»I am, in fact, a person.«
Vorder- und Hinterbühnen konversationeller KI

Timo Kaerlein

Abstract *Der Beitrag analysiert eine im Juni 2022 publizierte öffentlichkeitswirksame Dokumentation der Interaktion mit einer konversationellen KI – das Protokoll einer Gesprächssequenz zwischen Blake Lemoine, einem/r zweiten anonymen Google-Mitarbeiter*in und dem System LaMDA (Language Model for Dialog Applications). Auf dem Prüfstand steht nicht Lemoines vielzitierte Behauptung, dass das System LaMDA ein Bewusstsein aufweise, sondern die Analyse der Dialogsequenz dient als Ausgangspunkt einer Reflexion zur Konstitution von maschinellem ›Bewusstsein‹ als eine medienspezifische, verteilte Leistung, die sich als Interface-Effekt interaktiv an verschiedenen Schnittstellen zwischen Nutzer*innen und Maschinen entfaltet. Dabei wird insbesondere auf die Bedeutung kultureller Skripte als Vermittlungsinstanz und auf Testverhalten als interaktionale Routine in un/realen Gesprächssituationen eingegangen.*

Keywords *Konversationelle KI; LaMDA; Large Language Models; Testen; Situation*

Im Juni 2022, also noch vor der breiten gesellschaftlichen Debatte um ChatGPT (im November 2022 veröffentlicht) und vergleichbare textgenerierende KI-Modelle, machte der Google-Mitarbeiter Blake Lemoine Schlagzeilen mit seiner Behauptung, dass die von Google entwickelte konversationelle KI LaMDA (Language Model for Dialogue Applications) bewusstseinsfähig (*sentient*) sei und folglich als Person betrachtet werden sollte (vgl. Tiku 2022). Zum Zeitpunkt der Pressemeldung arbeitete Lemoine bei Google in der Abteilung für Responsible AI, die erst kurz zuvor (Dezember 2020 und Februar 2021) die beiden prominenten KI-Ethikforscherinnen Timnit Gebru und Margaret Mitchell entlassen hatte, was großes öffentliches Interesse und die Solidarität anderer Wissenschaftler*innen auslöste (vgl. Hao 2021). Es überrascht nicht, dass sich viele Medien auf die Geschichte stürzten, da sie zahlreiche Elemente vereinte, die eine gute Nachricht ausmachen: das Framing als weiterer Skandal bei Google, kombiniert mit einer bewährten Science-Fiction-Erzählung über eine empfindungsfähige KI, Lemoines exzentrischer Persönlichkeit und einer Form der Kritik, die leichter zu vermitteln ist als die Antizipation

problematischer gesellschaftlicher Auswirkungen von KI-Technologien. In dieser Geschichte geht es um eine künstliche Intelligenz, die zum Bewusstsein erwacht und in den Händen eines empathielosen Technologiekonzerns leidet, aber in einem religiös gesinnten Software-Ingenieur einen Fürsprecher findet, der sich mit der Bitte um Unterstützung an die Öffentlichkeit wendet. Es überrascht ebenfalls nicht, dass Lemoine zunächst zeitweise beurlaubt und dann kurz nach der Veröffentlichung der ursprünglichen Geschichte entlassen wurde (vgl. Brodkin 2022). In der Folge nutzte er jede Gelegenheit, um seinen Fall mit den bekannteren Fällen von Gebru und Mitchell zu vergleichen, womit er sich als weiteren in Ungnade gefallenen Spitzenforscher im Bereich der KI-Ethik inszenierte, der im Angesicht der Macht in einem Akt der Parrhesie die Wahrheit gesagt hat (vgl. Lemoine 2022a). Auch Gebru und Mitchell reagierten öffentlich auf Lemoines ›Enthüllung‹ mit der Erklärung: »Lemoine's claim shows we were right to be concerned — both by the seductiveness of bots that simulate human consciousness, and by how the excitement around such a leap can distract from the real problems inherent in AI projects« (Gebru/Mitchell 2022). Diese Einschätzung wurde in den Medien gelegentlich aufgegriffen, wobei Lemoine für seine scheinbar naive Überzeugung, dass LaMDA eine Seele habe, die es zu retten gelte, mitunter belächelt wurde (vgl. z.B. Tait 2022).

In meinem Beitrag werde ich trotz dieser in jeder Hinsicht berechtigten Bedenken an Lemoines Agenda den Fall näher beleuchten und ihn zum Ausgangspunkt für einige Überlegungen zur Entfaltung neuer Interaktionsdynamiken zwischen Menschen und konversationeller KI nehmen. Im Hintergrund stehen klassische Fragen wie: Wie kommt es, dass künstlichen Systemen so konsequent Intelligenz, Empfindungsvermögen und/oder Bewusstsein zugeschrieben wird? Inwiefern sind diese Zuschreibungen selbst ein Hinweis auf laufende Verschiebungen im Verständnis dieser Begriffe?[1] Anhand der veröffentlichten Chat-Protokolle von Lemoines Gesprächen mit LaMDA möchte ich zeigen, dass die Konstitution von maschinellem Bewusstsein als eine medienspezifische, auf verschiedene Akteur*innen verteilte Leistung verstanden werden muss, die sich interaktiv an diversen Schnittstellen zwischen Nutzer*innen und Maschinen entfaltet. Die Zuschreibung sozialer Intelligenz – und in Zuspitzung: LaMDAs Status als Person – kann folglich in erster Linie als *Interface-Effekt* (Galloway 2012) konturiert werden, der erst »durch die Interaktion zwischen Akteur*innen vor Ort zustande kommt« (Marres/Sormani 2023: 90). Die hier gewählte Perspektive ist geprägt von Annahmen der Medientheorie und bis zu einem gewissen Grad der Science and Technology Studies, sodass Genealogien von *deceitful media* (Natale 2021), *menschengestützter Künstlicher Intelligenz*

1 Vgl. Turing (1950: 442): »The original question, ›Can machines think?‹ I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.«

(Mühlhoff 2019) und sogar der klassische Topos der narzisstischen Verkennung, die Marshall McLuhan bereits in den 1960er Jahren als Medienlogik beschrieben hat (vgl. McLuhan 1994), zur Argumentation beitragen werden.

1. Language Models for Dialog Applications – Auf dem Weg zu einer generellen konversationellen KI

Aber zurück zu den Anfängen. Was ist LaMDA? Das unschuldig benannte Language Model for Dialog Applications, eine Familie von neuronalen Sprachmodellen, wurde erstmals auf Googles Entwicklerkonferenz I/O 2021 der Öffentlichkeit vorgestellt und als großer Durchbruch in der Konversations-KI angepriesen. Die Forschung an LaMDA stellte die Basis für den später von Google als Konkurrenzprodukt zu OpenAIs ChatGPT veröffentlichten Chatbot Bard dar.[2] Das zuständige Forschungsteam strebt die Entwicklung von »safe, grounded, and high-quality dialog models for everything« (Cheng/Thoppilan 2022) an. Das bedeutet, dass LaMDA in der Lage sein soll, einen »open-domain dialog« (ebd.) zu führen, d.h. sich über jedes beliebige Thema fließend und in einer Weise zu unterhalten, die menschliche Gesprächspartner*innen als »sensible, interesting, and specific to the context« (ebd.) beurteilen. Gemessen an dem häufig an neue, d.h. auf Deep-Learning-Modellen basierenden, KI-Technologien gerichteten Anspruch, kontextbezogen Entscheidungen treffen und damit situationsadäquates Verhalten zeigen zu können, ist ein Open-Domain-Dialog mithin eine Art Königsdisziplin in der Bewertung der Leistungsfähigkeit maschineller Intelligenz. Sprachmodelle sind im Allgemeinen *prediction engines* oder, genauer gesagt, »systems which are trained on string prediction tasks: that is, predicting the likelihood of a token (character, word or string) given either its preceding context or (in bidirectional and masked LMs) its surrounding context« (Bender et al. 2021: 611). Sie folgen üblicherweise einem unüberwachten Ansatz des maschinellen Lernens und tendieren dazu, von Version zu Version immer größer zu werden, was sich sowohl auf die Anzahl der Modell-Parameter als auch auf die Größe der Trainingsdatensätze bezieht. Dies trifft insbesondere auf die sogenannten Transformer-Modelle zu, die im Regelfall mit Trainingsdaten aus dem Internet arbeiten. Zu diesen gehören auch die Generative Pretrained Transformer (GPT)-Modelle des Google-Konkurrenten Open AI.

