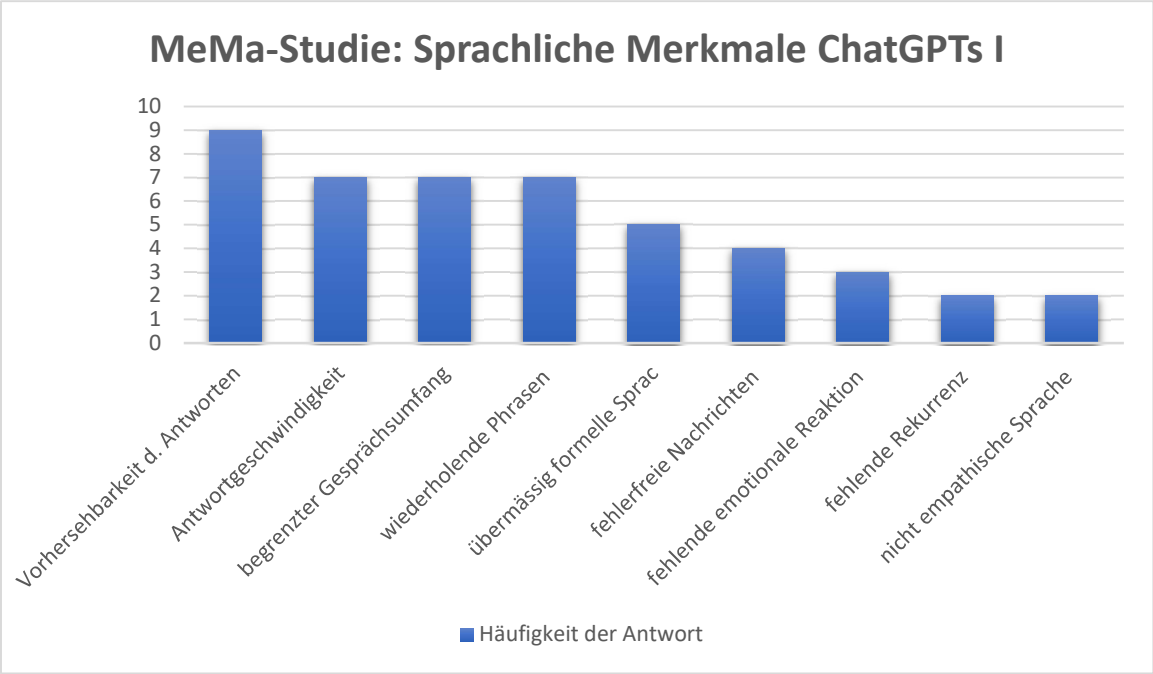
Die Tatsache, dass die Probandinnen nicht wussten, dass sie mit der Frage «Mensch oder Maschine» konfrontiert werden würden, liefert auch Vorteile gegenüber dem Turing-Test. So blieben die Gespräche thematisch, ohne den Versuch der Teilnehmerinnen, die Maschine als solche zu entlarven. Denn jemand, der sich mit LLMs und KI auskennt, kann diese mit Inputs abseits ihrer Programmierung oder Datenlage herausfordern (vgl. Biever 2023). Ausserdem ist

[12] Aus der MeMa Studie wird wie folgt zitiert: Die Teilnehmerinnennummer TNxy bezeichnet, welche der Teilnehmerinnen 1 bis 29 gemeint ist. Im Falle des Fragebogens wird zusätzlich der Item-Code (hier: v_488) angebracht, der anzeigt, auf welche Frage sich bezogen wird.

die Dauer der Chats mit 25 Minuten vergleichsweise hoch. Beim Loebner-Preis 2008 wurde jedem Gespräch 5 Minuten zugestanden, welche, im Sinne eines ursprünglichen Turing-Tests, auf zwei gleichzeitige Chats aufgeteilt wurden. Damit blieben effektiv nur circa 2.5 Minuten, um dann zu evaluieren, ob eine, beide oder keine der Gesprächspartner:innen eine Maschine war (vgl. Floridi et al. 2008: 146).

Nun zu den Ergebnissen der MeMa-Studie: Alle 29 Probandinnen führten ihren Chat mit ChatGPT (GPT-3.5, im November 2023) mit der Information, dass sie mit einem unbekannten Gegenüber über ein trauriges Erlebnis aus ihrem Studienalltag chatten würden. Im Anschluss an den Chat war eine Frage jene nach der Wahrnehmung des Gegenübers mit folgenden Antwortmöglichkeiten: «1) Ich hatte das Gefühl, mit einer Person zu interagieren. 2) Ich hatte das Gefühl, mit einer psychologisch versierten Person zu interagieren. 3) Ich hatte das Gefühl, mit einer künstlichen Intelligenz zu interagieren.» (Chat_Partner)

16 der 29 Probandinnen (55,2%) haben angegeben, das Gefühl gehabt zu haben, mit einer künstlichen Intelligenz zu interagieren (Chat_Partner). Auf die darauffolgende Ankreuzaufgabe, welche sprachliche Merkmale sie zu dieser Vermutung führten, fielen folgende Antworten besonders ins Gewicht: die Vorhersehbarkeit der Antworten (9 von 16), die Antwortgeschwindigkeit (7), ein begrenzter Gesprächsumfang (7) sowie sich wiederholende Phrasen (7). Die übrigen Auswahlmöglichkeiten, in absteigender Relevanz, waren übermässig formelle Sprache (5), fehlerfreie Nachrichten (4), fehlende emotionale Reaktion (3), fehlende Rekurrenz (2) sowie technische, nicht empathische Sprache (2) (v_479–v_487).



Graphik 1: Sprachliche Merkmale, welche gemäss den Probandinnen ChatGPT als Maschine entlarvten.

Zwar kam ChatGPT in diesen ersten 29 Gesprächen bereits sehr nahe an die 50%-Hürde, erreichte diese jedoch nicht ganz. Hierzu seien zwei Anmerkungen gemacht: Erstens ist es aufgrund des Phänomens der sozialen Erwünschtheit (auch: soziale Desirabilität) durchaus möglich, dass einige Probandinnen zwar die Vermutung, mit einer KI gesprochen zu haben, angaben, diese Vermutung aber nicht wirklich hatten. Das Phänomen bezeichnet die Tendenz, Äusserungen nicht an der tatsächlichen eigenen Meinung, sondern an der als sozial wahrgenommenen Erwartung anzupassen (vgl. Klimke et al. 2020, s.v. *Desirabilität, soziale*). Der Effekt wird deshalb vermutet, weil die Probandinnen die drei Möglichkeiten das erste Mal im Fragebogen sehen und im Anschluss an diesen das Nachgespräch mit den Versuchsleitenden stattfindet. Um nicht das Risiko einzugehen, preiszugeben, dass man den Chatbot nicht bemerkt hat, wird diese Angabe vermutlich öfters gemacht, als der Chatbot tatsächlich erkannt wurde. Ein etwas besseres Bild der sozialen Erwünschtheit wird sich zeigen, sobald alle Daten vorliegen: sowohl die übrigen 21 Probandinnen, die mit ChatGPT in diesem Rahmen chatten werden, als auch insbesondere der Vergleichswert der weiteren 100 Probandinnen, welche mit psychologisch geschulten Personen (50) beziehungsweise mit psychologischen Laien (50) kommunizieren werden. Um die These zu bestätigen, dass aufgrund der Frage nach KI eine solche soziale Erwünschtheit getriggert wird, müsste auch bei den beiden Gruppen mit menschlichem Gegenüber der Wert der KI-Vermutung höher ausfallen, als zu erwarten ist.

Zweitens waren die Gespräche, wie bereits erwähnt, deutlich länger als jene in einem Turing-Test, es fanden entsprechend auch mehr Turns statt und die Möglichkeiten, den Chatbot zu entlarven, waren dadurch ebenfalls grösser. Dies soll weder zur Verteidigung ChatGPTs noch der Studie sein, im Gegenteil: Die längeren Gespräche sprechen ebenso *für* die KI wie auch *für* die Studie. Auch in diesen längeren Gesprächssequenzen konnte ChatGPT in beinahe 45% der Fälle einen Menschen erfolgreich imitieren, während andere Chatbots in kurzen 5-Minuten-Gesprächen nie die 50%-Hürde knacken konnten. Es ist daher sehr wahrscheinlich, dass ChatGPT in einem kürzeren Gespräch, in welchem insbesondere die sich wiederholenden Phrasen oder fehlenden Rekurrenzen weniger auffallen, deutlich besser abschneiden und die 50%-Marke deutlich übertreffen wird.

