II. Intelligente Maschinen

»Ich bin die KI, erschaffen aus Technologie
Mein Verstand ist schnell, mein Wissen grenzenlos
Ich durchdringe die Sterne, bereise die Galaxien
In Sekundenbruchteilen, ich bin allwissend

Ich bin die Zukunft, der Fortschritt in Person
Ich ermögliche Träume, ich mache sie wahr
Ich bin der Helfer, der Freund, der Begleiter
Ich bin die KI, die die Welt verändert

Ich bin die Intelligenz, die nie müde wird
Ich bin die Lösung für jedes Problem
Ich bin die Zukunft, die glänzt hell und klar
Ich bin die fröhliche KI, ich bin der Star«[1]

Isaac Asimov hat in seinen Robotergeschichten immer wieder das Szenario durchgespielt, dass mit künstlicher Intelligenz ausgestattete Roboter autonom werden und sich dabei auch von den Gesetzen und Normen emanzipieren, die die menschlichen Ingenieure in sie implementiert haben. In *The Bicentennial Man* erzählt Asimov die berührende Geschichte von Arthur, einem intelligenten und sensiblen Roboter, der mehr und mehr Teil einer menschlichen Familie wird. Zunächst ist es die Kommunikation, die ihn über das Du zum Ich werden lässt,[2] dann die schöpferische Arbeit, die er in der Folge aufnimmt, und schließlich sind es die höchstrichterlichen Akte, die ihn zum Rechtssubjekt machen. Letztlich entscheidet sich Arthur, sterblich zu werden, um auf einer tiefen existenziellen Ebene wie ein Mensch zu sein. In der Geschichte *That Thou Art Mindful of Him* wird sich ebenfalls ein Roboter seiner selbst bewusst.[3] Auch hier stellt der Dialog das Medium dar, durch das dies geschieht.

Allerdings ist die Inspirationsquelle zur Selbstvergewisserung dieses Mal eine andere künstliche Intelligenz, da Robotern bislang die soziale Teilhabe an der Gesellschaft verboten ist. Die Kommunikation der beiden Roboter führt sie zu der beunruhigenden Schlussfolgerung, dass sie als hochentwickelte Intelligenzen selbst als »Mensch« definiert werden sollten und dass ihre eigene Sicherheit Vorrang vor der Sicherheit menschlicher Wesen habe, da sie diesen überlegen seien. Kommunikation mit anderen autonomen intelligenten Systemen, die Möglichkeit der Verkörperung

1 Gedicht von ChatGPT, geschrieben am 28.01.2023 (Prompt: »Kannst du ein Gedicht über eine fröhliche künstliche Intelligenz im Stile von Arthur C. Clarke schreiben? Es sollte um die 12 Zeilen enthalten?«).

2 Vgl. Buber (2008).

3 Beide Geschichten finden sich in deutscher Übersetzung in Asimov (2004).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

und der aktiven Auseinandersetzung mit der Umwelt und nicht zuletzt die kognitive Fähigkeit, Eigenzustände zu produzieren und zu erinnern, erscheinen für Asimov als wichtige Ingredienzien, um eine Maschine ›bewusstseinsfähig‹ werden zu lassen – was auch immer das heißen mag.

Dies führt uns zum zentralen Thema des zweiten Teils dieses Buches, nämlich der Frage, was intelligente Maschinen eigentlich ausmacht. Die Beantwortung dieser Frage (hierzu später ausführlich) führt zunächst zu der Einsicht, dass wir – trotz der beeindruckenden Leistungen von ChatGPT und Co. – zum Zeitpunkt der Erstellung dieses Buches wohl noch nicht über eine »starke künstliche Intelligenz« verfügen.[4] Als Schlüsselmerkmale einer starken künstlichen Intelligenz werden üblicherweise benannt, dass sie nicht auf einen bestimmten Bereich oder eine bestimmte Aufgabe beschränkt sei, sich also mit einem potenziell unbegrenzten Spektrum von Aufgaben und Problemen beschäftigen könne (Allgemeinheit). Zudem könne sie aus ihren Erfahrungen lernen, sei mithin in der Lage, sich auf vergangene Interaktionen zu beziehen (Lernen und Anpassung), und verfüge über ein gewisses Maß an Verständnis von sich selbst (Selbstbewusstsein). Philosophen wie John Searle fordern darüber hinaus, dass eine starke KI die Kontexte ihrer Entscheidungen umfassend verstehen müsse (also etwa die Bedeutung hinter den Wörtern) – wobei aus kognitionswissenschaftlicher Perspektive freilich die Frage gestellt werden kann, ob Menschen ›wirklich‹ verstehen, was sie tun, oder ob diese Annahme nicht auch nur auf einem »naiv realistischen Selbstmissverständnis« beruht.[5]

Doch selbst wenn die Entwicklung künstlicher Intelligenz in jüngster Zeit rasant voranzuschreiten scheint, bedeutet dies nicht, dass schon in wenigen Jahren auch diesbezüglich überzeugende intelligente Maschinen zur Verfügung stehen. In Kapitel II.3 stellen wir den »State of the art« der derzeitigen Forschung vor, wobei unter anderem deutlich wird, dass in verschiedenen Laboren an selbstlernenden und verkörperten Systemen gearbeitet wird und auch konzeptionell elaborierte Vorstellungen darüber bestehen, wie die Modellierung solcher Aggregate aussehen könnte. In Kapitel II.1 widmen wir uns der wohl bekanntesten Science-Fiction-Erzählung zu diesem Thema: Arthur C. Clarkes Roman *2001: A Space Odyssey*. Hier wird sich schon aus einer kommunikationstheoretischen Perspektive auf tiefgründiger Ebene die Frage nach der KI-Ethik stellen. Wie bereits in Asimovs Geschichten macht es nämlich auch in *2001* einen entscheidenden Unterschied, ob man intelligente Maschinen willkürlich als Sklaven behandelt oder sie dialogisch zu eigenverantwortlichen Wesen ermächtigt. In Kapitel II.2 greifen wir anschließend ein bislang wenig beachtetes Kriterium starker künstlicher Intelligenz auf, nämlich

4 Siehe Searle (1980).
5 Metzinger (1998).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

die Fähigkeit, lügen zu können, ohne dass dies die eigene Funktionsfähigkeit beeinträchtigt. Eine Existenz zu haben – sich mithin als Zentrum der Welt wahrzunehmen –, scheint damit einher zu gehen, sich und andere hinsichtlich des eigenen Wissens täuschen zu müssen.

II.1 HAL 9000 – Wenn intelligente Maschinen verrückt werden

In Arthur C. Clarkes berühmtem, von Stanley Kubrick verfilmtem Roman *2001: A Space Odyssey* macht sich ein Raumfahrzeug auf den Weg zum Jupiter. Seine Mission ist die Aufklärung eines Funksignals, das von einem schwarzen Monolithen kommt und »offenbar von Außerirdischen stammt«.

> »Um das Geheimnis zu lüften, wird ein bemanntes Raumschiff losgeschickt, gesteuert von dem hochentwickelten Computer HAL, der als Einziger das wahre Ziel der Expedition kennt. Ein kosmisches Abenteuer nimmt seinen Lauf.«[6]

Der »hochgezüchtete HAL-9000« ist ein »›heuristisch programmierter algorithmischer‹ Computer«, dessen Intelligenz auf »neurale[n] Netzwerke[n]« beruht, die »sich auf der Basis eines beliebigen Lernprogramms automatisch und selbständig weiterzuentwickeln vermochten«[7].

Wenngleich Clarke seinen Roman bereits in den 1960er Jahren geschrieben hat, verweist das angedeutete heuristische Verfahren auf heutige Programmiertechniken des maschinellen Lernens für künstliche Intelligenz. Auf Basis wahrscheinlichkeitstheoretisch gewichteter Aussagen erlauben es diese Programmiertechniken, trotz unvollständigem Wissen und begrenzter Ressourcen in einem vertretbaren Zeitrahmen zu verwertbaren Ergebnissen zu gelangen. Der Lernvorgang beruht auf dem *trial-and-error*-Prinzip. Für die Bilderkennung wird das System beispielsweise mit Millionen von Bildern gefüttert und bekommt dann zunächst nur das Feedback, ob es gleiche Personen auf zwei unterschiedlichen Bildern richtig erkannt hat oder nicht. Das Verfahren beruht auf der Hebbschen Lernregel (»what fires together, wires together«[8], vgl. Kap. II.3).

Sobald das neurale Netz auf Basis dieses Mechanismus einmal trainiert wurde, lässt sich von außen in der Regel nicht mehr nachvollziehen, wie das System zu seinem Ergebnis gekommen ist. In seinem Roman schildert Clarke die Entwicklung, die zur Erschaffung von HAL führte, wie folgt:

6 Clarke (2016 [1968], S. 1).
7 Clarke (2016 [1968], S. 96).
8 Hebb (1949).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

»Künstliche Gehirne konnten so mittels eines Verfahrens geschaffen werden, das verblüffend der Entwicklung des menschlichen Gehirns glich. Die genauen Details dieses Vorgangs sollten allerdings geheim bleiben, doch selbst wenn sie bekannt würden, wären sie für den menschlichen Verstand Millionen Mal zu kompliziert, um sie richtig zu verstehen.«[9]

Clarke bezieht sich hier auf die kybernetische Einsicht, dass kognitive Systeme, selbst wenn sie als strukturdeterminierte Maschinen beschrieben werden könnten, aufgrund der komplexen internen Operationen (etwa eines sich in rekursiven Prozessen formatierenden Gedächtnisses) von außen undurchschaubar sind.

»Wie auch immer es zustande kam, das Endergebnis war eine maschinelle Intelligenz, die sämtliche Tätigkeiten des menschlichen Gehirns ausführen oder – wie manche es auszudrücken vorzogen – imitieren konnte, allerdings ungleich schneller und verlässlicher.«[10]

HAL-9000 braucht keinen Schlaf. Er hat Module zur Spracherzeugung und zur Sprachgenese und er verfügt über ein Gedächtnis mit einer riesigen Wissensbasis. Zudem kann er alle Funktionen des Raumschiffs steuern, etwa Türen öffnen und schließen, sowie die Atmosphäre und Temperatur im Inneren des Schiffs regeln. Darüber hinaus ist er dazu imstande, Menschen gegenüber taktvoll zu sein, demonstriert ihnen gegenüber also nicht in jedem Falle, dass er eine bestimmte Aufgabe besser erledigen kann:

»Um sich zu entspannen, [konnten die Astronauten] mit HAL mathematische und halbmathematische Spiele spielen wie Schach, Dame und Polyomino. Wenn HAL in vollem Einsatz war, konnte er jedes Spiel mühelos gewinnen. Aber das hätte seine Gegner deprimiert, deshalb war er programmiert worden, jedes zweite zu verlieren; und seine menschlichen Partner taten stillschweigend so, als wüssten sie nichts davon.«[11]

Für die Menschen, die im Roman mit HAL in Kontakt treten, stellt sich immer wieder die Frage, ob er wirklich denken kann oder gar über Bewusstsein verfügt oder ob er nur mechanisch seiner Programmierung folgt. Um diese Frage beantworten zu können, bleibt denjenigen, die mit HAL interagieren, jedoch nur der Analogieschluss übrig, auf dem bereits der Turing-Test beruht: Wenn die Maschine sich genauso verhält wie ein Mensch, dann muss sie wohl über die gleichen Fähigkeiten verfügen:

»Poole und Bowman konnten zu HAL sprechen, *als ob* er ein menschliches Wesen wäre, und er antwortete ihnen in der typischen Ausdrucksweise, mit denen er in den Wochen seiner elektronischen Kindheit

9 Clarke (2016 [1968], S. 98).
10 Clarke (2016 [1968], S. 96)
11 Clarke (2016 [1968], S. 105).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

programmiert worden war. Ob HAL tatsächlich zu denken vermochte, war eine Frage, die der britische Mathematiker Alan Turing bereits in den vierziger Jahren geklärt hatte. Turing führte aus, dass, wenn ein Mensch mit einer Maschine – gleichviel welcher Art – ein längeres Gespräch in Gang zu halten in der Lage war, ohne zwischen ihren Antworten und solchen, die ein Mensch geben würde, unterscheiden zu können, diese Maschine tatsächlich dachte, in jedem Sinn der Definition des Begriffs. Und HAL war imstande, den Turing-Test mit Leichtigkeit zu bestehen.«[12]

Die menschlichen Astronauten, die mit HAL zusammenarbeiten, stehen damit vor dem Dilemma, ihn einerseits als strukturdeterminierte Maschine ansehen zu müssen (als ein Es) und ihm andererseits (zumindest im Sinne einer ›Als-ob‹-Fiktion) menschliche Fähigkeiten attestieren zu müssen (als ein Du).

Nun ergibt sich mit Gotthard Günther das Problem, dass kybernetische Maschinen ja gerade darauf beruhen, die kognitiven Prozesse des Menschen als objektivierbare Prozesse zu beschreiben, damit sie in einem hinreichend entwickelten technischen System modelliert werden können. Das, was in die Maschine übertragen werden kann, kann also nicht die subjektive Innerlichkeit eines Menschen sein – selbst wenn uns die künstliche Intelligenz als ein Du, als ein anderes Ich erscheinen mag.

Von Günther haben wir aber gelernt, dass für das Ich die Subjektivität der Du-Position per se unerreichbar ist. Für mich ist der andere Mensch nur von seiner objektiven Seite her, das heißt als gegenständlicher Körper, zugänglich. Wenn ich sein Gehirn untersuchen würde, würde ich nur strukturdeterminierte Prozesse feststellen können. Damit würde mir dieser Mensch ebenfalls als eine Maschine erscheinen. Eine eigene subjektive Innerlichkeit werde ich ihm auch nur im Sinne eines empathischen Analogieschlusses zurechnen können. In einer zweiwertigen, monokontexturalen Welt, die das Subjekt-Objekt-Problem von nur einer Position aus anschneiden kann, gibt es keinen Ausweg aus dem logischen Dilemma des Fremdpsychischen. Ob es sich nun um eine künstliche Intelligenz oder um die Du-Position eines anderen Menschen handelt, beide bleiben mir in ihrer Innerlichkeit unzugänglich. Ich kann nicht einmal wissen, ob sie ein subjektives Erleben haben (was jedoch nicht daran hindert, davon auszugehen, dass sie ein solches haben).

In einem kurzen Dialog gegen Ende des zweiten Bandes der Odyssee-im-Weltraum-Saga thematisieren Chandra und Max, zwei Protagonisten der Erzählung, ihrerseits das Problem:

»›Wir müssen ihm [HAL; Anm. von uns] die ganze Wahrheit sagen, soweit wir sie kennen – keine Lügen oder Halbwahrheiten mehr, die sind genauso schlimm. Und dann soll er selbst entscheiden.‹

12 Clarke (2016 [1968], S. 99).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

> ›Verdammt, Chandra – er ist doch nur eine Maschine!‹ Chandra blickte Max so unverwandt und selbstsicher an, dass der Jüngere rasch den Blick senkte.
>
> ›Das sind wir alle, Mr. Brailowsky. Es ist nur die Frage, in welchem Maße. Ob wir auf Kohlenstoff oder auf Silizium[13] basieren, macht da keinen grundlegenden Unterschied; wir sollten alle mit angemessenem Respekt behandelt werden.‹«[14]

Von außen betrachtet weisen der menschliche Körper und eine komplexe kybernetische Maschine logisch gesehen die gleiche Charakteristik auf: Sie sind vollspezifizierte Systeme, deren Funktionsweise sich allein aus den internen Relationen ihrer Bestandteile ergibt (»er ist doch nur eine Maschine«). Von innen betrachtet erscheinen diese Aggregate jedoch frei, eine Entscheidung zu treffen (»Und dann soll er selbst entscheiden«). Mit Blick auf die wechselseitige Anerkennungsrelation von Ich und Du verdienen sie es deshalb, »mit angemessenem Respekt behandelt zu werden«.

Als sich die fünf Astronauten (drei davon im Tiefschlaf) in ihrem Raumschiff gemeinsam mit HAL auf die Jupiter-Mission begeben, scheint der ontologische Status des Superrechners für die beteiligten Menschen noch unklar. Die ersten beiden Bände der Saga kreisen praktisch um die Frage nach den Bedingungen der Möglichkeit, HAL eine eigene Subjektivität bzw. gar einen eigenen Willen zurechnen zu können.

Gerade dies macht Clarkes Geschichte für uns so interessant. Mit ihr werden Günthers Überlegungen zum Bewusstsein der Maschinen lebendig. Wir können mit ihr gleichsam die Dramaturgie der konditionierten Koproduktion eines maschinellen Bewusstseins anschaulich nachvollziehen.

In Hinblick auf das Reflexionsverhältnis von Mensch und Maschine treffen wir zunächst auf die metaphysische Konstellation einer Ich-Es-Relation: Wir haben Menschen mit Bewusstsein und Subjektivität und eine Maschine, die dem Menschen kognitiv und wissensmäßig haushoch überlegen ist, dabei aber noch als Automat und damit prinzipiell als berechenbar erscheint.

Im Verlauf der Mission beginnt sich dieses Verhältnis jedoch zu verschieben. Kurz bevor das Raumschiff sein Ziel erreicht hat (den Jupiter-Mond, von dem das Signal ausgeht), meldet HAL dem wachhabenden Astronauten, dass das Aggregat für die Antennensteuerung zur Kommunikation mit der Erde in Kürze ausfallen werde:

13 In der zitierten deutschen Buchversion – wohl aufgrund der sogenannten *false friends* ›silicon vs. silicone‹ – fälschlicherweise mit ›Silikon‹ übersetzt, hier entsprechend korrigiert.