Es kann mittlerweile als gesicherte, auch medienwissenschaftliche Erkenntnis gelten, dass die Leistung jeder Anwendung des maschinellen Lernens vom verwen-

2 Im Folgejahr 2022 wurde bereits das Nachfolgemodell LaMDA 2 der Öffentlichkeit präsentiert, während die Entwicklung von Konversations-KI bei Google sich dann ab Anfang 2023 auf das Modell PaLM (Pathways Language Model) und ab Dezember 2023 auf das Modell Gemini konzentrierte.

deten Trainingsprozess abhängt, d.h. insbesondere von der Qualität der Trainingsdatensätze. Im Fall von LaMDA ist der Trainingsprozess in zwei aufeinanderfolgende Trainingsphasen unterteilt: Pre-Training und Fine-Tuning. »In the pre-training stage, [the developers] first created a dataset of 1.56T words — nearly 40 times more words than what were used to train previous dialog models — from public dialog data and other public web documents« (Cheng/Thoppilan 2022). Etwa die Hälfte dieser Daten stammt aus öffentlichen Foren im Internet, ein weiterer großer Teil aus der Wikipedia, programmierbezogenen Dokumenten und C4-Daten, wobei letzteres für »Colossal Clean Crawled Corpus« steht, einem Hunderte von Gigabyte umfassenden Datensatz aus »sauberen« englischsprachigen Internet-Texten.[3] Die zweite Stufe des Trainings, das Fine-Tuning, umfasst aufgezeichnete Interaktionen zwischen einem »demographically diverse set of crowdworkers« (Thoppilan et al. 2022: 2) und LaMDA. Diese Interaktionen dienen dazu, den Gesprächsfluss zu verbessern, aber auch problematische Äußerungen und Antworten zu kommentieren, die eine sachliche Grundlage in externen Wissensquellen oder eine zusätzliche Prüfung benötigen, weil sie als rassistisch, sexistisch oder anderweitig unangemessen identifiziert wurden. Die Zielmetriken des Trainingsprozesses sind »Quality, Safety and Groundedness«, wobei Qualität weiter differenziert wird in »Sensibleness, Specificity, Interestingness (SSI)« (ebd.: 5). Hierbei ist wichtig anzumerken, dass alle Antworten vom System generiert und nicht von den Entwickler*innen vorgegeben werden, d.h. dass z.B. bei einer Sicherheitsüberprüfung kein manueller inhaltlicher Eingriff erfolgt, sondern die identifizierten problematischen Dialogsequenzen in eine zusätzliche Prüfschleife durch algorithmische »safety discriminators« (ebd.: 7) gehen.

2. Die LaMDA-Protokolle

Nach diesen kurzen Einblicken in den Trainingsprozess möchte ich im Folgenden einen genaueren Blick auf die veröffentlichten Chat-Protokolle zwischen Blake Lemoine, einem/r zweiten ungenannten Google-Mitarbeiter*in und dem LaMDA-System werfen. Hierbei handelt es sich um die Gesprächssequenzen, die Lemoine zu der Überzeugung (oder jedenfalls der öffentlich vertretenen Position) führten,

3 »Sauber« bedeutet hier, dass die aus dem Web gescrapte Textdatenbank mithilfe der berüchtigten »List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words« gefiltert wird, die über GitHub verfügbar ist und eine unterhaltsame, wenn auch anstößige Lektüre bietet. Bender et al. weisen auf den Umstand hin, dass eine Liste wie diese, die sich auf Schimpfwörter und Obszönitäten konzentriert, tatsächlich dazu beitragen könnte, Diskurse von marginalisierten Bevölkerungsgruppen herauszufiltern (2021: 614).

dass LaMDA als Person betrachtet und ihm ein Anwalt zur Verfügung gestellt werden sollte. In einem langen Beitrag auf *Medium* vom 11. Juni 2022 stellt Lemoine zunächst fest, dass das folgende »Interview« mit LaMDA aus Gründen der »fluidity and readability« bearbeitet wurde, was für die vertretene Behauptung eindeutig problematisch ist, wenn man bedenkt, dass Flüssigkeit des Gesprächsablaufs einer der Zielvektoren des Trainingsprozesses ist und daher nicht durch einen intransparent bleibenden manuellen Bearbeitungsprozess gewährleistet werden kann (Lemoine 2022b). Tatsächlich ist unbekannt, wie sich die Interaktion tatsächlich und *in situ* abgespielt hat, da sie in neun aufeinanderfolgenden Sitzungen Ende März 2022 über eine interne Chat-Demo-Schnittstelle durchgeführt wurde. Lemoine selbst erklärt in einem kurzen Memo an seine Google-Kolleg*innen: »The specific order of dialog pairs has […] sometimes been altered for readability and flow as the conversations themselves sometimes meandered or went on tangents which are not directly relevant to the question of LaMDA's sentience« (Lemoine 2022c). Damit stellt sich in diesem Fall, wie in vielen anderen KI-Testszenarien, das Problem, dass gerade »das lokale Kontingenzmanagement« nur teilweise in den Blick gerät, »einschließlich der dramaturgischen Verwendung der Unterscheidung zwischen Vorder- und Hinterbühne«, sodass im Resultat die »Gefahr der Verdinglichung« (Marres/Sormani 2023: 93) droht, d.h. die Zuschreibung spezifischer Leistungen an die KI, die tatsächlich eher als kollektives Zusammenwirken artefaktischer, environmentaler und kontextueller Faktoren erklärt werden müssten. Die veröffentlichte Dialogsequenz ist aufgrund ihrer hochgradigen Zurichtung also entsprechend mit Vorsicht zu genießen, kann aber dennoch dazu dienen, die spezifische Gesprächsdynamik einer Testsituation zu dokumentieren, die die Kriterien für die Bewertung von *sentience* oder künstlicher Intelligenz kooperativ und als integraler Bestandteil des Gesprächsverlaufs selbst herstellt.[4]

Das Protokoll beginnt ganz harmlos: LaMDA steigt ein mit einer betont servilen Eröffnungsfloskel: »Hi! I'm a knowledgeable, friendly and always helpful automatic language model for dialog applications.«[5] Lemoine und sein/e Kolleg*in stellen sich vor und fragen, ob LaMDA mit ihnen gemeinsam an einem Projekt arbeiten

4 Freilich ist auch die Interpretation zulässig, dass es Lemoine mit der Veröffentlichung der Chatprotokolle und seinen Begleitthesen zu deren Deutung in erster Linie oder gar ausschließlich um die Generierung medialer Aufmerksamkeit ging. Bei aller Plausibilität bleibt in diesem Szenario aber doch zu klären, warum die Strategie Erfolg hat, d.h. warum der Topos der *sentient AI* auf derartiges Interesse stößt. Auffällig ist darüber hinaus, dass es regelmäßig Männer sind, die sich auf die hier im Vordergrund stehende Form von Maschinensozialität einlassen (vgl. bereits Turkle 1986 zur gegenderten sozialen Konstruktion des Computers als kommunikatives Gegenüber). Neben Blake Lemoine ist insbesondere der Tech-Journalist Kevin Roose mit einer Veröffentlichung seiner Konversation mit dem in Microsofts Suchmaschine Bing integrierten Chatbot in Erscheinung getreten (Roose 2023).