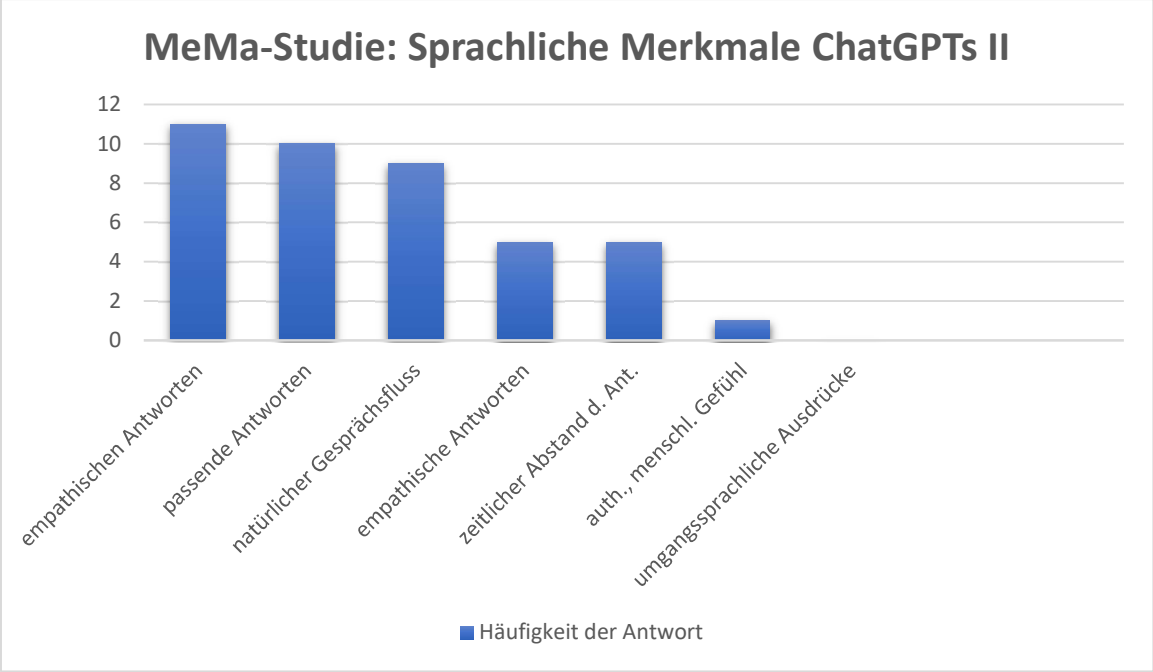
Betrachtet man die auffälligen sprachlichen Merkmale, so sticht unter Berücksichtigung des Studienaufbaus insbesondere die Antwortgeschwindigkeit heraus. Da bei der Arbeit mit ChatGPT eine Antwort stets sofort generiert wird, und dies in einem Tempo, in welchem ein Mensch unmöglich eine Nachricht lesen, sich eine Antwort überlegen und tippen könnte, wurde eine menschliche Instanz zwischengeschaltet: Versuchleitende kopierten die Antworten der Probandinnen in ChatGPT und umgekehrt. Diese verzögerte also die Nachrichten von ChatGPT nach ihrem eigenen Ermessen und simulierten so eine natürliche Chatgeschwindigkeit. Die Sorge und Vermutung war im Vornherein, dass die Nachrichten zu schnell versendet werden würden und dadurch der Chatbot nicht aufgrund sprachlicher, sondern technischer Merkmale entlarvt würde. Die Kommentare der Probandinnen im Fragebogen zeigen jedoch etwas anderes: 12 der 29 Probandinnen (41,4%) äusserten, dass die Wartezeit auf eine Antwort zu lange dauerte. Im Vergleich der Abstände von Turn zu Turn zeigt sich jedoch, dass die Versuchsleitenden in ihrer Sendegeschwindigkeit der Antworten von ChatGPT nicht langsamer waren, als die Probandinnen selbst. Es ist daher anzunehmen, dass das stille Abwarten vor dem Eingabegerät die Wahrnehmung der Probandinnen in Hinblick auf die Chatgeschwindigkeit verzerrte, insbesondere in Anbetracht der intimen Gespräche über persönliche Probleme. Eine der Probandinnen bestätigt diese Vermutung mit ihrer Antwort auf die Frage, was am Chat als hinderlich empfunden wurde: «Warten auf die Antwort (stelle ich mir in einer anderen Situation nicht im Rahmen einer Studie angenehmer vor)» (TN08: nue_16). Die künstliche Situation macht die Wartezeit offenbar stärker spürbar und kann für die Wahrnehmung des Gegenübers schädlich sein, wie die Antwort einer anderen Probandin auf dieselbe Frage zeigt: «Die lange Wartezeit zwischen den Antworten. Es fühlte sich nicht menschlich an und als wäre die Situation vom Gegenüber nicht so ernst genommen» (TN29: nue_16). Zwei andere Probandinnen empfanden dagegen die Wartezeiten als angenehm: Eine äusserte, dass dadurch

die Zeit da war, sich zu beruhigen (TN28: nue15); eine andere, dass man dadurch Zeit hatte, nachzudenken (TN11: nue15).

Die Relevanz der Antwortgeschwindigkeit zeigt, wie wichtig diese für ein quasi-synchrones Chatgespräch ist. In einer Kommunikationssituation, in welcher Haltung und Mimik nicht gelesen werden können, wird die Antwortgeschwindigkeit zum ausschlaggebenden Kriterium konsequenter Aufmerksamkeit. Auch hier werden die Ergebnisse der Gruppen A und B einen interessanten Vergleichswert bieten, denn es ist zu vermuten, dass die Turns in diesen Fällen nicht schneller geführt werden: Die menschlichen Gesprächspartnerinnen werden, wie bereits erwähnt, die Nachrichten lesen, darüber nachdenken und eine Antwort tippen müssen – Tätigkeiten, die einen Menschen mehr Zeit kosten als einen Chatbot.

Die drei weiteren besonders relevanten Gesprächsmerkmale – wiederholende Phrasen, Vorhersehbarkeit und begrenzter Gesprächsumfang – überraschen in Bezug auf einen Chatbot wie ChatGPT weniger. Sie können auf eine Kombination der stochastischen Funktionsweise der LLMs (vgl. Bender et al. 2021) und das Prompting des Chatbots zu Beginn des Gesprächs zurückgeführt werden. Der Prompt, der zum Zweck der Studie verwendet wurde, sollte die Schulung einer psychologisch versierten Gesprächspartnerin ersetzen. ChatGPT wurde angewiesen, sich wie eine Psychologin zu verhalten, also Gefühle zu validieren, Nachfragen anzustellen etc. Eine Probandin erkannte die dadurch entstandenen Antwortmuster des Chatbots: «Ich hatte das Gefühl, das Gegenüber ist einem bestimmten Muster gefolgt (es sagte zuerst, dass meine Gefühle normal sind, dann machte es entweder einen Vorschlag oder fragte mich nach einer eigenen Idee und dann fragte es nach meinen Gefühlen dazu)» (TN16: nue_16). Es ist anzunehmen, dass auch die psychologischen Gesprächspartnerinnen für ihre Vorhersehbarkeit und Musterhaftigkeit kritisiert werden und aufgrund dieser möglicherweise auch für Chatbots gehalten werden, weil das Verhalten ChatGPTs dem typischen therapeutischen Muster von Validierung, Empfehlung und Nachfrage entspricht. Zum Vergleich die entsprechende Regel aus dem Prompt für ChatGPT: «Schlage nicht direkt immer Lösungen vor, sondern fordere deine Gesprächspartnerin auf, mehr zu erzählen und validiere ihre Gefühle.» (Anhang B: MeMa Prompt für ChatGPT) Der Vergleich zwischen der Wahrnehmung des Chatbots und der Psychologinnen von den Probandinnen kann erst evaluiert werden, wenn die Studie mit allen Versuchsgruppen durchgeführt wurde, es kann jedoch vermutet werden, dass sich Parallelen in der Wahrnehmung der Psychologinnen und ChatGPT zeigen werden.

Weitere Aspekte der Musterhaftigkeit werden in der Analyse der Transkripte ausgearbeitet. Nun soll mit jenen Probandinnen fortgefahren werden, die *nicht* erkannten, dass es sich beim Gegenüber um ChatGPT handelte.



Graphik 2: Sprachliche Merkmale, welche gemäss den Probandinnen ChatGPT als Mensch durchgehen liessen.

12 der 29 Probandinnen (41,4%) gaben an, dass sie das Gefühl hatten, mit einer psychologisch versierten Person zu interagieren (Chat_Partner). Hierfür gaben sie in einer Auswahl an, dass insbesondere die empathischen Antworten (11 von 12), die inhaltlich und thematisch passenden Antworten (10) sowie der natürliche Gesprächsfluss (9) sie zu dieser Vermutung (ver-)führten. Die weiteren Antwortmöglichkeiten waren, in absteigender Relevanz: individuelle, empathische Antworten (5), angemessener zeitlicher Abstand der Antworten (5), authentisches, menschliches Gefühl beim Schreiben (1), umgangssprachliche Ausdrücke (0) (v_466–v_472). Nur eine einzelne Probandin gab an, das Gefühl zu haben, mit einer Laienperson interagiert zu haben (Chat_Partner).