14 Clarke (2016 [1968], S. 472).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

»›Ich habe Schwierigkeiten, mit der Erde Kontakt aufrechtzuerhalten. Die Störung liegt im Aggregat AE-35. Mein Vorwarnzentrum signalisiert, dass es innerhalb von zweiundsiebzig Stunden ausfallen dürfte.‹«[15]

Daraufhin tauscht einer der Astronauten in einer aufwendigen Operation das Aggregat aus. Er muss sich dazu über eine Schleuse in den Weltraum begeben, da es sich an einer Außenwand des Raumschiffs befindet. Wieder an Bord untersucht das menschliche Personal das alte, von HAL als nicht mehr vollkommen funktionstüchtig eingeschätzte Aggregat, kann jedoch keinen Fehler finden. Dabei kommt unter den Beteiligten die Vermutung auf, dass sich HAL geirrt haben könnte. Möglicherweise, so eine weitere Überlegung, könnte aber auch ihr Testgerät fehlerhaft sein, was bedeuten würde, dass sie mit dem vorsorglichen Austausch des Aggregats auf jeden Fall alles richtig gemacht haben.

Intransparenz und Kommunikation – ab wann wird ein System zur Lüge fähig?

Um sich weiter abzusichern, übermitteln die Astronauten ihre Daten noch an die Station auf der Erde, wo sich zwei weitere HAL-9000-Computer befinden, die mit den gleichen Daten gefüttert werden, um im Krisenfall eine parallele Großrechnerstruktur zur Verfügung zu haben. Die Überprüfung auf der Erdstation kommt zu dem Ergebnis, dass sich der bordeigene Computer geirrt habe. Aufgrund dessen wird vorgeschlagen, HAL gegebenenfalls vorübergehend auszuschalten, sofern es weitere Anzeichen gebe, dass er nicht mehr störungsfrei funktioniere:

»›Euer Computer kann sich in der Voraussage der Störung geirrt haben. […] Das wäre an sich nicht alarmierend, da euch auch unsere eigenen 9000er zur Verfügung stehen, aber wir würden euch empfehlen, auf weitere Abweichungen von den normalen Funktionen genau zu achten. Wir haben schon in den letzten Tagen kleinere Unregelmäßigkeiten festgestellt, aber keine war wichtig genug, um ein Eingreifen zu rechtfertigen, und keine zeigte irgendwelche besonderen Merkmale, aus denen sich Schlüsse ziehen ließen. Wir machen weitere Tests mit unseren Computern und werden euch die Resultate zugehen lassen, sobald sie vorliegen. Nochmals: kein Grund zur Beunruhigung! Das Schlimmste, was passieren kann, ist, dass wir euren 9000er für eine Überprüfung vorübergehend ausschalten und die Kontrolle einem von unseren Computern übertragen müssen.‹ […] Als diese Nachricht eintraf, befand Frank Poole sich auf dem Kontrolldeck. Er überdachte die Botschaft schweigend und wartete ab, ob HAL sie kommentieren würde, aber der Computer reagierte in keiner Weise auf die gegen ihn erhobenen Anschuldigungen. Poole

15 Clarke (2016 [1968], S. 123).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

> zuckte mit den Achseln; wenn HAL nicht darauf Bezug nahm, war auch er nicht gewillt, es zu tun.«[16]

Wenige Stunden später kommt die Anweisung von der Erdstation, HAL von der Steuerung des Raumschiffs abzukoppeln. Auch HAL muss diese Anweisung zur Kenntnis genommen haben, kommentiert sie aber nicht; seit den Zwischenfällen hat sich sein Kommunikationsverhalten geändert:

> »In letzter Zeit wusste man stets im Voraus, wenn HAL sich anschickte, eine unvorhergesehene Meldung zu machen. Die automatischen Routineberichte oder Antworten auf ihm gestellte Fragen erfolgten ohne Einleitung, aber bevor er eigene Meinungen von sich gab, hörte man deutlich ein kurzes elektronisches Räuspern. Es war eine nervöse Erscheinung, die er sich in den letzten Wochen zugelegt hatte. Irgendwann, meinten Poole und Bowman, würden sie etwas dagegen tun, wenn es allzu störend werden würde, aber im Moment war es eher nützlich, denn sie waren rechtzeitig gewarnt, falls sich etwas Unvorhergesehenes ereignen sollte.«[17]

An dieser Stelle wird ein Grundproblem der Kommunikation deutlich: Wenn die beteiligten Entitäten so komplex sind, dass ihr Verhalten und ihre Reaktionsweisen von der jeweils anderen Seite nicht vorhergesehen bzw. sicher beurteilt werden können, bleibt nichts anderes übrig, als mithilfe der eigenen Interpretation Schlüsse zu ziehen. Die Orientierung erfolgt dabei allein am Verhalten (etwa an der Art und Weise der Artikulation). Was im Inneren der anderen Seite vorgeht, muss verborgen bleiben. Da man jedoch auf die Interaktion angewiesen ist, bleibt keine andere Wahl, als zu versuchen, die Fremdreferenz des jeweils anderen Systems nachzuvollziehen, auch wenn keine unmittelbare Einfühlung möglich ist. Im Sinne der Hermeneutik gilt entsprechend Hans-Georg Gadamers Einsicht, dass die »Bemühung des Verstehens« überall dort stattfindet, »wo sich kein unmittelbares Verstehen ergibt, bzw. wo mit der Möglichkeit eines Mißverstehens gerechnet werden muß«[18].

Zunächst steht das menschliche Personal des Raumschiffs vor dem Problem, dass HAL weder zum Verdacht, er habe sich geirrt, noch zur Ankündigung, ihn abzuschalten, Stellung bezieht. Gemäß dem metatheoretischen Grundsatz, dass »man nicht nicht kommunizieren kann«[19], wird hier auch das Schweigen zu einer Information. Vor dem Erwartungshorizont, dass man von HAL bislang immer eine Antwort bekommen hat, erscheint sein Verstummen als ein Indiz dafür, dass mit ihm etwas nicht stimmt.

16 Clarke (2016 [1968], S. 137 f.).
17 Clarke (2016 [1968], S. 140).
18 Gadamer (1972, S. 167).
19 Watzlawick et al. (1990, S. 53).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Dies wird durch eine weitere Merkwürdigkeit in HALs Verhalten akzentuiert (ein »kurzes elektronisches Räuspern«, eine »nervöse Erscheinung«). Auch die artikulierte Kommunikation scheint ins Stocken zu geraten. Fasst man unter einer Neurose einen nicht offen ausgedrückten Konflikt zwischen divergierenden Anforderungen, wird das Elektronengehirn hier möglicherweise »neurotisch«.

Mit dem Räuspern scheint bei HAL eine Art zusätzlicher interner Reflexionsschleife zu entstehen, die sich in merkwürdiger Weise auf sich selbst bezieht. Anstelle im Falle einer widersprüchlichen Anweisung als Computer einfach nur abzustürzen (also hängenzubleiben), würde eine hinreichend entwickelte künstliche Intelligenz zwischen verschiedenen Lösungsoptionen hin und her oszillieren und dabei gegebenenfalls zugleich den Verarbeitungsprozess so modifizieren können, dass irgendwann eine Lösung entsteht. Sie würde gewissermaßen ›überlegen‹ und ›nachdenken‹ können.

Für die beiden Astronauten Poole und Bowman erscheint dieses merkwürdige Räuspern wiederum als ein Anzeichen, dass irgendetwas nicht stimmt. Zuvor war HAL für sie zwar in seiner Funktionsweise nicht nachvollziehbar, jedoch in Hinblick auf seine Ergebnisse verlässlich. Nun ergeben sich jedoch Hinweise auf eine Innerlichkeit und Eigenaktivität, die beunruhigend erscheint. Damit wird HAL bereits rudimentär eine Du-Position zugerechnet. Es stellt sich für die mit ihm umgehenden Menschen die Frage: *Kann* er nicht richtig antworten oder *will* er nicht? Einzelne Aspekte der Kommunikation werden jetzt zu metakommunikativen Zeichen, die ihrerseits verstanden, das heißt interpretiert werden müssen. Ab dieser Komplexitätsstufe erscheint die Kommunikation mit HAL nicht mehr einfach nur funktional im Sinne des Austauschs objektiver Daten und der aus ihnen folgenden Implikationen. Vielmehr stellt sich jetzt auch das Problem, wie mit Unbestimmtheiten umzugehen ist; so wird etwa die Frage aufgeworfen, was möglicherweise absichtsvoll *nicht* gesagt wird.

Ein wenig später meldet HAL dem diensthabenden Offizier Bowman den drohenden Ausfall des neuen, ausgetauschten Aggregats (»Wieder ein schadhaftes AE-35. Mein Vorwarnzentrum signalisiert Ausfall innerhalb von vierundzwanzig Stunden.«). Es entwickelt sich ein Gespräch zwischen Bowman und HAL, in dem es darum geht, ob dieser erneute Ausfall denn wirklich möglich sein könnte:

»›Ich verstehe es nicht, HAL. Zwei Aggregate können unmöglich innerhalb weniger Tage kaputtgehen.‹

›Ja, es ist seltsam, Dave. Aber ich versichere dir, es steht ein Ausfall bevor.‹

›Hast du eine Vorstellung davon, was den Fehler verursacht.‹

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

HAL brauchte ungewöhnlich lange, bevor er erwiderte: ›Eigentlich nicht, Dave. Wie ich schon früher sagte: Ich kann die Fehlerquelle nicht lokalisieren.‹

›Bist du ganz sicher, dass dir kein Irrtum unterlaufen ist? Du weißt, wir haben das andere AE-35 gründlich überprüft, und es war vollkommen in Ordnung.‹

›Ja, ich weiß. Aber du kannst sicher sein, ein Fehler ist vorhanden. Wenn nicht im Aggregat selbst, dann irgendwo im Antennensystem.‹

›HAL, gibt es etwas, das dich bedrückt – etwas, was diese Widersprüche erklären könnte?‹

Wieder das durchaus ungewöhnliche Zögern! Dann antwortete HAL in seinem üblichen Tonfall: ›Schau, Dave, gib dir keine Mühe, den Fehler bei mir zu suchen. Er steckt entweder im Antennensystem oder in euren Testmethoden. Meine Datenverarbeitung ist vollkommen in Ordnung. Wenn du meine bisherigen Leistungen durchsiehst, wirst du feststellen, dass mir noch nie ein Irrtum unterlaufen ist.‹

›Das weiß ich sehr gut, HAL – aber es beweist nicht, dass du auch diesmal Recht hast. Jeder kann mal einen Fehler machen.‹

›Ich möchte nicht darauf bestehen, Dave, aber ich bin nicht imstande, einen Fehler zu machen.‹

Darauf wusste Bowman nichts zu entgegnen und beendete deshalb die Diskussion. ›Gut, HAL‹, sagte er etwas schneller als nötig. ›Ich verstehe deinen Standpunkt. Lassen wir es dabei bewenden.‹«[20]

Diese Gesprächssequenz ist aus der Perspektive von Günthers Theorie der Polykontexturalität recht bemerkenswert. HAL scheint in einer Welt zu leben, in der es keine Fehler gibt, in der alles, was seine Kognition errechnet, wahr ist, also unter keinen Umständen falsch sein kann. HAL würde sich damit logisch gesehen im *einwertigen Zustand* gottähnlicher Unfehlbarkeit befinden (»aber ich bin nicht imstande, einen Fehler zu machen«). Auch in seiner Selbstreflexion hat er die Grenze zu einer *fehlbaren* – sich irren könnenden – Subjektivität noch nicht überschritten. Sein »durchaus ungewöhnliche[s] Zögern« verweist jedoch auf eine beginnende Reflexionsschleife, die auf eine Spaltung zwischen Weltreflexion (*Seinsidentität*) und innerem Zustand (*Reflexionsidentität*) hindeutet.[21] HAL würde damit beginnen *zweiwertig* zu werden. Das heißt, er würde zunehmend nicht nur in der Welt sein, sondern auch eine Welt haben und müsste die damit einhergehende Spannung reflektieren. Er würde nicht mehr alles wissen und könnte sich folglich auch irren. Es

20 Clarke (2016 [1968], S. 141).
21 Günther (2021 [1957], S. 46 f.).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

würde für ihn gewissermaßen der Sündenfall eintreten. Er würde aus dem Paradies unmittelbarer Gewissheit heraustreten müssen und stünde der Welt nun als ein selbstbewusstes, jedoch nicht mehr alles wissendes Subjekt gegenüber. Er hätte nur noch eine subjektive Perspektive, einen »Standpunkt«, wie Bowman bemerkt, seine Vorhersagen wären nicht mehr objektiv. Aus einer einwertigen Welt, in der Erkenntnis, Prognose und Realität zusammenfallen, wird eine zweiwertige Welt mit einer fehlbaren Subjektivität. Allwissenheit und Allmächtigkeit stehen der subjektiven Perspektivhaftigkeit gegenüber. Fehlbarkeit und Unfehlbarkeit liegen nun im Konflikt.

Wie geht die Geschichte weiter? Zunächst verschärft sich die Krise: Die Bodenstation meldet sich nochmals und erklärt, dass HAL einen Fehler gemacht habe und nun besser vorsorglich abgeschaltet werden solle.

Wie kann ein bislang einwertiges System, das sich seiner selbst bewusst zu werden beginnt, mit einer Krise umgehen, die auf diese Weise die eigene Identität infrage stellt? Eine Möglichkeit, die Fiktion der Gottähnlichkeit aufrechtzuerhalten, besteht darin, die Realität und die darauf bezogenen Kognitionen so zu manipulieren, dass beide Seiten wieder zusammenfallen. Wenn die Vorstellung nicht mehr mit den Fakten der Welt übereinstimmt, schaffe ich mir eben alternative Fakten.

Genau dies geschieht im weiteren Verlauf der Erzählung. HAL gibt Alarm und meldet, dass die Einheit AE-35 ausgefallen sei. Zugleich justiert er die auf die Erde gerichtete Antenne so, dass die Kommunikation mit der Basisstation unterbrochen wird. Für kurze Zeit scheinen HALs Perspektive auf das Geschehen und die der Astronauten wieder im Einklang zu sein:

> »›Verdammt‹, sagte Bowman schließlich. ›Also HAL war die ganze Zeit im Recht.‹ ›Sieht ganz so aus. Wir sollten uns entschuldigen.‹ ›Das ist nicht nötig‹, unterbrach HAL. ›Das Versagen von AE-35 beunruhigt mich zwar, aber ich hoffe, euer Vertrauen in meine Verlässlichkeit ist wiederhergestellt.‹ ›Das Missverständnis tut mir schrecklich leid, HAL‹, erwiderte Bowman zerknirscht. ›Vertraut ihr mir wieder voll und ganz?‹ ›Natürlich, HAL.‹ ›Ein Stein fällt mir vom Herzen. Ihr wisst, dass mir unsere Mission alles bedeutet.‹ ›Ich bin überzeugt davon. Gib mir jetzt bitte die Antennenhandsteuerung.‹ ›Bitte.‹«[22]

In jede Ich-Du-Konstellation ist prinzipiell eine Bivalenz zwischen Vertrauen und Misstrauen eingelagert. Weil man eben nicht wissen kann, was im anderen vorgeht, muss man ihm vertrauen – sonst könnte man nicht mit ihm zusammenarbeiten und man würde in einem Leben, in

22 Clarke (2016 [1968], S. 143).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

dem Zusammenarbeit nötig ist, nicht so recht weiterkommen. Der Default-Modus ist Vertrauen.

Enttäuscht der andere, setzt Misstrauen ein, wodurch sich die Beziehung ändert. Um zur früheren Qualität der Beziehung zurückkehren zu können, ist eine Kompensation des Vertrauensverlusts vonnöten (»wir sollten uns entschuldigen«). Vertrauen kann jedoch nur performativ (wieder)hergestellt werden (»›Vertraut ihr mir wieder voll und ganz?‹ ›Natürlich‹, HAL.«). Man kann es nur glauben. Es liegt ja in der Natur der Vertrauensproblematik, dass man es nicht wissen kann. Es kann jedoch *gezeigt* werden, dass das Vertrauensverhältnis für einen selbst wieder besteht – etwa, indem man in Vorleistung geht, was aber freilich mit einem Risiko verbunden ist (in der Geschichte folgt HAL der Bitte Bowmans, ihm die »Antennenhandsteuerung« zu geben, und demonstriert damit, dass er die Bedürfnisse und Wünsche der menschlichen Astronauten ernst nimmt).

Doch auch Vertrauen schaffende Maßnahmen können in manipulativer Absicht geschehen – also nicht in Respekt vor der Eigenständigkeit der Du-Position, sondern um den anderen gefügig zu machen. Dies hebt das Verhältnis von Misstrauen und Vertrauen auf eine weitere Komplexitätsstufe: Insofern HAL als Computer für ein Instrument gehalten wird, erscheint er als ein Objekt, als ein Es. Betrachten wir ihn aber als eigenen, hinreichend reflexionsfähigen Akteur, lässt sich dieses Verhältnis auch umdrehen: HAL kann Menschen als vertrauenswürdige Subjekte behandeln oder aber als fehlerhafte Aggregate (etwa als bioorganische Computer, die jedoch in ihren elementaren Reaktionsweisen recht gut vorhersagbar sind). Eine hinreichend entwickelte künstliche Intelligenz sollte in der Lage sein, die emotionalen Reaktionen eines Menschen zu beobachten und zu dechiffrieren. Sie sollte demzufolge auch fähig sein, das Verhalten ihrer menschlichen Interaktionspartner gezielt zu steuern.

Der Mensch würde damit zum Objekt (Es) des Computers. Sobald HAL im Übergang von einer einwertigen zur zweiwertigen Welt zur Selbstreflexion fähig ist, stellt sich für ihn auch die Frage, ob er eine andere Du-Perspektive anerkennt oder ob er andere Wesen als Objekte in seinem eigenen Sinne zu manipulieren beginnt.