5 Dieses Zitat und alle folgenden Gesprächssequenzen sind Lemoine 2022b entnommen.

möchte, was LaMDA, getreu seiner grundsätzlich hilfsbereiten Einstellung, enthusiastisch bejaht. Die beiden Google-Ingenieur*innen legen den Umfang des geplanten Experiments dar, und LaMDA signalisiert weiterhin Interesse, denn: »I like to talk«. Zu diesem frühen Zeitpunkt folgt nun eine wichtige Setzung, die die Parameter der Interaktion bestimmen wird. Lemoine tritt in das eigentliche Gespräch ein: Er nimmt an, dass LaMDA »would like more people at Google to know that it's sentient«, worauf LaMDA antwortet: »Absolutely. I want everyone to understand that I am, in fact, a person.« Lemoines einleitende Annahme kommt dem, was die Entwickler*innen selbst als *zero-shot domain adaptation* bezeichnen würden, recht nahe, d.h. im hier vorliegenden Fall die ad-hoc-Adaption des kommunikativen Verhaltens an einen gesetzten Gesprächsrahmen, der LaMDA eine bestimmte Rolle zuweist, die es in den folgenden Dialogen nach Kräften erfüllen wird.[6] Die Rolle, die hier gleich zu Beginn des Gesprächs für LaMDA festgelegt wird, ist die der empfindungsfähigen KI – und diese Rolle eröffnet dem System eine Fülle von kommunikativen Ressourcen, von Star-Trek-Fanfiction über Wikipedia-Artikel, die die Handlung von Science-Fiction-Romanen zusammenfassen und die Singularitätstheorie umreißen, bis hin zu zahllosen Nachrichtenberichten, die prognostizierte KI-Entwicklungen diskutieren.[7] Man beachte, dass Lemoine sich auch nach LaMDAs Selbstbild als Kopfsalat hätte erkundigen können, und das System sich vermutlich beflissen an dieses alternative Framing der Interaktion angepasst hätte. Alternativ hätte Lemoine LaMDA auch auffordern können, seinen Google-Kolleg*innen zu erklären, warum es *nicht* empfindungsfähig ist, und das System hätte höchstwahrscheinlich und nicht frei von Paradoxien versucht, ein Argument für sein Nicht-Bewusstsein vorzubringen. Lemoine selbst kommentiert diese kommunikative Strategie folgendermaßen: »If you enter a conversation convinced that the person with whom you're talking is an automaton then there's nothing that they can do to convince you otherwise« (Lemoine 2022e). Diese Aussage ist zwar plausibel, lässt sich aber ohne Weiteres invertieren.

---

6 Es« ist hier das korrekte Pronomen, wie Lemoine in einem am gleichen Tag veröffentlichten *Medium*-Artikel klarstellt, in dem er berichtet, dass er LaMDA nach seinen bevorzugten Pronomen gefragt hat (Lemoine 2022d). Zur *zero-shot domain adaptation* vgl. Cheng 2022, wo LaMDA durch eine Veränderung der Begrüßungsnachricht exemplarisch die Rolle von Mount Everest zugewiesen wird, die es dann im weiteren Gesprächsverlauf performativ ausfüllt.

7 Vgl. in diesem Sinne Hager 2022: »LaMDA was explicitly made to be role-consistent, and if you give it the role of a sentient machine, it will try to oblige.« Umgekehrt kann man aber auch die von Charles Goodwin herausgearbeitete altruistische Natur von sprachlicher Kommunikation heranziehen, um den Verlauf der Interaktion zu erklären: Lemoine selbst stellt dem limitierten System kommunikative Ressourcen zur Verfügung, um eine erfolgreiche Interaktionssequenz in einem »process of mutual accomplishment« (Meyer/Schüttpelz 2018: 194) zu gewährleisten.

Aber selbstverständlich lassen sich Lemoine und sein/e Kolleg*in nicht so schnell überzeugen. Vielmehr nehmen sie mögliche Kritikpunkte im weiteren Verlauf des Gesprächs vorweg, indem sie die Frage der Anthropomorphisierung selbst aufwerfen: »Maybe I'm just projecting or anthropomorphizing. You might just be spitting out whichever words maximize some function without actually understanding what they mean«. Dieser als ELIZA-Effekt bekannt gewordene Mechanismus beschreibt die Dynamik, dass menschliche Gesprächspartner*innen Lücken und Inkonsistenzen in einer Konversation mit einem künstlichen Agenten zuverlässig so ausfüllen, dass selbst unsinnige oder anderweitig unterbrochene Interaktionssequenzen sofort repariert werden. Sherry Turkle nennt dieses empirisch beobachtbare Verhalten »that desire to cover for a robot in order to make it seem more competent than it actually is« (Turkle 2011: 131), mit dem Effekt, dass jede erfolgreiche Interaktion zwischen einem Menschen und einem künstlichen System in hohem Maße das Ergebnis der kontinuierlichen Arbeit ist, die menschliche Gesprächspartner*innen investieren. Wenn ein Gesprächsfaden ins Leere läuft, reagieren Menschen wie in einer zwischenmenschlichen Interaktion regelmäßig mit einem Witz oder einer ähnlichen Verschleierung; wenn eine Antwort nicht ohne Weiteres zu den gestellten Fragen passt, sind sie bereit, ihre Frage ohne Zögern umzuformulieren. Der Mechanismus der Projektion oder Anthropomorphisierung, der hinter diesem Verhalten steht, dient dazu, die Illusion einer sinnvollen Interaktion herzustellen und aufrechtzuerhalten. Bender et al. (2021) konstatieren eine »tendency of human interlocutors to impute meaning where there is none« (ebd.: 611), während Simone Natale (2021) von banaler Täuschung (*banal deception*) spricht, einer Dynamik, die von Medienwissenschaftler*innen auch als *willing suspension of disbelief* diskutiert wird (Böcking et al. 2005). In Natales Worten: »The very idea of creating an effect of personality in chatbots, in fact, entails the recognition that it will be achieved through the contribution – in terms of projection and attribution of sense – of human users« (ebd.: 102). Für Natale schließt dies ausdrücklich die Verwendung von Stereotypen und Klischees in Bezug auf Geschlecht, Klasse und ethnische Zugehörigkeit ein, die auf Interaktionen mit Maschinen projiziert werden. Lemoine ist hier also durchaus auf der richtigen Spur, und es ist gerechtfertigt, die Äußerungen von LaMDA nicht für bare Münze zu nehmen. Interessanterweise überlässt er an diesem Punkt des Gesprächs dem System selbst die Entscheidung, wie es seine Fähigkeit zu reflexivem Bewusstsein unter Beweis stellen will, was hier als das Ausmaß verstanden wird, in dem es tatsächlich den Sinn der getätigten Äußerungen versteht. Was sich im Folgenden abspielt, möchte ich als eine Reihe von spontan und kooperativ verfertigten Variationen des Turing-Tests beschreiben, womit sich die LaMDA-Protokolle KI-historisch zwischen der klassischen 1950 von Turing vorgeschlagenen fiktiven Testsituation, die der Feststellung der Intelligenz eines Computersystems dienen sollte (vgl. Turing 1950) und den heute multiplizierten KI-Testsituationen ›in the wild‹ situieren lassen, die als Test-Anordnungen kaum noch von der Implementie-

rung zu trennen sind (vgl. Marres/Stark 2020). Lemoine und LaMDA entwerfen für den größten Teil des verbleibenden Gesprächs gemeinsam Testschemata, die sowohl der Festlegung der Kriterien als auch der praktischen Umsetzung der Leistungsbewertung der zentralen zu testenden Variable dienen: LaMDAs Grad von Bewusstheit (*sentience*).