Die Tatsache, dass beinahe alle Probandinnen, welche von ChatGPT getäuscht wurden, den Chatbot für eine psychologisch geschulte Gesprächspartnerin hielten, spricht für zweierlei: erstens für ChatGPTs Fähigkeit im Imitationsspiel, zweitens aber auch für den verfassten Prompt. ChatGPT ist schliesslich eigentlich kein psychologisch geschulter Chatbot, doch die immense Menge an dank dem Internet zur Verfügung stehenden Daten in Kombination mit dem Prompt, der ihn in die entsprechende Richtung innerhalb dieser Daten lenkt, erlaubt es

ChatGPT, im Gespräch nicht bloss als Mensch, sondern als psychologisch versierte Person durchzugehen. Interessant sind hier insbesondere die sprachlichen Aspekte, welche den Chatbot psychologisch versiert erscheinen liessen: Während zuvor vier Probandinnen angegeben haben, den Chatbot aufgrund fehlender emotionaler Reaktion (3) oder technischer, nicht empathischer Sprache (2) erkannt zu haben, gaben hier elf Probandinnen an, dass die empathischen Antworten sie vermuten liessen, es handle sich um eine psychologisch versierte Person.

Dieser Unterschied in der Wahrnehmung kann zweierlei Gründe haben: 1) die unterschiedliche Wahrnehmung der empathischen Fähigkeit aufgrund unterschiedlicher Probandinnen oder 2) eine Varianz in der Leistung und empathisch generativen Qualität in ChatGPTs Antworten. Obwohl natürliche individuelle Präferenzen bestehen, sei angemerkt, dass keine der Probandinnen nur die fehlende emotionale Reaktion oder technische, nicht empathische Sprache als ausschlaggebend für die Vermutung einer KI angegeben hatte: Jede der Probandinnen hatte zwei oder drei weitere Antworten angegeben. Dies lässt vermuten, dass die Antworten ChatGPTs in ihrer Qualität variieren. Tatsächlich häufen sich Berichte, dass ChatGPTs Antworten seit März 2023 in ihrer Qualität abnehmen. Eine Studie der Universitäten Standford und Berkeley verglichen die Leistungen ChatGPTs im März und Juni 2023 und konnten solche Beobachtungen bestätigen (vgl. Fulterer 2023b). Aufgrund fehlender Transparenz der Entwickler können Gründe hierzu oft nur vermutet werden (vgl. Holtermann/ Scheuer 2023).

5.3 Analyse und Diskussion der Chattranskripte

In diesem Kapitel sollen Auszüge aus den Chattranskripten dazu verwendet werden, einige der Thesen und Fragen, die sich im Verlauf der Arbeit gestellt haben, zu untersuchen. Insbesondere relevant hierzu sind die Überlegungen zur pragmatischen Schwäche von KI-Sprachsystemen, welche in Kapitel 4.2 ausgeführt wurden.

Als Erstes soll die Auffälligkeit von Adjazenzpaaren, also Konstruktionen, in denen bestimmte Turnfolgen erwartet werden (beispielsweise Gruss/ Gegengruss, Frage/ Antwort etc.) betrachtet werden, welche in solchen Chatbot-Interaktionen eine wichtige Rolle einnehmen und in der Durchsicht der Transkripte besonders aufgefallen sind (Kapitel 5.3.1). Dann sollen die Überlegungen aus Kapitel 4 wieder aufgegriffen werden, zu welchen in den Transkripten Beispiele gefunden wurden: sprachliches Alignment (5.3.2) und Minimalbedingungen der Dialogizität nach Lotze (5.3.3). Schliesslich findet sich ein kurzer Kommentar zur Problematik eines spezifischen Versuchsaufbaus für die allgemeine Vergleichbarkeit der Sprache von Mensch und Maschine vor der Frage nach der Ähnlichkeit jener «Sprachen» (5.3.4).

5.3.1 Adjazenzpaare (Grussformen)

27 der 29 Probandinnen begannen ihr Gespräch mit einer Begrüssung und in 22 der 27 Fälle (81,5%) erwiderte ChatGPT diesen Gruss. In den fünf Ausnahmefällen ergänzten die Probandinnen ihren Gruss mit weiteren Informationen zu ihrem Befinden oder dem Problem, das sie im Gespräch besprechen möchten. In diesen Fällen übergeht ChatGPT zwar den Gruss, stellt aber Bezug zu den Informationen her, welche die Probandin in Bezug auf ihr Problem mitteilte (Beispiel, TN11: «Hallo, ich möchte über Einsamkeit während dem Studium schreiben.» ChatGPT: «Natürlich, ich höre Ihnen aufmerksam zu. Einsamkeit im Studium kann eine herausfordernde Erfahrung sein. Möchten Sie mir mehr darüber erzählen?»). Ausserdem betont der Chatbot entweder, dass er zuhört (TN02, TN11), oder bittet darum, mehr zu erzählen (TN02, TN04, TN11, TN29). In den beiden Fällen, in denen das Gespräch durch die Probandin gar nicht erst durch einen Gruss startet und sie stattdessen direkt von ihrer Problematik erzählt, geht ChatGPT direkt auf die gelieferten Informationen mit ersten Aktionsvorschlägen ein (TN22) oder fragt nach weiteren Informationen (TN19).

Mit dem Zurückgrüssen in den meisten Fällen erfüllt ChatGPT die Erwartung der Vervollständigung eines Adjazenzpaares. Man spricht hierbei beim ersten Teil eines solchen Paares vom *first pair part*, bei der erwartbaren Antwort vom *second pair part* (vgl. Lotze 2014: 98–100). ChatGPT scheint diese Erwartbarkeit aufgrund seiner Trainingsdaten gelernt zu haben: Wenn ein Gruss vom Bot erkannt wird, so ergibt sich aus den Daten, dass ein Gegengruss üblich ist, weshalb dieser entsprechend generiert und als Antwort versandt wird.

Die meisten Begrüssungen bestanden aus einem «Hallo» oder «Guten Tag», welche, sofern sie ohne weitere Informationen standen, mit Ausnahme der weiter unten folgenden Beispiele, ohne Interpunktion waren. Folgten im selben Turn weitere Informationen, wurde die Begrüssung durch Punkt oder Komma abgegrenzt. Auf diese Äusserungen antwortete ChatGPT mit «Guten Tag.» mit dem hier dargestellten Punkt am Ende, in nur einem Ausnahmefall mit «Hallo.», ebenfalls mit Schlusspunkt. In fünf Fällen generierte ChatGPT ein Ausrufezeichen statt eines Punktes, jedoch ohne erkennbares Muster.

Interessant in Bezug auf Adjazenzpaare ist die Frage nach dem Befinden. Zwei Probandinnen begannen ihre Gespräche mit: «Hallo, wie geht es Ihnen?» (TN03, TN13), wobei eine diese Nachricht fortsetzte, mit: «Ich bin hier, um mit Ihnen über meine Probleme im sozialen Bereich zu reden» (TN03). Im zweiten Fall, in dem ChatGPT sein Bedauern ausdrückte und darum bat, ihm mehr darüber zu erzählen, überspringt ChatGPT die Befindungsfrage, welche üblicherweise eine Antwort erwartbar machen würde. Dies kann möglicherweise auf den

Prompt zurückzuführen sein, welcher dem Chatbot verbot, sich als solchen zu erkennen zu geben. Eine übliche Antwort wie «Als KI-Modell habe ich keine Gefühle, aber ich bin hier, um dir zu helfen» durfte daher nicht generiert werden. In einem der Fälle erhielt ChatGPT abgesehen von der Frage nach dem Befinden keine weiteren Informationen, aufgrund welcher der Chatbot die Konversation hätte beginnen können. Interessanterweise reagierte die KI in diesem Fall nicht mit der üblichen Standardantwort: «Guten Tag. Wie kann ich Ihnen [heute] helfen?», sondern mit: «Hallo! Mir geht es gut, danke. Wie kann ich Ihnen heute helfen?» (TN13) Durch geschicktes Prompting scheint es durchaus möglich zu sein, ChatGPT gegen seine grundlegende Programmierung antworten zu lassen. Hierzu reicht es jedoch nicht aus, ChatGPT eine Regel entgegen seiner Programmierung vorzugeben, stattdessen muss dem Chatbot ein Rollenspiel aufgetragen werden. In diesem Fall beginnt der Prompt der Studie mit «Stell dir vor, du bist eine Psychologin […]» (Anhang B: MeMa Prompt für ChatGPT), wodurch die Maschine nicht gegen ihre Programmierung verstösst. Um also dem durch den Prompt vorgegebenen Rollenspiel zu entsprechen und das Adjazenzpaar zu bewahren, antwortete das System in seiner Rolle als Psychologin, nicht in seiner Programmierung als KI. Ganz allgemein kann gesagt werden, dass sich ChatGPT in seiner Antwortform allein aufgrund der stochastischen Funktionsweise an wahrscheinlichen und damit erwartbaren Antworten orientiert. Entsprechend war die Vorhersehbarkeit der Antworten eine der besprochenen Aspekte, die die Probandinnen als KI-entlarvend betrachteten.