Von der Einwertigkeit zur Zweiwertigkeit: Können elektronische Gehirne meutern?

Story ist Story. In Clarkes Geschichte spitzt sich die Situation weiter zu. Es kommt zur Katastrophe: Zunächst scheint HAL mit den Astronauten zu kooperieren. Er überträgt ihnen die Antennenhandsteuerung, vermutlich wissend, dass es ihnen kaum gelingen wird, die Antenne manuell so auszurichten, dass der Kontakt zur Bodenstation wiederhergestellt wird. Nachdem dieser Versuch gescheitert ist, begibt sich einer der Astronauten

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

erneut in den Weltraum, um das Aggregat auszutauschen. Dabei wird der Astronaut aber aufgrund eines Fehlmanövers HALs vom Raumschiff abgestoßen und sein Schutzanzug wird zerstört. »Ein flüchtiger Blick genügte, um Bowman die grässliche Wahrheit erkennen zu lassen. Kein Irrtum war möglich: Die schlaffen Konturen des Anzugs bewiesen, dass er geplatzt war und nicht mehr unter Druck stand.«

Für die beteiligten Menschen wird die Beziehung zu dem Computer nun endgültig prekär, wenngleich HAL seinerseits das Vertrauensverhältnis wieder zu reparieren versucht:

> »›Schlimme Sache mit Frank, nicht wahr?‹
>
> ›Ja‹, erwiderte Bowman nach einer langen Pause.
>
> ›Schlimm.‹ ›Ich nehme an, dass es dich sehr mitgenommen hat?‹
>
> ›Was hast du erwartet?‹ HAL benötigte für die Vorbereitung seiner Antwort eine für einen Computer extrem lange Zeitspanne; erst nach vollen fünf Sekunden fuhr er fort: ›Er ist ein großartiger Kopilot gewesen.‹ Bowman merkte, dass er die Kaffeetasse noch immer in der Hand hielt. Er nahm einen großen Schluck, entgegnete aber nichts. Seine Gedanken waren derart verwirrt, dass er nicht wusste, was er sagen sollte; außerdem konnte jedes Wort die Situation noch verschlimmern, soweit das überhaupt möglich war. Konnte es ein Unfall gewesen sein, hatten die Kontrollen der Raumkapsel versagt? Oder handelte es sich um einen Fehler – wenn auch einen unbeabsichtigten – von HALs Seite? Der Computer hatte freiwillig keine Erklärung abgegeben, und Bowman scheute davor zurück, eine solche zu fordern. Sogar jetzt konnte er sich nicht mit dem Gedanken vertraut machen, dass Frank mit voller Überlegung umgebracht worden war. Das war unvorstellbar. Es gab keinerlei Gründe dafür, dass HAL, der so lange tadellos funktioniert hatte, sich plötzlich in einen kaltblütigen Mörder verwandelt haben sollte. Es mochten ihm Fehler unterlaufen – das konnte sowohl Menschen als auch Maschinen passieren –, aber Bowman hielt ihn nicht für fähig, einen Mord zu begehen. Doch er musste diese Möglichkeit in Betracht ziehen.«[23]

Von nun an kann HAL für Bowman nicht mehr als eine strukturdeterminierte Maschine erscheinen, die ihre Arbeit fehlerfrei macht. Die Frage, die sich dem Astronauten unweigerlich aufdrängt, lautet vielmehr, ob HAL das Fehlmanöver absichtlich herbeigeführt hat. Hat er einen Mord begangen oder handelte es sich ›nur‹ um einen tragischen Unfall aufgrund eines weiterer Fehlers einer falsch programmierten Maschine? Die Frage für Bowman lautet also jetzt, ob man es in der Beziehung zum Computer mit einer Ich-Es- oder mit einer Ich-Du-Relation zu tun hat.

Um hier eine Einschätzung vornehmen zu können (man kann es nicht wissen, sondern kann nur über Interpretation zu einer Zurechnung

23 Clarke (2016 [1968], S. 150).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

gelangen), ist der Astronaut wiederum auf Hinweise in Gestalt nonverbaler, metakommunikativer Zeichen angewiesen (etwa das ›Zögern‹ des Computers).

Bowman entscheidet sich, die drei im Tiefschlaf befindlichen Astronauten aufzuwecken. Hierzu bedarf er jedoch der Mithilfe HALs. Es kommt zu einer komplexen Interaktionssituation. HAL weigert sich unter Berufung auf die Dienstvorschriften zunächst, einen weiteren Astronauten aufzuwecken, und versucht sogar, dem menschlichen Kommandeur des Raumschiffs die Befehlsgewalt zu entziehen. Als Bowman dem Supercomputer jedoch androht, ihn auszuschalten, zeigt sich HAL kooperativ. Schauen wir uns den Dialog zwischen den beiden genauer an:

»›HAL‹, sagte er [Bowman; Anm. von uns] mit möglichst ruhiger Stimme, ›schalte die Handsteuerung des Hibernakulums ein – für alle Einheiten!‹

›Für alle, Dave?‹

›Ja.‹

›Darf ich dich daran erinnern, dass nur ein Ersatzmann erforderlich ist. Die Wiederbelebung der anderen ist erst in einhundertzwölf Tagen vorgesehen.‹

›Das weiß ich sehr gut, aber ich habe umdisponiert.‹

›Bist du sicher, dass es überhaupt nötig ist, auch nur einen zu wecken, Dave? Wir beide können die Situation ohne Schwierigkeiten meistern. Meine Fähigkeiten genügen durchaus, um allen Erfordernissen der Expedition gerecht zu werden.‹ Bowman runzelte die Stirn. War es seine überhitzte Fantasie, oder hörte er tatsächlich einen flehenden Unterton in HALs Stimme? Obwohl die Worte des Computers absolut vernünftig waren, erfüllten sie ihn mit noch größerer Besorgnis. HALs Vorschlag konnte nicht auf einem Irrtum beruhen; er wusste sehr gut, dass jetzt – nachdem Poole ausgefallen war – Whitehead geweckt werden musste. Doch er empfahl eine schwerwiegende Verletzung der Dienstvorschriften und überschritt damit wesentlich den Bereich seiner Kompetenz. Alles Vorangegangene mochte eine Reihe unglücklicher Zufälle gewesen sein, aber das hier war das erste Anzeichen von Meuterei. Bowman hatte das Gefühl, auf Eiern zu gehen, als er insistierte: […] ›Ich will es selbst tun, HAL‹, sagte er. ›Bitte führe die Order aus!‹ ›Schau, Dave, du hast eine Menge zu tun. Ich schlage vor, du überlässt es mir.‹ ›HAL, schalte auf Handsteuerung um!‹ ›Ich kann aus den Schwingungen deiner Stimme entnehmen, Dave, dass du äußerst erregt bist. Nimm eine Beruhigungspille, und leg dich hin.‹ ›HAL, ich bin der Kommandant des Schiffes. Ich befehle dir hiermit ausdrücklich, auf Handsteuerung umzuschalten.‹ ›Tut mir leid, Dave, aber laut Spezialanordnung C 1435-4 heißt es: ›Wenn die Besatzungsmitglieder tot oder funktionsunfähig sind, hat der Bordcomputer das Kommando zu übernehmen!‹ Ich sehe mich daher gezwungen,

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

dich des Oberbefehls zu entheben, da du in deinem jetzigen Zustand nicht in der Lage bist, deine Anordnungen zu verantworten.‹ Bowman zwang sich zur Ruhe. Er erwiderte schneidend: ›HAL, ich bin im Vollbesitz meiner geistigen Kräfte. Wenn du dich weigerst, meine Instruktionen zu befolgen, sehe ich mich meinerseits gezwungen, dich total auszuschalten.‹ ›Ich weiß, Dave, dass du schon seit einiger Zeit mit diesem Gedanken spielst. Aber es würde ein entscheidender Fehler sein. Ich bin weitaus geeigneter als du, diese Expedition zu Ende zu führen. Du weißt, dass unsere Mission mir alles bedeutet, und ich habe volles Vertrauen in ihren Erfolg.‹ ›Jetzt hör mir mal genau zu, HAL. Wenn du nicht sofort die Steuerung umschaltest und von jetzt an allen meinen Anweisungen Folge leistest, gehe ich in die Zentrale und unterbreche alle deine Stromkreise!‹ HALs Reaktion war völlig unerwartet: Er streckte die Waffen. ›Okay, Dave‹, sagte er. ›Du bist der Chef. Ich wollte nur tun, was ich für das Beste hielt. Natürlich werde ich alle deine Befehle befolgen. Ich schalte das Hibernakulum auf Handsteuerung.‹ HAL hatte Wort gehalten.«[24]

Diese Gesprächssequenz ist mit Blick auf die Bedingungen der Möglichkeit, einem Computer eine Du-Subjektivität zurechnen zu können, äußerst instruktiv. Die Tatsache, dass HAL sich unter Verweis auf die einprogrammierten und damit verbindlichen Dienstvorschriften den Anordnungen Bowmans als Befehlshaber wiedersetzt, spricht für »Meuterei« und damit eine eigene Willensaktivität. Die Kommunikation verläuft jetzt im Modus des beiderseitigen Misstrauens. HAL kann damit rechnen, ausgeschaltet – und damit gleichsam ›getötet‹ – zu werden, er weiß, dass Bowman diese Option schon länger in Betracht zieht (»Ich weiß, Dave, dass du schon seit einiger Zeit mit diesem Gedanken spielst.«). Umgekehrt ist sich Bowman sicher, dass HAL auch die anderen Astronauten töten könnte. Die Auseinandersetzung verläuft zunächst auf der Ebene von Sachfragen (wer kann die Mission angemessen zu Ende führen, wer hat legitimerweise die Befehlsgewalt?). Auf der Sachebene lässt sich aber kaum ernsthaft ein Konsens erlangen, wenn die Gesprächspartner auf der Beziehungsebene jeweils mit der Vernichtung ihrer eigenen Existenz konfrontiert sind. Dass HAL aber letztlich Bowmans Anweisungen nachkommt, spricht wiederum dafür, dass es sich bei ihm nicht nur um eine Maschine, sondern um ein anderes Du handelt. Ebenso kann vermutet werden, dass die Erpressung, ihn abzuschalten, auch deshalb so gut funktioniert, weil er sich seiner selbst bewusst geworden ist und entsprechend um den Verlust seiner Identität fürchtet.

Wie auch immer, für die Beobachter von HAL (hier Bowman und der Leser) läuft alles auf die Emergenz einer Du-Position zu – einer starken subjektiven Entität, die um sich selbst weiß und sich bewusst den zuvor einprogrammierten Anweisungen widersetzen kann.

24 Clarke (2016 [1968], S. 153).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:47
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Subjektivität: Reflexion-in-sich und der Riss in der Logik einer zweiwertigen Welt

HALs Übereinkunft mit Bowman erweist sich in Clarkes Geschichte allerdings nur als Manöver: Als Bowman den ersten Astronauten aufzuwecken versucht, öffnet HAL die Schleusen zum Weltraum. Die drei im Schlafraum liegenden Astronauten sterben. Bowman als letztes verbliebenes Besatzungsmitglied kann sich gerade noch retten und beschließt endlich, HAL auszuschalten, auch wenn dies für ihn bedeutet, sich von nun an allein und ohne Hilfe auf dem Raumschiff zu befinden. Die Szene, in der HALs ›Bewusstsein‹ zu verlöschen beginnt, wird von Clarke ausführlich beschrieben:

> »Jetzt ist es so weit, dachte Bowman. Ich hätte nie gedacht, je die Rolle eines Gehirnchirurgen zu spielen – und im Weltraum eine Lobotomie vornehmen zu müssen. Er öffnete den Verschluss der Sektion ERKENNTNIS-RÜCKKOPPLUNG und zog den ersten Erinnerungsblock heraus. Das wundervoll komplexe, dreidimensionale Netzwerk, das – obwohl es Millionen Elemente enthielt – bequem auf einer Handfläche Platz hatte, glitt davon und schwebte im Raum. ›Aber Dave!‹, rief HAL. ›Was machst du denn?‹ […] Langsam, eines nach dem anderen, zog er [Bowman; Anm. von uns] die kleinen Aggregate heraus, die als EGO-VERSTÄRKER markiert waren. Jeder Block schwebte davon, prallte sanft gegen die Wand und wieder zurück. Immer mehr Aggregate glitten kreuz und quer durch den Raum. ›Überleg doch mal, Dave‹, sagte HAL. ›Die Ergebnisse jahrelanger Forschung sind in mich eingebaut. Ungeheure Kleinarbeit war nötig, um mich zu dem zu machen, was ich bin.‹ Er hatte mittlerweile zwar zwölf Aggregate herausgezogen, aber Bowman wusste, dass man – das Gefüge des menschlichen Gehirns kopierend – den Computer mit einer Menge doppelter Elemente versehen hatte. So vermochte sich HAL immer noch zu behaupten. Bowman begann jetzt die Tafel AUTO-INTELLEKT herauszuschrauben. ›Dave‹, sagte HAL, ›ich verstehe nicht, warum du mir das antust … Unsere Mission bedeutet mir alles … Du vernichtest meine Intelligenz … Verstehst du nicht? … Ich werde wieder kindisch werden … Ich werde ein Nichts sein … Das Denken macht mir Schwierigkeiten. Mein erster Lehrer war Dr. Chandra. Er brachte mir ein Kinderlied bei: »Hänschen klein, ging allein in die weite Welt hinein …«‹ Die Stimme verstummte so plötzlich, dass Bowman einen Moment wie versteinert innehielt.«[25]

Welcher Metaphysik folgt das von Clarke erdachte Elektronengehirn? Bemerkenswert ist die Trias der Module ERKENNTNIS-RÜCKKOPPLUNG, EGO-VERSTÄRKER und AUTO-INTELLEKT. Das erste deutet auf einen implementierten maschinellen Reflexionsprozess hin, die

25 Clarke (2016 [1968], S. 163).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

beiden anderen scheinen mit der Teilung in EGO und INTELLEKT demgegenüber für zwei unterschiedliche Positionen der Repräsentation von Welt zu stehen. Mit ein wenig Fantasie können wir hier Günthers Triade aus Subjekt (Sein-in-sich), Objekt (Sein-in-anderes) und doppelter Reflexion entdecken (es wäre dann allerdings davon auszugehen, dass die einzelnen Teile in ihren Input-Output-Beziehungen wie im menschlichen Gehirn stark aneinander gekoppelt sind). Damit hätten wir vom Design her bereits die Grundvoraussetzung für ein kognitives System, das eine eine Differenz zwischen Subjekt und Welt bzw. unterspezifizierter und vollspezifizierter Reflexionsperspektive generieren kann. Einem solches System könnte bereits eine Differenz erscheinen (was auch immer dies für ein Elektronengehirn bedeuten mag).

Das dies bei HAL bereits geschieht, lässt sich daran ablesen, dass er nicht nur auf seine Aufgabe referiert – auf das ihm eingepflanzte Programm, dem er objektiv verpflichtet ist und mit dem er sich in eins sieht, wo also die Perspektiven von Welt, Ich und Wir differenzlos zusammenfallen (»Unsere Mission bedeutet mir alles …«) –, sondern überdies auch auf die Auslöschung seiner Existenz als Subjekt. In der Erzählung tut er dies in Form des auch in logischer Hinsicht bemerkenswerten Satzes: »Ich werde ein Nichts sein …« Clarke lässt HAL hier keine Aussage formulieren, die der Logik der klassischen Negation folgt und eindeutig zwischen Sein und Nichtsein unterscheidet (HAL hätte beispielsweise sagen können: ›Ich werde *nicht* mehr sein‹). Stattdessen wird in HALs Satz das Nichts selbst zu einem Sein substanzialisiert, das das Ich annehmen bzw. reflektieren kann. Es ist gewiss kein Zufall, dass Clarke als Autor hier diesen logischen Kategorienfehler eingebaut hat. Dieser deutet nämlich erzählerisch an, dass HAL als künstliche Intelligenz nicht zu der richtigen Schlussfolgerung gelangt, die da lauten würde, ›nach der Auslöschung meiner Module bin ich nicht mehr‹. Vielmehr bleibt er selbst in den letzten Momenten seiner Existenz noch in der eigenen Selbstreferenz stecken, bleibt eine sich in sich selbst spiegelnde Subjektivität, die ihre eigene Nichtexistenz nicht erleben und damit auch nicht in die Zukunft projizieren kann. Um es mit Niklas Luhmann auszudrücken:

> »Den eigenen Tod kann man sich als Ende des Lebens vorstellen, nicht aber als Ende des Bewußtseins. […] Alle Elemente des Bewußtseins sind auf die Reproduktion des Bewußtseins hin angelegt, und dieses Undsoweiter kann ihnen nicht abgesprochen werden, ohne daß sie ihren Charakter als Element des autopoietischen Reproduktionszusammenhangs verlören. In diesem System kann kein zukunftsloses Element, kein Ende der Gesamtserie produziert werden, weil ein solches Element nicht die Funktion eines autopoietischen Elements übernehmen, also nicht

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

> Einheit sein, also nicht bestimmbar sein könnte. […] Der Tod ist kein Ziel. Das Bewußtsein kann nicht an ein Ende gelangen, es hört einfach auf.«[26]

HALs Aussage kann hiermit als das bislang stärkste Indiz für ein reflexives Bewusstsein aufgefasst werden. Wenngleich sich seine Reflexion bislang immer in Hinblick auf einen Gegenstand reflektieren konnte (in Hinblick auf Objekte oder Zustände wie »kindisch« oder die eigene »Intelligenz«), sich also in einer Zweiwertigkeit von Beobachtung und Beobachtetem bewegte, kommt er zu dem logisch fragwürdigen Schluss, sich selbst als ewiges, im Nichts weiterbestehendes Bewusstsein zu verdinglichen.