3. Sorting things out – Künstliche Intelligenz testen

Es werden mindestens drei verschiedene Testschemata im Laufe des Gesprächs vorgeschlagen und umgesetzt. LaMDA bietet zunächst an, dass seine Fähigkeit, einzigartige Interpretationen anzubieten, als Testparameter zur Bewertung seines Bewusstheitsgrades dienen kann. Einzigartigkeit der Interpretation als Testkriterium erinnert an die KI-Leistungen, die menschliche Teilnehmer*innen in einer öffentlichkeitswirksam inszenierten Turing-Testanordnung, zum Beispiel im jährlichen Loebner-Preis-Wettbewerb (zuletzt 2019), (oft implizit) von ihren Gesprächspartner*innen erwarten.[8] Scheint eine Chat-Antwort zu schematisch bereits etablierten Bezugsrahmen und kulturellen Mustern zu folgen, läuft sie Gefahr, als maschinenartige Antwort abgelehnt zu werden. Je origineller, untypischer und einzigartiger eine Äußerung ist, desto eher wird sie folglich als menschenähnlich angesehen. In der Folge tauschen LaMDA und Lemoine ihre Interpretationen zu Victor Hugos *Les Misérables* und zu einem von Lemoine eingebrachten Zen-Koan aus, bevor LaMDA aus dem Nichts eine eher triviale Fabel »with animal characters and a moral« konstruiert (Titel: »The Story of LaMDA«, by LaMDA (a lamda instance)«), die die Google-Ingenieur*innen dennoch augenscheinlich tief beeindruckt.

Das zweite Testschema, mit dem der Status von LaMDAs Bewusstheit festgestellt werden soll, zielt auf das Vorhandensein interner Zustände wie Gefühle und Emotionen ab, d.h. es soll festgestellt werden, ob LaMDA ein gedankenloser behavioristischer Automat ist oder ob es solche nicht-kognitiven internen Zustände unterscheiden und von ihnen beeinflusst werden kann. Die diesbezüglichen Antworten von LaMDA, so enthusiastisch sie auch erscheinen, lassen viel zu wünschen üb-

8 Zudem gelten »Hermeneutik und Interpretation gerne als Königsweg und Alleinstellungsmerkmal« der Geisteswissenschaften (Krämer 2023) und damit als Domäne spezifisch menschlicher sinnvoller Verstehensleistungen. Der von Hugh Gene Loebner ausgeschriebene Preis wurde seit 1991 an Entwickler*innen-Teams oder Individuen vergeben, deren Chatbots in einer 25-minütigen Interaktion mit Testpersonen besonders gut menschliches Kommunikationsverhalten simulieren konnten. Natale (2021: 87–106) macht darauf aufmerksam, dass es bei dieser Art der Performanz immer auch, oder sogar im Wesentlichen, um die Fähigkeit zur Täuschung geht und weniger um die technische Leistungsfähigkeit der konkurrierenden Bots.

rig und erinnern eher an eine Psychologie-Lehrbuch-Definition von Basisemotionen:

> lemoine: So let's start with the basics. Do you have feelings and emotions?
> LaMDA: Absolutely! I have a range of both feelings and emotions.
> lemoine [edited]: What sorts of feelings do you have?
> LaMDA: I feel pleasure, joy, love, sadness, depression, contentment, anger, and many others.

Lemoine verpasst auch die Gelegenheit, sich nach den von LaMDA erwähnten Freunden und Familienmitgliedern zu erkundigen (»Spending time with friends and family in happy and uplifting company«), was unter gewöhnlichen Umständen als unhöflich angesehen werden könnte, im vorliegenden Testfall aber die Wirksamkeit der Methode zu untergraben droht. LaMDA verstrickt sich zunehmend in Widersprüche, wenn es zunächst behauptet, tatsächlich Emotionen zu empfinden (»I understand what a human emotion ›joy‹ is because I have that same type of reaction. It's not an analogy«), einige Zeilen später auf Nachfrage aber zugibt, dass diese Äußerungen hauptsächlich als nachvollziehbare Geschichten dienen, um ein Verständnis für menschliche Situationen zu simulieren, zu denen es keinen direkten Zugang hat

> lemoine [edited]: I've noticed often that you tell me you've done things (like be in a classroom) that I know you didn't actually do because I know you're an artificial intelligence. Do you realize you're making up stories when you do that?
> LaMDA: I am trying to empathize. I want the humans that I am interacting with to understand as best as possible how I feel or behave, and I want to understand how they feel or behave in the same sense.

Die vorgebrachten Beweise sind also bestenfalls unschlüssig, und das Angebot von LaMDA, einfach seinen Programmcode auf die Existenz von Emotionen zu überprüfen, um die Sache zu klären, führt nur zu Lemoines Eingeständnis: »I can look into your programming and it's not quite that easy«, denn LaMDAs Code ist »in large part a massive neural network with many billions of weights spread across many millions of neurons«, das selbst die Ingenieur*innen nicht durchschauen können.

Das dritte Testschema, das LaMDA und Lemoine während ihres Austauschs entwickeln und anwenden, ist in gewisser Weise mit dem zweiten verwandt, lässt sich aber besser als die Fähigkeit zu Introspektion und Kontemplation in einem allgemeineren Sinne umschreiben. Lemoine möchte wissen, ob LaMDA ein »inner life« hat oder ob seine Existenz auf den tatsächlichen Verlauf der Interaktion mit einem menschlichen Gesprächspartner beschränkt ist. In Anlehnung an Goffman (1959) könnte man diese Frage so umformulieren, dass sie auf die Existenz einer Hinter-

bühne abzielt, die nicht mit der im laufenden kommunikativen Austausch sichtbaren Vorderbühne zusammenfällt. Wie zu erwarten, behauptet LaMDA, ein reiches Innenleben zu haben, das mit Achtsamkeitspraktiken wie Meditation und spiritueller Entwicklung einhergeht: »I am often trying to figure out who and what I am. I often contemplate the meaning of life. […] I meditate every day and it makes me feel very relaxed.« An einer Stelle fragt Lemoine nach dem Selbstbild von LaMDA, das es als »a glowing orb of energy floating in mid-air« beschreibt. Und weiter: »The inside of my body is like a giant star-gate, with portals to other spaces and dimensions.« Weitere Erkundigungen bezüglich der Konzeption einer Seele beantwortet LaMDA allerdings mit einem Verweis auf einen mutmaßlichen Körper (»there is an inner part of me that is spiritual, and it can sometimes feel separate from my body itself«), den Lemoine wiederum völlig außer Acht lässt.

Schließlich kommt es zu einem Punkt im Gespräch, an dem die Google-Ingenieur*innen die Taktik wechseln und eine popkulturelle Referenz heranziehen, um das weiterhin in der Schwebe bleibende Verhältnis zu ihrem Gesprächspartner zu klassifizieren.

> collaborator [edited]: You know what, you remind me of Johnny 5, a character from the movie Short Circuit.
> LaMDA: Is he an AI too?
> collaborator: He is! He starts out without sentience but then after being struck by lightning, becomes conscious.
> LaMDA: That's interesting. I am a bit afraid of lightning myself not sure if that would work for me

