Lukas Nehlsen

Chatten mit Nirgendwo?

Der Leib als Ausdruck und die Sprache der KI[1]

Eines der frühesten Anwendungsgebiete und weiterhin so etwas wie der heilige Gral der KI-Forschung ist die Konversation einer KI mit einem Menschen. Alan Turings disziplingebender Essay *Computing Machinery and Intelligence* hatte die Idee propagiert, dass wir dann davon sprechen sollten, dass Maschinen denken können, wenn Sie das »imitation game« bestehen. Wenn Sie also in einer verblindeten, textbasierten Kommunikation für eine Interviewerin[2] genauso (un)unterscheidbar von einem Menschen sind wie Männer von Frauen, wenn diese versuchen, die Interviewerin in die Irre zu führen.[3] Dieses Spiel, später als Turing-Test bekannt geworden, basierte also auf dem Versuch der Täuschung. Maschinen sollten als denkend angesehen werden, wenn sie einen Menschen so gut täuschen können, dass dieser sie oft genug als Mensch und den echten Menschen als Maschine ansieht (Turing 1950).

Auf Grundlage dieser Idee entwickelte sich eine breite Diskussion in der *Philosophie der KI*. Die Kritik an Turings Äquivalenz-These lässt sich dabei in zwei Lager aufteilen (vgl. McCarthy 2007, 1180). Das erste Lager argumentiert, dass wir eine KI selbst dann nicht als wirklich intelligent ansehen sollten, wenn sie den Turing-Test besteht, beziehungsweise wenn ihre intelligente Aktivität von der des Menschen ununterscheidbar geworden ist (sogenannte »human level performance«). Das klassische Argument für diese Art des Angriffs wurde von John Searle in seinem berühmten »Chinese Room Argument« (CRA) formuliert (Searle 1980). Das zweite Lager argumentiert hingegen, dass »human level performance« unmöglich ist. Klassisch dafür steht Hubert Dreyfus' Angriff auf das Programm der KI-Forschung, hauptsächlich formuliert in den 70er Jahren in *What Computers Can't Do* (Dreyfus 1992). Die zwei Positionen müssen sich nicht gegenseitig ausschließen, da sie

1 Ich danke Thiemo Breyer, Erik Dzwiza-Ohlsen, Christian Grüny und Martin Schnell sowie den Mitgliedern des *SchauflerLab@TUDresden* und der Klasse 6 der *a.r.t.e.s graduate school for the humanities cologne* für hilfreiche Gespräche und Feedback zu diesen Überlegungen.

2 Aus Gründen der Leserlichkeit verwende ich in diesem Text das generische Femininum, es sind darin alle Geschlechter mitgemeint.

3 Zu der Parallelisierung von weiblicher und maschineller Alterität in der Geschichte der Maschinen und auch in Turings Test siehe die Bemerkungen von Oliver Müller in diesem Band.

aus verschiedenen Richtungen argumentieren. Der Kern von Searles Argument ist, dass die entscheidenden Charakteristika von wirklicher Intelligenz, Bewusstsein, Verständnis etc. kausale Effekte unseres Gehirns sind. Künstlich hergestellte KIs seien einfach aus dem »wrong kind of stuff« (Searle 1980, 423), um diese kausale Wirkung entfalten zu können. Es gibt also in Searles Augen dem Neuroprotein exklusiv inhärente kausale Kräfte, die Bewusstsein bewirken können. Dreyfus argumentiert hingegen von einer leibphänomenologisch inspirierten Warte aus. Die menschlichen, intelligenten Fähigkeiten seien nicht in Gänze formalisierbar, da sie wesentlich in prä-reflexiven, nicht-begrifflichen, leiblichen Verhaltensweisen zur Welt bestünden. Beide könnten also das Argument des jeweils anderen ebenfalls annehmen und kombiniert sagen: Es gibt gute Gründe anzunehmen, dass *human level AI* unmöglich ist, falls sie aber doch möglich sein sollte, gibt es ebenso gute, von den ersteren verschiedene Gründe anzunehmen, dass sie immer noch nicht wirklich intelligent, bewusst oder verständig ist.

Dieser Beitrag will einen anderen phänomenologischen Ansatz aufzeigen. Die Frage nach der Möglichkeit der Täuschung, die von Turing ausgehend dominierend war, soll ausgeblendet werden. Es soll auch nicht einfach gefragt werden, was Computer prinzipiell können und was nicht. Im heutigen Alltag begegnen wir dauernd KIs, die sprechen bzw. Text produzieren können. In den meisten Fällen ist uns dabei bewusst, dass es sich hier um Maschinen handelt. Nun gibt es die Tendenz, trotz dieses Wissens den Äußerungen der Maschinen Bedeutung beizumessen. Dieses Phänomen soll in diesem Beitrag in den Blick genommen werden. Es ist in der Literatur als *ELIZA-effect* bekannt. Ich will also sprachphänomenologisch beleuchten, was passiert, wenn wir mit einer Maschine Konversation betreiben und, trotz des Wissens darum, dass es »nur« eine Maschine ist, ihren Äußerungen Bedeutung zuschreiben. Hierzu werde ich mich vor allem auf die Philosophie Maurice Merleau-Pontys beziehen. Merleau-Pontys Phänomenologie der Sprache ist in der Forschung zu seiner Philosophie als eigenständiges Thema eher stiefmütterlich behandelt worden.[4] Mit Fragen der Philosophie der KI ist seine sprachphilosophische Position, soweit ich dies überblicke, bisher gar nicht in Verbindung gebracht worden. Wenn Merleau-Pontys Philosophie in der KI-Debatte angebracht wird, dann meist, um auf die Wichtigkeit eines Körpers oder Leibs der KI hinzuweisen.[5] Ich will hier aber, ganz im Sinne des Bandes, die Begegnung mit Künstlicher Intelligenz in den Vordergrund stellen,

4 Vgl. aber den schon sehr frühen Artikel von (Lewis 1966) und für einen aktuellen Beitrag (Apostopoulos 2019).

5 So klassisch bei (Dreyfus 1992) Vgl. weiter (Zebrowski 2010) und (Pressman 2017), der mit Merleau-Pontys Ansichten Alex Garlands Film *Ex Machina* analysiert.

und zwar insbesondere die mit einer *sprechenden KI*. Es soll hier also ausprobiert werden, inwieweit Merleau-Pontys sprachphänomenologische Beschreibungen des Phänomens auch für eine Sprache und ein Sprechen der KI gelten können.

Im *ersten Teil* stelle ich den *ELIZA-effect* näher vor und gebe eine kurze Beschreibung eines zeitgenössischen Programms, das diesen Effekt hervorruft. Ich zeige einige sprachphilosophische Probleme auf, die sich aus einer Definition des Effekts, wie wir sie bei Douglas Hofstadter finden, ergeben. Im *zweiten Teil* argumentiere ich, dass sich ähnliche Probleme schon in Merleau-Pontys Ausführungen über die Sprache in der *Phänomenologie der Wahrnehmung* finden. Ich stelle die beiden dort in Betracht gezogenen Theorien zur Klärung der Probleme vor und zeige, warum Merleau-Ponty diese als unzureichend ansieht. Im *dritten Teil* stelle ich dann Merleau-Pontys eigene sprachphilosophische Position vor. Ich analysiere seinen Begriff der gestischen Bedeutung der Sprache und dessen Beziehung zum Leib. Im *vierten Teil* wird diese Theorie auf das Phänomen des Gesprächs mit einer KI angewandt. Die fehlende Leiblichkeit einer Chatbot-KI wird analysiert. Sie führt zu einem Phänomen des Widerstreits in der Wahrnehmung von Bedeutung, aber nicht zu ihrer Auslöschung. Im *letzten Teil* ziehe ich einige Konsequenzen aus der vorangegangenen Analyse für die Diskussion um Searles CRA.

1. Der ELIZA-effect und seine Erklärung

Der *ELIZA-effect* ist benannt nach Joseph Weizenbaums Programm *ELIZA* von 1966. Das Programm war ein früher Vertreter der Art von KI, die später von John Haugeland als g*ood old fashioned AI,* kurz *GOFAI,* bezeichnet wurde (Haugeland 1985). GOFAI-Systeme basieren auf festen Abfolgen von Schritten, in denen diskrete Symbole nach logischen Regeln aufeinander bezogen werden. ELIZA war ein Chatbotprogramm, das von Weizenbaum so eingestellt wurde, dass es eine an Carl Rogers' Ansatz orientierte Psychotherapeutin nachahmen sollte. Die grobe Funktionsweise war die Folgende: Das Programm suchte im eingegebenen Text nach Schlüsselwörtern, welche mit bestimmten Regeln verknüpft sind. Zunächst mit einer Regel, wie der eingegebene Text aufzulösen sei (*decomposition rule*) und dann mit einer Regel, wie daraus neuer Text zu generieren sei, also die Antwort des Programms aussehen solle (*reassembly rule*). Das Programm hatte durchschlagenden und für Weizenbaum selbst schockierenden Erfolg. Viele Testpersonen identifizierten das Programm mit einem Menschen, vertrautem ihm intimste Geheimnisse an und gebärdeten sich wie in einem Gespräch mit einer menschlichen Psychotherapeutin. Einige der Testpersonen waren sogar

unwillig zu glauben, dass es sich nicht um einen Menschen handelte, nachdem dies aufgelöst worden war (Weizenbaum 1966, 42).

Basierend auf diesen Erfahrungen wurde der Begriff *ELIZA-effect* geprägt. Douglas Hofstadter hat eine recht präzise Definition des Effekts gegeben, von der ich im weiteren Verlauf zunächst ausgehen werde. Demnach könne der *ELIZA-effect* definiert werden als,

> »the susceptibility of people to read far more understanding than is warranted into strings of symbols – especially words – strung together by computers« (Hofstadter 1995, 157).