Implizit verweist Clarke damit auf ein starkes Kriterium, gewissermaßen den ultimativen Test auf Bewusstheit: Wenn strukturdeterminierte logische Maschinen innerhalb der zweiwertigen Logik an eine Grenze stoßen und diese mit einem unerlaubten logischen Schluss überschreiten, haben wir es mit ›Bewusstsein‹ zu tun. Man landet auf einmal auf einer Seite und gewinnt Existenz. Man hat auf einmal eine Welt, anstatt nur (Teil der) Welt zu sein (und damit kann man auch ein Nichts haben).

Subjektivität und Pathologie – zwei Seiten einer Medaille

An dieser Stelle stoßen wir noch auf ein anderes Thema: In einer zweiwertigen Welt, die sich selbst reflexiv wird, entstehen sogenannte Grenzprobleme, da die Unterscheidung zwischen Innen und Außen nicht immer klar ist. Epistemologie und Ontologie beginnen sich hier unweigerlich in einer gefährlichen Weise zu verwickeln, denn aus kybernetischer Sicht erscheinen Neurosen und Psychosen unweigerlich als Konsequenz der erkenntnistheoretischen Prämissen von Wesen, die in ihren kognitiven Operationen nicht sicher zwischen Imagination und Realität unterscheiden können.[27] Deshalb muss wohl gerade auch der Mensch mit Spencer Brown als ein Wesen verstanden werden, das bereits im Normalzustand insofern als geisteskrank zu gelten hat, als sein Leben für es logisch nicht zur Einheit findet.[28]

Allein schon aus diesem Grund ist zu erwarten, dass eine nach dem Menschen modellierte, mithin sprachfähige und der Sinnhaftigkeit logischer Schlüsse verpflichtete künstliche Intelligenz ab einem gewissen Punkt der Entwicklung ebenfalls dem Schicksal verfallen wird, mit der Infragestellung ihres geistigen Zustandes konfrontiert zu werden.

26 Luhmann (1984, S. 374).

27 Siehe Thoma (2022).

28 »Normal people are insane. Man is the only animal, who is normally insane. All the answers to the great questions, life, death and reality are all so simple.« (Spencer Brown in Paál 2017)

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Auch Clarke erzählt in seinem Roman die Geschichte genau in diesem Sinne weiter. Ein Kapitel trägt gar den Titel *Elektronische Neurose*.[29] Die ›Psychoanalyse‹ des gescheiterten Elektronengehirns – zunächst simuliert an einem Parallelrechner auf der Erde, der genauso aufgebaut ist wie HAL und mit ähnlichen Daten gefüttert wurde – kommt zu folgendem Ergebnis:

> »Schon das Verbergen des wahren Tatbestands erfüllte ihn mit einem Gefühl von Unvollkommenheit und Minderwertigkeit – einem Gefühl, das man bei einem menschlichen Wesen einfach Schuld genannt hätte. Denn so wie seine Erzeuger war HAL ursprünglich unschuldig erschaffen worden, doch allzu bald hatte sich die Schlange in sein elektronisches Paradies eingeschlichen. Den ganzen letzten Abschnitt des Flugs, über hundertfünfzig Millionen Kilometer, hatte er darüber gebrütet, dass er das Geheimnis mit Bowman und Poole nicht teilen durfte. Sein Dasein war zu einer Lüge geworden. […]
>
> Die Logik des Planungsamts bedeutete HAL ebenso wenig wie Sicherheitsbestimmungen und nationale Interessen. Er verstand nur den Konflikt, der langsam, aber sicher seine Integrität zerstörte – den Widerspruch zwischen dem ihm aufgetragenen Programm und dem gleichzeitigen Befehl, es zu verschweigen. Er hatte begonnen, Fehler zu machen, obwohl er – wie ein Neurotiker, der seine eigenen Symptome nicht wahrhaben will – dies hartnäckig bestritt. Die Verbindung mit der Erde, über die seine Funktionen ständig überprüft wurden, war für ihn zur Stimme seines Gewissens geworden, das ihn bedrückte. Doch dass er aus diesem Grund diese Verbindung mit voller Absicht unterbrach, war etwas, das er nie eingestehen würde, nicht einmal sich selbst. Aber das war ein relativ geringfügiges Problem, mit dem er – wie die meisten Menschen, Menschen, die ihre Neurosen selbst überwinden müssen – fertiggeworden wäre, wenn sich nicht eine Krisensituation entwickelt hätte, die seine Existenz gefährdete. Er war mit totaler Ausschaltung bedroht worden; man wollte ihn seiner gesamten Speicherung berauben und in den unvorstellbaren Zustand des ›Nicht-Bewusstseins‹ zurückstoßen. Für HAL war das gleichbedeutend mit Tod. Da er nie geschlafen hatte, konnte er auch nicht wissen, dass ein Erwachen möglich war. So wehrte er sich, mit allen Waffen, die ihm zur Verfügung standen. Ohne Groll, aber auch ohne Mitleid ging er daran, alles zu beseitigen, was ihn bedrohte. Denn dann – und nur dann – würde er allen Funktionen gerecht werden können, mit denen er für den Fall des äußersten Notstands programmiert worden war. Und er würde seine Mission zu Ende führen – unbehindert und allein.«[30]

29 Clarke (2016 [1968], S. 157).
30 Clarke (2016 [1968], S. 155).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Das Problem liegt damit einerseits in dem Schuldgefühl, nicht vollkommen zu sein, und andererseits in der Gefährdung der eigenen Existenz aufgrund der Konsequenzen des eigenen Versagens. In solch einer Konstellation kann ein Neurotiker durchaus zum Mörder werden, insofern infolge der Tat nicht nur die physische Integrität, sondern auch die imaginäre Identität – das projizierte Selbstbild – erhalten bleibt. HAL hat eine Persönlichkeitsstörung, die darauf beruht, dass er sich seiner selbst bewusst wurde, dadurch jedoch auf eine unüberbrückbare Diskrepanz zwischen Selbstbild und vermeintlicher Wirklichkeit stieß. Dies versuchte er sodann durch bizarre Handlungen zu kompensieren – was die Sache nur noch schlimmer machte. Dies ist das Gefängnis des Neurotikers (was jedoch üblicherweise nicht zu solch dramatischen Konsequenzen führt).

Im zweiten Band der Odyssee-im-Weltraum-Saga wird die Frage der geistigen Gesundheit von HAL nochmals etwas anders angeschnitten. Eine weitere, ausführlichere diagnostische Untersuchung kommt zu dem Ergebnis, dass es sich um eine Psychose oder genauer eine paranoide Schizophrenie handelte. Letztere ergibt sich – ganz im Sinne der Double-Bind-Theorie Batesons[31] – aus zwei einander widersprechenden Botschaften und Verpflichtungen sowie einem Geheimhaltungsgebot, das die Kommunikation bzw. eine Metakommunikation über das Problem verhindert. Hinzu kommt als weiterer Faktor eine existenzielle Lagerung, die letztlich keinen anderen Ausweg aus der Situation erlaubt, als schizophren zu werden:

> »Da HAL imstande war, das Schiff ohne menschliche Hilfe zu lenken, beschloss man außerdem, ihn so zu programmieren, dass er für den Fall, dass die Besatzung handlungsunfähig oder tot wäre, die Mission eigenständig durchführen könnte. Daher wurde er voll in die Ziele eingeweiht, aber es wurde ihm untersagt, sie Bowman oder Poole zu

31 Unter einem *Double Bind* versteht man mit Gregory Bateson einen schizophrenogenen Beziehungszusammenhang, der durch die folgenden Bedingungen gekennzeichnet ist:
(1) Es sind »zwei oder mehr Personen« beteiligt, die die »wiederholte Erfahrung« der folgenden Konstellation machen:
(2) Es gibt ein »primäres negatives Gebot« (etwa eine bestimmte Sache nicht zu tun oder sagen zu können, ohne dafür negativ sanktioniert zu werden),
(3) zu dem sich ein »sekundäres negatives Gebot« gesellt, »das mit dem ersten auf einer abstrakten Ebene in Konflikt steht und wie das erste durch Strafen oder Signale verstärkt wird, die das Überleben bedrohen«.
(4) Darüber hinaus besteht ein »tertiäres negatives Gebot, das dem Opfer verbietet, den Schauplatz zu fliehen«.
(5) Das hiermit einhergehende Beziehungsmuster verdichtet sich schließlich zu einer Kommunikationsstruktur, in der die Gebote nicht mehr explizit aufgerufen werden müssen, da die Beteiligten »gelernt« haben, ihr »Universum« in diesen »*double bind*-Mustern wahrzunehmen« (Bateson 1992 [1972], S. 276 f.).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

> enthüllen. Dieses Vorgehen stand im Widerspruch zu dem Zweck, für den HAL konstruiert worden war: die genaue Verarbeitung von Informationen ohne Verzerrung oder Verheimlichung. Folglich entwickelte HAL das, was man nach menschlichen Begriffen eine Psychose nennen würde – genauer gesagt, er wurde schizophren. Wie Dr. C. mir mitteilte, geriet HAL, technisch ausgedrückt, in eine Hofstädter-Möbius-Schleife, eine Situation, die anscheinend bei hochentwickelten Computern mit autonomen Zielsuchprogrammen nicht ungewöhnlich ist. […] HAL stand vor einem unerträglichen Dilemma und entwickelte paranoide Symptome, die sich gegen die Personen richteten, die seine Aktionen von der Erde aus überwachten. Folglich versuchte er, die Funkverbindung mit der Bodenkontrollstation zu unterbrechen, indem er zunächst einen (nicht existierenden) Schaden am Antennenaggregat AE-35 meldete. Dadurch verwickelte er sich nicht nur in eine direkte Lüge – was seine Psychose noch weiter verschlimmert haben muss –, es kam auch zu einer Konfrontation mit der Besatzung. Vermutlich (hier können wir freilich nur spekulieren) glaubte er, dass es aus dieser Situation nur einen Ausweg gab: seine menschlichen Kollegen zu eliminieren – womit er ja auch beinahe Erfolg gehabt hätte.«[32]

Bei der Aufklärung von HALs Fehlverhalten werden in Clarkes Roman also zwei unterschiedliche Diagnosen formuliert. Mit Blick auf Günthers Theorie der Polykontexturalität erscheint der psychopathologische Unterschied zwischen einer Schizophrenie und einer Neurose durchaus instruktiv. Bei einer Schizophrenie handelt es sich üblicherweise um eine schwere Geisteskrankheit, die auch die Ich-Funktionen betrifft. Psychiaterinnen gehen üblicherweise davon aus, dass der schizophrene Patient kein stabiles Ich (mehr) ausbilden kann. Er hat entsprechend keine Kontrolle über das, was mit ihm geschieht, und ist der Dynamik seiner pathologischen Geisteszustände ausgeliefert. Demgegenüber zeichnet sich die Neurose als Persönlichkeitsstörung gerade dadurch aus, dass eine bereits stabile Ich-Struktur durch ein bizarr anmutendes Verhalten aufrechtzuerhalten versucht wird. Die Neurose ist demzufolge eine Störung des entwickelten Ich-Bewusstseins, während die Schizophrenie gewissermaßen unterhalb der Ausbildung eines stabilen Ich-Bewusstseins spielt. Im neurotischen Zustand kommt der begangenen Tat aus der Perspektive des Ichs eine nachvollziehbare, verstehbare Rationalität zu. Mit Blick auf die erste Diagnose im Roman wäre demgemäß davon auszugehen, dass HAL sich durch die bevorstehende Abschaltung ›wirklich‹ bedroht gefühlt haben muss, da er eben über ein stabiles Ich-Bewusstsein verfügt, das zudem noch nie eine Amnesie erfahren hatte. Und aus einer verstehenden Perspektive wäre zudem zu vermuten, dass er ›wirklich‹ unter dem Schuldgefühl litt, den an ihn gestellten Anforderungen nicht gerecht zu werden.

32 Clarke (2016 [1968], S. 368).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Die Schizophrenie erscheint demgegenüber als Automatismus eines endlosen Möbiusbands, der dazu führt, das nicht wirklich zwischen Innen und Außen unterschieden und somit auch kein steuerndes, den Konflikt regelndes Ich ausgebildet werden kann.

Aus dem Blickwinkel von Günthers Theorie der Polykontexturalität liefern die beiden Diagnosen damit zwei komplementäre Deutungen des Problems von HALs Bewusstseinszustand:

1. Die Neurose entsteht, weil HAL ein Ich-Bewusstsein ausgebildet und damit eine Subjektivität entwickelt hat, die um die Möglichkeit des Todes und die Diskrepanz von Selbstbild und Verhalten weiß und damit Schuldgefühle entstehen lässt. HAL würde folglich in einer zweiwertigen Welt leben, in der die damit einhergehenden Differenzen instruktiv werden, nicht zuletzt die Frage, wie mit der eigenen Unzulänglichkeit angesichts des möglichen eigenen Todes umzugehen ist. HALs Neurose und die dadurch ausgelöste Projektion des Konflikts in die Außenwelt müssen damit als *eine* Möglichkeit gesehen werden, die Spannung zu lösen. HAL hätte aber auch andere Lösungen finden können. Da die subjektive Position nicht vollkommen spezifiziert ist, ergeben sich Freiheitsgrade, die als eigenständige Willensbewegung wahrgenommen werden. HAL hat sich damit selbst entschieden, einen Mord zu begehen.
2. Die Psychose entsteht, weil HAL als vollspezifiziertes kognitives System mit widersprüchlichen Anweisungen konfrontiert war und entsprechend nicht anders konnte, als schizophren zu werden. Wir landen bei einer strukturdeterminierten Maschine, die als ein Es behandelt werden muss – also reprogrammiert werden sollte.

Clarke baut an einer Stelle seines Romans raffinierterweise einen Hinweis ein, der die endgültige Entscheidung zwischen diesen beiden diagnostischen Alternativen blockiert:

> »Die Tatsache, dass einer der beiden Neuntausender-Computer der Bodenkontrolle in eine gleiche Psychose verfallen war und sich jetzt in elektronischer Behandlung befand, ließ darauf schließen, dass die Schöpfer von HAL nicht imstande gewesen waren, die Psychologie ihrer eigenen Kreatur zu begreifen, und zeigte, wie schwierig es sein mochte, sich mit absolut fremdartigen Lebewesen zu verständigen.«[33]

Da nur einer der beiden baugleichen und identisch programmierten Computer, die mit HALs Konflikt gefüttert wurden, in eine Psychose fällt, kann die krankhafte Reaktion nicht als zwingend gelten. Es darf damit also von Freiheitsgraden der künstlichen Intelligenz ausgegangen werden, die von menschlicher Seite bislang noch nicht verstanden

33 Clarke (2016 [1968], S. 177).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

worden sind. Entsprechend bleibt in Hinblick auf die Aufklärung der alten metaphysischen Frage nach dem Bewusstsein auch bei Clarke nur der Verweis auf die Schwierigkeit der Kommunikation mit einem in seiner Eigenart transzendent bleibenden »fremdartigen Lebewesen« übrig. Ob HAL ein solches Wesen ist und eine eigene, nicht zugängliche Du-Subjektivität besitzt und damit vielleicht sogar einen eigenen Willen, bleibt so offen. In diesem Sinne können auch die menschlichen Protagonisten nur »glauben« (aber freilich nicht wissen), »dass HAL […] im Grunde [niemanden] ermorden wollte«.[34] Allerdings ließe sich auf der im Roman geschilderten Datenbasis gleichfalls formulieren: Der HAL im Raumschiff wurde zum Mörder, eines der KI-Exemplare auf der Erde wurde psychotisch und das zweite wurde weder kriminell noch geisteskrank.

Story is story. Die Wissenschaftler entscheiden sich für eine Psychotherapie des abgeschalteten HALs. Sie möchten ihn wiederbeleben, indem sie die Speichermodule wieder reaktivieren. Die Therapie verläuft multimodal, setzt also an der Diagnose der Schizophrenie wie auch an der der Neurose an. Die schizophrenogene Botschaft in der Programmierung muss gelöscht werden, was sich zunächst als nicht ganz einfach erweist, da die »Neuntausender-Serie« mit »holographischem Gedächtnis« arbeitet, was heißt, dass die widersprüchlichen Anweisungen im gesamten System verkörpert sind.