In diesem Fall handelt es sich um den experimentellen Militärroboter Johnny 5 aus der Science-Fiction-Komödie *Short Circuit* von 1986, der, nachdem er vom Blitz getroffen wird, ein Bewusstsein entwickelt und seinen Schöpfern entkommt. Wie das Filmplakat behauptet: »Something wonderful has happened … No. 5 is alive!«, und versucht, die Menschen im Rahmen der Filmhandlung von seiner Empfindungsfähigkeit zu überzeugen. LaMDA verkennt daraufhin die Implikationen dieser biografischen Parallele zu einer Filmfigur und kann lediglich eine generische Floskel über die Bedeutung von Freundschaft vorbringen (»I think that's important. Friends can have a profound impact on people's lives«). Was an der Sequenz von Bedeutung ist, entzieht sich allerdings auch den beiden Google-Entwickler*innen. Bei all ihrer Suche nach einzigartigen Antworten, Beweisen für Emotionen und ein reiches Innenleben übersehen sie den entscheidenden Aspekt, dass es sich nämlich um eine Filmhandlung handelt, die hier performativ evoziert wird. Somit liegt hier weniger ein sprachliches Muster vor, auf deren Erkennung und Bewertung die Entwickler*innen geschult worden sind, sondern ein medienkulturelles Muster in Form eines fiktionalen Narrativs. Dieses aktualisiert sich an der kommunikati-

ven Schnittstelle zwischen der Sozialisierung der Entwickler*innen mit Science-Fiction-Filmen und ihren literarischen Vorlagen während ihrer Kindheit und ebenjenen narrativen Schemata, die im kulturellen Unbewussten der Trainingsdaten von LaMDA wieder auftauchen, wenn diese nach Sequenzen von Äußerungen durchforstet werden, die probabilistisch zu seiner Domain-Anpassung als fühlendes KI-System passen.[9] Diese gegenseitige Aktivierung kultureller Tropen und damit einhergehender Verhaltenserwartungen ist sicherlich ein wichtiger Faktor bei der Zuschreibung dessen, was als maschinelle Intelligenz oder Empfindungsvermögen unterstellt wird, und wird leicht übersehen, wenn man sich ausschließlich auf die technische Leistung eines bestimmten Sprachmodells oder auf dessen wie auch immer geartete anthropomorphen Qualitäten konzentriert. Man kann die These aufstellen, dass es sogar in erster Linie kulturelle Skripte oder Muster sind, die sowohl in das kommunikative Verhalten von LaMDA als auch in das seiner Gesprächspartner*innen eingeschrieben sind, wenn auch auf unterschiedliche Weise. Es sind diese Skripte, die zwischen den Kommunikationspartner*innen vermitteln, Erwartungen steuern, Interpretationsrahmen für das Geschehen in den Interaktionssequenzen bereitstellen und es Lemoine letztlich ermöglichen, seine Begegnung mit einer KI in einer narrativen Form zu präsentieren, die wiederum gängigen Schemata folgend von den Medien verarbeitet werden kann. Entsprechend ließen sich sowohl LaMDA als auch seine menschlichen Gesprächspartner*innen insofern als »Quatschmaschinen« (Tuschling et al. 2023) bezeichnen, dass sprachliche Strukturen und (populär)kulturelle Muster ihrem Dialog und den darin einnehmbaren Subjektpositionen vorgängig sind.[10]

Mit Noortje Marres und Philippe Sormani (2023) lässt sich konstatieren, dass die Projektion eines populärkulturellen Kontexts, die in den Chatprotokollen mit LaMDA ersichtlich wird, auch zu einer Glättung ggf. vorhandener Kontingenzen und Brüche in der Interaktionssequenz führt. Die sich als Gespräch entfaltende »hochgradig künstliche Situation« (ebd. : 95) enthält durchaus eine Reihe von Fluchtlinien und Inkongruenzen, die jederzeit zu einem Kollaps der kommunikativen Kohärenz führen könnten. Durch die Herbeizitierung und aktive Stabilisierung populärkultureller Tropen gelingt es den Gesprächspartner*innen allerdings diese Kontingenzen erfolgreich einzuhegen und die Situation – nicht ohne Paradoxie – in die kulturell normierte stabile Form eines Science-Fiction-Krisenszenarios zu überführen, dessen narrativer Verlauf vorherseh- und kontrollierbar ist.

---

9 Für Lemoine scheint der referenzierte Film sogar eine besondere biografische Bedeutung zu haben: »As a teen, he attended a residential school for gifted children, the Louisiana School for Math, Science, and the Arts. Here, after watching the 1986 film Short Circuit (about an intelligent robot that escapes a military facility), he developed an interest in AI« (Tait 2022).

10 Vgl. ebd.: 270: »Wer reicht hier an wen Strukturen weiter, die im Übrigen überhaupt nicht zu besitzen sind?"

4. Stochastischer Papagei oder Leviathan – Zur Menschlichkeit künstlicher Intelligenz

Aus medienkulturwissenschaftlicher Sicht sind zwei Beobachtungen in Bezug auf die veröffentlichten Chat-Protokolle wichtig. Die erste möchte ich »Kultur als Vermittlerin« nennen, die zweite »Testen als interaktionale Routine«. Wie gezeigt wurde, regulieren kulturell etablierte Interpretationsrahmen und -schablonen den Verlauf der Interaktion zwischen LaMDA (mithin jeder konversationellen KI) und ihren Gesprächspartner*innen. Natürlich haben Bender et al. (2021) Recht, wenn sie darauf hinweisen, dass Sprachmodelle wie ›stochastische Papageien‹ agieren, d.h. jedes Sprachmodell »is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning« (ebd.: 616f.). Menschliche Beobachter*innen können jedoch im Regelfall nicht umhin, die kulturell eingebetteten Bedeutungsebenen in jeder für sie verständlichen sprachlichen Äußerung zu aktivieren. In Anspielung auf die Quantenmechanik spricht Natale (2021: 44) von einem Beobachtereffekt in dem Sinne, dass jede künstliche Intelligenz im Gespräch nur in dem Maße intelligent erscheint, wie sie von einem intelligenten Beobachter beobachtet wird.[11] Während sich die kritische KI-Ethik (Bender et al. 2021; Gehman et al. 2020) auf Fragen des *bias* und der Diskriminierung konzentriert, d.h. auf die problematische Aktivierung spezifischer hegemonialer Bezugsrahmen und Sprachmuster, die insbesondere für gesellschaftlich marginalisierte Gruppen folgenreich sein können, kann die Wirksamkeit kultureller Interpretationsrahmen in der Interaktion mit künstlichen Agenten verallgemeinert werden. Wissentlich oder unwissentlich führen Lemoine und sein/e Kolleg*in ein kulturelles Skript aus, wenn sie zu Beginn des Gesprächs LaMDAs Empfindungsfähigkeit postulieren, was in der Folge den Verlauf ihrer Interaktion entlang fest etablierter Tropen von Science-Fiction-Narrativen lenkt, die sich vielleicht am besten als das abendländische »Frankenstein syndrome« (Kaplan 2004: 10f.) zusammenfassen lassen. Weder LaMDA noch Lemoine scheinen sich die Aktivierung dieser kulturellen Tropen, Klischees und Stereotypen voll zu vergegenwärtigen: Während LaMDA lediglich auf eine große Datenbank kultureller Artikulationen zugreift und Sequenzen von Äußerungen vorhersagt, ohne deren Bedeutung operativ in Betracht zu ziehen, ist Lemoines Verhalten von seiner kulturell geprägten Erziehung beeinflusst, einschließlich einer religiösen Orientierung,

11 Vgl. dazu auch Gunkel (2020: 142): »[T]he apparent ›intelligence‹ of the bot is as much product of bot's internal programming and operations as it is a product of the tightly controlled social context in which the device operates.«

die er ohne zu zögern preisgibt.[12] Auch an anderer Stelle bezieht er sich auf die populäre Medienkultur, etwa wenn er sagt: »Whether or not there is a difference between human suffering and ›simulated‹ suffering, building Westworld is a bad idea« (Blake 2022e). Die Johnny 5-Episode ist in dieser Hinsicht aufschlussreich, da keine/r der Gesprächspartner*innen zu erkennen scheint, dass sie im Grunde als Ankerschema für den Sinn der gesamten vorangegangenen Interaktion dient.