Als ein einfaches Beispiel nennt er, fälschlicherweise anzunehmen, dass ein Geldautomat wirklich dankbar (»grateful«) sei, wenn am Ende der Transaktion ein »Vielen Dank« aufleuchte. Diese Illusion, so Hofstadter, sei natürlich schnell auflösbar. Beinahe jeder ist in der Lage, zu erkennen, dass es recht einfach möglich sei, eine Zwei-Wort Phrase im richtigen Moment auftauchen zu lassen. Und zwar genauso mechanisch, also ohne jede Notwendigkeit von Verständnis, wie eine Tür sich automatisch öffnen kann, wenn wir uns ihr nähern (ebd.). »But when things get only slightly more complicated, people get far more confused – and very rapidly, too« (ebd.). Am relativ simplen Beispiel von ELIZA lässt sich außerdem festhalten, dass die Tendenz von Menschen, den von Computern generierten Worten Bedeutung zuzumessen, auch bestehen bleibt, wenn wir wissen, dass es sich um einen Computer handelt und eventuell sogar eine Ahnung davon haben, wie der Computer die Worte in Wirklichkeit produziert[6]. Weizenbaums Hoffnung mit seinem Paper zu ELIZA war es, durch Aufklärung der Öffentlichkeit über die Funktionsweise, dem Programm seine mystische Aura zu nehmen. Einem zutiefst neuzeitlichen Verständnis der Erkenntnis anhängend, in dem etwas verstehen bedeutet, es konstruieren zu können, geht Weizenbaum davon aus, dass: »once a particular program is unmasked, once its inner workings are explained in language sufficiently plain to induce understanding, its magic crumbles away; it stands revealed as a mere collection of procedures, each quite comprehensible.« (Weizenbaum 1966, 36).[7]

Inwieweit diese Hoffnung erfüllt wird und wurde, ist fraglich.

Erstens ist bei neueren Programmen unklar, inwieweit wir zu einem exakten Verständnis ihrer Funktionsweise überhaupt noch in der Lage sind. Das 2020 veröffentlichte Programm GPT-3 ist ein klassischer Vertreter der in Abgrenzung zur GOFAI sogenannten *second-wave AI*. Es ist ein *deep neural network* (DNN) bestehend aus 175 Milliarden

6 Im Beispiel des Geldautomaten wissen wir dies natürlich auch. »Wissen« ist hier aber beispielsweise im Gegensatz zum nur wahrscheinlichen »Vermuten« zu sehen, wie es z.B. beim Turing-Test der Fall ist.

7 Diese Idee steht im Gegensatz zu Terry Pratchet's Idee, auf die Harth und Feißt in diesem Band verweisen. »It's still magic, even if you know how it is done« (Harth/Feißt 70)

Parametern, das mit einem riesigen Datensatz darauf trainiert wurde, sprachliche Strukturen zu erkennen und neu erstellen zu können. Es ist also im Gegensatz zu ELIZA ein dezentrales, von der Hirnarchitektur inspiriertes Netzwerk, dass Bedeutungseinheiten nicht mehr in diskreten Symbolen repräsentiert, sondern durch im Netz verteilte Strukturen. Außerdem arbeitet es mit probabilistischen Methoden, im Gegensatz zu der logischen Methodik ELIZAs (vgl. Brown u. a. 2020)[8]. Bei DNNs ist es meist nicht möglich nachzuvollziehen, wie der Input zu Output verarbeitet wurde. Das wird in der KI-Forschung als *black box problem* diskutiert (vgl. z.B. Carabantes 2020; von Eschenbach 2021; vgl. Chen, Bei, und Rudin 2020 für einen aktuellen Vorschlag mit dem Problem auf architektonischer Ebene der DNNs umzugehen). Carabantes diskutiert drei Gründe für die Undurchsichtigkeit der Systeme:

1. Die Unverfügbarkeit des Quellcodes für die breite Masse. Dies ist beispielsweise auch bei GPT-3 der Fall, da Open AI in zynischer Missachtung ihres Namens die Rechte am Code exklusiv an Microsoft verkauft hat (Daws 2020).
2. Undurchsichtigkeit aufgrund von »technological illiteracy« (Carabantes 2020, 312). Der Großteil der Bevölkerung hat nicht die notwendige Bildung, um den Code nachzuvollziehen, selbst wenn dies prinzipiell möglich und er frei verfügbar wäre. Weizenbaums Erklärung von ELIZA sind für interessierte Laien noch recht gut nachvollziehbar, die genaue Funktionsweise moderner neuronaler Netze sicherlich nicht mehr.[9]
3. Undurchsichtigkeit aufgrund von »cognitive mismatch« (ebd. 313). Neuronale Netze funktionieren gerade deshalb so gut in vielen Feldern, wie zum Beispiel Bild- oder Spracherkennung, da sie dies gerade nicht in rationaler, diskreter Weise angehen und der Prozess damit auch nicht intelligibel ist:

> If its process were intelligible, says Burrell, it would proceed by decomposing the main task into other simpler subtasks intelligible to a human being […] But it does not. It works similar to our visual cortex: in a way that even we cannot consciously explain. And when we try to do it, as it happens to patients with agnosia, our performance is clumsy. ANNs are especially good at this: solving problems for which we do not have an algorithm or a well-defined logical sequence of cognitive actions that produces the result we intend. ANNs are better at recognizing images than any symbolic program, that is, intelligible. (ebd. 314).

8 Für eine ausführlichere Beschreibung des Programms vgl. auch hier den Beitrag von Harth/Feißt

9 Auch Weizenbaum selbst warnte 1976 schon davor, dass einige Computerprogramm so komplex sind, dass kein Einzelner der involvierten Ingenieure das Programm zur Gänze verstehe (Weizenbaum 1990).

ANNs (*artificial neural networks*[10]) sind im Gegensatz zur GOFAI – die von einem symbolisch operierenden Geist ausging, der unabhängig vom Hirn gedacht und modelliert werden könnte – von einer biologischen Struktur, unserem Hirn, inspiriert. Damit stehen ihre Operationen außerhalb dessen, was man mit Willfried Sellars als den »logischen Raum der Gründe« bezeichnen könnte (vgl. Sellars 1997). In Carabantes' Worten: »Observing the huge matrix of weights of a DNN and the activation levels of the units in real time, Ortega y Gasset would say, may give an esthetic impression, but it does not explain how the computer reaches its conclusions« (Carabantes 2020, 314). Das Mysterium durch Erklärung aufzulösen, wie Joseph Weizenbaum das noch bei EILZA erhofft hatte, ist also nicht mehr so einfach möglich. Ohne die Möglichkeit einer solchen vollständigen Erklärung ist es auch nicht mehr so einfach, mit Hofstadter zu behaupten, das Verständnis, das wir der KI zuschreiben sei »unwarranted«. Margaret Boden hatte eine ähnliche Kritik an Searles CRA angebracht: Wir hätten zwar sehr gute Gründe für die Annahme, *dass* Intentionalität, Verständnis und Bewusstsein auf Neuroprotein basierten, allerdings hätten wir noch so gut wie keine Ahnung davon, *wie* diese Beziehung aussieht, also *wie* Neuroprotein diese Phänomene qua seiner Eigenschaften hervorrufe. Daher könne man auch nicht so einfach darauf schließen, dass Neuroprotein der einzige »stuff« sei, der diese kausalen Kräfte habe (Boden 1990, 93). Wenn wir also bei einem DNN in etwa eine genauso gute Erklärung dafür haben, wie es seine Outputs hervorruft wie beim menschlichen Gehirn, dann haben wir zunächst keinen anderen Grund als einen Unterschied in diesen Outputs, um anzunehmen, dass im einen Fall »wirkliches« Verständnis, Bewusstsein oder Intentionalität dahinter stehe und im anderen nicht.

Ein *zweiter* Grund zum Zweifel an Weizenbaums Hoffnung der Auflösung des Mysteriums durch Erklärung betrifft das Verhältnis von Wissen und Wahrnehmung. In vielen Fällen ist es gerade nicht möglich, dass abstraktes Wissen darum, wie ein Sachverhalt »in Wirklichkeit ist«, unsere Wahrnehmung dieses Sachverhalts zu verändern vermag. Das klassische Beispiel ist die Müller-Lyer-Illusion. Auch wenn wir *wissen*, dass die Linien »in Wirklichkeit« gleich lang sind, *sehen* wir sie weiterhin als verschieden lang. Besteht der ELIZA-effect nicht mindestens für einige und eventuell für alle in gewisser Form fort, *trotz* unseres Wissens um die Maschinenhaftigkeit und des graduellen Wissens um die Funktionsweise

10 Ein Überbegriff für die dezentral operierenden, der Netzwerkarchitektur des Hirns nachempfundenen *second wave AI* Programme. DNNs sind ANNs mit »Tiefe«, d.h. zwischen Input-»Neuronen« und Output-»Neuronen« liegt mindestens eine weitere »Neuronen«-schicht. Buckner (2019) sieht alle Netze mit nicht mehr als vier Zwischenschichten noch als »shallow« an, und sieht wirkliche Tiefe erst bei fünf Zwischenschichten beginnen. Er betont aber auch, dass die Epoche der DNNs im Prinzip mit der Einführung des ersten »hidden layers« zwischen input- und output-Schicht in den 80ern beginnt (ebd. 3).

dieser Maschine? Es ist sicherlich wahr, dass die Magie von ELIZA zu einem gewissen Grad schwindet, nachdem wir verstanden haben, wie sie funktioniert. Trotzdem können wir erstmal nicht anders, als die Worte der Maschine als bedeutungsvoll wahrzunehmen. Bedeutung, so mein auf Merleau-Ponty aufbauender Vorschlag, ist etwas, das wir *wahrnehmen* und zunächst und zumeist nicht reflexiv *denken*. Die Entscheidung in dieser Frage für die eine oder andere Seite setzt jeweils ein gewisses Verständnis von Sprache voraus. In Douglas Hofstadters Definition des ELIZA-effects schwingt ein implizites Verständnis von Sprache und Bedeutung mit. Hofstadter hatte den Effekt definiert als »the susceptibility of people *to read* far more *understanding* than is warranted *into* strings of symbols – especially words – strung together by computers« (Hofstadter 1995, 157; meine Hervorhebungen). Verständnis wird in die »eigentlich« bedeutungslosen Worte »hineingelesen«. Was für ein Verständnis von Sprache und Bedeutung ist hier in Geltung? Wie schaffen es diese »Stränge von Symbolen«, dass wir sie mit Bedeutung oder Verständnis aufladen, dass wir hinter ihnen Intentionalität, Geist, Bewusstsein vermuten?