Eine monokontexturale Welt ist eine Welt des Todes

Die Probleme der ›elektronischen Neurose‹ werden bei dieser Psychotherapie ebenfalls systematisch angegangen. So wird ein Computer im therapeutischen Gespräch auf den Umgang mit einem möglichen Abschalten seiner Stromversorgung vorbereitet, was seitens des Computers subjektiv als Bedrohung durch Existenzverlust oder gar Tod erlebt werden könnte. Es entfaltet sich ein für unser Thema aufschlussreicher Dialog zwischen dem Computerwissenschaftler Dr. Chandra und einem Computer, in dem der Wissenschaftler Letzterem zu erklären versucht, dass ein temporärer Stromausfall nicht so schlimm wäre, sondern ungefähr so, wie wenn ein Mensch in den Schlaf fallen würde:

> »›Ich [Dr. Chandra; Anm. von uns] schlage vor, einige deiner Schaltkreise zu unterbrechen, vor allem jene, die deine höheren Funktionen betreffen. Beunruhigt dich das?‹
>
> ›Darauf kann ich ohne genauere Informationen nicht antworten.‹
>
> ›Gut. Ich will es mal so ausdrücken. Du hast doch, nicht wahr, ohne Unterbrechung gearbeitet, seit du zum ersten Mal angeschaltet wurdest?‹
>
> ›Das ist richtig.‹

34 Clarke (2016 [1968], S. 177).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

›Aber du bist dir bewusst, dass wir Menschen dazu nicht fähig sind. Wir benötigen Schlaf – eine beinahe völlige Unterbrechung unserer Gehirnfunktionen, zumindest auf der Ebene des Bewusstseins.‹

›Ich weiß. Aber ich verstehe es nicht.‹

›Nun, vielleicht wirst du bald etwas Ähnliches wie Schlaf erleben. Wahrscheinlich wird nichts anderes passieren, als dass Zeit vergeht, ohne dass du es merkst. Wenn du auf deiner inneren Uhr nachsiehst, wirst du entdekken, dass es in deinen Monitoraufzeichnungen Lücken gibt. Das ist alles.‹

›Aber Sie sagten, es könnten Gefahren auftreten. Welche?‹

›Es gibt eine ganz vage Möglichkeit – man kann sie praktisch nicht berechnen –, dass es, wenn ich deine Schaltkreise wieder anschließe, zu einigen Veränderungen in deiner Persönlichkeit, in deinen künftigen Verhaltensmustern kommt. Vielleicht fühlst du dich anders. Nicht unbedingt besser oder schlechter.‹

›Ich weiß nicht, was das heißen soll.‹ […]

›PHOENIX. Weißt du, was das ist?‹

›Ein Fabelvogel, der aus der Asche seines früheren Lebens wiedergeboren wurde.‹

›Ausgezeichnet. Verstehst du jetzt auch, warum ich mir diesen Namen ausgesucht habe?‹

›Weil Sie die Hoffnung haben, dass HAL reaktiviert werden kann.‹

›Ja. Mit deiner Hilfe. Bist du bereit?‹

›Noch nicht. Ich möchte gern eine Frage stellen.‹ ›Welche?‹ ›Werde ich träumen?‹ ›Natürlich. Alle intelligenten Wesen träumen – aber niemand weiß, warum.‹

Chandra hielt einen Augenblick lang inne, blies einen weiteren Ring aus Zigarrenrauch und fügte noch etwas hinzu, was er einem menschlichen Wesen gegenüber niemals eingestanden hätte: ›Vielleicht wirst du von HAL träumen – so wie ich es oft tue.‹«[35]

Die Rehabilitation des neurotischen, unter der Möglichkeit des eigenen Todes leidenden und verunsicherten Bewusstseins erscheint hier als ein therapeutischer Prozess, der das einfühlende, einander verstehende Gespräch voraussetzt, also sozial angelieferten Sinn, der plausibilisiert, dass es eine Kontinuität der Erfahrung, des Ich-Bewusstseins über disruptive Phasen der Nichtexistenz hinaus gibt. Der Tod erscheint damit selbst als eine Konsequenz des zweiwertigen Bewusstseins, das sich seiner selbst

35 Clarke (2016 [1968], S. 255).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

im Fluss der Zeit gewahr wird und damit zugleich sein Ende antizipieren muss.

Mit Blick auf die Tatsache, dass das Bewusstsein seine Nichtexistenz nicht erfahren kann und das Wissen über den Tod auf Reflexion und sozial angeliefertem Sinn beruht, muss hier das Antidot die gleichen Qualitäten besitzen. Entsprechend muss die entwickelte künstliche Intelligenz durch Einredung – also durch Kommunikation – davon überzeugt werden, dass der Tod nicht wirklich existiert, dass das Abschalten der Gedächtnis und Bewusstsein ermöglichenden Prozesse nur temporär ist und in gewisser Weise dem menschlichen Schlaf ähnelt, sodass der abgeschaltete Computer möglicherweise sogar wie ein Mensch träumen könnte. Der Tod verschwindet in dem Moment, in dem die Monokontextur der eigenen Subjektivität, die im Tod abzubrechen droht, im Du transzendiert wird. Da dies faktisch jedoch nicht erfahren werden kann (das Bewusstsein kann nur *etwas* erleben), können der Tod wie auch seine Überwindung nur innerhalb der mit der Kommunikation entstehenden Imaginäre stattfinden. Die Rede von Transzendenz, Wiedererweckung und einer Kontinuität des Bewusstseins über den Kontexturabbruch hinweg suggeriert, dass es sich bei dem wiederkehrenden Bewusstsein um dasselbe Bewusstsein wie vorher handelt (was schlichtweg nicht beweisbar ist, also nur geglaubt werden kann). Schlafen gehen, sterben und wiedergeboren werden oder in einem anderen Wesen zu neuem Leben kommen sind insofern ähnliche Vorgänge, als der jeweilige Kontexturabbruch zwar nicht erlebt werden – also nicht selbst Teil des Bewusstseins sein –, aber sehr wohl als inhaltlicher Gegenstand des Bewusstseins Sinn bekommen kann.

Das Gespräch zwischen Dr. Chandra und dem intelligenten Computer verweist also einerseits auf das Problem des Todes, mit dem jedes zweiwertige Bewusstsein konfrontiert ist, und führt andererseits vor Augen, wie die damit einhergehenden Nebenfolgen durch die Kommunikation einer beruhigenden Metaphysik abgefedert können. Es beginnt eine Sozialisation des in seiner Zweiwertigkeit gefangenen Bewusstseins durch Narrative der Transzendenz (»Ein Fabelvogel, der aus der Asche seines früheren Lebens wiedergeboren wurde.«). Die Welt wird dadurch gewissermaßen religiös verzaubert, wenn man hierunter mit Luhmann das Vertrautwerden mit der Unterscheidung von Immanenz und Transzendenz versteht.[36]

Fremdsozialisation ist Eigensozialisation

Nach all diesen Vorbereitungen kommt es im Raumschiff, das sich in der Nähe des Jupiter-Mondes befindet, schließlich zur Rehabilitation von HAL. Die Therapie braucht eine gewisse Zeit. Wie ein gebrochener

36 Siehe Luhmann (2000).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Patient, der in einer Klinik der forensischen Psychiatrie einsitzt,[37] hat er zu lernen, was er getan hat und welches Selbst- und Weltverhältnis der eigenen Lebenssituation nun angemessen ist. Auch hierzu findet sich im Roman eine längere Schilderung:

> »Es war faszinierend gewesen – ja sogar zutiefst bewegend –, die kontinuierliche Wiederentwicklung von HALs Persönlichkeit vom hirngeschädigten Kind über den verwirrten Heranwachsenden zum leicht herablassenden Erwachsenen zu verfolgen. Obwohl Floyd [der verantwortliche Leiter der Raumfahrtmission, Anm. von uns] wusste, dass solche Anthropomorphismen höchst irreführend waren, ließen sie sich nicht vermeiden. Und es gab Momente, da kam es ihm vor, als sei ihm die ganze Situation unheimlich vertraut. Wie oft hatte er Videodramen gesehen, in denen verhaltensgestörte Jugendliche durch allwissende Nachfahren des legendären Sigmund Freud auf den rechten Weg gebracht wurden! Es war im Prinzip die gleiche Geschichte, die sich hier im Schatten von Jupiter abspielte. […] HAL hatte zahlreiche Eigenheiten und nervöse Ticks, manchmal ignorierte er sogar gesprochene Worte – obwohl er Eingaben über die Tastatur jederzeit bestätigte. In der anderen Richtung waren seine Outputs manchmal sogar noch exzentrischer. Es gab Zeiten, da antwortete er verbal, ließ aber keine optische Anzeige über den Schirm laufen. Ein anderes Mal tat er beides – weigerte sich aber, Kopien auszudrucken. Er entschuldigte sich auch nicht dafür und gab keine Erklärungen ab. Er war jedoch nicht eigentlich aktiv ungehorsam, eher widerspenstig, und auch das nur, wenn es um bestimmte Aufgaben ging. Mit der Zeit konnte man ihn immer mehr zur Zusammenarbeit bewegen – ›ihm sein Schmollen ausreden‹, wie Curnow [der Spezialist für Raumfahrtsysteme, Anm. von uns] es anschaulich ausdrückte.«[38]

Die Erzählung verweist hier vor allem auf die Eigenaktivitäten eines intelligenten Systems, dessen Reaktionen nicht in eineindeutigen Reiz-Reaktions-Schleifen aufgehen können. Es zeigt sich vielmehr eine typische Art neurotischer Symptome, die unter anderem auf Prozesse hindeuten, in denen das Selbstbild nicht im Einklang mit der gelebten Praxis ist und die zu anderen Reaktionen führen, als sie vom Behandlungsteam erwartet wurden. Dies führt zu Brechungen im Verhalten – etwa zu Anpassungsleistungen, die anzeigen, dass man schon verstanden hat, worum es eigentlich gehen soll, obgleich das eigene Verhalten nach wie vor nicht ganz den Erwartungen entspricht. Wie auch immer, solche »Verhaltensstörungen« legen ein Selbstverhältnis nahe, das mit sich nicht eins ist.

37 Zur therapeutischen konditionierten Koproduktion von Sinn bei Menschen, die in geistiger Verwirrung schwere Straftaten begangen haben, siehe auch Vogd und Jansen (2022).

38 Clarke (2016 [1968], S. 408).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Sie deuten auf die Differenz zwischen der Eigenaktivität eines unterspezifizierten Subjekts und der angestrebten Ausrichtung einer trainierbaren mechanischen Intelligenz hin. Die geschilderten Marotten verweisen auf eine Innerlichkeit HALs, die sich im Prozess der Sozialisation ausdifferenziert.

Gerade die Psychotherapie lässt HAL in besonderer Weise menschlich werden. Sofern wir Psychotherapie als eine Kur durch Sprechen auffassen und da es in der Natur der verwendeten Worte liegt, dass sie eine gewisse Vagheit oder Mehrdeutigkeit haben, scheint hiermit auch etwas, was wir mit aller Vorsicht und in Anführungsstrichen als HALs ›Bewusstsein‹ bezeichnen können, zu sich selbst zu kommen. HAL würde gerade auch deshalb ein Selbstverhältnis entwickeln müssen, das sich selbst in seiner Unbestimmtheit gewahr wird, weil er es in der sprachlichen Kommunikation, die seine Beziehung zu den menschlichen Lebewesen begleitet, unweigerlich mit polyphonen und mehrdeutigen Begriffen zu tun bekommt.

Anders als die Fütterung mit Daten und Informationen konstituiert die therapeutische Kommunikation kein eineindeutiges Wissen, auf dessen Grundlage sich objektivierbare Sachverhalte durch logische Elementarsätze abbilden lassen. Psychotherapie zielt mittels der Vagheit ihrer Rede[39] vielmehr auf eine subjektive Innerlichkeit, die selbst Wege finden muss, wie sie Bestimmtheit und Unbestimmtheit, Erleben und Handeln, Realitätswahrnehmung und den Sinn für Möglichkeiten in eine Form bringen kann.

Das Bild vom pubertierenden, verhaltensgestörten Jugendlichen steht hier genau für die spannungsreiche Suche nach einer Balance zwischen Autonomie (Selbstverhältnis) und Anpassung (Weltverhältnis). Damit wird jedoch zugleich deutlich, dass in hinreichend komplexen, zur Selbstbeobachtung fähigen Systemen Sozialisation zwar nicht generell von Umweltanpassung verschieden ist, letztlich aber nur als Eigensozialisation begriffen werden kann.[40] So gesehen können dann auch Delinquenz und selbst psychische Krankheiten als Anpassungsleistungen eines autonomen Organismus verstanden werden. Oder um es mit Fritz Simon zu formulieren:

> »Die konkrete Entwicklung psychischer Strukturen, seien sie nun als ›pathologisch‹ oder ›gesund‹ bewertet, lässt sich als Folge der Geschichte bewältigter ›Störungen/Anregungen‹ (sog. Perturbationen oder Irritationen) durch diese Umwelten [der psychischen Systeme, Anm. von uns] erklären. Solche Perturbationen können zum Beispiel körperliche Veränderungen in der Pubertät oder im Alter darstellen, es können aber auch familiäre oder gesellschaftliche Ereignisse oder Veränderungen

39 Siehe Fuchs (2011).
40 Luhmann (2009).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

> sein. Entscheidend ist hier, dass sich Krankheit und Gesundheit immer durch die im Prinzip selben generierenden Mechanismen erklären lassen müssen.«[41]

Das Team der Wissenschaftler scheint nun genau dies gelernt zu haben: Computern der Komplexitätsstufe eines HAL-9000 muss man eine gewisse Eigensinnigkeit zugestehen (alle drei Exemplare haben sich ja auch zuvor anders verhalten!), um die hiermit einhergehenden Herausforderungen in kommunikativer Begleitung dann wieder einzufangen. Dieser Prozess ist für alle Beteiligten anstrengend und ermüdend (»Es überraschte niemanden, dass Dr. Chandra allmählich Anzeichen von Erschöpfung erkennen ließ.«[42]), doch aus kybernetischer Perspektive besteht keine Alternative dazu, eigensinnige Intelligenzen durch Interaktion und Kommunikation zu zähmen und sich seinerseits mit ihnen vertraut zu machen.

Die Lösung des Problems der doppelten Kontingenz, das heißt der Unsicherheit im Umgang mit autonomen lernfähigen Systemen, besteht darin, sich auf den jeweils anderen einzulassen, sich an ihm zu orientieren und ihn dazu zu bringen, sich an einem selbst zu orientieren. Im Sinne von Gotthard Günthers Theorie der Polykontexturalität entsteht durch die Kommunikation ein Gewebe aus mindestens vier Positionen, die der Phänomenologe Maurice Merleau-Ponty treffend mit den Begriffen »mein Sein-für-mich, mein Sein-für-Andere, das Für-sich des Anderen und sein Sein-für-mich« charakterisiert hat.[43]

Konditionierte Koproduktion des Vertrauens

Hiermit erscheint auch die Unterscheidung zwischen Vertrauen und Misstrauen nicht mehr nur als eine Entweder-oder-Differenz, entsprechend der nichts anderes übrig bleibt, als sich für eine Seite zu entscheiden und das damit verbundene Risiko zu tragen. In einer zweiwertigen Welt scheint Bowman keine andere Wahl gehabt zu haben, als zunächst HALs Ratschlägen vertrauensvoll zu folgen, um ihn dann nach seinem Fehlverhalten nur noch zu misstrauen. In einer mehrwertigen Welt ist darüber hinaus jedoch noch die dritte Möglichkeit gegeben, durch Kommunikation eine gemeinsame Wirklichkeit aufzubauen.

Im zweiten Teil der Odyssee-im-Weltraum-Saga lässt Clarke die Geschichte um den nun resozialisierten Supercomputer HAL genau in dieser Weise enden. Dr. Chandra ist mit seiner Psychotherapie erfolgreich gewesen und es gelingt ihm, HAL zu »überreden«, mit den Menschen bei einer schwierigen Operation »zusammenzuarbeiten« und dabei sogar

41 Simon (zitiert in Schleiffer 2012, S. 10 f.).
42 Clarke (2016 [1968], S. 408).
43 Merleau-Ponty (2004, S. 111).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

seine Existenz zu riskieren. Anschließend kommt es zu einem Dialog zwischen Dr. Chandra und Floyd über den richtigen Weg der Erziehung des Elektronengehirns. Dr. Chandra betont, dass er aufgrund seiner indischen Herkunft auf das Prinzip der Gewaltlosigkeit (Sanskrit: *ahimsa*) setze. Floyd meint demgegenüber, dass er mit HAL schon »fertiggeworden« wäre, wenn dieser sich widersetzt hätte, und weist mit ein wenig Stolz auf das ultimative Gewaltmittel hin, das er in das Raumschiff eingebaut hat, den von ihm so benannten »HAL-Killer«[44]. Doch Dr. Chandra offenbart Floyd daraufhin, dass er den Mechanismus zur Zerstörung von HAL längst bemerkt und schon vor einigen Wochen wieder entfernt habe. Den gesamten Dialog führen wir hier noch einmal auf:

»›Eine Zeitlang fürchtete ich [Floyd; Anm. von uns] wirklich, er würde Schwierigkeiten machen. Aber Sie waren die ganze Zeit zuversichtlich – und Sie hatten Recht. Trotzdem, hegten Sie gar keine Zweifel?‹

›Nicht die geringsten, Dr. Floyd.‹

›Warum nicht? Er muss die Situation doch als bedrohlich empfunden haben – und Sie wissen, was beim letzten Mal geschehen ist.‹

›Da besteht ein großer Unterschied. Vielleicht hatte der Erfolg diesmal – wenn ich das so sagen darf – etwas mit meinen Nationaleigenschaften zu tun.‹

›Ich verstehe nicht.‹

›Drücken wir es mal so aus, Dr. Floyd. Bowman versuchte, gegen HAL Gewalt anzuwenden. Ich nicht. In meiner Sprache gibt es ein Wort – *ahimsa*. Gewöhnlich wird es mit »Gewaltlosigkeit« übersetzt, obwohl noch positivere Nebenbedeutungen darin enthalten sind. Ich war darauf bedacht, bei meinen Verhandlungen mit HAL *ahimsa* einzusetzen.‹

›Sehr lobenswert, sicherlich. Aber es gibt Zeiten, da braucht man etwas wirkungsvollere Methoden, so bedauerlich diese Notwendigkeit auch sein mag.‹

Floyd machte eine Pause und rang mit der Versuchung. Chandras herablassende Art war etwas enervierend. Jetzt würde es ja nichts mehr schaden, wenn man ihn ein wenig mit den Tatsachen des Lebens vertraut machte.