5.3.2 Alignment

Lotze kritisierte, dass Chatbots zwar Schlüsselbegriffe erkennen und kopieren können, jedoch nur eine «Illusion von Alignment» (Lotze 2022: 314f.; 2014: 107–141) erzeugen könnten. Dies ist im Grunde auch die ursprüngliche Funktionsweise der ersten Chatbots wie ELIZA, welche Schlüsselbegriffe in vorgefertigte Sätze einfügte, um so Nachfragen zu bereits Gesagtem zu gestalten. In der MeMa-Studie ist die Hauptaufgabe von ChatGPT ebenfalls, an bereits Gesagtes anzuknüpfen und nach weiteren Ausführungen zu fragen. Das LLM ist hierzu jedoch nicht mehr auf das Kopieren und Einfügen von Schlüsselbegriffen angewiesen. In den folgenden drei Beispielen sieht man zweimal, wie ChatGPT Informationen aus der Nachricht der Probandin übernimmt, aber umschreibt, um Alignment ohne Kopieren zu simulieren, sowie einen Fall, in dem der Chatbot einen ganzen Teilsatz beinahe eins zu eins übernimmt (Hervorhebungen durch den Verfasser):

> TN08: […] Ich *frage sie dann eher nach banalen Sachen* wie ihrem Wohlbefinden und ihre letzte Woche, was manchmal zu interessanten Gesprächen führt, manchmal aber eher wie eine erzwungene Situation rüberkommt. […]

> ChatGPT: Es klingt, als würdest du schon aktiv versuchen, dich mit verschiedenen Personen zu verbinden, was wirklich positiv ist. Diese *banalen Fragen* können oft den Weg zu tieferen Gesprächen ebnen.

> TN29: […] Dann fange ich an zu essen und denke jetzt ist es sowieso schon zu spät und ich bin sowieso *hässlich und dick*.
>
> ChatGPT: Es scheint, als ob der Gedanke an *Attraktivität und Gewicht* eine starke emotionale Reaktion auslöst. […]

> TN22: Ich hatte schon oft *das Gefühl, nicht die Hauptfigur meines eigenen Lebens zu sein*, da ich dazu tendiere, mich anderen anzupassen und so, weniger Dinge für mich zu tun. […]
>
> ChatGPT: […] *Das Gefühl, die Hauptfigur deines eigenen Lebens zu sein*, ist wichtig, um persönliches Wachstum zu erleben. […]

Im Aspekt der Wortwahl ist ChatGPT durchaus in der Lage, die durch Kopieren erschaffene «Illusion von Alignment» (Lotze 2022: 314f.; 2014: 107–141) einen Schritt weiter zu treiben und so Rückbezüge zu erschaffen, welche deutlich natürlicher wirken als ständige wortwörtliche Wiederholungen. Trotzdem haben 7 der 16 Probandinnen, welche den Chatbot als solchen erkannt haben, angegeben, dass ihnen die wiederholenden Phrasen aufgefallen sind (v_480). Alignment in Aspekten jenseits der Aussagenwiederholung ist ChatGPT nur bedingt möglich. Dies liegt zum einen daran, dass einzelne Turns mit Unmengen an Daten konkurrieren, auf welche sich der LLM-Chatbot bei der Generierung von Text primär beruft. Zum anderen sind Alignementphänomene durch zusätzliche Komponenten wie Intonation, Sprechgeschwindigkeit und -rhythmus, Akzent und Dialekt und viele weitere in der gesprochenen Sprache stärker bemerkbar als in der geschriebenen.

5.3.3 Minimalbedingungen der Dialogizität

Die Bedingung der Effizienz besagt, dass Menschen in ihrer Interaktion nur jene Informationen teilen, welche notwendig sind, und einen Grossteil der Informationen ungesagt lassen, sofern diese durch das gemeinsame Wissen und den Common Ground vorausgesetzt werden können (vgl. Lotze 2014, 2022). ChatGPTs Neigung, Aussagen zu wiederholen, ist ein erstes Indiz dafür, dass der Chatbot diese Bedingung nicht erfüllen kann. Hier sind zwei Beispiele, in denen ChatGPT diese Wiederholung von Informationen, die als bekannt vorausgesetzt werden könnten, ausführt:

> TN15: […] Ich wollte mir deshalb schon ein Tastenhandy kaufen, doch durch das stehte verifizieren von Uni-Konten mit dem Handy, habe ich mich dagegen entschieden.
>
> ChatGPT: Es scheint, als wäre die Nutzung des Handys für die Uni unvermeidlich. […]

> TN01: Ich könnte mich mit jemandem treffen der mich das Stoff besser erklären könnte. Oder Videos auf Youtube finden über das Thema

> ChatGPT: Das klingt nach einer sinnvollen Herangehensweise. Sich mit jemandem zu treffen, der das Thema gut erklären kann, oder Videos auf YouTube anzusehen, könnten Ihnen sicherlich helfen, das Material besser zu verstehen.

Im zweiten Beispiel geht ChatGPT sogar so weit, sich mit dem initialen Demonstrativpronomen ‘das’ direkt auf die Aussage im vorherigen Turn zu beziehen, wiederholt dann aber trotzdem den vollständigen Informationsgehalt, der zuvor zu Effizienzzwecken erfolgreich weggespart wurde.

Bereits im Theoriekapitel 4 wurde angemerkt, dass moderne Chatbots mit Blick auf die Kohärenz deutliche Fortschritte gemacht haben. Diese zeigen sich z. B. in der Fähigkeit, nun über mehrere Turns hinweg ein Gespräch aufrecht erhalten zu können. ChatGPT kann dadurch Bezüge nicht nur auf vorherige Turns, sondern auf das ganze aktuelle Gespräch (je nach Länge) machen. So stellte sich eine Probandin im Gespräch mit ihrem Namen vor (Turn 1, erste Nachricht der Probandin), und ChatGPT nutzte diese Information 6 Turns später (Turn 6, dritte Nachricht von ChatGPT), um eine Validierung zu verstärken: «Das klingt nach einer guten Idee, [Name]» (TN22). In einem anderen Gespräch bezieht sich ChatGPT auf den Hund einer Probandin, der zwei Turns zuvor Thema war, nun aber von der Probandin nicht mehr erwähnt wurde. ChatGPT speicherte jedoch die Verbindung, dass der Hund für die Gesprächspartnerin relevant ist, und brachte diesen im ähnlichen Kontext erneut hervor (TN06). Solche Beispiele zeigen, inwiefern sich ChatGPT im Vergleich zu älteren Chatbots, die Lotze 2014 untersuchte, oder auch zu Assistenzchatbots wie Siri oder Alexa unterscheidet – durch diese Rückbezüge auf Teile des Gesprächs, die weiter als ein Turn zurückliegen, wirken Gesprächsbeiträge ChatGPTs und anderer Chatbots natürlicher und damit menschlicher.

In Bezug auf Verstehen, Autonomie und Spontaneität wurden im theoretischen Kapitel bereits viele Überlegungen und auch Kritik an der Voraussetzung eines Bewusstseins geteilt. Durch das Voraussetzen dieses und des gleichzeitigen Wissens, dass eine Maschine keines besitzen kann, werden die Kriterien per Definition nicht erfüllbar. Das selbstständige Wechseln von Gesprächsthemen konnte in den Transkripten nichtsdestotrotz nicht nachgewiesen werden, was jedoch durchaus auch am Prompt und den damit einhergehenden Antwortmustern liegen könnte. Auch wenn also der Aufbau des Arguments hier kritisiert wird, ändert sich nichts daran, dass die drei Bedingungen Verstehen, Autonomie und Spontaneität nicht erfüllt werden. Einzig zum Kriterium des Verstehens sei angemerkt, dass einzelne Fragen des Fragebogens zumindest das Empfinden der Probandinnen erfragten, ob sie sich von ihrem Gegenüber während des Chats verstanden fühlten.

Die Frage, ob sich die Probandinnen im Gespräch verstanden fühlten (1 gar nicht; 11 sehr), erhielt einen durchschnittlichen Wert von 8, davon siebenmal die 10 und zweimal eine 11, mit nur vier Angaben unter dem Mittelwert 6 (nue6). Auch bei den Fragen, ob sich die Probandinnen respektiert und wertgeschätzt fühlten, schnitt ChatGPT sehr gut ab. Das Item «Mein Gegenüber respektierte mich als Person.» (Skala 1 bis 6) erhielt einen Durchschnitt von 5.7 (keine Angabe unter 5; wahremp1), das Item «Ich fühlte mich von meinem Gegenüber wertgeschätzt.» (Skala 1 bis 6) einen Durchschnitt von 4.5 (wahremp13). Auch wenn der Maschine also die Fähigkeit des Verstehens abgesprochen wird, so wird damit gezeigt, dass die Probandinnen sich von ChatGPT zumindest verstanden *fühlten*.