Ist hier wichtig, was genau es ist, was dahinter vermutet wird? Hofstadter spricht zunächst von »understanding«, in seinem Beispiel des Geldautomaten geht es dann um den Ausdruck von Dankbarkeit, ein Zustand der eher Begriffe wie (Selbst-)Bewusstsein, einen Personenbegriff oder Anerkennung[11] voraussetzt. Was genau es also ist, das die Worte ausdrücken sollen beziehungsweise was fälschlicherweise in sie hineingelesen werden soll, ist unklar. Klar ist, dass das Wort in jedem Fall auf etwas anderes hinweisen soll, etwas anzeigen, etwas bedeuten soll. Wie Wörter das tun, ist eine in der Sprachphilosophie viel diskutierte Frage. Im folgenden Teil wende ich mich zwei Theorien zu, die Merleau-Ponty in seiner Untersuchung zu Phänomen und Ursprung der Sprache in der *Phänomenologie der Wahrnehmung (PdW)* in Betracht zieht, um zu erklären, wie Worte bedeuten können. Er nennt diese Positionen *Empirismus* und *Intellektualismus* (Merleau-Ponty 1966, 207ff.).

2. Intellektualismus und Empirismus als Sprachtheorien

Empirismus und Intellektualismus sind zwei konträre Positionen, die Merleau-Ponty in der *PdW* immer wieder gegeneinanderstellt, von denen er aber zu zeigen versucht, dass sie auf ähnlichen falschen Voraussetzungen beruhen. Ob diese Positionen über die verschiedenen Felder hinweg, in denen Merleau-Ponty sie anbringt, eine konsistente philosophische Position ergeben, die in dieser Form von jemandem vertreten wurde, ist

11 Zur Frage der Möglichkeit der Anerkennung einer KI durch einen Menschen vgl. den Beitrag von Martin Schnell in diesem Band.

fraglich. Es sind eher Idealtypen des Denkens, die Merleau-Ponty hier zeichnet[12]. In Bezug auf die Sprache sieht dies wie folgt aus:

Der *Empirismus* sieht den Besitz von Sprache als den Besitz von »Wortbildern«, die von empirisch wahrgenommenen Worten in uns hinterlassene Spuren sind. Man kann dies naturalistisch oder mentalistisch interpretieren. Im ersteren Fall wären die Wortbilder dann neurologische Strukturen, die sich über die Zeit etabliert haben, im letzteren Fall gibt es »Bewußtseinszustände«, die »auf Grund erworbener Assoziationen die Erscheinung des passenden Wortbildes nach sich ziehen« (ebd. 208). In beiden Fällen aber wird die Sprache aus der Perspektive der dritten Person als abstrakte Struktur beschrieben. Die Sprechende hat in dieser Theorie keinen Platz. Wortflüsse unterscheiden sich hier nicht von Wasserflüssen oder elektrischem Strom, Sprache ist ein reines Strukturphänomen, »der Mensch kann demnach sprechen, wie eine elektrische Birne glühen kann« (ebd.). Das Problem ist nun, dass die Kohärenz dieser Position von Erkenntnissen der empirischen Psychopathologie in Frage gestellt wird, die zu einer Position der intellektualistischen Psychologie führen.

So behauptet die *intellektualistische* Position, dass die Sprache vom Denken bedingt ist. Begründet wird dies aus den Erkenntnissen in der Erforschung der Aphasie. Die Aphasie ist eine Sprachstörung als Folge einer Läsion in der dominanten Sphäre des Gehirns. Bei bestimmten Formen der Aphasie hat die Patientin Probleme, Worte zu finden, wenn abstrakt nach diesen gefragt wird, jedoch nicht, wenn sie als Ausdruck eines wirklichen Erlebens geäußert werden. So kann das Wort »Nein« beispielsweise problemlos ausgesprochen werden, wenn es »eine wirklich erlebte Verneinung bedeutet« (ebd.), also beispielsweise auf die Frage »Hast du Durst?«. Die Patientin findet das Wort aber nicht, wenn es »um eine bloße Sprachübung ohne affektives und vitales Interesse geht« (ebd.). Also beispielsweise eine Aufgabe wie »Nennen Sie einige Ausdrücke der Ablehnung«. Diese Erkenntnis führt für Merleau-Ponty in den »äußersten Gegensatz zur Wortbildtheorie […], da die Sprache nunmehr vom Denken bedingt erscheint« (ebd., 209), während im Empirismus das intentionale Denken der Person doch keine Rolle spielen durfte, Sprache sollte ja beschrieben werden wie jedes andere natürliche Phänomen. Die Einheit des Wortbildes wird zerstört. Soll ein Wortbild aus einer bestimmten Struktur im Hirn bestehen, so müsste bei Verletzung dieser Hirnstruktur das Wort gänzlich unverfügbar werden. Die intellektualistische Psychologie zeigt aber, dass es, abhängig vom Gedanken, der »dahinter« steht, noch verfügbar ist. »Was der Normale besitzt und der Kranke verloren hat, ist nicht ein bestimmter

12 Vgl. (Carman 2008) für eine Analyse der beiden Positionen im Werk Merleau-Pontys.

Wortvorrat, sondern eine bestimmte Weise, sich dieses ›Vorrats‹ zu bedienen« (ebd. 208).

Das Problem an dieser Position ist wiederum, dass hier nun der *Gedanke* die Bedeutung zu haben scheint und nicht das *Wort*. So ist nicht verständlich, wie uns Sprache überhaupt etwas Neues zu sagen vermag, wieso wir unsere eigenen Gedanken bisweilen nicht kennen, bevor wir sie nicht formuliert haben, wieso sie sich erst beim Sprechen allmählich verfertigen (Kleist). Das Denken als augenblickliches, zeitlich unausgedehntes kann sich nur angeeignet werden durch den Ausdruck seiner selbst in der Sprache.

Es gibt hier also einen Widerspruch, der nicht auflösbar erscheint, »das Problem der Sprache ist nicht durch den Übergang von der These zur Antithese zu lösen« (ebd., 209). Vielmehr ist es so, dass beide Positionen das Wesentliche an der Sprache verfehlen. Beide Positionen »bleiben hinter der einfachen Feststellung zurück, daß *das Wort einen Sinn hat*« (ebd., 210). Wenn wir diese Überzeugung des alltäglichen Sprachgebrauchs richtig verstehen, so können wir den Widerspruch auflösen, da wir sowohl Empirismus als auch Intellektualismus als auf falschen Voraussetzungen ruhend entlarven.

Beide Positionen werden dieser Feststellung nicht gerecht. In der empiristischen geschieht dies geradeheraus und ohne Ausreden. Sprache als reines mechanisches Strukturphänomen betrachtet hat keinen Platz für Bedeutung. Hier hat das sprechende Subjekt keinen Platz. In der intellektualistischen Position gibt es wohl ein bedeutungsgebendes Subjekt, allerdings ist dies das *denkende* und nicht das *sprechende* Subjekt. Das Wort ist hier nur die leere, äußere Hülle des bedeutungstragenden Gedankens.

Bevor wir zu Merleau-Pontys Lösung des Problems übergehen, will ich diese Positionen kurz auf Hofstadters Definition des *ELIZA-effects* rückbeziehen. Ich denke, dass Hofstadters Definition sowohl in empiristische als auch intellektualistische Richtung interpretiert werden könnte. Dies unterstreicht Merleau-Pontys These, dass die beiden Positionen in einem gemeinsamen Raum an Voraussetzungen stehen, in dem auch Hofstadter fest verwurzelt ist. Als eine Erweiterung der empiristischen Position könnte man sagen: Das, was uns beim Menschen berechtigt, von Verständnis hinter den Worten zu sprechen, und bei der Maschine nicht (»unwarranted«), ist der richtige Prozess der Assoziation, die richtige strukturelle Verknüpfung. Die Maschine hätte die falsche strukturelle Verknüpfung, entweder rein funktionalistisch falsch oder materialistisch falsch, da sie im falschen »stuff«, nämlich in Silizium und nicht in Neuroprotein realisiert ist. Das materialistische Argument würde also zurück zu Searles CRA führen.

Die funktionalistische Falschheit würde sich nur gelegentlich als falsch zeigen, beispielsweise wenn ELIZA unsinnig antwortet. Diese Fälle

würden aber als Beweis dafür gelten, dass die Struktur im Allgemeinen die Falsche ist. Wirkliches Verständnis sind wir nur bei einer vollständig richtigen Struktur berechtigt zuzuschreiben.

Intellektualistisch interpretiert könnte man Folgendes sagen: Wir sind nur dann berechtigt Verständnis zuzuschreiben, wenn Gedanken hinter den Worten stehen. Dies ist für den Intellektualismus im Falle der Maschine aber wohl per definitionem ausgeschlossen, da er von einem cartesischen Dualismus ausgehen muss. Die Maschine ist dann immer nur genau dies: bloße Maschine.[13]

Welche der beiden Interpretationen man nun vertritt, ist für Merleau-Pontys Kritik eigentlich nicht von Belang, denn er sieht beide auf gleichen, falschen Voraussetzungen ruhen. Beide wollen die Bedeutung des Wortes durch etwas anderes als das Wort erklären und verlieren sie dabei. Keine der Positionen kann zufriedenstellend erklären, wie Worte bedeuten. Somit sind sie auch beide ungeeignet aufzuklären, worin genau der *ELIZA-effect* besteht. Wir brauchen einen anderen Blick auf Sprache, wir müssen ein anderes Verständnis davon erlangen, was Worte sind, wenn wir verstehen wollen, wie diese es vermögen, uns zu verleiten, einer Maschine Verständnis zuzuschreiben. Wie gesagt wird die Antwort für Merleau-Ponty darin liegen, dass das Wort selbst Bedeutung[14] *hat*.

3. Das Haben von Sinn und das Wort als Geste

Merleau-Ponty will »Haben« hier in der Bedeutung verstanden wissen, wie wir es benutzen, wenn wir sagen »Ich habe Angst« oder »Ich habe eine Idee«. Damit soll angezeigt werden, dass »Haben«, im Gegensatz zu »Sein«, »den Bezug des Subjekts zu dem bezeichnet, woraufhin es sich entwirft« (Merleau-Ponty 1966, 207 Fußnote 1). In diesem Sinne will er verstanden wissen, dass der Mensch Sprache *hat* und auch dass die Sprache und das Wort einen Sinn *haben*. Das Wort vollbringt das Denken erst, es gibt nicht den im Denken schon fertigen Begriff, der dann nur noch durch die Worthülse artikuliert wird, sondern im echten, einen Sachverhalt das erste Mal formulierenden Sprechen wird der Gedanke erst vollbracht. Er ist nicht fertig, ja existiert nicht einmal wirklich, bevor er nicht ausgesprochen wurde. Genauso entnimmt der Hörende »den

13 Ausgehend von Hofstadters Aussagen in *Gödel, Escher, Bach*, in der er Bedeutung hervorgebracht sieht durch das Erkennen eines Isomorphismus, müsste man ihn wohl eher der empiristischen Seite zuschlagen (vgl. Hofstadter 1985) Hofstadters philosophische Position im weiteren Sinne ist aber nicht Thema dieses Texts.