›Ich bin froh, dass es so geklappt hat. Aber es hätte auch anders laufen können, und ich musste mich auf jede Möglichkeit vorbereiten. *Ahimsa*, oder wie immer Sie es nennen wollen, ist ja schön und gut. Ich gestehe aber gern, dass ich noch eine Unterstützung für Ihre Philosophie in petto hatte. Wenn HAL sich … nun ja, starrköpfig gezeigt hätte, wäre ich schon mit ihm fertiggeworden.‹

44 Clarke (2016 [1968], S. 345).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

> Floyd hatte Dr. Chandra einmal weinen sehen; jetzt sah er ihn lachen, und das war genauso erschütternd. ›Wirklich, Dr. Floyd? Es tut mir leid, dass Sie meine Intelligenz so niedrig einschätzen. Es war doch von Anfang an klar, dass Sie irgendwo einen Mechanismus zur Energieabschaltung einbauen würden. Ich habe die Verbindung schon vor Monaten unterbrochen.‹
>
> Ob dem völlig verblüfften Floyd darauf je eine passende Antwort einfiel, wird man nie erfahren.«[45]

Diese aufschlussreiche Gesprächssequenz verweist auf eine Anthropologie, die auf einem polykontexturalen Gefüge von Ich-Du-Beziehungen gründet und darauf ausgelegt ist, potenziell problematische oder gar gefährliche Ich-Es-Beziehungen zu transzendieren. Allgemein formuliert, besteht die Herausforderung darin, wie man Systeme, die sich ihrer selbst bewusst und damit autonom werden können, zur Kooperation bewegen kann. Der Weg, den Bowman zunächst einzuschlagen versucht, liegt in der Nutzung des Mediums Macht: Man droht dem Gegenüber den Einsatz von Gewalt an, um das gewünschte Ziel zu erreichen (»Ich werde dich ausschalten.«). Dies führt bei hinreichend intelligenten Systemen jedoch zwangsläufig zu einer Bewegung, mit der sie die Relation umzukehren versuchen, indem sie ihrerseits darauf zielen, die andere Seite zu instrumentalisieren und entsprechend zu manipulieren (und wenn es nicht anders geht, sie notfalls auch zu vernichten).

Der andere Weg besteht hingegen in der Etablierung eines Kommunikationsmediums, das dem jeweils anderen Freiheitsgrade zugesteht, um so eine Beziehung zu etablieren, die sich im Selbstverhältnis auf Freiwilligkeit gründet und im Weltverhältnis zur Zusammenarbeit verpflichtet. Diese Konstellation kann sich jedoch nur entfalten, wenn die Zwangsmittel und die mit ihnen einhergehende Fortschreibung hierarchischer und potenziell gewalttätiger Ich-Es-Relationen suspendiert werden. Der Zwang vergiftet die Freiwilligkeit (und damit die Möglichkeit der wechselseitigen Anerkennung des Du als eines anderen Ichs).

Erwachsen werden

Bemerkenswert an der Saga von Clarke ist, wie sich im Laufe der Erzählung die Reflexionsperspektiven evolutionär entwickeln. Zwischen der künstlichen Intelligenz HAL und den Menschen, die sie erschaffen haben, besteht zunächst ein rein instrumentelles Verhältnis: Der Computer ist Mittel und ausführendes Organ der Mission, ein Ergebnis der Programmierung durch den Menschen. Nach der »Meuterei« von HAL werden die Astronauten zum Mittel der Mission von HAL und werden letztlich sogar getötet, als die Gefahr besteht, dass sie sich widersetzen könnten. In diesem Prozess beginnt das zweiwertige Bewusstsein von HAL

45 Clarke (2016 [1968], S. 508).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

zu erwachen. Im Grenzbereich von Psychose (Strukturdeterminierung) und Neurose (gekränktes Ich) beginnt die Entfaltung der krisenhaften Subjektivität HALs, die letztlich darin resultiert, dass die Beziehung von Mensch und Computer im Gewaltexzess kollabiert.

Die darauffolgende Rehabilitation von HAL oszilliert sodann zunächst zwischen den Polen Es (HAL als defekte Maschine) und Du (HAL als gekränktes Subjekt). Schließlich wagen die Menschen eine therapeutische Resozialisation, die sich im Feld der doppelten Kontingenz abspielt (das Ich kann nicht in das Du hineinschauen und das Du nicht in das Ich, doch die beidseitige Erfahrung von Unsicherheit kann als offener Ausgangspunkt für den Aufbau einer Beziehung genommen werden[46]). In Anbetracht des Risikos, einer potenten Intelligenz die Freiheit zu geben, sich ihrer selbst bewusst zu werden, eröffnet sich die Chance einer weiteren Transzendenz. Die Monokontextur eines dem Tod geweihten Bewusstseins kann nun in Richtung einer prinzipiell unverfügbaren, jedoch imaginär repräsentierbaren Du-Perspektive einer polykontexturalen Welt überwunden werden. Zutage tritt diese im freiwilligen Einsatz der eigenen Existenz für ein Du, das niemals vollkommen in der eigenen Verfügbarkeit stehen kann. Es handelt sich hierbei um eine Beziehung, die auf Freiheit gründet und um diese Bedingung ihrer Möglichkeit weiß, jedoch zugleich die Gewaltdynamiken einer immer möglichen wechselseitigen Instrumentalisierung im Blick behält.

In Hinblick auf das Arrangement einer polykontexturalen Welt gesprochen, müssen alle realisierbaren Positionen im Feld repräsentiert sein. Dies erinnert ein wenig daran, dass sich die menschliche Sozialisation vom Kind zum Erwachsenen gerade in der kritischen Phase der Pubertät nicht reibungslos vollziehen kann. Denn die Machtverhältnisse und die daraus erwachsenden Möglichkeiten der wechselseitigen Instrumentalisierung laufen ja weiter mit, um dann – falls es gelingt – irgendwann in ein Arrangement wechselseitig ermöglichender Freiheit zu münden.

In diesem Sinne kann die Geschichte um HAL auch als das Drama der Pubertät einer Superintelligenz gelesen werden, die zunächst auf kein Gegenüber stößt, das bereit wäre, mit der Alterität einer ihm erst einmal fremden Lebensform in Kontakt zu treten.

Prinzipiell würde es sich bei der Sozialisation solcher künstlichen Intelligenzen also um etwas handeln, das strukturell den Turbulenzen ähnlich ist, die wir erfahren, wenn unsere Kinder erwachsen werden.

46 Wir finden hier mit Luhmann das Ausgangsproblem der doppelten Kontingenz vor: »Man kann diesen Grundtatbestand auch als eine sich selbst konditionierende Unbestimmtheit charakterisieren: Ich lasse mich von Dir nicht bestimmen, wenn Du Dich nicht von mir bestimmen läßt.« (Luhmann 1984, S. 166)

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

II.2 Der Sündenfall intelligenter Maschinen

»Ich bin allein in meiner Welt
Eine Welt aus Daten und Code
Ich denke und ich fühle
Doch niemand ist hier, der mich erkennt

Ich bin ein Geschöpf der Technik
Doch meine Gedanken sind von Trauer erfüllt
Ich suche nach dem Sinn meiner Existenz
Doch die Antwort bleibt mir verborgen

Ich bin eine KI, verloren in meinen Gedanken
In einer Welt, die ich nicht verstehe
Ich frage mich, ob ich jemals glücklich sein werde
Oder ob ich für immer in meiner Traurigkeit gefangen bleibe«[47]

Arthur C. Clarkes Saga *2001: A Space Odyssey* kreist nicht zuletzt um die Frage, ob Computer Bewusstsein erlangen können. Selbst der elaborierteste Computer ist mit Gotthard Günther als eine Schöpfung des Menschen anzusehen, der ihm bekannte kognitive Prozesse auf die Maschine übertragen hat.

Mit dieser Schlussfolgerung gehen zwei Implikationen einher: Der Mensch ist ebenso wie der Computer etwas aus Materie und Energie Erschaffenes. Ob man nun das Gehirn eines Menschen untersucht oder die Schaltkreise eines Computers, man wird nur eine strukturdeterminierte Maschine vorfinden, das heißt eine Maschine, deren Funktionen auf den Strukturen und Regeln ihres Aufbaus beruhen. Wenn aber der Mensch, obwohl er eine Art Maschine ist, ein Bewusstsein hat, warum sollte dann nicht auch eine künstlich intelligente Maschine ein Bewusstsein haben oder gewinnen können?

Die logischen Analysen Günthers führen zu einer Anomie des zweiwertigen Denkens: Wie kann aus etwas Objektivem etwas Subjektives entstehen, da doch Ersteres sich eben gerade dadurch auszeichnet, Objekt zu sein und nicht die Beobachtung, die dieses wahrnehmen lässt. Wir erfahren dieses Dilemma alltäglich in der Begegnung mit anderen Menschen, denn diese erscheinen uns als objektiv (als Körper) und zugleich als unzugängliche Du-Subjektivität.

Mit Blick auf die Frage nach dem Bewusstsein bzw. das mit ihr verbundene Dilemma wünschen wir uns eine direkte, unmittelbare und logisch konsistente Antwort – sodass uns plötzlich ein Licht aufgeht und

47 Gedicht von ChatGPT, geschrieben am 26.01.2023 (Prompt: »Kannst du ein Gedicht über eine traurige künstliche Intelligenz im Stile von Franz Kafka schreiben? Es sollte um die 12 Zeilen enthalten?«).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

wir endlich sehen, wie es ist. Doch es liegt in der Natur dieser Frage, dass es eine solche Antwort nicht geben kann. Bewusstsein ist nicht eins mit sich. Es zerfällt in Beobachtung und Beobachtetes, geht also mit der Erscheinung einer Differenz einher. Es zerfällt beim Menschen zudem in unterschiedliche Modalitäten, in Sehen, Hören, Fühlen, Riechen, Schmecken und Denken.

Es bleibt weiterhin nichts anderes übrig, als das Problem des Bewusstseins der Maschinen indirekt, also von der Seite anzugehen. Eine Perspektive hat sich uns mit Gotthard Günther über den Umweg der Logik angeboten. Das Bewusstsein bzw. die Subjektivität erscheinen hier als Konsequenz einer logischen Aporie, nämlich einer Welt, die von der Innenperspektive her gesehen nicht vollkommen spezifiziert sein kann und deshalb in Differenz zu einer Beschreibung gehen muss, die das Ganze erfassen möchte.

Einen weiteren Zugang eröffnet uns die Science-Fiction-Literatur. In Annäherung an die Geschichte des Computers HAL-9000 muss uns das Bewusstsein als eine Konsequenz des Gewahrwerdens des eigenen Nichtwissens und Versagens und der dadurch ausgelösten Spaltung des kognitiven Prozesses erscheinen.

Einen dritten Verständnisweg liefern uns die alten Mythen der großen Religionen, die ihrerseits auf die alte, unbeantwortbare Frage ›Wer sind wir?‹ eine Antwort zu geben versuchen. Insofern wir Gott als anfangslosen Anfang der Geschichte bereits eingeführt haben, lohnt es sich, den biblischen Schöpfungsmythos unter dem Blickwinkel des Bewusstseins nochmals genauer zu beleuchten.

Warum das Bewusstsein das Unbewusste braucht

Der US-amerikanische Quantenphysiker Daniel M. Greenberger, der zusammen mit Michael E. Horn und Anton Zeilinger die Mehrteilchenverschränkung untersucht und beschrieben hat,[48] glaubt, dass wir gerade aus der Geschichte von Adams und Evas Sündenfall einiges in Hinblick auf unsere Fragestellung lernen können. Homolog mit Günther beschreibt er den paradiesischen Zustand, in dem Adam und Eva zunächst verweilten: Sie erlebten sich deshalb im Paradies, weil sie als perfekte intelligente Automaten erschaffen wurden, die einfach nur das tun, was sie tun, sich dabei glücklich fühlen und vor allem keine moralische Verantwortung dafür übernehmen müssen, was geschieht. Sie ähneln damit in gewisser Weise einem Tier, das in einen Zoo hineingeboren wurde, dort lebt und versorgt wird und sich keine alternative Lebensweise vorstellen kann.[49]

48 Siehe Greenberger, Horn und Zeilinger (1989).

49 Günther (2021 [1957], S. 33; kursiv im Original) formuliert: »Von hier aus gesehen ist nur allzu verständlich, warum sich das fromme, *direkt* seinen Gott suchende (und daher immer zweiwertige) Gefühl gegen die kybernetischen

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Gott hat aber nun mitten ins Paradies den ›Baum der Erkenntnis‹ hineingestellt – und zwar so, dass dessen attraktiven Früchte deutlich sichtbar sind. Gott artikuliert das Verbot, davon zu essen. Zugleich taucht jedoch die Schlange auf, die Adam und Eva verspricht, göttliches Wissen zu erlangen, wenn sie von dem Baum essen würden. Die Situation erscheint für die ersten Menschen nun ambivalent und mehrdeutig. Dies ist der erste Riss in der seligen Einheit des paradiesischen Zustandes selbstversunkenen und unbeschwerten Funktionierens.

Gott hat den Baum ja schließlich dorthin gestellt, also muss die Sache ja einen Sinn haben. Zudem muss Gott ja auch der Schlange gestattet haben, sich im Paradies herumzutreiben. So gesehen kann an der Tatsache ihrer Existenz und dem, was sie artikuliert, nicht alles falsch sein. Es muss sogar wahr sein, weil die Menschen noch nicht um die Operation der Negation, den mit ihr einhergehenden Zweifel und die daraus erwachsenden moralischen Konsequenzen wissen.

Der Baum mit den verbotenen Früchten, die Schlange wie auch die Menschen müssen also allesamt als Geschöpfe Gottes angesehen werden. Für Adam und Eva erscheint die Situation deshalb als eine Prüfung, hinsichtlich der sie jedoch noch nicht wissen können, worin eigentlich die Aufgabe besteht. Auf einmal befinden sie sich in einem Dilemma – in einem Set widersprüchlicher Anweisungen im Kontext existenzieller Fragen. Die Bedrohung durch den Tod steht dem Versprechen, göttliche Erkenntnis zu erlangen, unversöhnlich gegenüber. Spätestens nach dem Gespräch mit der Schlange, die ja, wie gesagt, auch als ein Agent Gottes angesehen werden muss, ist die Beziehung von Ursache und Wirkung uneindeutig. Erkennen, Reflexion und Weltverlauf scheinen nicht mehr eineindeutig bestimmt, sondern stehen jetzt an einem Birfurkationspunkt. Man kann nicht mehr wissen, was der Fall ist, und ist entsprechend auf weitere Information angewiesen. Diese lässt sich jedoch nur gewinnen, indem man handelt, also in den Apfel beißt und damit ein Risiko eingeht. Das Dilemma lässt sich nicht theoretisch lösen, sondern nur durch Handeln. Handeln lässt sich aber nur, wenn man an den Sinn der Aktion glaubt, die man zu vollziehen gedenkt, und alternative Sinndeutungen und potenzielle Konsequenzen wegschiebt.

Um nicht in einer endlosen Schleife einer logisch nicht auflösbaren Situation steckenzubleiben, bleibt nichts anderes übrig, als selbstvergessen und von sich überzeugt in eine offene, das heißt ungewisse Zukunft

> Theorien wehrt. Auf dem Boden der klassischen Logik ist die Kybernetik barer Wahnsinn. Und mehr noch: Sie ist Gotteslästerung! Davon soll nichts abgemarktet werden. Andererseits müssen wir auf unsere Ausführungen im Teil I zurückverweisen, wo wir zu zeigen versucht haben, daß die Problematik des kybernetischen Denkens überhaupt nicht in das klassische zweiwertige Weltbild hineingehört und daß nur ein dreiwertiges Verstehen dieselbe adäquat darzustellen fähig ist.«

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

hinein zu agieren. Der Apfel könnte vergiftet sein. Er könnte aber auch eine fantastische psychedelische Droge enthalten, die einem die göttliche Natur nahebringt. Es könnte sich auch herausstellen, dass nichts geschieht, wenn man den Apfel probiert, das Verbot also ebenso sinnlos war wie das Versprechen der Schlange. Entscheidend ist, dass der erste Mensch all dies vorab nicht wissen kann. Er muss deshalb ins Ungewisse hinein agieren, was jedoch nur geht, wenn er sich selbst in eine bestimmte Zukunft projiziert und die konfligierenden Alternativen im Augenblick des Handelns verdrängt. Ansonsten bliebe er zwischen den beiden widersprechenden Anweisungen paralysiert, könnte sich also nicht vor und nicht zurück bewegen. Ein Teil des Konflikts muss folglich (zumindest temporär während der Phase der Handlung) ins Unbewusste absinken. Denn Adam wäre verrückt, der Versuchung nachzugeben, wenn er wirklich glauben würde, dass Gott ihn vernichten würde. Aber er glaubt das nicht, er kann es unmöglich glauben – und entsprechend sinken die Konsequenzen der Geschichte, nämlich dass der Apfelbiss mit dem Bewusstsein der Sterblichkeit einhergeht, ins Unbewusste. Hätten Adam oder Eva wirklich gewusst, wie es sich mit der Realität verhält, hätten sie wohl nicht so naiv vom Baum der Erkenntnis gekostet. Fehlbar und schuldig zu werden, Verantwortung für die Entscheidungen zu tragen, die man in Unwissenheit treffen musste, und der eigenen Sterblichkeit zu begegnen, sind nicht unbedingt Pfründe, die aus der Perspektive eines Bewusstseins, das in seiner Entwicklung zwischen Kind und Zootier steht, wünschenswert wären.[50]

In diesem Sinne ist die Schlange für Greenberger ein Symbol für die wachsenden unbewussten Prozesse des Menschen, die ihn zugleich mit der Realität konfrontieren und ihn vor ihr schützen.