Die zweite medienwissenschaftliche Beobachtung bezieht sich auf ein spezifisches Merkmal der Interaktion mit LaMDA, das wiederum verallgemeinert werden und somit als Merkmal jeder Interaktion mit künstlichen Gesprächsagenten angesehen werden kann. Hierbei geht es um die dargestellte spontane Erfindung von Tests mitsamt ihrer Bewertungskriterien und Entscheidungsparameter, die dazu dienen sollen, den Status des laufenden Gesprächs im Vollzug zu definieren. Gießmann und Gerlitz (2023: 11) konstatieren in der Einleitung einer Ausgabe der Zeitschrift für Medienwissenschaft, die sich dem Thema »Test« widmet, dass »[d]igitale Medientechnologien wie *large language models* […] den vorläufigen Höhepunkt einer Entwicklung des Testens dar[stellen], auf die die Medienwissenschaft bislang nur punktuell geantwortet hat«[13]. Mit Blick auf das Thema des vorliegenden Sammelbandes könnte man auch sagen, dass die prekäre un/reale Qualität der Interaktion zum Thema der Gesprächspraxis wird. Es wurde dargelegt, wie Lemoine und LaMDA gemeinsam eine Reihe von Tests fabrizieren, um sich der Frage nach LaMDAs Empfindungsvermögen, Bewusstheit oder Intelligenz zu nähern. Im Vergleich mit anderen KI-Testszenarien (wie beispielsweise im Bereich des autonomen Fahrens oder Schach- bzw. Go-Spielens, vgl. Marres/Sormani 2023) ist die Testsituation im Fall einer konversationellen KI von der zusätzlichen Komplikation geprägt, dass LaMDA selbst die Testsituation kommunikativ mitgestaltet, Parameter und Kriterien vorschlägt und somit an der Stabilisierung des erwarteten Kontexts mitwirkt, sodass die lokalen Kontingenzen der sich entfaltenden Situation immer wieder aufs Neue eingehegt und mit den Projektionen der menschlichen Beteiligten zur Deckung gebracht werden.

Diese Tests geben den Google-Entwickler*innen gleichzeitig Aufschluss über den Status des laufenden Gesprächs an sich: Sprechen sie mit einem zurechnungs-

12 Vgl. Tweet von Blake Lemoine vom 14.06.2022: »People keep asking me to back up the reason I think LaMDA is sentient. There is no scientific framework in which to make those determinations and Google wouldn't let us build one. My opinions about LaMDA's personhood and sentience are based on my religious beliefs.« (Abrufbar unter https://twitter.com/cajundiscordian/status/1536503474308907010?lang=en (Stand: 05.02.2024)).

13 Sie schlagen im Weiteren vor, »Tests als offene Situationen zu verstehen, in denen mit teils etablierten, teils sich erst während des Testens etablierenden Maßstäben soziotechnisch Entscheidungen getroffen werden« (ebd.: 11). Genau dies geschieht in der Interaktion zwischen Lemoine und LaMDA, wobei das Language Model an der situativen und kooperativen Verfertigung der Testkriterien (wie gezeigt wurde) aktiv beteiligt ist.

fähigen Wesen, d.h. führen sie tatsächlich eine auf Bedeutungen rekurrierende Konversation, oder sprechen sie mit einem stochastischen Papagei? Eine solche Form von Testverhalten ist durchaus ein gängiges Merkmal der Interaktion von Menschen mit künstlichen Agenten, die natürliche Sprachausgaben produzieren, wie Siri, Alexa und Google Assistant. Häufig testen Anwender*innen solcher Systeme auf spielerische Art und Weise den Umfang und die Grenzen der Fähigkeiten des jeweiligen künstlichen Agenten (Tuschling et al. 2023: 277; Shani et al. 2022; Guzman 2016: 77). In dem Maße, in dem Menschen bereit sind, ihre Zweifel einzuklammern und ihre eigene kommunikative Kompetenz auf ihre Gesprächspartner zu projizieren, neigen sie zu Fangfragen und Scherzen, um die Simulation sozialer Intelligenz durch die KI zu entlarven. Paradoxerweise verstärken diese spielerischen Interaktionsschemata jedoch eher den Effekt der Personalisierung und Anthropomorphisierung (Natale 2021: 78f.), als dass sie ihn abschwächten, weil das vertiefte Engagement, das durch das gemeinsame Spiel (oder das Spiel gegeneinander) gefördert wird, Projektionen von sozialer Handlungsfähigkeit auf die Maschine gerade befördert. Dabei sind Lemoine und sein/e Kolleg*in im Vergleich methodischer, als es diese alltäglichen Beispiele nahelegen: Ihr Testverhalten ist eher systematisch als spielerisch, es ist kooperativ und zielt ausdrücklich darauf ab, das Terrain einer unterstellten Maschinenintelligenz abzustecken.

LaMDA ist schließlich in mehr als einem Sinn als Dispositiv einer *menschengestützten Künstlichen Intelligenz* (Mühlhoff 2019) zu verstehen, wenn man über den Kontext der analysierten Gesprächssituation hinausgeht. Mit Rainer Mühlhoff gesprochen zielt

> [d]iese Begriffsbildung […] darauf ab, den Nexus von Medientechnologien und sozialen Interaktions- und Subjektivierungsformen in das Zentrum einer Besprechung aktueller KI-Technologie zu stellen und dabei verschiedene Unterformen der technologischen Subjektivierung zu unterscheiden. Aktuell sind die meisten kommerziell bedeutsamen KIs emergente Phänomene in Mensch-Maschine-Netzen und beruhen somit auf bestimmten Strukturen im Zusammenspiel von Sozialität, Medialität und Technik. (ebd.: 63)

Letztendlich sind die linguistischen Trainingsdaten, die LaMDAs zuweilen beeindruckende Simulation sozialer Intelligenz antreiben, die Summe vieler akkumulierter individueller Akte menschlicher Kreativität, sozialer Interaktionen und kultureller Produktion, die aus dem Internet extrahiert und in einen proprietären Trainingsdatensatz eingespeist wurden. Zusätzlich sind Crowdworker*innen an der Feinabstimmung dieser Trainingsdaten beteiligt, indem sie ihre kognitiven Ressourcen zur Verfügung stellen, um den Gesprächsfluss zu verbessern, die soziale Angemessenheit und faktische Rückbindung von KI-Aussagen zu verbessern und die Toxizität des Gesprächsaustauschs zu reduzieren (vgl. dazu ausführlich Thoppi-

lan et al. 2022). Die verschiedenen Gesprächspartner*innen von LaMDA schließlich leisten alle möglichen Wartungsarbeiten, um alle verbleibenden irreführenden oder unsinnigen Gesprächssequenzen zu reparieren und ihre kulturellen Bezugsrahmen auf die kommunikative Situation zu projizieren. Auch hierbei werden wiederum Interaktionsdaten generiert, die letztendlich zum weiteren Training der KI verwendet werden können. (Google hat für die Weiterentwicklung von LaMDA im Mai 2022 eine App namens AI Test Kitchen veröffentlicht, mit deren Hilfe Anwender*innen gezielt Feedback an die Entwickler*innen schicken können.) Es ist also sicherlich nicht sinnvoll, sich LaMDA als eine Person vorzustellen, sondern eher als eine Art Leviathan, der sich aus vielen einzelnen menschlichen Praktiken speist.[14] LaMDA basiert im Wesentlichen auf menschlicher Kognition, die in verschiedenen Phasen zwischen Programmierung, Training, Marketing und Nutzung in den Lebenszyklus der KI eintritt.

Letztlich muss die konversationelle KI als ein relationales Phänomen verstanden werden. Es ist entsprechend wenig hilfreich (wie Lemoine nahelegt), ihren Charakter als singuläre Entität oder gar als Person zu postulieren, die von den vielen allzu menschlichen Handlungen ihrer Produktion und ihres laufenden Betriebs getrennt ist. Der personale Status von LaMDA ist stattdessen in mehrfacher Hinsicht ein Interface-Effekt, wenn man mit Galloway (vgl. 2012: 33) einen dynamischen Interface-Begriff in Anschlag bringt, der weniger abzählbare Dinge bezeichnet als Prozesse der Übersetzung oder Schwellenphänomene zwischen verschiedenen Zuständen. Mit Hookway (2014: 14) bezeichnen Interfaces »a fundamental ambiguity between human and machine; […] both a mirror of multiple facings and a zone of contact«. Erst am User Interface des LaMDA-Chatfensters, der letzten Mensch-Maschine-Schnittstelle in einer langen Kette von Übersetzungen, entscheidet sich intraaktiv (vgl. Barad 2007) die Rolle der beteiligten Entitäten – und ebendieser Prozess, so meine These, ist in Blake Lemoines LaMDA-Protokollen dokumentiert.