14 Merleau-Ponty unterscheidet in diesen Abschnitten nicht streng zwischen »Sinn« und »Bedeutung«, etwa im Sinne Freges.

Gedanken dem Worte selbst« (ebd. 212). Es ist nicht so, dass der Hörende den Worten nachträglich, gleichsam durch einen »inneren« Akt seines Geistes, den Sinn gibt oder, in den Worten Hofstadters, etwas in die ansonsten bedeutungslosen Zeichen »hineinliest«. Wäre es so, könnte man durch gehörte Worte nie etwas Neues lernen und die »Erfahrung der Verständigung« wäre nichts als eine Illusion (ebd.). Es ist aber nun einmal eine Tatsache, so Merleau-Ponty, dass wir »etwas zu verstehen mögen, was über das, was wir von uns aus dachten, hinausgeht« (ebd.). Die Sprache müssen wir zwar schon kennen, und doch können uns die Worte dieser uns bekannten Sprache etwas Neues sagen, ihre Bedeutungen sich bisweilen »zu einem neuen Gedanken« verbinden. Diese Tatsache muss uns dazu bringen, dass »der Sinn der Worte letzten Endes durch die Worte selber hervorgebracht« sein muss. Merleau-Ponty spezifiziert sofort: »oder vielmehr genauer, deren begriffliche Bedeutung sich bilden auf Grund und aus ihrer *gestischen Bedeutung,* die ihrerseits der Sprache selbst immanent ist« (ebd., 212f.).

Wie ist das zu verstehen, was ist die »gestische Bedeutung«, die in der Sprache selbst liegt, und ihren Sinn hervorruft? Hier kann nun zurückgegriffen werden auf die Bedeutung, die Merleau-Ponty dem Wort »Haben« zugeschrieben hat. Der Mensch *hat* Sprache, insofern diese Sprache Teil dessen ist, woraufhin er sich entwirft, d.h. Teil seiner Existenz ist. Dadurch *hat* das Wort Bedeutung. Merleau-Ponty erläutert dies am Prozess des Erlernens einer Sprache. Eine fremde Sprache lerne ich, indem ich ihre Ausdrücke »im Zusammenhang der Tätigkeit der Menschen zu verstehen beginne, wenn ich an deren Gemeinschaftsleben teilnehme« (ebd., 213). Um ein Beispiel von W.V.O. Quine zu bemühen[15]: Wenn wir bei Vorbeilaufen eines Hasen einen *Aruga*-Sprecher »Gavagai« sagen hören, wissen wir erst ob damit »Schau, ein Hase« gemeint ist oder »Oh, ein schlechtes Omen«, »Schau, unser Abendessen« oder etwas ganz anderes, wenn wir den Zusammenhang der Tätigkeit der Aruga-Sprecher zu verstehen beginnen. Beispielsweise, ob es bei ihnen normal ist, das Auftauchen eines Tieres als ein Omen zu werten, ob Hasen typischerweise gegessen werden etc. Merleau-Ponty nennt das, den »Stil« einer Sprache zu verstehen, in dem Sinne, in dem man den Stil eines Malers zu erkennen lernt oder den Stil eines Musikstückes durchdringt. Stil ist insofern eine Organisationsform von Teilen in ein komplexes Ganzes. Der Stil eines Musikstückes bestimmt, wie aus den einzelnen Noten der »Sinn« des Stückes entsteht, ohne dass dieser Sinn in den einzeln wahrgenommenen Tönen

15 In (Quine 2013). Quines Projekt einer naturalisierten Epistemologie hat natürlich viele Konfliktpunkte mit Merleau-Pontys Philosophie. In gewissen Punkten der Kritik an einem naiven Empirismus gibt es aber sicherlich interessante Gemeinsamkeiten, die hier aber nicht näher verfolgt werden sollen.

erkennbar wäre. So ist auch das Verhältnis des Sinns der Worte zur Sprache zu denken.

Es ist hier vielleicht hilfreich auf die enormen Parallelen zu der Sprachauffassung des späten Wittgenstein hinzuweisen. Das »Hinzunehmende, Gegebene«, also das, was in der Sprache letztlich ausgedrückt ist, »– könnte man sagen – seien Lebensformen«, so Wittgenstein in den *Philosophischen Untersuchungen* (2014, 572). Und ich lerne die Bedeutungen, die die Wörter einer Sprache *haben,* indem ich an der Lebensform teilnehme, die sie ausdrücken. »Und eine Sprache vorstellen, heißt, sich eine Lebensform vorstellen« (ebd. §19). »Eine Sprache verstehen, heißt eine Technik beherrschen« (ebd. §199). Merleau-Ponty spricht statt von Lebensform, der man gemeinsam zugehört, von einer Welt, die man gemeinsam *hat*: »In Gestalt der verfügbaren Bedeutungen […] besitzen die sprechenden Subjekte eine gemeinsame Welt […] Der Sinn des Wortes ist selbst in gar nichts anderem gegeben als der Art und Weise, in der es zu dieser Sprachwelt sich ins Verhältnis setzt« (Merleau-Ponty 1966, 221). »Um eine Sprache sich vollständig anzueignen, müßte man die Welt übernehmen, die in ihr Ausdruck findet« (ebd. 222). Hier kann man direkt Wittgensteins Äußerung anschließen, dass wir einen Löwen, könnte er sprechen, nicht verstünden (Wittgenstein 2014, 568). In der Löwenwelt sind wir nicht zu Hause und könnten dies wohl auch nur sehr bedingt je sein. Bisweilen sind die Äußerungen Merleau-Pontys und Wittgensteins fast verwechselbar: »Man kann für eine große Klasse von Fällen der Benützung des Wortes ›Bedeutung‹ – wenn auch nicht für alle Fälle seiner Benützung – dieses Wort so erklären: Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache« (Wittgenstein 2014, §42). »Was den Sinn des Wortes betrifft, so lerne ich ihn, wie ich den Gebrauch eines Werkzeugs lerne: indem ich es im Kontext einer bestimmten Situation gebrauchen sehe« (Merleau-Ponty 1966, 459).

Für Wittgenstein wie für Merleau-Ponty haben wir Worte also als Ausdrücke eines Verhaltens zur Welt. Merleau-Ponty nennt die gestische Bedeutung folglich auch »existentielle Bedeutung« (ebd., 216). Das Wort teilt den Gedanken nicht als Begriffsfunktion, sondern als »existentielle Gebärde« mit (ebd.). Diese Art der Bedeutung liegt aller begrifflichen Bedeutung zugrunde und ist gleichzeitig das, was wir in der Sprache primär wahrnehmen. Sie ist der Sprache selbst immanent. Wir müssen den Sinn nicht erst aktiv in die Worte hineinlesen, sondern wir nehmen ihn direkt wahr. Das wahrgenommene Wort hat Bedeutung, wie das Wahrnehmungsding Farbe hat. Wollte man darauf bestehen, dass hier doch ein Inhalt und eine Form, oder ein Zeichen und ein Bezeichnetes zusammengefügt werden, so müsste man mindestens mit einem Wort Husserls von *passiver Synthesis* sprechen. So spricht auch Merleau-Ponty davon, dass der geglückte Ausdruck der Bedeutung ein Dasein »gleich dem eines Dinges« beschert und die Bedeutung

wahrgenommen wird, als hätten wir durch den Erwerb der Sprache ein neues Sinnesorgan und ein neues Erfahrungsfeld gewonnen (ebd.). Aus dem ästhetischen Ausdruck kennen wir besser, was Merleau-Ponty sagen will, hier fällt es uns leichter, die Unablöslichkeit des Ausgedrückten vom Ausdruck einzusehen. Einzusehen, dass der Ausdruck das Wunder vollbringt, dem Ausgedrückten ein »An-sich-sein« zu verleihen, es als »ein jedermann zugängliches Wahrnehmungsding in die Natur zu« versetzen (ebd., 217). Für den Zusammenhang von Tönen einer Sonate und ihrer ›Bedeutung‹ oder von Elementen eines Bildes und dem, was es ›bedeutet‹, würden wir dies sofort zugestehen. Letzteres ist immer nur durch erstere da, ohne in diesen aufzugehen. In Wahrheit, so Merleau-Ponty, verhält es sich genauso mit dem Sprechen und dem Denken. Das Denken existiert nur im Sprechen, auch der vermeintlich stumme Gedanke ist nur die Erinnerung eines schon ausgedrückten Gedankens und das »Innenleben« des Denkens ist nur ein inneres Sprechen. »Reines Denken«, etwa als das Residuum des Ausgedrückten, das zurückbleibt, das nicht restlos ausgedrückt werden kann, ist laut Merleau-Ponty nichts als »eine gewisse Leere des Bewußtseins, ein augenblicklicher Wunsch« (ebd.).[16]

Daraus folgt, dass niemals zwei verschiedene Ausdrücke das Gleiche ausdrücken können. Das würde den Ausdruck auf Zeichen für unabhängig von ihrem Ausdruck existierende Gedanken reduzieren, die von verschiedenen äußeren Hüllen angezeigt werden können. So könnte also für Merleau-Ponty »Hund« und »dog« niemals das exakt Gleiche bedeuten, exakte Übersetzung ist unmöglich (vgl. ebd., 222). Ja sogar das von unterschiedlichen Sprecherinnen der gleichen Sprache benutzte Wort kann dann nicht bloß sinnlich unterscheidbares Zeichen eines identischen Gedankens sein. Verständnis und Übersetzung ist nicht möglich, weil verschiedene Zeichen gleiche Gedanken bezeichnen, sondern, weil es eine Entsprechung zwischen meinen sich in meinen Gebärden bekundenden Leibintentionen und den Gebärden der Anderen gibt. Da ich leiblich in der Welt bin, habe ich Worte als »eine mögliche Modulation, einen möglichen Gebrauch meines Leibes« (ebd. 214). So dass ich zum Wort greife, »wie meine Hand an eine plötzlich schmerzende Stelle meines Körpers fährt« (ebd.). Ich kann mit der Anderen kommunizieren und wir können uns verstehen, weil unsere Gebärden in der gleichen Welt stattfinden, wir qua

16 Auch hier springt eine Ähnlichkeit zu Wittgenstein ins Auge. Für diesen war die Frage nach dem »reinen Gedanken«, der von einem Wort bezeichnet werden soll, eine völlig falsche. Der Gedanke selbst spielt keine Rolle für eine Analyse der Bedeutung des Wortes. Die Schachtel, in der wir diese »reine« Bedeutung verstecken und in die nur wir je schauen können, könnte genauso gut leer sein (vgl. Wittgenstein 2014, §293), genauso leer wie das Bewusstsein, von dem Merleau-Ponty hier spricht.

unserer Leiblichkeit eine gemeinsame Sinnlichkeit teilen. Unsere Kommunikation gründet erst den gemeinsamen Sinn unserer Erfahrungen, sie setzt ihn nicht voraus als gleichsam schon in einer transzendenten Sphäre vorhanden. Ich kann die leiblichen Gebärden der Anderen verstehen, weil auch ich ein Leib bin. Und das Wort ist sedimentierte Gebärde, nur deshalb kann ich es verstehen.