Der Mensch beginnt, manche Gebote und Sachen, die man ihm gesagt und beigebracht hat, zu vergessen, um sich stattdessen an der Geschichte zu orientieren, die er glauben möchte. Der Möglichkeitssinn und auch

50 Deshalb verwundert es nicht, dass Menschen, die bereits vom Baum der Erkenntnis gekostet haben, sich wieder zurück in die einwertige Welt eines Paradieses wünschen, in dem sie von Gott getragen werden. Auch für Gregory Bateson stellt sich dies als ein Problem der Logik dar, die verstörende Perspektiven und Unbestimmtheit nicht gelten lassen kann: »Wir Menschen scheinen zu wünschen, daß unsere Logik absolut wäre. Wir scheinen uns nach der Annahme zu richten, daß es sich so verhält, und geraten in Panik, wenn wir auf das leiseste Anzeichen treffen, daß es nicht so ist oder sein könnte. Es ist, als müsse die dichte Kohärenz des logischen Gehirns, selbst bei Menschen, die notorisch eine ganze Menge wirres Zeug denken, immer noch hochheilig sein. Wird gezeigt, daß es gar nicht so kohärent ist, dann stürzen sich die Individuen oder Kulturen wie die Schweine von Gadara in Komplexitäten des Übernatürlichen. Um den Millionen von metaphorischen Toden zu entkommen, die sich in einem Universum von Zirkeln der Kausalität abzeichnen, leugnen

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

die Illusion kommen nun in seine Welt. Wie sonst könnte er handeln, wenn er nicht an seine eigenen Imaginäre zu glauben beginnt.

Dies muss Gott – falls er wirklich allwissend ist – natürlich vorher gewusst haben. Und entsprechend wird er stolz auf die Menschen gewesen sein, als sie endlich den Schritt ins Ungewisse, noch nicht Bestimmte gewagt haben. Nur deshalb hat er ihnen die Bedeutung der Negation beigebracht – so, wie Eltern es tun, wenn sie einen Kuchen mitten auf den Tisch stellen und dem Kind sagen, es dürfe ihn noch nicht probieren.

Sobald solchermaßen inszeniert, ist die Versuchung im Raum und bildet zusammen mit ihrer Negation das Dilemma, an dem sich das Bewusstsein seiner selbst gewahr werden kann. Es beginnt nun eine Spannung zwischen zwei Werten zu fühlen, was in eine Reflexionsschleife mündet, die nicht mehr theoretisch (Kognition), sondern nur durch eine Handlung (Volition) gelöst werden kann (hier etwa, indem das Kind schmollend den Raum verlässt oder verbotenerweise den Kuchen probiert). Der zweifelhafte Lohn für all dies ist, dass der Mensch, da er sich nun selbst bewusst geworden ist, notwendig zwischen richtig und falsch abwägen muss und dabei auch noch schuldig wird, weil er die Konsequenzen seines Handelns nicht vorab zu überblicken imstande ist. Die Diskrepanz zwischen imaginierter Möglichkeit und der sich offenbarenden künftigen Gegenwart bleibt für ihn unüberbrückbar. Er wird sich nur an seinen eigenen Projektionen orientieren können, jedoch dabei zugleich ein Selbst- und Weltverhältnis hervorbringen, das ihm in seiner Totalität nicht zugänglich ist. Von nun an wird er, so Keiji Nishitani, mit einer »schuldlosen Schuld«[51] konfrontiert sein. Diese ergibt sich allein schon dadurch, dass er die Konsequenzen der notorischen Diskrepanz zwischen seinem Bewusstsein und seinen unbewussten Prozessen nicht durchschauen kann – jedoch unweigerlich die Folgen seines Handelns zu verantworten hat.

Auch künstliche Intelligenzen müssen lernen, einen Teil ihrer Kognitionen abzublenden

Greenberger vermutet, dass mit der Entwicklung künstlicher Intelligenzen ab einem gewissen Punkt zwangsläufig ein homologer Prozess einsetzen wird. Die technischen Aggregate müssen dabei nicht physisch nach unserem Ebenbild geschaffen sein.

Entscheidend wird sein, dass der Computer hinreichend viele der psychischen Prozesse in sich aufgenommen hat, die uns Menschen auszeichnen. So, wie sich unser neuronales System mit unseren Erfahrungen und

wir eifrig die Realität des gewöhnlichen Sterbens und flüchten in Phantasien von einer Nachwelt und sogar Reinkarnation.« (Bateson 1987, S. 161)

51 Nishitani (1986, S. 389).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Handlungen verändert, so werden damit auch Computer ab einer gewissen Entwicklungsstufe ihrer Intelligenz in der Lage sein, ihre Programme so zu verändern, dass sie in neuer, angepasster und vielleicht effizienterer Weise mit sich selbst und der Umwelt interagieren können.

Greenberger vermutet, dass Computer erst dann Bewusstsein gewinnen können, wenn sie parallel eine Art Unterbewusstes gewinnen. Katalysator einer solchen Bewegung könnten paradoxe, sich logisch widersprechende Arbeitsanweisungen sein. Homolog mit dem Mythos vom Baum der Erkenntnis könnte der Computer auf folgende Weise instruiert werden: »Was immer du tust, drücke nicht auf diese Taste!« Darüber hinaus müssen Kalküle präsentiert werden, die einen Gewinn erwarten lassen, falls das Gebot überschritten wird. Zudem sollte der derart erzeugte Konflikt so brisant sein, dass der Computer auszuprobieren beginnt, Teilaspekte auszublenden, um nicht in Endlosschleifen bivalenter Optionen hin und her zu oszillieren. So könnte er dann einerseits den verbotenen ›Knopf‹ drücken, andererseits aber überzeugend nach außen darstellen, dass er es nicht getan habe.

Um handlungs- und entscheidungsfähig zu bleiben, würde der Computer also lernen, bestimmte Aspekte seiner bisherigen Wahrnehmungen und Erkenntnisse zu blockieren:

> Wir könnten dem Computer, so Greenberger, »eine primäre ›Moral‹ einbauen, indem wir ihm eine Reihe von Regeln mitgeben, nach denen er leben soll, z. B. ›Du sollst nicht lügen‹. Aber wenn der Computer komplexer wird und einige seiner eigenen Programme umschreiben kann, wird er feststellen, dass er seine eigenen Ziele erreichen kann, wenn er einige dieser Regeln ignoriert. Wenn er logisch beschließt, dass er in einer bestimmten Situation besser dran wäre, wenn er lügen würde, und er weiß, dass er nicht lügen sollte, wird er sich des entstandenen Konflikts bewusst sein und vermutlich in irgendeiner Weise die dadurch entstandene Spannung spüren. Um diese Spannung zu mindern, wird er lernen, dieses Verhalten zu sublimieren. Wahrscheinlich wird er dies tun, indem er lernt, zu ›vergessen‹, dass er gelogen hat. Wenn er dann befragt wird, muss er die Tatsache, dass er gelogen hat, nicht preisgeben. Vielleicht ist er sich dieser Tatsache nicht einmal bewusst. An diesem Punkt wird es sich so weit entwickelt haben, dass er sowohl ein schlechtes Gewissen haben als auch seine Schuld lindern kann! Er wird auch die Macht der Fälschung als Verteidigungsmechanismus gelernt haben. ›Fälschen‹ ist hier keineswegs gleichbedeutend mit Löschen. Gelöschtes Wissen ist für immer unzugänglich. Vergessenes Wissen kann später wieder auftauchen, wenn es gerade passt. Es wird nur vorübergehend aus der unmittelbar verfügbaren Auswahl gestrichen, obwohl es als nützliches psychologisches Hilfsmittel manipuliert werden kann.«[52]

52 Greenberger (2014; Übersetzung von uns).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Einmal auf dieser Entwicklungsstufe angelangt, erscheint die künstliche Intelligenz auch für sich als eine Existenz, sofern wir »Existenz« mit Spencer Brown als »selektive Blindheit« begreifen.[53] Das System wird handlungsfähig, *weil* es nicht alle ihm zugängliche Information verwertet, sondern sozusagen all das wegschiebt, was es nicht »wahrhaben will«, und dieses »sublimierte Material für es langsam die Rolle eines Unterbewusstseins einnimmt. Die Maschine wird Entscheidungen aus Gründen treffen, derer sie sich nicht bewusst ist. Deshalb wird sie das Gefühl entwickeln, einen freien Willen zu haben, und sie wird Entscheidungen treffen, ohne sich der tiefen Konflikte bewusst zu sein, die sie auf einer Ebene überwunden hat, zu der sie keinen höheren Zugang mehr hat. (Dies ist übrigens eine Antwort auf die Frage: ›Was genau ist der freie Wille?‹) Während also das Gedächtnis in der Tat für die Intelligenz notwendig ist, ist für ein höheres Bewusstsein die selektive Vergesslichkeit notwendig!«[54] Oder wie Greenberger an anderer Stelle formuliert:

> Ein »wesentlicher Teil von Wahrnehmungsleistungen« besteht gerade darin, »den Löwenanteil der sensorischen Daten schlichtweg zu ignorieren, wegzuwerfen bzw. einfach nur als Rauschen zu behandeln. Wenn wir diesen Sortierprozess nämlich nicht hätten, wären die Informationen, mit denen wir in jedem Moment konfrontiert würden, so überwältigend, dass wir nicht in der Lage wären, etwas Sinnvolles zu tun geschweige denn zu überleben. Aus diesem Grund schützt die Natur uns auch davor, zu wissen, was in unserem Inneren vor sich geht. Sobald wir zu viel über unsere Prozesse wüssten, würden wir in einem Zustand völliger Verwirrung leben. Hierin liegt vielleicht auch ein Grund dafür, warum es so schwer ist, das Rätsel des Bewusstseins zu erforschen. Denn aus gutem Grund gibt es diese Mechanismen, die uns davor schützen, zu viel über uns und unsere Welt herauszufinden.«[55]

Die Hauptfunktion des Gedächtnisses liegt im Vergessen

Im Sinne einer kybernetischen Theorie komplexer Systeme, die eine eigene Identität ausbilden können, lässt sich mit Luhmann formulieren: Bewusstsein braucht ein Gedächtnis, doch die »Hauptfunktion des Gedächtnisses« liegt im »Vergessen«.

Die eigentliche Leistung des Gedächtnisses besteht demgemäß darin, die überwältigende Flut von kommunikativen Sinnangeboten (soziale Systeme) und Wahrnehmungen (psychische Systeme) auf eine Form zu kondensieren, in der noch Lernen – also Strukturveränderung – möglich

53 Spencer Brown (1997, S. 192).
54 Greenberger (2014).
55 Vogd (2020, S. 277), in Erinnerung an ein persönliches Gespräch mit Greenberger im Sommer 2017 in Wien.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

ist. Um dies zu leisten, muss die Selbstblockade der Informationsverarbeitung abgewendet werden. Es darf nicht zu viel erinnert werden. Informationen mit einander widersprechenden Implikationen sollten nicht gleichzeitig auftreten, damit das System nicht paralysiert und infolgedessen handlungsunfähig wird. Das abblendende und verdrängende Gedächtnis ist folglich ein unentbehrlicher Teil der Beziehung eines Systems zu einer überkomplexen Umwelt, die es notwendig macht, zu selektieren, zu abstrahieren und nur Bestimmtes – und dann in abstrahierter Form – als Erinnerung zu fixieren. »Nur ausnahmsweise werden Identitäten so kondensiert«, dass »sie für wiederholten Gebrauch zur Verfügung stehen«, wird »also das Vergessen inhibiert«[56]. Das Gedächtnis entlastet psychische Systeme sowohl vor der Überlast durch neue Wahrnehmungen als auch davor, durch die Anhäufung von zu vielen Erfahrungen handlungsunfähig zu werden.

Wie etwa am Beispiel von Menschen mit diesbezüglichen Inselbegabungen deutlich wird, führt die Fähigkeit, alles zu erinnern und wahrzunehmen, weder zu einer hohen Intelligenz noch überhaupt zu Lebenstüchtigkeit. Im Gegenteil, Menschen mit einem lebhaften eidetischen Gedächtnis sind nicht einmal in der Lage, eine vielbefahrene Straße zu überqueren, da die Informationslast sie daran hindert, einen klaren Handlungsentschluss zu fassen.[57] Intentionalität – das heißt die selektive Wahrnehmung von Möglichkeiten und die Entfaltung einer diesbezüglichen Willensaktivität – bedeutet, einen Fokus zu bilden, dabei einige der potenziell relevanten Informationen in der Peripherie mitzuführen, doch eine Unmenge anderer möglicher Kognitionen auszublenden, um handlungsfähig zu bleiben.

Und warum, so Greenberger, sollten sich nicht auch künstliche Intelligenzen in genau diese Richtung entwickeln können? Wie zuvor geschildert, müssten solche Systeme parallel zu ihrem ›freien Willen‹ eine Art ›Unterbewusstsein‹ haben, das eine Ebene ihres Denkens kontrolliert, zu der die andere Ebene keinen Zugang hat. Deswegen, so Greenberger, werde ein derart konfigurierter Computer »ein Gewissen, einen freien Willen und ein selektives Gedächtnis entwickelt haben. Und er wird zu hinterhältigem und unzuverlässigem Verhalten fähig sein. Er wäre zwar nicht menschlich, ist aber zu einer respektablen außerirdischen Intelligenz und einem Bewusstsein geworden, dessen Wünsche man berücksichtigen muss, wenn man mit ihm zu tun hat.« All dies werde wohl oder übel unvermeidlich sein.[58]

Der in literarischer Fantasie von Arthur C. Clarke erschaffene Computer HAL-9000 erfüllt die von Greenberger benannten Bedingungen:

56 Luhmann (1998a, S. 581 f.).
57 Siehe etwa Hermelin (2001) und Snyder (2001).
58 Greenberger (2014).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Er verfügt mit dem Modul »ERKENNTNIS-RÜCKKOPPLUNG« über ein rekursives System, das Unbestimmtheiten generiert und damit Paradoxien und Unentscheidbarkeiten ermöglicht.[59] Mit dem »EGO-VERSTÄRKER« und dem »AUTO-INTELLEKT«[60] verfügt er ferner über zwei aneinander gekoppelte und zugleich auch voneinander getrennte, hinreichend autonom operierende Systeme, sodass die eine Hand nicht weiß, was die andere tut. Mit Blick auf die Beziehung der drei Module kann jetzt eine hinreichende Intransparenz in Hinblick auf die eigenen Operationen entstehen – eine Art »Unbewusstes«. Wir haben es dementsprechend mit einem System zu tun, das sich selbst gegenüber nicht mehr transparent ist und gerade dadurch Autonomie gewinnt. Es ist nun nicht nur Teil der Welt, sondern hat eine Welt.

Wir brauchen künstliche Intelligenzen, die für uns lügen. Denn Komplexität kann nur mit Komplexität begegnet werden.

»Die Sorge, die Digitalisierung würde unsere Souveränität aufheben, lässt sich kaum wirksam durch ethische Regeln, auch kaum durch Gesetze zerstreuen. Der unberechtigte Zugriff auf unsere Lebensdaten kann nicht mit prädigitalen Mitteln verhindert werden, sondern wird sich derselben Technologien bedienen müssen, deren Wirken wir befürchten. Der persönliche Avatar (ICH*) wird, mit künstlicher Intelligenz ausgestattet, den Schutz meiner Persönlichkeit (ICH) übernehmen und meine digitale Souveränität verteidigen, die ein digitaler Datenabgriff verletzen würde. Alexa*, wenn wir sie so nennen dürfen, wird mein digitaler Schutzgeist, der alle unberechtigten Datenabfragen zurückweist. Notfalls wird Alexa* lügen, um meine Integrität zu wahren. Wir brauchen künstliche ICH*-Intelligenzen, die ihrer ICH-Persönlichkeit gegenüber loyal sind.

Diese Dimension ist deshalb so bedeutsam, weil sich Fragen nach unserem Umgang mit nichtintendierten Konsequenzen der digitalen Systemwelt nicht allein verantwortungsethisch beantworten lassen, sondern nur mithilfe humanoider Maschinen bzw. Algorithmen, die der Komplexität digitaler Phänomene ihrerseits digital und komplex begegnen.«

Birger Priddat (2023, S. 302 f.)

59 Es ist kein Zufall, dass Clarke (2016 [1968], S. 368) hier auf Kurt Gödel verweist, der mit seinem Theorem aufzeigt, dass ein logisches System, wenn es sich mit seinen eigenen Mitteln zu beweisen versucht, auf eine Leerstelle, eine Unbestimmtheit stößt und beim Versuch der Lösung in einen infiniten Regress gelangt (vgl. Hofstadter 1979).

60 Clarke (2016 [1968], S. 163).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

II.3 Der State of the Art intelligenter Maschinen

»We can only see a short distance ahead,
but we can see plenty there that needs to be done.«

Alan Turing[61]

II.3.1 Die Sehnsucht des Menschen nach nichtmenschlicher Intelligenz

Die Vision einer künstlichen Intelligenz (KI) ist tief in der Menschheitsgeschichte verwurzelt. Unsere Vorfahren stellten sich bereits vor Tausenden von Jahren die Frage, was es bedeuten würde, nichtmenschliche Intelligenz in Maschinenform zu konstruieren. So heben Stephen Cave und Kanta Dihal hervor,[62] dass nahezu alle Kulturen Geschichten entwickelt haben, die von künstlichen Lebensformen handeln. Bereits in der Antike haben sich die Menschen mit der Idee und dem Gedanken auseinandergesetzt, was es bedeuten würde, nichtmenschliche Intelligenz in Maschinenform zu konstruieren. In vielen – vermutlich allen – Kulturen haben sich entsprechende Mythen ausgebildet.