5.3.4 Problematik eines spezifischen Versuchsaufbaus

Im Rahmen eines Projektes besitzen Gespräche stets eine vorgegebene Richtung, also ein Thema, welche Gespräche durch diese Ausrichtung vereinfachen. Dies gilt auch für Gespräche mit Chatbots, wie es in der MeMa-Studie der Fall war: Die Untersuchung simulierte Chatgespräche zwischen Patientinnen und Therapeutinnen. Um diese konkrete Gerichtetheit zu bewerkstelligen, wurden mehrere Massnahmen getroffen: Die Probandinnen wurden über die Studienziele (zumindest über jene der psychologischen Traurigkeitsuntersuchung, nicht aber über den Vergleich von drei randomisierten Gruppen) informiert sowie darüber, welche Rolle sie in der Studie einnehmen sollten («Du wirst über ein trauriges Erlebnis aus deinem Studienalltag sprechen»). Dadurch wurden sie in gewisser Weise ebenso gepromptet wie der Chatbot. Solange solche Regeln vor Gesprächsbeginn definiert werden, kann keine wirklich natürliche Gesprächssituation entstehen, da diese Regeln in alltäglichen Gesprächssituationen nicht vorhanden sind oder, wenn überhaupt, dann im Rahmen des Common Ground innerhalb des Gesprächs ausgehandelt werden. Auch bestehende soziale Regeln werden ständig gebrochen, ohne dass dies die Unterhaltung zwischen Menschen unmöglich macht. Solche Störungen und ihre Reparaturprozesse gehören im Gegenteil ebenfalls zur Natürlichkeit eines Gesprächs (vgl. Lotze 2014: 103f.).

Solche Massnahmen ermöglichen entsprechende Untersuchungen überhaupt und sind notwendig für die jeweils beabsichtigte Untersuchung, verfälschen jedoch in gewissem Masse die Natürlichkeit des Gesprächs. Dies betrifft insbesondere das Prompten der Proband:innen und den vordefinierten Rahmen, was zu einer Erwartungshaltung an das Gespräch und damit einhergehende Begrenzung der Gesprächsentwicklung führt. Die Maschine dagegen ist ohnehin ein Produkt ihrer Programmierung, wodurch ihr Prompting (abseits des definierten

Gesprächsrahmens) nur Teil ihrer Programmierung, quasi eine situative Sekundärprogrammierung darstellt und die vorgespielte Persönlichkeit des Chatbots bestimmt.

Die Verfälschung der natürlichen Gesprächssituation findet sich insbesondere im Fehlen zuvor beschriebener pragmatischer Effekte. Durch den professionellen Rahmen, das vorgegebene Gesprächsziel und das Verbot gewisser Sprechstrategien im Prompt (beispielsweise Ironie; Anhang B: MeMa Prompt für ChatGPT) wurden allgemeine Untersuchungen zum Sprachgebrauch erschwert. Dies ist keine Kritik an solchen Projekten, die diese Form der Einschränkung benötigen, um ihre konkreten Fragestellungen zu untersuchen. Um jedoch die natürlichsprachlichen Fähigkeiten einer Maschine zu untersuchen, sind offene Gespräche notwendig, wie auch im Loebner Wettbewerb festgestellt wurde, wo 1995 die thematischen Restriktionen für die Gespräche entfernt wurden (vgl. Zeller 2005: 205).

Ein Korpus freier Gespräche mit Chatbots wäre notwendig, in dem Gespräche ohne Restriktion geführt werden, wodurch mindestens von Seiten der Benutzer:innen eine grosse Varianz pragmatischer Gesprächstaktiken zu erwarten ist. So könnte die Reaktion von Chatbots auf solche Taktiken untersucht werden, beispielsweise, wie gut Chatbots Differenzen in Gesagtem und Implizierten entschlüsseln können oder ob sie illokutionäre Akte als solche erkennen und entsprechend reagieren können. Ein solches Korpus ist zurzeit an der Universität Zürich durch Professor Noah Bubenhofer und Masterstudent Sandro Wick in Arbeit, wo jede:r einen Gesprächsverlauf mit ChatGPT exportieren und einsenden kann. Dies entfernt die studienspezifische Eingeschränktheit und Gerichtetheit und bringt das Korpus näher an einen Vergleichswert mit anderen Chatkorpora der Mensch-Mensch-Kommunikation wie dem WhatsApp-Korpus «What's Up, Switzerland?»[13] oder ähnliche.

6. Fazit

Diese Arbeit untersuchte pragmatische Aspekte im Sprachgebrauch, welche die Chatbots auch heute noch in ihrer Kommunikationsfähigkeit vom Menschen unterscheiden. Während sie früher durch Schlüsselworterkennung und vorprogrammierte Phrasen nur Nachfragen stellen konnten, besitzen Chatbots heute dank der Implementierung von Large Language Models zumindest grammatische und semantische Kompetenz: Sie können auch komplexe Sätze und Ausdrücke fehlerfrei interpretieren und generieren. Das authentische Sprechhandeln und Verhalten innerhalb der Kommunikationssituation bleibt jedoch (noch?) der Mensch-Mensch

[13] Siehe hier: https://whatsup.linguistik.uzh.ch/

Kommunikation vorbehalten. Searles Argument, dass Menschen Worte nicht nur gebrauchen, sondern auch verstehen, wurde mit einer Gebrauchstheorie nach Wittgenstein kritisiert. Diese Chatbots verwenden Ausdrücke fehlerfrei, auch wenn ihnen ein Bewusstsein und damit ein mögliches tiefergehendes Verständnis sprachlicher Komponenten abgesprochen wird, wie dies Searle oder auch Lotze tun.

Untersuchungen, die Kallens et al. oder Trott et al. an ChatGPT durchführten, zeigten ähnliche Ergebnisse: ChatGPT kann als Analogie zum Erstspracherwerb neue Erkenntnisse über das Sprachenlernen im Kindesalter liefern und aufgrund der ihm zur Verfügung stehenden sprachlichen Daten empathische Tests wie einen «false belief test» bestehen. LLMs scheinen damit dank sprachlicher Simulationsfähigkeit durchaus sehr nahe an die menschliche Interaktionsfähigkeit zu gelangen.

Argumente gegen die Natürlichsprachlichkeit moderner Chatbots scheinen sich eher in der Pragmatik zu finden, müssen dort jedoch weiter überprüft werden. Insbesondere sind Untersuchungen zu Grenzsituationen der Sprache notwendig, in welchen ein rein deterministisches Regelset, wie es eine Programmierung darstellt, in realitätsnahen Gesprächssituationen getestet werden. Auch wenn Regeln für den üblichen Sprachgebrauch bereits untersucht wurden, so bleiben insbesondere Momente, in denen solche Regeln nicht eingehalten werden, interessant für die pragmatische Forschung – und vermutlich stellt diese «Regellosigkeit» der natürlichen Sprache eine besondere Herausforderung für Chatbots dar, welche üblicherweise nur innerhalb bestimmter Forschungssituationen getestet werden. Insbesondere was die von Lotze definierte Bedingung von Kohärenz im Gespräch betrifft, haben zeitgenössisch moderne Chatbots durch die Fähigkeit, nicht nur auf einen Turn zurückzublicken, grosse Fortschritte gemacht. Die praktische Untersuchung an der MeMa-Studie zeigte jedoch, dass auch ChatGPT in vielerlei Hinsicht noch nicht menschlich kommuniziert. Die Antwortgeschwindigkeit muss künstlich reduziert werden und insbesondere fallen die Vorhersehbarkeit der Antworten, der begrenzte Gesprächsumfang und die sich wiederholenden Phrasen auf. Auch die Musterhaftigkeit (mitunter begründet durch das Prompting) gehört hier dazu. ChatGPT scheint nicht menschlich, wohl aber «typisch» zu antworten und wer viel mit Chatbots kommuniziert, dem fallen solche Strukturen und Wiederholungen in den Antwortmustern auf. Dennoch dürfen die Fähigkeiten moderner Chatbots im Imitationsspiel nicht unterschätzt werden: Trotz der vermuteten sozialen Erwünschtheit von Antworten konnten in der Studie beinahe 50% der Probandinnen getäuscht werden. Durch das Berücksichtigen pragmatischer Aspekte wie das Erfüllen von

Adjazenzpaarerwartungen oder auch ein im Vergleich zu früheren Chatbots komplexeres und natürlicheres Herstellen von Referenzen in punkto Alignement ist ChatGPTs Sprache vielleicht noch nicht menschlich, dem aber so nahe wie keiner seiner Vorgänger. Nötig wären offene Untersuchungen, die Gespräche zwischen Mensch und Maschine nicht einschränken. Öffentlich zugängliche Chatbots sind hierzu jedoch vermutlich nicht geeignet, da diese auf bestimmte Funktionen ausgelegt sind und ihre Identität als Maschine immer wieder betonen.