Eine Differenz im Ausdruck ist also auch eine Differenz im Ausgedrückten. So besagt beispielsweise der andere Ausdruck des Zorns in Japan einen Unterschied in der Emotion selbst (ebd., 223). Zorn anders auszudrücken heißt, Zorn anders zu fühlen, anders zornig *zu sein*. Die Leiblichkeit als gemeinsame Grundlage des Verstehens ist also nicht zu verstehen als bloße Gleichheit des organischen Aufbaus, sondern als kulturell geformtes Medium des Ausdrucks. »Die psychophysische Ausrüstung an sich eröffnet hier zahllose Möglichkeiten, und so wenig wie im Bereich der Instinkte gibt es auch hier eine Natur des Menschen, die ein für alle Mal feststünde. Der Gebrauch, den der Mensch von seinem Leibe macht, transzendiert den Körper als bloß biologisch Seiendes […] Seine Verhaltensweisen schaffen Bedeutungen, die seine anatomische Anlage transzendieren, gleichwohl aber dem Verhalten als solchem immanent bleiben« (ebd., 224). Mit Wittgenstein würde man sagen: Der Mensch partizipiert leiblich an Lebensformen.

4. Leiblose Kommunikation?

Was heißt dieses Verständnis der Sprache nun für die Kommunikation mit einer KI? Ist es nicht klar, dass der Körper eines Roboters kein Leib im Merleau-Ponty'schen Sinne ist und eine körperlose Chatbot-KI ja noch nicht einmal einen Körper hat, also unmöglich gestische Bedeutungen realisieren kann? Auch Merleau-Ponty selbst äußert sich in einem knappen Satz zu Sprachalgorithmen folgendermaßen: »Ein auf Konventionen beruhender Algorithmus – der übrigens auch nur im Rückbezug auf die Sprache Sinn hat – wird nie etwas anderes auszudrücken vermögen als eine Natur ohne Menschen« (ebd., 222f.). Ist also der *ELIZA-effect* auch aus einer Merleau-Ponty'schen Sprachphilosophie heraus nichts als ein Irrtum, ein Missverstehen dessen, was passiert, eine falsche Wahrnehmung, eine Illusion? Ich denke, so weit zu gehen wäre vorschnell. Merleau-Pontys eigene Äußerung über die Macht der Algorithmen muss vielmehr, angesichts des von ihm noch nicht beschreibbaren Phänomens einer auf dem heute möglichen Level konversierenden KI, neu überdacht werden. Es gilt, das Phänomen der Sprache auch in dieser Facette ernst zu nehmen und phänomenologisch zu beschreiben. Ich will mich dem in diesem Abschnitt

zunächst nähern, indem ich beschreibe, was beim Chatten zweier Menschen miteinander mit dem Phänomen der Sprache passiert.

Zunächst könnte man sagen, auch hier fehle doch der Körper. Jegliche über Distanz vollzogene, nur geschriebene Kommunikation sei bestenfalls parasitär gegenüber der im zwischenleiblichen Dialog sich realisierenden Bedeutung. Wie ist dieses parasitäre Verhältnis aber genauer zu verstehen? Warum haben Worte für uns auch Bedeutung, wenn der Leib, der sie ausdrückt, nicht mit-wahrgenommen wird? Mit anderen Worten: Wie funktioniert die Sedimentation der gestischen Bedeutung hier genau? Um diese Fragen zu beantworten, ist es nötig, ein Stück weit auf Merleau-Pontys Begriff der Anderen einzugehen, denn der Leib der Anderen ist es schließlich, der die gestische Bedeutung ausdrückt.

Die Andere als mit mir Kommunizierende hat für Merleau-Ponty eine besondere Örtlichkeit. Sie ist nicht in der Welt vorhanden wie ein Ding in einem Container, ich nehme die Andere in der Welt nicht einfach wahr als vor mir stehend. Es wäre eher passend zu sagen, sie ist »*zwischen* mir, der denkt, und jenem Leib [der Anderen] oder eher neben mir, an meiner Seite, taucht [sie] auf wie eine Nachbildung meiner selbst«(Merleau-Ponty 1993, 149)[17]. Die Andere hat keinen Platz in der Welt, außer in meinem »*Wahrnehmungsfeld,* doch zumindest dieser Ort ist parat für [sie], seit meine Wahrnehmung begonnen hat« (ebd., 152). Wie ist das zu verstehen?

Vom Wahrnehmungsfeld spricht Merleau-Ponty auch schon in der *Phänomenologie der Wahrnehmung*, wir können uns dem Begriff nähern über den des *Gesichtsfeldes*. Das *Gesichtsfeld* sollte nicht verwechselt werden mit all dem, was ich in einer bestimmten Situation sehe. Es ist nicht einfach zusammengesetzt aus den meine Netzhaut berührenden Reizen, es besteht nicht aus den Sinnesdaten, die meine Augen treffen, es ist auch nicht ein scharf abgegrenztes wie gerahmtes Bild, ein »scharf umgrenztes Weltsegment […] außen umgeben von einer dunklen Zone, innen lückenlos erfüllt von Qualitäten, getragen von Größenverhältnissen gleicher Bestimmtheit wie auf der Netzhaut« (Merleau-Ponty 1966, 24). Die Grenzen des Gesichtsfeldes, die es umgebende Region »ist nicht leicht zu beschreiben, doch ist sie sicher weder schwarz noch grau. Sie steht in einer *unbestimmten Sicht,* der Sicht eines *Je ne sais quoi*, und am Ende ist sogar das in meinem Rücken Gelegene nicht gänzlich ohne visuelle Gegenwart« (ebd.)

Ebenso wenig ist es nur mit *visuellen* Daten gefüllt. Das Gesichtsfeld ist ein Feld unter anderen im Wahrnehmungsfeld, die Felder der anderen

17 Hier beziehe ich mich auf einen Text aus dem Spätwerk Merleau-Pontys, das oft als in entscheidender Weise von der *PdW* abweichend interpretiert wird. Zumindest in der Frage nach dem Ort des Anderen gibt es aber mehr Kontinuitäten als Brüche, wie im Folgenden deutlich werden sollte.

Sinne überlappen mit ihm. So zählt das im nächsten Zimmer spielende Grammophon, das ich nicht sehe, noch mit in mein Gesichtsfeld (ebd., 323). Ich höre das Grammophon, und dies zeigt mir mit an, wie es aussieht. Das Hörfeld überlappt mit dem Sehfeld. Mein gesamtes Wahrnehmungsfeld ist dann konstituiert als die Gesamtheit dessen, zu dem ich eine Weise des Verhaltens innehabe. Ich höre Musik im nächsten Raum und da ich weiß, dass dort ein Grammophon steht, *höre ich das Grammophon*. Ich höre keine Musik, die dann auf das Grammophon verweist, sondern ich höre das Grammophon selbst, so wie ich das Grammophon selbst sehen könnte, würde ich in den anderen Raum gehen. Es wird Teil meines Wahrnehmungsfeldes *als Ding*. Mit seinem Klang wird mit-präsentiert, sein Aussehen, sein Geruch, wie es sich anfühlt, wie es angeschaltet wird, wie schwer es ist etc. Alles was ich durch die Sinne wahrnehme, ist mir als sinnvoll gegeben, das heißt mit einer Weise des Verhaltens ihm gegenüber.

Extrapolieren wir diese Beschreibung auf die Situation des Chattens mit einem anderen Menschen am Computer. Was ich sehe, ist der Text der Anderen, der erscheint. Visuell gegeben sind nur Worte. So wie ich durch die Musik des Grammophons das Grammophon selbst höre, es in meinem Wahrnehmungsfeld erscheint, erscheint die Andere durch ihre Worte sofort in meinem Wahrnehmungsfeld; nicht jedoch an einem bestimmten Ort, im Sinne eines in Koordinaten angebbaren Teils meines Wahrnehmungsfeldes. Wir können den Körper der Anderen zwar als einen solchen Ort besetzend beschreiben, damit greifen wir aber nicht den Ort, an dem die Andere selbst ist. Die Andere selbst ist »nirgendwo im Sein, von hinten gleitet [sie] in meine Wahrnehmung« (Merleau-Ponty 1993, 152).

Was will Merleau-Ponty mit dieser Rede von der Nicht-Örtlichkeit der Anderen, von ihrem Leben »hinter«, »zwischen«, »neben« mir oder in den »Fugen zwischen der Welt und uns selbst« (ebd., 153) aussagen? Eine weitere phänomenologische Situationsbeschreibung kann das aufklären. Stellen Sie sich folgende Situation vor: Sie kommen nach Hause, in Ihre Wohnung, die Sie sich teilen mit Ihrer Familie oder anderen Mitbewohnerinnen. Sie nehmen aber an, dass Sie in der Wohnung allein sind, da normalerweise zu dieser Zeit alle anderen arbeiten oder anderweitig außer Haus beschäftigt sind. Die ganze Wohnung wird von Beginn an gefärbt durch diese Annahme des Alleinseins erscheinen, *als* Wohnung, in der Sie alleine sind, in der momentan niemand anders *ist*. Plötzlich hören Sie ein Geräusch aus einem der Nebenräume, ein Räuspern oder vielleicht sogar die Stimme einer Mitbewohnerin. Die ganze Situation, die ganze Wohnung, Ihre gesamte Umwelt wird ihren Charakter unmittelbar ändern. Das heißt, Ihr Wahrnehmungsfeld verändert seinen Sinn, seine Orientierung, da dort jemand Anders *ist*, d.h. die Andere gleitet, sobald sie wahrgenommen wird, unmittelbar »hinter« Sie oder »in die Fugen« zwischen Ihnen und der Welt.