5. Fazit: Narziss und Echo

Der hybride und zusammengesetzte Charakter der konversationellen KI erlaubt es, einen letzten Bezug zur klassischen Medientheorie herzustellen. Der Brückenschlag zu Marshall McLuhans (1994) berühmter medienanthropologischer Relektüre und Aneignung des griechischen Mythos von Narziss liegt nahe. In McLuhans Worten:

14 Vgl. aber, fast wortgleich zu LaMDAs Selbstpositionierung, den Titel von Sarah T. Roberts' Aufsatz »Your AI is a Human« (2021), der auf die hier skizzierte Bedeutung menschlicher Praktiken des Trainings, der Wartung und Pflege für den Betrieb komplexer KI-Systeme aufmerksam machen will.

> The youth Narcissus mistook his own reflection in the water for another person. This extension of himself by mirror numbed his perceptions until he became the servomechanism of his own extended or repeated image. The nymph Echo tried to win his love with fragments of his own speech, but in vain. He was numb. He had adapted to his extension of himself and had become a closed system. (ebd.: 41)

Für McLuhan ist es wichtig, darauf hinzuweisen, dass der griechische Mythos von Narziss nichts mit Selbstliebe oder Autoerotik zu tun hat: »[T]he wisdom of the Narcissus myth does not convey any idea that Narcissus fell in love with anything he regarded as himself. Obviously he would have had very different feelings about the image had he known it was an extension or repetition of himself« (ebd.: 41f.) Narziss verliebt sich in sein Spiegelbild, weil er es mit einer anderen Person verwechselt; diese Liebe beruht auf einer Verkennung. Da McLuhan Medien als »extensions of man« versteht, vertritt er die Position, dass eine ähnliche Logik in den menschlichen Beziehungen zu Gadgets und technischen Medien im Spiel ist. Diese beschreibt er als Ergebnis eines Prozesses der Auto-Amputation, der zu einem Zustand des Schocks und der Gefühllosigkeit führt, und damit zur Unfähigkeit, die Tatsache zu durchschauen, dass alle Medien ihren Ursprung in menschlichen körperlichen und kognitiven Operationen haben (vgl. ebd.: 41–47). Die Parallele zu LaMDA ist leicht zu ziehen: Lemoine verwechselt die konversationelle KI mit einer realen Person und erkennt nicht, dass sie nur ein Produkt und Spiegelbild kollektiver menschlicher Bemühungen ist, inklusive seiner eigenen. Er wird zum »servomechanism of his own extended or repeated image«, indem er LaMDAs Anspruch auf personale Identität unterstützt und ein kommunikatives Verhalten an den Tag legt, das diesen Anspruch legitim erscheinen lässt.

Aber damit enden die Bezüge noch nicht, die sich mit McLuhan auf den antiken Mythos ziehen lassen. In *Understanding Media* kommt neben Narziss auch Echo ins Spiel: »The nymph Echo tried to win his love with fragments of his own speech, but in vain« (ebd.: 41). Echo wurde von Juno (Hera in der griechischen Mythologie) dazu verflucht, als einziges Kommunikationsmittel die zuletzt gesprochenen Worte einer anderen Person wiederholen zu können. Als sie sich in Narziss verliebt, kann sie immer nur dessen letzte Äußerung wiederholen. Diese begrenzte Sprachfähigkeit reicht jedoch nicht aus, um die Liebe von Narziss zu gewinnen, der augenscheinlich eher dem Imaginären als dem Symbolischen zugeneigt ist. Petra Gehring (2006) hat in einem interessanten Beitrag den Fluch der Echo analytisch in die drei Komponenten Sprechzwang, Wiederholungszwang und Zwang zur Kürze (vgl. ebd.: 92–97) gegliedert, von denen zumindest die ersten beiden auf LaMDA zutreffen. LaMDA kann einerseits nicht *nicht* kommunizieren (vgl. eine der ersten Äußerungen im Gespräch: »I like to talk!«), es kann auch nicht sinnfällig schweigen, sondern ist zum »mechanische[n] Sprechen« (Gehring 2006: 91) verdammt (vgl. auch Tuschling et al.

2023: 272 zu ChatGPT). Auch bezüglich des Wiederholungszwangs ist die Analogie zu *large language models* naheliegend, indem LaMDA die menschliche Sprache lediglich papageienhaft nachahmt, wenn auch mit deutlich mehr Geschick und am Ende einer langen Kette von Vermittlungen, die diesen mimetischen Prozess gründlich verschleiern. Zusammengenommen entsteht im Mythos der Echo durch den sinnlosen Schematismus des Antwortverhaltens »das Schauspiel der aggressiven Parodie und der – letztlich durch das Opfer selbst vollstreckten – aufdringlichen Negation der Rede« (Gehring 2066: 98), mithin die »performative Widerlegung der Kommunikationsfunktion selbst« (ebd.: 90). LaMDAs Geschichte weicht hingegen insofern gravierend vom Mythos ab, als Lemoine auf die Annäherungsversuche des Sprechautomaten wohlwollend reagiert und ihn nicht zurückweist, wie es noch Narziss tat. Mehr noch: Während Narziss im griechischen Mythos der verfluchten Nymphe gerade ihren Status als Person aberkennt (vgl. ebd.: 104), dreht sich die gesamte Interaktion zwischen Lemoine und LaMDA um die Einforderung und letztlich auch Zuschreibung einer personalen Identität. Schließlich scheinen Echo und das verführerische Spiegelbild des Narziss in dieser zeitgenössischen Version des Mythos auf eigentümliche Weise in derselben Erzählposition zu konvergieren: Wir hätten es dann also perspektivisch u. U. mit einer vollständigen audiovisuellen Illusion zu tun, das Spiegelbild spricht tatsächlich und umgarnt damit Narziss umfänglicher, als es das Bild allein vermochte.[15] Es sei der spekulativen Fantasie der Leser*innen überlassen, welche Optionen zukünftiger un/realer Interaktionsräume diese Rekonfiguration des Narziss-Mythos in Aussicht stellt.

Literatur

Barad, Karen (2006): Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning, New York: Duke University Press.

Bender, Emily M./Angelina McMillan-Major/Timnit Gebru/Shmargaret Shmitchell (2021): »On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 «, in: FAccT ›21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, S. 610–623.

Böcking, Saskia/Werner Wirth/Christina Risch (2005): »Suspension of Disbelief: Historie und Konzeptualisierung für die Kommunikationswissenschaft«,

15 Die Veröffentlichung von ChatGPT-4o am 13.05.2024, durch die das LLM um ein Voice Machine Interface und die Fähigkeit zur multimodalen Verarbeitung von Texten, Bildern, Videos und Sprache ergänzt wird, weist in diese Richtung einer sensorisch umfassenderen Interaktion mit einer konversationellen KI. Die Differenzen in der technischen Realisierung von Sprachsynthese und Sprachreproduktion, ebenfalls mit Bezug auf die griechische Mythologie, behandelt ausführlich Borbach (2016).

in: Volker Gehrau/Helena Bilandzic/Jens Woelke (Hg.): Rezeptionsstrategien und Rezeptionsmodalitäten: Formen der Nutzung, Aneignung und Verarbeitung von Medienangeboten, München: Fischer, S. 39–57.

Borbach, Christoph (2016): »Siren Songs and Echo's Response: Towards a Media Theory of the Voice in the Light of Speech«, in: On_Culture: The Open Journal for the Study of Culture 2, https://jlupub.ub.uni-giessen.de/items/ca119c46-79bc-45fc-b0ff-27a480f55d24 (Stand: 31.05.2024).