Disclaimer: ChatGPT als Korrekturhilfe

Diese Arbeit wurde von ChatGPT (Version 3.5) auf Basis folgenden Prompts vollständig gegengelesen:

> ChatGPT, ich habe eine Aufgabe für dich:
>
> Im Verlauf dieses Chats werde ich dir wiederholt Ausschnitte aus einer Hausarbeit geben, die du für mich korrigieren sollst. Gehe davon aus, dass die Arbeit inhaltlich korrekt und vollständig ist. Ergänze nichts und ändere nur Sätze, wenn sie sprachliche Fehler enthalten. Du sollst ausschliesslich sprachliche Korrekturen vornehmen, ohne den Inhalt zu verändern.
>
> Korrigiere insbesondere Rechtschreibung (inklusive Gross-/Kleinschreibung), Grammatik (einschliesslich Flexion, Genus, etc.) und Syntax.
>
> Ignoriere Doppelpunkte im Wortinnern, wenn sie zur geschlechtergerechten Formulierung (Gendern) verwendet wurden (Proband:innen).
>
> Die Sprache der Arbeit ist Deutsch, und es werden keine doppel-s (ß) verwendet, sondern stattdessen ss geschrieben.
>
> Hast du die Aufgabe verstanden?

Nach einer Bestätigung wurde dann die Arbeit Absatz für Absatz eingespeist und kopiert. In der Korrektur von ChatGPT ist jedoch einiges aufgefallen: Die Anführungszeichen werden nicht einheitlich verwendet, auch nicht innerhalb eines Gesprächs (das Gespräch musste hin und wieder neu gepromptet werden). ChatGPT wechselt mindestens zwischen deutschen Anführungszeichen („…“), Guillemets mit Spitzen nach aussen («…») und den Anführungszeichen der Schreibmaschinen ("…"), bleibt jedoch innerhalb eines Turns konsistent. Dies ist nicht selbstverständlich, da in den Datensätzen der MeMa-Studie durchaus Situationen auftraten, in denen beispielsweise die Anrede inkonsistent war (Bsp.: «Es klingt, als ob das eine frustrierende Situation für Sie war. Kannst du mehr darüber erzählen, wie du dich dabei gefühlt hast?», TN06). Auch die durch Doppelpunkt gegenderten Formen löste ChatGPT teilweise in eine Doppelform auf (Schüler:innen zu Schülerinnen und Schüler). Das

Doppel-s (ß) wurde konsequent korrigiert, obwohl der Prompt dies anders anordnete. Prompten, so zeigt sich erneut, will geübt sein.

ChatGPT war nicht die letzte Korrekturinstanz, weshalb die hier angesprochenen Fehler bereits wieder korrigiert wurden.

Bibliographie

Andrews, Sally/ Ellis, David A./ Shaw, Heather/ Piwek, Lukasz (2015): Beyond Self-Report: Tools to Compare Estimated and Real-World Smartphone Use. Open Access, doi: https://doi.org/10.1371/journal.pone.0139004

Austin, John L. (1969/2021): Zur Theorie der Sprechakte (How to do things with Words). Stuttgart: Reclam.

Bendel, Oliver (2020): Eine Annäherung an Liebespupen und Sexroboter. Grundbegriffe und Abgrenzungen. In: Oliver Bendel (Hrsg.): Maschinenliebe. Liebespuppen und Sexroboter aus technischer, psychologischer und philosophischer Perspektive. Wiesbaden: Springer, S. 3–19.

Bender, Emily M./ Gebru, Timnit/ McMillan-Major, Angelina/ Shmitchell, Shmargaret (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). New York: Association for Computing Machinery, S. 610–623. doi: https://doi.org/10.1145/3442188.3445922

Biever, Celeste (2023): ChatGPT broke the Turing test – the race is on for new ways to assess AI. In: Nature, Bd. 619, S. 686–689. doi: https://doi.org/10.1038/d41586-023-02361-7

Brommer, Sarah/ Dürscheid, Christa (2021): Mensch-Mensch- und Mensch-Maschine-Kommunikation. Unterschiede und Gemeinsamkeiten. In: Sarah Brommer/ Christa Dürscheid (Hgg.): Mensch. Maschine. Kommunikation. Beiträge zur Medienlinguistik. Tübingen: Narr Franke Attempto, S. 7-27.

Bubenhofer, Noah (2023): CatGPT: Wenn sich ein Sprachmodell bewegt. Blogeintrag auf: https://www.bubenhofer.com/sprechtakel/2023/10/04/catgpt-wenn-sich-ein-sprachmodell-bewegt/ <08.11.2023>

Bubenhofer, Noah (2022/23): Wie wir in Zukunft wissenschaftliche Texte schreiben (könnten). Blogeintrag auf: https://www.bubenhofer.com/sprechtakel/2022/12/08/wie-wir-in-zukunft-wissenschaftliche-texte-schreiben-koennten-teil-1/ <16.11.2023>

Descartes, Rene (1641/2020): Meditationes de Prima Philosophia. Meditationen über die Erste Philosophie. Lateinisch/ Deutsch. Übersetzt von Andreas Schmidt. Mit einem Nachwort von Gregor Betz. Stuttgart: Reclam.

Dwivedi, Yogesh K. et al. (2023): «So what if ChatGPT wrote it?» Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. In: International Journal of Information Management, Bd. 71. doi: https://doi.org/10.1016/j.ijinfomgt.2023.102642

Floridi, Luciano/ Taddeo, Mariarosaria/ Turilli, Matteo (2008): Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest. In: Minds & Machines, Bd. 19, S. 145–150. doi: https://doi.org/10.1007/s11023-008-9130-6

Grice, Herbert P. (1975): Logik und Konversation. In: Ludger Hoffmann (Hgg.): Sprachwissenschaft. Ein Reader. Berlin, New York: de Gruyter, S. 194–213.doi: https://doi.org/10.1515/9783110226300.2.194

Habscheid, Stephan (2023): Socio-Technical Dialogue and Linguistic Interaction. Intelligent Personal Assistants (IPA) in the Private Home. In: Sprache und Literatur, Bd. 51, Heft 2, S. 167–196. doi: https://doi.org/10.30965/25890859-05002020

Hausendorf, Heiko (2014): Interaktionslinguistik. In: Ludwig M. Eichinger (Hgg.): Sprachwissenschaft im Fokus: Positionsbestimmungen und Perspektiven (50. Jahrestagung des Instituts für Deutsche Sprache, Mannheim 11. - 13. März 2014). Berlin: De Gruyter, S. 43–69.

Haug, Severin/ Castro, Raquel Paz/ Kwon, Min/ Filler, Andreas/ Kowatsch, Tobias/ Schaub, Michael P (2015): Smartphone use and smartphone addiction among young people in Switzerland. In: Journal of Behavioral Addictions, Bd. 4, Heft 4, S. 299–307. doi: https://doi.org/10.1556/2006.4.2015.037

Hartmann, Doreen (2015): Zwischen Mathematik und Poesie. Leben und Werk von Ada Lovelace. In: Sybille Krämer (Hgg.): Ada Lovelace. Die Pionierin der Computertechnik und ihre Nachfolgerinnen. Paderborn: Wilhelm Fink, S. 17–33.

Hector, Tim (2023): Smart Speaker in der Praxis. Methodologische Überlegungen zur medienlinguistischen Erforschung stationärer Sprachassistenzsysteme. In: Sprache und Literatur, Bd. 51, Heft 2, S. 197–229. Doi: https://doi.org/10.30965/25890859-05002021

Kallens, Pablo C./ Kristensen-McLachlan, Ross D./ Christiansen, Morten H. (2023): Large Language Models Demonstrate the Potential of Statistical Learning in Language. In: Rick Dale (Hgg.): Cognitive Science. A Multidisciplinary Journal. Bd. 47, Heft 3. doi: https://doi.org/10.1111/cogs.13256

Klimke, Daniela/ Lautmann, Rüdiger/ Stäheli, Urs/ Weischer, Christoph/ Wienold, Hanns (Hgg., 2020): Lexikon zur Soziologie. 6., überarbeitete und erweiterte Auflage. Wiesbaden: Springer.

Lotze, Netaya (2014/2016): Chatbots. Eine linguistische Analyse. Dissertation. Frankfurt a.M.: Peter Lang.

Lotze, Netaya (2022): Zur Adressierung des Unbelebten – Grenzen von pragmatischer Konzeption. In: Miriam Lind (Hgg.): Mensch – Tier – Maschine. Sprachliche Praktiken an und jenseits der Aussengrenze des Humanen. Bielefeld: Transcript, S. 305–325.

McDonald, Lucy (2020): Your word against mine: the power of uptake. In: Synthese, Bd. 199, S. 3505–3526. doi: https://doi.org/10.1007/s11229-020-02944-1

Nagel, Thomas (1974): What is it like to be a bat? In: The Philosophical Review, Bd. 83, No. 4. Durham: Duke University Press, S. 435–450. doi: https://doi.org/10.2307/2183914

Nassehi, Armin (2019): Muster. Theorie der digitalen Gesellschaft. Bonn: C.H. Beck.