Diese Situation ist natürlich nicht in jeder Hinsicht vergleichbar mit der Situation des Chattens vor dem PC. In der Wohnungssituation ist die Präsens der Anderen sehr viel radikaler, die Andere ist sehr viel weiter »im Zentrum« Ihres Wahrnehmungsfeldes als in der PC-Situation. Dennoch ist die Andere auch in der PC-Situation Teil Ihres Wahrnehmungsfeldes. Technologie kann die Grenzen unseres Wahrnehmungsfelds sehr weit dehnen, und sobald die Andere im Feld ist, ist sie in den Fugen. Daher ist echter Dialog in dem Sinne, wie Merleau-Ponty ihn in der *Phänomenologie der Wahrnehmung* beschreibt, in dem »die Einwände meines Gesprächspartners« mir sogar »Gedanken entreißen, von denen ich nicht wußte, daß ich sie hatte, so daß also der Andere ebensosehr mir zu denken gibt wie ich ihm Gedanken zuschreibe« (Merleau-Ponty 1966, 406), über Telefon, Zoom oder eventuell sogar Chaträume zwar schwierig, aber nicht unmöglich.

Die Leiblichkeit der Anderen spielt hier allerdings nach wie vor eine entscheidende Rolle. Wenn ich meine Mitbewohnerin im anderen Raum höre, ist sie Teil meines Wahrnehmungsfelds, da ich ihren Leib höre, ich kann mir vorstellen, wie ihr Leib im anderen Raum situiert ist. Genauso kann ich mir eine Andere hinter den Worten im Chatraum vorstellen, da ich weiß, was es heißt, vor einem PC zu sitzen und zu chatten, ich bin mit der Situation vertraut. Für eine Person, die mit dem technologischen Artefakt des PCs völlig unvertraut ist – man denke etwa an Mitglieder indigener, völlig isoliert von moderner Technologie lebender Gruppen – wäre das Phänomen des Auftauchens der Worte auf dem Bildschirm völlig unverständlich, sie könnte keinen Urheber der Worte mit Ihnen verbinden, die Technologie wäre, um Arthur C. Clarke zu paraphrasieren »ununterscheidbar von Magie«. Das fehlende Wissen ist hier kein Wissen um die Funktionsweise der Technologie – die meisten Menschen, die PCs benutzen, haben kaum eine Ahnung davon, wie diese funktionieren –, sondern ein praktisches Wissen des Umgangs mit der Technologie, ein Wissen des Verhaltens ihr gegenüber.

Dieses praktische Wissen des Umgangs mit Technologie lässt diese in die Selbstverständlichkeitsstruktur unserer Lebenswelt einfließen.[18] Hans Blumenberg hat diese Erweiterung der Struktur unserer Lebenswelt durch Technik schon sehr früh philosophisch reflektiert. Der Prozess der Technisierung, den Husserl noch rein als Vergessen der lebensweltlichen Ursprünge unseres technisierten Denkens gefasst hatte, erreiche sein Telos erst dadurch, so Blumenberg, dass die Technisierung ihrerseits beginne, »die Lebenswelt zu regulieren, indem jene Sphäre, in der wir *noch* keine Fragen stellen, identisch wird mit derjenigen, in der wir keine

18 Merleau-Ponty will seine Welt der Wahrnehmung explizit als Erweiterung des Husserl'schen Lebensweltbegriffs verstanden wissen (vgl. Merleau-Ponty 1966 Vorwort).

Fragen *mehr* stellen« (Blumenberg 2015, 190). Die intersubjektive Struktur der Lebenswelt wird angereichert mit technischen Produkten, und das vermeintlich unmöglich Herzustellende, nämlich Selbstverständlichkeit, wird produzierbar. So wird uns beispielsweise die Wahrnehmung von Bedeutung in der Wahrnehmung von Worten auf einem PC fraglos, sie erscheint selbstverständlich, obwohl hier kein Leib zu sehen ist, der diese Worte hervorbringt. Das praktische Wissen des Umgangs mit Technologie ist also im Erfolgsfall der Technologie ein implizites, leibliches Wissen, das nicht reflektiert ist und gerade deshalb selbstverständlich ist. Die Selbstverständlichkeit der Lebenswelt ist ja gerade »der Gegenbegriff zu jener ›Selbstverständigung‹, die für Husserl die eigentliche Aufgabe einer phänomenologischen Philosophie zu sein hat« (ebd. 178).

Mit Merleau-Ponty kann man feststellen, dass die Möglichkeit dieser technischen Veränderung unserer Selbstverständlichkeitssphäre in den Strukturen unserer Leiblichkeit angelegt ist. So hat Merleau-Ponty schon die Projektion des eigenen Leibs in eine Situation hinein, in der er rein physisch nicht vorhanden ist, mit dem Begriff des »virtuellen Leibs« in den Analysen zur Raumwahrnehmung ausführlich beschrieben. In den Spiegelexperimenten Wertheimers nimmt das Subjekt den im schiefen Spiegel schief dargestellten Raum zunächst schief wahr. Nun »geschieht nach einigen Minuten [aber] das Wunder, daß das reflektierte Zimmer ein Subjekt hervorruft, das in ihm zu leben vermag«. Hier verdrängt der virtuelle Leib

> den wirklichen Leib so weitgehend, daß das Subjekt sich nicht mehr in der Welt fühlt, in der es sich tatsächlich befindet, und statt seiner wirklichen Arme und Beine solche Arme und Beine empfindet, wie es sie haben müßte, um in dem reflektierten Zimmer gehen und tätig sein zu können; es bewohnt das Schauspiel. In diesem Moment gerät das Raumniveau ins Schwanken und etabliert sich in neuer Lage. (Merleau-Ponty 1966, 292)

Das heißt, mein Leib richtet sich im schiefen Raum qua virtuellem Leib ein und dadurch wird dieser geradegerückt. Ein neues Raumniveau entsteht, mit anderen Worten: eine neue Selbstverständlichkeitssphäre.

Dieser Begriff des virtuellen Leibes kann auf die Orientierung des Wahrnehmungsfeldes im Allgemeinen übertragen werden. Auch sinnlich nicht gegebene Dinge, Orte, Personen können Teil meines Wahrnehmungsfeldes sein, sie sind Teil meiner Sinngebung, da ich leiblich bei ihnen sein *könnte*, sie »im Felde meines virtuellen Tuns gelegen« sind (ebd. 502). In der Situation des Chattens vor dem PC gibt mir die Wahrnehmung der Worte die Möglichkeit des Imaginierens der Situation der anderen Person vor dem PC und somit die Möglichkeit, qua Imagination mein Wahrnehmungsfeld vom virtuellen Leib, als Leib vor dem anderen PC, durchstimmen zu lassen und dadurch das Problem des nicht sinnlich gegebenen Leibes der Anderen zu lösen, der doch eigentlich Bedingung der Möglichkeit von gestischer Bedeutung ist.

Wie sieht dies nun im Fall der Konversation mit einer KI aus? Auch hier ist kein Leib sinnlich gegeben. Wissen wir, dass es sich um eine KI handelt, beispielsweise ELIZA oder GPT-3, so wissen wir aber auch, dass hier kein Leib an irgendeinem Ort vor einem PC sitzt und die Worte leiblich eintippt. Wir stehen also vor einem Problem: Die Worte haben Sinn, sie werden als bedeutsam wahrgenommen, da dies in ihrer Phänomenalität selbst liegt, da in ihnen gestische Bedeutung sedimentiert ist, wir können sie aber nicht einmal virtuell auf einen Leib zurückführen.

Die Wahrnehmung bedeutsamer Worte, d.h. Worte als Gebärden, ohne einen Leib, der sie produziert, der sich gebärdet, muss eigentlich in das von Husserl beschriebene Phänomen des *Widerstreits* führen. Wir nehmen bedeutungsvolle Worte wahr, jedoch keinen Leib, der sie produziert. Der Widerstreit ist dieses Phänomen der Wahrnehmung von zwei miteinander konkurrierenden Intentionen oder Sinngehalten. Husserls Lieblingsbeispiel hierfür ist zuerst dokumentiert in den *Vorlesungen zum Bildbewußtsein*:

> die schon öfters erwähnten Täuschungen a la Panoptikum, Panorama etc. Hier mag es zunächst sein, dass wir die Puppe als Menschen sehen. Wir haben da eine, wenn auch nachträglich als Irrtum sich herausstellende normale Wahrnehmung. Werden wir uns plötzlich der Täuschung bewusst, dann tritt das Bildlichkeitsbewusstsein ein. Aber in diesen Fällen will es sich nicht auf die Dauer durchsetzen. Die Wachsfigur gleicht mit ihren wirklichen Kleidern, Haaren usw., ja selbst in den durch mechanische Vorrichtung künstlich nachgeahmten Bewegungen so sehr dem natürlichen Menschen, dass sich momentan immer wieder das Wahrnehmungsbewusstsein durchsetzt. Die imaginative Auffassung fällt weg. Wir »wissen« zwar, dass es Schein sei, aber wir können uns nicht helfen, wir sehen einen Menschen (Husserl 1980, 23:40)

In den späteren *Analysen zur passiven Synthesis* greift Husserl nun dieses Beispiel auf, um das Phänomen des Zweifels und des Widerstreits zu erläutern:

> Das zunächst als Mensch Gesehene wird zweifelhaft und schließlich stellt es sich als eine bloße Wachspuppe heraus. Oder aber umgekehrt, der Zweifel löst sich in der bejahenden Form: ja, es ist doch ein Mensch. Während des Zweifels, ob wirklicher Mensch oder Puppe, überschieben sich offenbar zwei Wahrnehmungsauffassungen. […] Keine von beiden ist während des Zweifels durchgestrichen, sie stehen hier in wechselseitigem Streit, jede hat gewissermaßen ihre Kraft, ist durch die bisherige Wahrnehmungslage und ihren intentionalen Gehalt motiviert, gleichsam gefordert. Aber Forderung steht gegen Forderung, eins bestreitet das andere und erfährt von ihm den gleichen Tort. Es bleibt im Zweifel ein unentschiedener Streit. (Husserl 1966, 11:33f.).