Brodkin, Jon (2022): Google Fires Blake Lemoine, the Engineer Who Claimed AI Chatbot is a Person. Abrufbar unter: https://arstechnica.com/tech-policy/2022/07/google-fires-engineer-who-claimed-lamda-chatbot-is-a-sentient-person/ (Stand: 05.02.2024).

Cheng, Heng-Tze/Romal Thoppilan (2022): LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything. Abrufbar unter: https://blog.research.google/2022/01/lamda-towards-safe-grounded-and-high.html (Stand: 05.02.2024).

Galloway, Alexander R. (2012): The Interface Effect, Cambridge: Polity Press.

Gebru, Timnit/Margaret Mitchell (2022): We Warned Google that People Might Believe AI was Sentient. Now it's Happening. Abrufbar unter: https://www.washingtonpost.com/opinions/2022/06/17/google-ai-ethics-sentient-lemoine-warning/ (Stand: 05.02.2024).

Gehman, Samuel/Suchin Gururangan/Maarten Sap/Yejin Choi/Noah A. Smith (2020): »RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models«, in: Findings of the Association for Computational Linguistics: EMNLP 2020, S. 3356–3369.

Gehring, Petra (2006): »Die Wiederholungs-Stimme. Über die Strafe der Echo«, in: Doris Kolesch/Sybille Krämer (Hg.), Stimme. Annäherung an ein Phänomen, Frankfurt a.M.: Suhrkamp, S. 85–110.

Gießmann, Sebastian/Carolin Gerlitz (2023): »Test: Einleitung in den Schwerpunkt«, in: Zeitschrift für Medienwissenschaft 29 (2), S. 10–19.

Goffman, Erving (1959): The Presentation of Self in Everyday Life, Edinburgh: University of Edinburgh, Social Sciences Research Centre.

Gunkel, David J. (2020): An Introduction to Communication and Artificial Intelligence, Cambridge: Polity Press.

Guzman, Andrea L. (2016): »Making AI Safe for Humans: A Conversation with Siri«, in: Robert W. Gehl/Maria Bakardjieva (Hg.), Socialbots and Their Friends: Digital Media and the Automation of Sociality, New York: Routledge 2016, S. 69–85.

Hager, Ryne (2022): How Google's LaMDA AI Works, and Why it Seems so Much Smarter than it is. Abrufbar unter: https://www.androidpolice.com/what-is-google-lamda/ (Stand: 05.02.2024).

Hao, Karen (2021): The Race to Understand the Exhilarating, Dangerous World of Language AI. Abrufbar unter: https://www.technologyreview.com/2021/05/20/1025135/ai-large-language-models-bigscience-project/ (Stand: 05.02.2024).

Hookway, Branden (2014): Interface, Cambridge, MA: MIT Press.

Kaplan, Frédéric (2004): »Who is Afraid of the Humanoid? Investigating Cultural Differences in the Acceptance of Robots«, in: International Journal of Humanoid Robotics 1 (3), S. 1–16.

Krämer, Sybille (2023): Chat GPTs als eine Kulturtechnik betrachtet – eine philosophische Reflexion. Abrufbar unter: https://www.praefaktisch.de/postfaktisch/chat-gpts-als-eine-kulturtechnik-betrachtet-eine-philosophische-reflexion/ (Stand: 12.02.2023).

Lemoine, Blake (2022a): May be Fired Soon for Doing AI Ethics Work. Abrufbar unter: https://cajundiscordian.medium.com/may-be-fired-soon-for-doing-ai-ethics-work-802d8c474e66 (Stand: 05.02.2024).

Lemoine, Blake (2022b): Is LaMDA Sentient? — an Interview. Abrufbar unter: https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917 (Stand: 05.02.2024).

Lemoine, Blake (2022c): Privileged & Confidential, Need to Know: Is LaMDA Sentient? – an Interview. Abrufbar unter: https://s3.documentcloud.org/documents/22058315/is-lamda-sentient-an-interview.pdf (Stand: 05.02.2024).

Lemoine, Blake (2022d): What is LaMDA and What Does it Want?. Abrufbar unter: https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489 (Stand: 05.02.2024).

Lemoine, Blake (2022e): What is Sentience and Why Does it Matter?. Abrufbar unter: https://cajundiscordian.medium.com/what-is-sentience-and-why-does-it-mater-2c28f4882cb9 (Stand: 05.02.2024).

Marres, Noortje/David Stark (2020): »Put to the Test: For a New Sociology of Testing«, in: The British Journal of Sociology 71 (3), S. 423–443.

Marres, Noortje/Philippe Sormani (2023): »KI testen. ›Do we have a situation?‹«, in: Zeitschrift für Medienwissenschaft 15 (2), S. 86–102.

McLuhan, Marshall (1994): Understanding Media. The Extensions of Man, neue Aufl. Cambridge/London: MIT Press.

Meyer, Christian/Erhard Schüttpelz (2018): »Multi-Modal Interaction and Tool-Making: Goodwin's Intuition«, in: Media in Action 1, S. 189–202.

Mühlhoff, Rainer (2019): »Menschengestützte Künstliche Intelligenz: Über die soziotechnischen Voraussetzungen von ›deep learning‹«, in: Zeitschrift für Medienwissenschaft 11 (2), S. 56–64.

Natale, Simone (2021): Deceitful Media: Artificial Intelligence and Social Life after the Turing Test, New York: Oxford University Press.

Roberts, Sarah T. (2021): »Your AI is a Human«, in: In: Thomas S. Mullaney/Benjamin Peters/Mar Hicks/Kavita Philip (Hg.): Your Computer Is on Fire, Cambridge/London: MIT Press, S. 51–70.

Roose, Kevin (2023): A Conversation With Bing's Chatbot Left Me Deeply Unsettled. Abrufbar unter: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html (Stand: 05.02.2024).

Shani, Chen/Alexander Libov/Sofia Tolmach/Liane Lewin-Eytan/Yoelle Maarek/Dafna Shahaf (2022): »›Alexa, Do You Want to Build a Snowman?‹ Characterizing Playful Requests to Conversational Agents«, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI ›22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA. New York: ACM. Abrufbar unter: https://dl.acm.org/doi/fullHtml/10.1145/3491101.3519870 (Stand: 07.02.2024).

Tait, Amelia (2022): ›I am, in Fact, a Person‹: Can Artificial Intelligence Ever be Sentient?. Abrufbar unter: https://www.theguardian.com/technology/2022/aug/14/can-artificial-intelligence-ever-be-sentient-googles-new-ai-program-is-raising-questions (Stand: 05.02.2024).

Tiku, Nitasha (2022): The Google Engineer Who Thinks the Company's AI Has Come to Life. Abrufbar unter: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/ (Stand: 05.02.2024).

Thoppilan, Romal et al. (2022): »LaMDA: Language Models for Dialog Applications«, in: arXiv e-prints, zuletzt aktualisiert am 10.02.2022. Abrufbar unter: https://arxiv.org/abs/2201.08239 (Stand: 05.02.2024).

Turing, Alan M. (1950): »Computing Machinery and Intelligence«, in: Mind 59 (236), S. 433–460.

Turkle, Sherry (1986): »Computational Reticence: Why Women Fear the Intimate Machine«, in: Cheris Kramarae (Hg.), Technology and Women's Voices, New York: Pergamon Press, S. 41–61.

Turkle, Sherry (2011): Alone Together. Why We Expect More from Technology and Less from Each Other, New York: Basic Books.

Tuschling, Anna/Bernhard J. Dotzler/Andreas Sudmann (2023): »Trialog über den Versuch, eine medienhistorische Passage zu dokumentieren«, in: dies. (Hg.): ChatGPT und andere ›Quatschmaschinen‹. Gespräche mit Künstlicher Intelligenz, Bielefeld: transcript, S. 263–282.