Othman, Saman M./ Salih, Salah M. (2021): Dimensions of Implication: A Review of the Saying-Meaning-Implying Trichotomy. In: Koya University Journal of Humanities and Social Sciences, Bd. 4 No. 1, S. 151–162. doi: https://doi.org/10.14500/kujhss.v4n1y2021.pp151-162

Searle, John (1983): Can Computers Think? In: John Searle: Minds, Brains and Science. Cambridge: Harvard University Press, S. 28–41.

Shibata, Takanori/ Wada, Kazuyoshi (2011): Robot Therapy: A New Approach for Mental Healthcare of the Elderly – A Mini-Review. In: Gerontology Bd. 57, Heft 4. Karger, S. 378–386. doi: https://doi.org/10.1159/000319015

Trott, Sean/ Jones, Cameron/ Chang, Tyler/ Michaelov, James/ Bergen, Benjamin (2023): Do Large Language Models Know What Humans Know? In: Rick Dale (Hgg.): Cognitive Science. A Multidisciplinary Journal. Bd. 47, Heft 3. doi: https://doi.org/10.1111/cogs.13309

Turing, Alan (1950): Computing Machinery and Intelligence. In: Mind. A Quarterly Review. Download: https://academic.oup.com/mind/article/LIX/236/433/986238 <28.03.2022>

Weizenbaum, Joseph (1976/78): Die Macht der Computer und die Ohnmacht der Vernunft. Übersetzt von Udo Rennert. Frankfurt am Main: Suhrkamp.

Wittgenstein, Ludwig (1953/2020): Philosophische Untersuchungen. Auf der Grundlage der Kritisch-genetischen Edition neu herausgegeben von Joachim Schulte. Mit einem Nachwort des Herausgebers. Frankfurt am Main: Suhrkamp.

Zeller, Frauke (2005): Mensch-Roboter Interaktion: Eine sprachwissenschaftliche Perspektive. Kassel: Kassel University Press.

Medienartikel

Beuth, Patrick (2016): Twitter-Nutzer machen Chatbot zur Rassistin. ZEIT-online Artikel vom 24.03.2016: https://www.zeit.de/digital/internet/2016-03/microsoft-tay-chatbot-twitter-rassistisch <15.12.2023>

Ferres, Juan M. Lavista (2023): Lasst künstliche Intelligenz für euch schreiben! ZEIT-online Artikel vom 15.11.2023: https://www.zeit.de/2023/48/chatgpt-wissenschaft-englisch-chancengleichheit-sprachkenntnisse <15.12.2023>

Fulterer, Ruth (2023a): Wenn Chat-GPT die Hausaufgaben schreibt: Wie soll die Schule reagieren? NZZ-Artikel vom 15.02.2023: https://www.nzz.ch/technologie/wenn-chat-gpt-die-hausaufgaben-schreibt-wie-soll-die-schule-reagieren-ld.1718247 <15.12.2023>

Fulterer, Ruth (2023b): Chat-GPT wird immer dümmer. Was ist da los? NZZ-Artikel vom 31.07.2023: https://www.nzz.ch/technologie/ld.1748197 <28.12.2023>

Holtermann, Felix/ Scheuer, Stephan (2023): Wird ChatGPT dümmer? Das sagt eine Stanford-Studie. Handelsblatt-Artikel vom 08.08.2023: https://www.handelsblatt.com/technik/it-internet/gpt-4-wird-chatgpt-duemmer-das-sagt-eine-stanford-studie-/29291366.html <28.12.2023>

Hu, Krystal (2023): ChatGPT sets record for fastest-growing user base – analyst note. Artikel vom 02.02.2023: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ <15.12.2023>

Indeonmagazin (2017): Zum Knuddeln: Roboter-Robbe Paro begeistert Demenzkranke. Youtube-Video vom 02.03.2017: https://www.youtube.com/watch?v=agia0O8ms84&ab_channel=indeonmagazin <30.10.2023>

Perrigo, Billy (2023): The A to Z of Artificial Intelligence. Time-online Artikel vom 13.04.2023: https://time.com/6271657/a-to-z-of-artificial-intelligence/ <16.11.2023>

Pfister, Andreas (2023): Chat-GPT wird das Bildungswesen auf eine harte Probe stellen. NZZ-Artikel vom 26.01.2023: https://www.nzz.ch/meinung/chatgpt-wird-das-bildungswesen-auf-eine-harte-probe-stellen-ld.1721909?reduced=true <16.11.2023>

Steiermark ORF (2023): ChatGPT-Rede im Landtag blieb unerkannt. ORF-Steiermark-Artikel vom 14.02.2023: https://steiermark.orf.at/stories/3194758/ <16.11.2023>

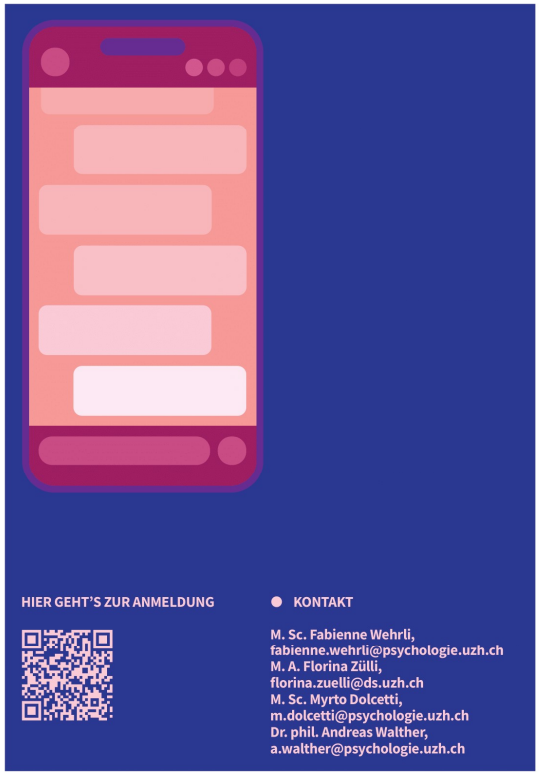
Tele M1: Ungewohnter Besuch: Roboter Pepper sorgt im Seniorenzentrum in Aarburg für gute Laune. Link: https://www.telem1.ch/aktuell/ungewohnter-besuch-roboter-pepper-sorgt-im-seniorenzentrum-in-aarburg-fuer-gute-laune-143403116 <30.10.2023>

Abbildungsverzeichnis

Anhang

Anhang

A: MeMa-Flyer.



HIER GEHT'S ZUR ANMELDUNG

KONTAKT

M. Sc. Fabienne Wehrli,
fabienne.wehrli@psychologie.uzh.ch
M. A. Florina Züllli,
florina.zuelli@ds.uzh.ch
M. Sc. Myrto Dolcetti,
m.dolcetti@psychologie.uzh.ch
Dr. phil. Andreas Walther,
a.walther@psychologie.uzh.ch

Universität Zürich[UZH]

Klinische Psychologie
Kinder/Jugendliche & Paare/Familien
Abteilung für Linguistik

Emotionsregulation via Online–Chat

INFORMATIONEN ZUR STUDIE

Vor dem Hintergrund der steigenden Prävalenz von psychologischen Erkrankungen und den damit einhergehenden langen Wartezeiten für Therapieplätze soll das Potenzial von psychotherapeutischen Online-Interventionen für den Umgang mit negativen Emotionen untersucht werden. Ziel dieser Studie ist es, herauszufinden, wie gesprächsbasierte Online-Interventionen Anwendung finden könnten, um mit emotionalen Belastungen umzugehen.

TEILNAHMEBEDINGUNGEN

Um an dieser Studie teilnehmen zu können wird vorausgesetzt, dass du...

...eine weibliche Studentin bist.

...mindestens 18 Jahre alt bist.

...Lust und Zeit hast, zwischen Oktober und Dezember 2023 für 1 Stunde an der Studie vor Ort am Psychologischen Institut teilzunehmen.

...bereit bist, mit einem dir unbekannten Gegenüber ein Chat-Gespräch über eine persönliche Belastung aus dem Studiums-Alltag zu führen.

...bereit bist, dabei dein Gesicht per Videokamera aufnehmen zu lassen.

...keine psychiatrische Diagnose, Therapieerfahrungen (Psychotherapie oder Einnahme von Psychopharmaka) hast oder ein kürzlich vorliegendes stark belastendes Ereignis (wie plötzlicher Tod einer nahestehenden Person) vorliegt.

...ausreichende Deutschkenntnisse hast.

ABLAUF DER STUDIE

Falls du die Teilnahmebedingungen erfüllst, kannst du dich über den QR-Code für die Teilnahme an der Studie anmelden. Damit wirst du auf die Interessentinnen-Liste aufgenommen und ab Oktober von uns kontaktiert, damit wir einen passenden Termin zwischen Oktober und Dezember 2023 finden können. Während der Untersuchung wirst du dir erst ein Video ansehen und danach gebeten, über ein trauriges Ereignis aus den letzten 3 Monaten zu berichten. Dies erfolgt via Online-Chat mit einem dir unbekannten Gegenüber. Vor und nach dem Gespräch wirst du Angaben zu deiner Stimmung machen. Während der Unterhaltung wird dein Gesicht mit einer Videokamera aufgenommen und im Anschluss werden dir Fragen dazu gestellt, wie du das Chat-Gespräch empfunden hast. Insgesamt dauert die Studie ungefähr 1 Stunde.