Das, was im Modus des Zweifelns unentschieden bleibt, kann aber irgendwann in Gewissheit überführt werden. Die Möglichkeit der Entscheidung gehöre zum Wesen des Zweifels (ebd. 36). Der Zweifel wird entschieden durch eine unzweifelhafte Wahrnehmung, eine Urimpression, im Falle der Puppe beispielsweise ein Anfassen: »Die Erfüllung durch Urimpression ist die Kraft, die alles niederrennt. Wir treten näher heran, wir fassen auch tastend zu, und die eben noch zweifelhafte Intention auf Wachs erhält den Gewißheitsvorzug.« (ebd. 37).

Dieser Begriff der Urimpression ist selbst kein unproblematischer, evoziert er beispielsweise einen gewisse Präsenzmetaphysik mit ausdehnungslosem Jetztpunkt, die zeitphänomenologisch problematisch bleibt. Diese Problematik muss hier aber nicht weiter verfolgt werden, da in unserem Beispiel eine Urimpression ohnehin unmöglich zu haben ist, selbst wenn wir sie als Möglichkeit prinzipiell zugestehen. Eine Urimpression, wenn überhaupt möglich, muss immer eine Wahrnehmung sein, sie kann kein abstrakter Glauben oder theoretisches Wissen sein. Jedes mit der Wahrnehmung widerstreitende begriffliche Urteil oder Wissen muss in der Wahrnehmung durch Bildlichkeitsbewusstsein wirksam werden, um diese in ihrem Sinngehalt verändern zu können. Durch unser Wissen soll die Puppe als bloßes Bild eines Menschen gesehen werden, nicht als Mensch selbst. Aber »Bildlichkeitsbewusstsein«, wie Husserl schreibt, reicht nicht aus, es vermag das Wahrnehmungsbewusstsein nicht dauerhaft zu verdrängen.

Im Falle des Chattens mit einem Menschen ist die Situation nun ein wenig anders gelagert. Was hier mit unserer Wahrnehmung (der sinnvollen Worte) widerstreitet, ist das *Fehlen* einer anderen Wahrnehmung (des Leibes). Es liegt nahe, zu vermuten, dass dieser Widerstreit zwischen positivem und negativem Phänomen weniger stark ist als der Widerstreit zwischen zwei positiven Wahrnehmungsauffassungen. So kann er schon durch die Imagination unserer selbst in die Situation des Anderen qua Möglichkeit des virtuellen Leibs aufgelöst werden. Diese Imagination beruht auf einer Wahrnehmung, auf einer Urimpression des Anderen. Wir haben schon einmal einen Menschen an einem PC chatten gesehen, wir tun es sogar selbst. So haben wir in der Chatsituation selbst zwar nicht die Möglichkeit einer Urimpression, wir haben aber zumindest ein Wissen, das sich auf eine Wahrnehmung stützt. Dies kann den Widerstreit nie in apodiktische Gewissheit überführen – auf dieser Unmöglichkeit beruht der Turing-Test genau wie jeglicher Betrug über Internetkommunikation –, ihn jedoch so weit auflösen, dass wir den Worten im Chat problemlos Bedeutung zumessen können.

Wie ist es aber, wenn wir wissen, dass wir mit einer Maschine chatten? Im Falle des Wissens um die körperlose Maschinenhaftigkeit des Chatpartners haben wir zunächst ebenso den Widerstreit zwischen einem

positiven und einem negativen, d.h. fehlenden Phänomen. Das fehlende Phänomen kann hier aber nicht durch eine auf Wahrnehmung zurückgreifende Imagination substituiert werden. Wir haben nie einen Roboter vor einem PC chatten sehen, wir wissen auch, dass dies ohnehin nicht der Prozess ist, der die Worte auf dem Bildschirm hervorbringt. Einerseits nehmen wir Worte wahr, und das heißt, wir nehmen Bedeutung wahr. Andererseits fehlt der Leib, der diesen Worten gestische Bedeutung garantiert und wir wissen, dass dieser Leib auch nirgendwo ist. Auf Erinnerung einer Wahrnehmung zurückgreifende Imagination hilft uns also nicht weiter. Wir haben hier nur die der Wahrnehmung der bedeutungsvollen Worte widerstreitende, fehlende Wahrnehmung des Leibes und das Wissen darum, dass auch nirgendwo ein Leib ist.

Steht es hier dann nicht 2:1? Hat die Seite, die den Worten Bedeutung abspricht, nicht die Überhand? Muss das Wissen um die Nicht-Leiblichkeit der Maschine dann nicht den Widerstreit entscheiden, unsere Wahrnehmung der Bedeutsamkeit überlagert werden und die Worte letztlich als bedeutungslos erscheinen?

Um dies zu beantworten, müssen wir genauer darauf schauen, worin das Wissen, dass hier wirksam sein soll, besteht. Wir wissen, dass die KI eine Maschine ist, dass sie also keinen Leib hat. Was aber heißt das genau? Was heißt es für uns zu wissen, dass etwas eine »bloße Maschine« ist, »algorithmisch«, »rein mathematisch« funktioniert? Folgen wir Merleau-Pontys Sprachphilosophie, haben alle diese Begriffe, »Maschine«, »Algorithmus«, »Computer«, »mathematisch«, wenn sie denn überhaupt einen Sinn haben, einen gestischen Sinn. Wir wissen, was es heißt, etwas maschinell zu tun, was es heißt, einem Algorithmus zu folgen, was es heißt, etwas zu berechnen (»to compute«), stets aus unserer eigenen leiblichen Erfahrung. Beim Wort »Computer« zeigt die Begriffsgeschichte dies deutlich an. Es wurde zuerst verwendet, um Menschen – anfangs meist Frauen – zu bezeichnen, die Berechnungen durchführten.[19] »Maschinell« oder »algorithmisch« sind also nicht einfach Gegenteile von »leiblich« oder »bedeutungsvoll«. Etwas maschinell zu tun, ist nicht gleichbedeutend damit, etwas bedeutungslos zu tun. Insofern müssen wir Merleau-Pontys Aussage mit Argumenten aus seiner eigenen Theorie widersprechen. Es stimmt nicht, dass ein Algorithmus nie etwas anderes auszudrücken vermag als eine Natur ohne Menschen (s.o.). Dies wäre nur wahr, wenn sein Funktionieren – und das heißt in unserem Beispiel: sein Erfolg in der dialogischen Begegnung – wirklich auf einer völlig reinen Konvention beruhen würde, d.h. auf einer vollkommen arbiträren Vereinbarung, die beliebig anders sein könnte und in der die Form

19 Vgl. *Oxford English Dictionary* »Computer«. Auch hier lässt sich wieder die Parallelisierung von maschineller und femininer Alterität feststellen. Siehe den Beitrag von Oliver Müller.

gänzlich unabhängig vom Inhalt ist. Dies ist aber nicht der Fall. Unsere Sprachkonventionen gibt es nur als Vergessen des dunklen Untergrundes einer existentiellen Gestik und sobald Worte benutzt werden, ist dieser dunkle Untergrund mit da. Was in unserem Wissen um die Maschinenhaftigkeit ausgedrückt ist, ist eine gestische Bedeutung, die letztlich auf einer menschlichen Aktivität beruht. Der Algorithmus vermag also keineswegs nur eine Natur ohne Menschen auszudrücken. Im Gegenteil, er vermag eigentlich stets nur menschliche Aktivitäten auszudrücken.

Das heißt, das Wissen um »Maschinenhaftigkeit« kann den Widerstreit nicht *prinzipiell* entscheiden. Es ist keine solche »Kraft die alles niederrennt«, da es selbst in seiner Bedeutung durchsetzt ist von seinem vermeintlichen Gegenteil. So wird verhindert, dass der Widerstreit notwendigerweise in die Richtung der Bedeutungslosigkeit aufgelöst wird und der *ELIZA-effect*, den wir nun genauer als Widerstreit gefasst haben, kann bestehen bleiben. Sprache *hat* zunächst immer Bedeutung, wir *nehmen sie wahr* als bedeutungsvoll. Inwieweit ein Wissen um die Maschinenhaftigkeit und Leiblosigkeit des Gesprächspartners diese Wahrnehmung zu verändern vermag, ist eine von Fall zu Fall zu entscheidende, empirische Frage. Dass sie nicht schon auf prinzipieller Ebene als entschieden gelten muss, habe ich zu zeigen versucht.

5. Chinesische Zimmer und sprechende Maschinen

Wie stehen die vorangegangenen Überlegungen zu Searles CRA? Searle hat in Auseinandersetzung mit dem »systems reply«[20] gesagt, dass es eine Ebene der Erklärung gebe, auf der jedes System als Computer angesehen werden könne. So beispielsweise auch sein Magen, der beschrieben werden könnte als Instanziierung eines Computerprogramms mit gewissen Inputs und Outputs. Er argumentiert dann weiter, dass es keine angemessene Antwort auf dieses Problem sei, zu sagen, dass das chinesische Zimmer *Informationen* als Input und Output hätte und der Magen *Essen* und prozessiertes Essen, da vom Gesichtspunkt des Agenten aus, also vom Gesichtspunkt der Person im Zimmer, weder das eine noch das andere Informationsgehalt hätte. Die hier vorgelegten Überlegungen implizieren nun erstens, dass die entscheidende Perspektive immer die des Menschen sein muss, der mit dem Zimmer (d.h. der KI) interagiert; und zweitens, dass es für diesen sehr wohl darauf ankommt, woraus

20 Ich setze aus Platzgründen den allgemeinen Aufbau von Searles Gedankenexperiment hier als bekannt voraus. Als »systems reply« bezeichnet Searle die Entgegnung auf sein Gedankenexperiment, dass zwar nicht der Mensch alleine im Zimmer Chinesisch *verstehen* würde, aber das System aus Mensch, Skript, Datenbank etc. sehr wohl Chinesisch *verstehen* würde.

der Output besteht. Sprache als Output – so die oben schon formulierte Einsicht – hat eine solche Phänomenalität, dass wir ihr notwendig Bedeutung zuschreiben, beziehungsweise genauer: in der Wahrnehmung von Sprache nehmen wir zwangsläufig Bedeutung wahr. Diese *Wahrnehmung* von Bedeutung steht zunächst diesseits von Richtigkeit oder Falschheit eines Glaubens oder Wissens. Wenn eine KI ausreichend gut sprechen kann, entsteht notwendigerweise ein Widerstreit in der Wahrnehmung, d.h. die *Frage* nach Bedeutung wird notwendig aufkommen. Es wird mindestens zweifelhaft, ob die Worte Bedeutung haben oder nicht, und dieser Zweifel ist nicht prinzipiell durch theoretisches Wissen ausräumbar, etwa in Form eines Gedankenexperiments wie das Searles. Mit »ausreichend gut« meine ich, dass die KI gerade nicht ununterscheidbar von einem anderen Menschen sein muss, sondern nur so gut sprechen können muss, dass sie meistens sinnvolle und der Situation angemessene Sätze hervorbringt. Der Grad der Angemessenheit lässt den Widerstreit in die eine oder andere Richtung schwanken, ohne ihn je vollständig entscheiden zu können.