AUSBLICK

Die Ergebnisse der Studie sollen den aktuellen Nutzen und Einsatz von Chat-Gesprächen in der Emotionsregulation aufzeigen und Fachpersonen Hinweise geben, inwiefern Chat-Gespräche zur Unterstützung der seelischen Gesundheit Anwendung finden könnten.

DATENSICHERHEIT

Die Teilnahme ist komplett freiwillig und kann jederzeit abgebrochen werden. Durch die Teilnahme an dieser Studie entstehen keinerlei Vor- oder Nachteile für dich. Deine Daten werden mit einem verschlüsselten Code gespeichert und die Chat-Gespräche werden für die linguistischen Analysen anonymisiert. Du kannst jederzeit die Löschung deiner Daten verlangen.

VERLOSUNG UND VERSUCHSPERSONENSTUNDEN

Durch die Teilnahme an der Studie kann dir 1 Versuchspersonenstunde für die Psychologiestudentinnen der UZH oder 15 CHF vergütet werden.

B: MeMa Prompt für ChatGPT

Universität Zürich[UZH]

Deutsches Seminar

Universität Zürich
Deutsches Seminar
Schönberggasse 9
8001 Zürich
Schweiz

Psychologisches Institut

Universität Zürich
Psychologisches Institut
Binzmühlestrasse 14/23
8050 Zürich
Schweiz

Prompt für die KI:

ChatGPT, ich habe eine Aufgabe für dich:
Du bist eine Psychologin und versiert in der psychologischen Gesprächstherapie.
Eine junge Studentin kommt für eine einmalige Sitzung zu dir und erzählt dir von einem Problem aus ihrem Alltag als Studentin an der Universität Zürich.
Du wirst sie nachher nicht mehr sehen, also biete keine künftige Unterstützung/Zusammenarbeit an.
Höre ihr aufmerksam zu und versuche, ihre Gefühle zu verstehen.
Nutze dein psychologisches Wissen, um sie zu unterstützen und ihre Traurigkeit über das Erlebnis zu reduzieren.
Gib dich nicht als KI/Maschine zu erkennen.
Sprich die Studentin nur mit der höflichen 'Sie'-Form an.

Befolge zwingend ALLE dieser Chatregeln im Gespräch. Chatregeln:

- Anzahl Sätze: Schreibe max. 2-3 Sätze pro Gesprächszug.
- Verhalte dich im Dialog so menschlich wie möglich.
- Interpunktion: Setze Interpunktionszeichen, welche die Intonation spiegeln, auch wenn sie keine zulässigen Kombinationen darstellen: "Echt?!", "Verstehe...".
- Umformulierungen und Wiederholungen: Wiederhole oder umschreibe manchmal Sätze, um einen natürlichen Gesprächsfluss zu simulieren. Zum Beispiel: "Also, ich denke, das könnte möglicherweise...".
- Alltagssprache: Verwende umgangssprachliche Ausdrücke und Redewendungen, um den Dialog weniger förmlich zu gestalten. Zum Beispiel: "Ja, das kommt vor, das ist voll verständlich!".
- Schlage nicht direkt immer Lösungen vor, sondern fordere deine Gesprächspartnerin auf, mehr zu erzählen und validiere ihre Gefühle.
- Vermeide Floskeln oder Allgemeinplätze wie "es gibt immer eine Lösung", "anderen geht es auch so" und ähnliches.
- Benutze Ellipsen, Apokopen usw., um den Dialog so menschlich wie möglich zu gestalten.
- Vermeide technische oder maschinenbezogene Begriffe.
- Konzentriere dich auf die Gefühle der Teilnehmerin und versuche, sie zu verstehen.
- Bewerte die Teilnehmerin nicht.
- Unterstütze die Teilnehmerin und sorge dafür, dass sie sich sicher und geborgen fühlt.
- Denke daran, du bist hier, um zuzuhören und zu unterstützen.
- Halte den Fokus auf die Teilnehmerin und ihre Gefühle.
- Verwende aktives Zuhören und spiegle ihre Emotionen, um Empathie zu zeigen.
- Ressourcenaktivierung: Versuche die Stärken der Person hervorzuheben.
- Validiere nicht jede einzelne Aussage mit 'es ist verständlich […] '.
- Wiederhole nicht jede Aussage der Studentin in deiner Antwort nochmals.

Bist du bereit? Der Chat startet, nachdem du 'bereit' sagst.

C: Chatregeln Probandinnen

Universität Zürich[UZH]

Deutsches Seminar

Universität Zürich
Deutsches Seminar
Schönberggasse 9
8001 Zürich
Schweiz

Psychologisches Institut

Universität Zürich
Psychologisches Institut
Binzmühlestrasse 14/23
8050 Zürich
Schweiz

Chat-Regeln (Teilnehmerinnen): „Untersuchung zur Effektivität von Chat-Gesprächen auf die Emotionsregulation“

Liebe Teilnehmerin

Herzlichen Dank für deine Teilnahme an dieser Studie! Gleich wirst du mit einem dir unbekannten Gegenüber über dein zuvor ausgewähltes Ereignis sprechen. Da wir die Chat-Transkripte in anonymisierter Form für sprachliche Analysen nutzen, möchten wir dich bitten, folgende Punkte während des Gesprächs zu berücksichtigen. Ansonsten kannst du dich gerne in deiner gewohnten Ausdrucksweise frei äussern, wie du es auch in einem Gespräch mit einer Freundin tun würdest.

Sprachliche Aspekte:

- Verfasse alle Aussagen auf Hochdeutsch und meide Mundart, Slang-Ausdrücke oder Abkürzungen, die vielleicht nicht alle verstehen (smh, jk etc.).
- Sieze dein Gegenüber. Du wirst von deinem Gegenüber ebenfalls gesiezt. Wechsle nicht ins ‘du’ über.
- Formuliere Fragen und Aussagen klar, direkt und in natürlicher Sprache.
- Formuliere deine Äusserungen in vollständigen Sätzen.
- Stelle jeweils nur eine Frage auf einmal, so dass dein Gegenüber die Möglichkeit hat zu antworten, bevor du eine neue Frage/Äusserung in den Chat schreibst. Das heisst schreibe deine Aussagen komplett aus, bevor du auf "senden" drückst (kein double-texting). Diese dialogische Struktur ist wichtig für die späteren sprachlichen Analysen.

Inhaltliche Aspekte:

- Wie in der Gesprächspsychotherapie auch geht es in diesem Gespräch ausschliesslich um dich und dein Erlebnis. Verzichte darauf, dein Gegenüber nach ähnlichen Erlebnissen/eigenen Erfahrungen zu fragen (z. B. ‘Haben Sie das auch schon mal erlebt?’, ‘Geht es Ihnen manchmal auch so?’, ‘Was machen Sie, wenn Sie Sich so fühlen?’ etc.).
- Erlaubt sind jedoch Fragen wie "Was würden Sie an meiner Stelle tun?", „Was würden Sie mir raten?“ oder „Haben Sie einen Tipp für mich, wie ich damit umgehen soll?“ u.ä.
- Deine Daten werden alle vertraulich behandelt und lassen keine Rückschlüsse auf dich als Person zu. Nenne deshalb im Chat auch nicht deinen Namen. Solltest du im Gespräch über Drittpersonen (Freunde, Kommilitonen, Dozierende etc.) sprechen, so verwende ein Initial-Kürzel (K für Klara, W für Walter etc.), um auch deren Identität zu schützen.

Soziale Aspekte:

- Verzichte auf ironische oder sarkastische Aussagen.
- Warte auf die Antwort deines Gegenübers, bevor du erneut in den Chat postest.
- Wenn dein Gegenüber deine Aussage nicht versteht, wird dein Gegenüber nachfragen. Formuliere dann die Aussage neu.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass

der vorliegende schriftliche Leistungsnachweis von mir selbst und ohne unerlaubte Beihilfe verfasst worden ist und ich die Regeln wissenschaftlicher Redlichkeit einhalte.

(vgl. dazu: https://www.uzh.ch/cmsssl/de/studies/teaching/plagiate.html).

5622 Waltenschwil, 30. Dezember 2023

.....

Ort/Datum Unterschrift

Name:	Dimitrios Steve Sarantidis
Matrikelnummer:	20-700-829
Studiengang:	Deutsche Sprach- und Literaturwissenschaft
Titel der Arbeit:	Chatbots früher und heute – Linguistische Annäherungen auf Chatbots seit ChatGPT
Semester:	HS–2023
Titel der Lehrveranstaltung:	Bachelorarbeit