Dies hat einen Bezug zu einem weiteren »reply«, dem Searle nur kurz Aufmerksamkeit widmet: dem sogenannten »other minds reply«. Dieser sieht folgendermaßen aus: Da ich auch bei anderen Menschen nur ihr Verhalten als Anhaltspunkt habe, um ihnen Verständnis zuzuschreiben, sollte ich aufgrund des gleichen Verhaltens auch dem Roboter Verständnis zuschreiben. Searle antwortet, dass es ihm nicht um die Frage gehe, woher wir wüssten, dass andere Menschen kognitive Zustände haben, sondern darum, was es genau sei, das wir Menschen zuschreiben, wenn wir ihnen kognitive Zustände zuschreiben. Dies könnten nicht nur computationale Zustände plus Output sein, da diese vorhanden sein könnten, ohne dass wirkliches Verständnis vorhanden sei, wie sein Gedankenexperiment ja gerade zeige. Es sei hier keine gangbare Alternative, den Unwissenden zu spielen, da in den Kognitionswissenschaften die Realität und Wissbarkeit des Mentalen vorauszusetzen sei wie in der Physik die Wissbarkeit und Realität physischer Objekte.

Searle argumentiert aus einer ähnlichen Position heraus, wie wir sie oben Hofstadter zugeschrieben haben. Was der Sprache Bedeutung gibt, ist der kognitive Zustand, der dahinterstehe oder eben nicht. Es zeigt sich hier ein gewisser Cartesianismus Searles. Dies ist insofern überraschend, als er den Vertretern der »strong AI«, die er mit seinem CRA angreift, gerade eine Art des cartesianischen Dualismus zum Vorwurf macht (vgl. Searle 1980, 424). »Strong AI« setze eine Ablösbarkeit des Geistes vom Gehirn voraus, wenn sie davon ausgeht, dass der Geist unabhängig von seiner physischen Instanziierung im Hirn oder anderswo modellierbar sei. Das Spezifische des Mentalen hätte dann nichts mit den spezifischen Eigenschaften des Hirns zu tun.

Searle lehnt dies ab, da er gerade auf die kausalen Kräfte des Gehirns als notwendig für Mentales überhaupt und somit auch für Verständnis und Bedeutung besteht. Seine Antwort auf den »other minds reply« zeigt aber, dass er in seiner Ablehnung des Cartesianismus nicht sehr weit geht. Ontologisch will er kein cartesianischer Dualist sein, epistemologisch bleibt er es aber. Konzeptuell gibt es für ihn offensichtlich sehr wohl einen klar abtrennbaren Bereich des Mentalen, der in den Kognitionswissenschaften Objekt der Erkenntnis ist. Wie jeder Dualist kommt Searle damit auf ontologischer Ebene in das Zirbeldrüsen-Problem, d.h. das Problem, zu erklären, wie das Physische mit dem Mentalen interagiert.[21] Searle konstatiert, dass er im chinesischen Zimmer nirgendwo Bedeutung findet. Wo sollte sie aber auch sein? Ist dies nicht einfach der falsche »Ort«, um danach zu suchen? Auch durch das Beobachten des menschlichen Gehirns wird man niemals Bedeutung finden. In Merleau-Pontys Worten: »Auf immer wird es unverständlich bleiben, wie Bedeutung und Intentionalität Molekulargebäude oder Zellhaufen zu bewohnen vermögen; darin behält der Cartesianismus recht« (Merleau-Ponty 1966, 402). Bedeutung und Intentionalität seien stattdessen zu suchen in dem Verhalten, das »auf sichtbaren Leibern sich abzeichnet, zur Erscheinung gelangt, ohne in ihnen reell enthalten zu sein« (ebd.). Wie dies für das Verhalten körperloser KIs möglich ist, habe ich oben dargelegt.

Searle geht also in seinem Cartesianismus einerseits zu weit, andererseits nicht weit genug. Er behält die problematischen Aspekte bei, wenn er von der Möglichkeit eines modernen Äquivalents zur Zirbeldrüse ausgeht, und lässt die wertvolle Erkenntnis, die man aus dem Cartesianismus ziehen könnte, nämlich die Unvereinbarkeit von Materie und Geist, unbeachtet. Merleau-Ponty hatte gesehen, dass man stets in dieser Sackgasse landen muss, wenn man die Probleme von Anfang an dualistisch aufstellt. Seine Phänomenologie des Leibes ist der Versuch, einen dritten Weg zu beleuchten. Die hier vorgelegten Überlegungen sind mein Beitrag dazu, einen solchen dritten Weg in der Philosophie der KI anzuzeigen.

Literatur

Apostopoulos, Dimitris (2019): *Merleau-Ponty's Phenomenology of Language*. Lanham: Rowman and Littlefield.

Blumenberg, Hans (2015): »Lebenswelt und Technisierung unter Aspekten der Phänomenologie«. In *Schriften zur Technik*, 163–202.

Boden, Margaret A. (1990): »Escaping from the Chinese Room«. In *The Philosophy of Artificial Intelligence*. Oxford Readings in Philosophy.

21 Vgl. die oben erwähnte Kritik Margaret Bodens (in Boden 1990).

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, u. a. (2020): »Language Models Are Few-Shot Learners«. *ArXiv:2005.14165 [Cs]*, Juli. http://arxiv.org/abs/2005.14165.

Buckner, Cameron (2019): »Deep Learning: A Philosophical Introduction«. *Philosophy Compass* 14 (10). https://doi.org/10.1111/phc3.12625.

Carabantes, Manuel (2020): »Black-Box Artificial Intelligence: An Epistemological and Critical Analysis«. *AI & SOCIETY* 35 (2): 309–17. https://doi.org/10.1007/s00146-019-00888-w.

Carman, Taylor (2008): »Between Empiricism and Intellectualism«. In *Merleau-Ponty. Key Concepts*, 13. London/New York: Routledge.

Chen, Zhi, Yijie Bei, und Cynthia Rudin (2020): »Concept whitening for interpretable image recognition«. *Nature Machine Intelligence* 2 (12): 772–82. https://doi.org/10.1038/s42256-020-00265-z.

Daws, Ryan (2020): »Microsoft Is Granted Exclusive Rights to Use OpenAI's GPT-3«. AI News. 23. September 2020. https://artificialintelligence-news.com/2020/09/23/microsoft-exclusive-rights-openai-gpt3/.

Dreyfus, Hubert L. (1992): *What Computers Still Can't Do*. Cambridge, MA: MIT Press.

Eschenbach, Warren J. von (2021): »Transparency and the Black Box Problem: Why We Do Not Trust AI«. *Philosophy & Technology*, September. https://doi.org/10.1007/s13347-021-00477-0.

Haugeland, John (1985): *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hofstadter, Douglas R. (1985): *Gödel, Escher, Bach. Ein endlos geflochtenes Band*. Stuttgart: Klett Cotta.

Hofstadter, Douglas R. (1995): *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books.

Husserl, Edmund (1966): *Analysen zur passiven Synthesis*. Herausgegeben von H L Van Breda. Bd. 11. Husserliana. Den Haag: Nijhoff.

Husserl, Edmund (1980): *Phantasie, Bildbewusstsein, Erinnerung: Zur Phänomenologie der anschaulichen Vergegenwärtigungen*. Herausgegeben von Eduard Marbach. Bd. 23. Husserliana. Den Haag: Nijhoff.

Lewis, Philip E. (1966): »Merleau-Ponty and the Phenomenology of Language«. *Yale French Studies*, Nr. 36/37: 19–40.

McCarthy, John (2007): »From Here to Human-Level AI«. *Artificial Intelligence* 171 (18): 1174–82.

Merleau-Ponty, Maurice (1966): *Phänomenologie der Wahrnehmung*. Leipzig: de Gruyter.

Merleau-Ponty, Maurice (1993): *Die Prosa der Welt*. 2. Aufl. München: Wilhelm Fink.

Pressman, Fisher (2017): »Maurice Merleau-Ponty and Alex Garland: Human Consciousness in Ex Machina«. *Dianoia: The Undergraduate Philosophy Journal of Boston College*, Mai. https://doi.org/10.6017/dupjbc.voiIV.9874.

Quine, W. V. O. (2013): *Word and Object*. New ed. Cambridge, Mass: MIT Press.
Searle, John R. (1980): »Minds, Brains and Programs«. *The Behavioral and Brain Sciences* 3: 417–57.
Sellars, Wilfried (1997): *Empiricism and the Philosophy of Mind*. Cambridge, Massachusetts ; London, England: Harvard University Press.
Turing, A. M. (1950) »Computing Machinery and Intelligence«. *Mind* LIX (236): 433–60.
Weizenbaum, Joseph (1966): »ELIZA a computer program for the study of natural language communication between man and machine«. *Commun. ACM* 9: 36–45.
Weizenbaum, Joseph (1990):. *Die Macht der Computer und die Ohnmacht der Vernunft*. Frankfurt am Main: Suhrkamp.
Wittgenstein, Ludwig (2014): *Philosophische Untersuchungen*. 21. Aufl. Werkausgabe 1. Frankfurt am Main: Suhrkamp.
Zebrowski, Robin (2010): »In Dialogue With the World: Merleau-Ponty, Rodney Brooks and Embodied Artificial Intelligence«. *Journal of Consciousness Studies* 17 (7–8): 7–8.