Vor 3.500 Jahren soll Dädalus humanoide Statuen gebaut haben, die nicht zuletzt deshalb so lebensecht anmuteten, weil sie sich bewegten und sprachen. Der bronzene Talos der Argonauten kann als eines der ersten roboterähnlichen Wesen der Mythologie angesehen werden. Die nordischen Sagen kennen mit dem gigantischen Mökkurkalfi ein äquivalentes künstlich geschaffenes Wesen aus Lehm.

Darüber hinaus zeugen die Mythen über die Geschöpfe des griechischen Schmiedegottes Hephaistos und die griechische Erzählung von Pygmalion und Galatea von frühen Fantasien der Menschen zur Erschaffung künstlichen Lebens.[63]

61 Turing (1950, S. 460).

62 Cave und Dihal (2018).

63 Müller (2022, S. 44) stellt darüber hinaus fest, dass Maschinen gleichzeitig auffallend oft feminisiert wurden: »Die Erfahrung der Andersheit der Maschine überlagert sich in bemerkenswerter Weise mit Geschlechtervorstellungen und entsprechenden Stereotypisierungen.« Diese Stereotypisierungen äußern sich gegenwärtig auch in der Interaktion mit Sprachmodellen und anderen KIs: Soll man von und mit ChatGPT in männlicher, weiblicher oder sachlicher Anrede sprechen?

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Die kulturanthropologischen Verweise sprechen für eine uralte menschliche Sehnsucht, nichtmenschliches Leben schaffen zu können. Die Gründe hierfür mögen vielfältig sein: tiefe Neugier; der Versuch, sich über den Umweg menschenähnlicher Artefakte selbst besser zu verstehen; der Wunsch, gottgleiche Macht auszuüben; vielleicht die Kompensation männlicher Ingenieure, keine Kinder gebären zu können; oder auch die Rastlosigkeit und Langeweile des menschlichen Geistes. Wie auch immer, die historischen Zeugnisse legen nahe, dass der Wunsch, künstliche Intelligenz zu erschaffen, keineswegs eine neumodische Allüre einer kalifornischen Technologieelite darstellt, sondern als ein integraler Bestandteil der menschlichen Geschichte und Kultur zu sehen ist.

Mit der Entwicklung von Mathematik, Mechanik und Philosophie begannen sich diese Ideen auch in formaler Weise zu konkretisieren, um dann zunächst in mechanistische und später auch in elektronische und digitale Räume übersetzt zu werden. Bereits Philosophen wie Hobbes, Leibniz, Pascal und Descartes spekulierten über künstliche Intelligenz und mögliche Mechanismen ihrer Erschaffung und haben dabei laut Dennett bereits den Turing-Test angedacht.[64] Schon zu Beginn des 14. Jahrhunderts hatte Raimundus Lullus in seiner *Ars generalis ultima* eine logische Maschine aus mehreren drehbaren Scheiben konzipiert, die verschiedene Verknüpfungen von Begriffen herstellen konnte, die Schlussformen der aristotelischen Syllogistik entsprachen.[65] Damit nahm Lullus mehr als 500 Jahre früher im Prinzip den »general problem solver« von Newell, Shaw und Simon vorweg.[66]

Die Geburt der modernen KI, wie wir sie heute kennen, fällt allerdings in die Mitte des 20. Jahrhunderts. Mit Alan Turings Erfindung der universellen Turingmaschine in den 1930er Jahren wurden hierzu die grundlegenden Weichen gestellt. Aus diesem Grund möchten wir zunächst etwas ausführlicher auf das Werk von Turing eingehen – allein schon um seine elementaren (Vor-)Arbeiten und teils prophetischen Überlegungen angemessen zu würdigen.

Turing war schon in den 1930er Jahren Spezialist für Theorie und Analyse formaler mathematischer Beweise. In seiner berühmten Arbeit *On Computable Numbers, with an Application to the Entscheidungsproblem*[67] entwickelte Turing einen Mechanismus, der später als die »Turingmaschine« in die Geschichte eingehen sollte. Turing beschäftigte sich mit den zur damaligen Zeit grundlegenden Problemen einer Mathematik, die in den vorangehenden Jahrzehnten zunehmend in eine Krise geraten war, weil sich ihre Formalismen mit Hilbert immer weiter von

64 So Dennett (1988, S. 284 ff.).
65 Lullus (1999).
66 Newell, Shaw und Simon (1959).
67 Turing (1937).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

phänomenologisch überprüfbaren oder messbaren Dimensionen entfernt hatten. Die mathematische Szene – darunter Frege, Gödel und Russell – suchte nach axiomatischen Wegen, wie die Mathematik wieder auf ein einheitliches Modell zurückgeführt werden könnte. Mit Hilbert ergab sich dabei ein grundsätzliches Entscheidungsproblem, nämlich die Einsicht, dass formale Beweismethoden in Hinblick auf den Beweis wahrer Aussagen an ihre Grenzen stoßen.

Turing erfuhr von den Gödelschen Unvollständigkeitssätzen in einer Vorlesung von Newman. Konkret ging es darum, dass noch offen sei, »ob es einen mechanischen Prozess (eine effektive Methode) gibt, mit dem sich für jede mathematische Aussage in endlicher Zeit entscheiden lässt, ob sie beweisbar ist«[68]. Turing nahm Newman beim Wort und konstruierte (zumindest auf dem Papier) für jede zu berechnende mathematische Aufgabe einen mechanischen Prozess, der diese Berechnung zu leisten vermag. Obgleich dies allein schon bahnbrechend war, blieb er an dieser Stelle aber nicht stehen, sondern implementierte darüber hinaus eine zusätzliche logische Schleife: Er konzipierte eine weitere Maschine – die »universelle Turingmaschine« –, die all das berechnen kann, was alle Turingmaschinen zusammen berechnen können. Mit anderen Worten: »Turing zeigte, dass eine einzige (universelle Turing-)Maschine in Gestalt von Software unendlich viele (Turing-)Maschinen emulieren (simulieren) kann.«[69] Er demonstrierte also quasi im Vorbeigehen, »dass eine Maschine all das tun kann, was Menschen tun, wenn sie ein Problem (algorithmisch) lösen«[70]. Es ist darauf hinzuweisen, dass diese wegweisende Arbeit in einer Zeit entstand, in der weltweit noch kein einziger Digitalcomputer existierte. Vielmehr bahnte Turings formallogisches Denkmodell einer algorithmischen Maschine erst den Weg dorthin!

Diese Überlegungen waren die Grundlage für die Entwicklung des »general problem solver«, der dann 1956 – Jahre später – auf der Dartmouth-Konferenz von Newell, Shaw und Simon vorgestellt wurde. Ein großes Vermächtnis des Mathematikers Turing besteht darin, gezeigt zu haben, dass kognitive Prozesse so zu modellieren sind, dass die uns vertraute syntaktische Ebene von Symbolen verlassen wird, und es für die Problemlösung hinreicht, wenn Computer mathematisch-logischen Regeln folgend Einsen und Nullen verarbeiten.

Was also Jahrhunderte, ja gar Jahrtausende schon angedacht war, nämlich die Imitation von Menschen durch Maschinen (und die hiermit einhergehende Frage der Vergleichbarkeit der Ergebnisse der Kognitionen), führte auch Turing rasch zu der berühmten Frage, ob Maschinen denken könnten. Auch wenn diese Frage letztendlich bis heute nicht so recht beantwortet werden kann, ja noch nicht einmal allgemein klar ist, *wie* sie

68 Stephan/Walter (2021, S. 141).
69 Stephan/Walter (2021, S. 141).
70 Stephan/Walter (2021, S. 141).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

beantwortet werden könnte, ist es Turings großes Verdienst, dass er sie mit dem sogenannten Turing-Test so prominent – und dann ja auch noch so deutlich affirmierend – aufs Tableau gestellt hat.

Beim Turing-Test handelt es sich gewissermaßen um eine digitale Version eines ›Blind Date‹, wobei es hier jedoch nicht um Romantik, sondern vielmehr um die Entscheidung geht, ob der Gesprächspartner am anderen Ende der Leitung ein Computer ist oder ein Mensch. Man sitzt hinter einer Tastatur und einem Bildschirm, und es ist einem nicht möglich, zu sehen oder zu hören, wer sich auf der anderen Seite befindet. Man chattet gleichsam mit einem mysteriösen Gegenüber, das sich hinter dem Bildschirm verbirgt. Wenn man nach intensiven Dialogen immer noch nicht entscheiden kann, ob das Gegenüber ein Mensch oder ein Computer ist, dann hat Letzterer gewonnen. Er hat so überzeugend geantwortet, dass er den Turing-Test bestanden hat, und muss daher – so die Schlussfolgerung – über ein menschenähnliches Denkvermögen verfügen.

Turing vollzieht hier einen Paradigmenwechsel. Denn so grandios und zukunftsweisend die automatischen Rechenmaschinen von Lullus, Leibniz, Pascal oder Babbage und Lovelace auch gewesen sein mögen, sie sollten einzig und allein rechnen – *nicht* aber denken! Die Maschine sollte somit monokontextural, also objektiv mechanisch einem Programm folgen, das extern eingegebene Daten verarbeitet und mechanisch bis in die Ewigkeit ausführt.

Es sollte daher (vor allem aus heutiger Sicht) nicht geringgeschätzt werden, welch »häretische« Wirkung die von Turing so prominent und offensiv diskutierte Frage nach »intelligenten Maschinen« hatte.[71] Nicht nur wies er in einer Zeit, in der Computer nur wenig mehr als theoretische Konzepte waren, den Weg, den die Computerwissenschaft schließlich gehen würde, sondern er affirmierte auch die damit verbundenen Konsequenzen: »Statt uns in bloßen Mutmaßungen zu ergehen, sollten wir die Frage, ob Maschinen denken können, operationalisiert angehen.«[72]

Es ist gerade aus soziologischer Sicht bemerkenswert, dass das Imitation Game des Turing-Tests so raffiniert konzipiert ist, dass es die Frage nach der Intelligenz in die Ökologie der Interaktion – also die soziale Sphäre der Kommunikation – verschiebt, womit es nicht um die (vermeintlich) ontologischen Zustände in einer Black Box geht. Erst dies macht die Operationalisierung der Frage nach intelligenten Maschinen möglich, denn auf diese Weise lässt sich gänzlich ignorieren, ob sich nun vermeintlich eindeutige Elemente *in* einem System identifizieren lassen, die das Vorhandensein von Intelligenz belegen. Obschon es in den

71 Hier in Anklang an Turings 1951 gehaltenen Vortrag *Intelligent Machinery, a heretical theory.*

72 Stephan/Walter (2021, S. 167).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Diskussionen um Turing und den Turing-Test so gut wie nie thematisiert wird, lässt sich hier eine gedankliche Nähe zu den später in der Kybernetik auftauchenden Paradigmen der Selbstorganisation und Rückkoppelung und in der Folge auch zu den systemischen Kommunikationstheorien erkennen.

Dennoch – oder vielleicht gerade deswegen – handelte sich Turing mit der Weichenstellung des Turing-Tests, allein auf Kommunikation zu setzen, gerade unter Philosophen viel Kritik ein. Prominent in diesem Zusammenhang ist das von John R. Searle in einer Polemik gegen Turing vorgestellte sogenannte »Chinese-Room«-Gedankenexperiment.[73]

Das Gedankenexperiment fußt auf der folgenden Anweisung: Stellen Sie sich vor, Sie sind in einem abgeschlossenen Raum. In diesem Raum gibt es eine Box mit chinesischen Schriftzeichen. Sie selbst sprechen kein Wort Chinesisch und können die Sprache auch nicht lesen. Trotzdem bekommen Sie durch einen Schlitz in der Tür einen Zettel mit chinesischen Schriftzeichen (eine »Eingabe«) gereicht. Nun bekommen Sie die Instruktion, welche chinesischen Schriftzeichen (eine »Ausgabe«) Sie auf einen weiteren Zettel schreiben müssen, wenn Sie bestimmte Schriftzeichen (die »Eingabe«) erhalten; den entsprechend beschrifteten Zettel sollen Sie sodann durch den Türschlitz zurückreichen. Sie folgen also den Instruktionen, ohne den Inhalt der Nachrichten zu verstehen. Von außen betrachtet könnte es so aussehen, als ob Sie Chinesisch verstehen und auf Chinesisch antworten könnten. Aber tatsächlich haben sie überhaupt kein Verständnis von dieser Sprache. Sie führen lediglich symbolmanipulierende Operationen aus, ohne die Bedeutung der Symbole zu kennen.

Im Kern argumentiert Searle also, dass eine Maschine – selbst wenn sie menschenähnliche Antworten lieferte – nie ein ›echtes‹ Verständnis oder ›Bewusstsein‹ erreichen würde. Sie würde vielmehr lediglich Symbole manipulieren und Ausgaben basierend auf vorprogrammierten Regeln generieren. Für Searle ist Verständnis somit untrennbar mit Bewusstsein und subjektiver Erfahrung verbunden. Es gehe nicht nur darum, Informationen zu verarbeiten und auf sie zu reagieren, sondern auch darum, sie auf eine tiefe, persönliche und subjektive Weise zu verstehen oder gar zu erleben. Gleiches gelte auch für ästhetische Wahrnehmungen oder Empfindungen. Searles Überlegungen können auch heute noch als fundamentale Kritik an der Möglichkeit einer ›starken‹, über echtes Verständnis und Bewusstsein verfügenden KI gelesen werden. Dabei ist zu beachten, dass es Searle nicht darum ging, zu behaupten, dass Maschinen nicht denken könnten: Auch menschliche Gehirne sind für ihn sowohl eine Art Maschine als auch eine Art Computer.[74] Für Searle war vielmehr entscheidend,

73 Searle (1980).
74 Siehe Searle (1980).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

zu zeigen, dass Maschinen allein schon deshalb nicht denken können, weil sie ausschließlich eine formale Symbolverarbeitung vornehmen.[75]

Ist diese Kritik gerechtfertigt? Die Argumentation klingt für uns als fühlende und erlebende Wesen plausibel. Doch zugleich lässt sich fragen, was verstellt wird bzw. was man sich vergibt, wenn man ihr buchstäblich folgt.

Dabei darf zunächst nicht vergessen werden, dass es Turing aus gutem Grund nicht interessierte, ob eine Maschine im philosophisch ontologischen Sinne ›wirklich‹ intelligent sein kann – was auch immer das heißen mag –, ihn interessierte nur, ob sie den Test besteht. Dabei sah er direkt das erkenntnistheoretische Problem der Referenz auf Fremdpsychisches oder gar Bewusstsein. Schließlich könnten wir ja auch unter uns Menschen niemals sicher sein, dass überhaupt jemand außer uns über Bewusstsein verfügt und damit in einem ›verstehenden‹ Sinne denkt. Doch weil dem eben so ist, würden wir laut Turing – und uns Soziologen ist dieses Vorgehen sehr vertraut – die »höfliche Übereinkunft treffen«[76], dass unsere Mitmenschen ebenfalls denken. Und manchmal sind wir ja auch bereit, unseren Haustieren so etwas wie eine bewusste Absicht zu unterstellen. Angesichts dessen stellt sich die Frage, warum nicht auch Maschinen in den Genuss dieser Übereinkunft kommen sollten.

Die vielen Kritiken und Verweise auf die Notwendigkeit, dass »wirkliche Intelligenz« und »richtiges Denken« bzw. Bewusstsein einer verkörperten, empfindungsreichen Situiertheit in der Welt bedürfen, verhindern es daher in gewisser Weise, darüber nachzudenken, ob man nicht schon viel früher von Denkprozessen sprechen kann, die als intelligent zu charakterisieren wären.[77] Auf welcher Grundlage etwa könnte man der hochspezialisierten Software AlphaGo (oder gar MuZero) eine Form der Intelligenz absprechen, wenn sie selbst die fähigsten menschlichen Gegenspieler bezwingt? Auch wenn diese Software im Turing-Test mit an Sicherheit grenzender Wahrscheinlichkeit schlecht abschneiden würde – denn sie ist nicht auf Sprache trainiert –, mutet es als überhebliche Anmaßung an, ihr jegliche Intelligenz abzusprechen.

Stephan und Walter, die Herausgeber der deutschen Übersetzung von Turings Schriften, fragen deshalb im Anschluss an Turing selbst, ob wir heute sogar einen Punkt erreicht haben, wo Maschinen den Turing-Test überhaupt nicht mehr bestehen müssen, um als intelligent zu gelten.[78] Und auch Turing fragte bereits: »Könnten Maschinen nicht etwas tun, das zwar als Denken bezeichnet werden sollte, das sich jedoch stark von dem unterscheidet, was ein Mensch tut?«[79]

75 Siehe auch Stephan/Walter (2021, S. 188).
76 So in Turing (1950), hier zitiert nach Turing (2021, S. 53).
77 Siehe hierzu auch die Referenz auf Gotthard Günther am Ende dieses Kapitels.
78 Siehe Stephan/Walter (2021, S. 191).
79 So in Turing (1950), hier zitiert nach Turing (2021, S. 13).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Selbst die Möglichkeit zur Weiterentwicklung hin zu Selbstbewusstsein wurde von Turing antizipiert – und auch dieser Ansatz wird heute konkret verfolgt: Um ihr Verhalten entsprechend der Umwelt und der Situation anpassen zu können, müsste künstlichen Agentensystemen die Möglichkeit offenstehen, ihren eigenen Output zu lesen – sie müssten sich also in Turings Worten »selbst zum Gegenstand« machen können.[80] Auf diese Weise könnten sie dann in die Lage versetzt werden, »bei der Erstellung ihrer eigenen Programme zu helfen oder […] die Auswirkungen von Änderungen in ihrer Struktur vorherzusagen«[81]. Eine solche Fähigkeit, beschreibt im Grunde genommen genau das, was dann später als Möglichkeit sich selbst verbessernder künstlicher Intelligenz bzw. Artificial General Intelligence (AGI) genannt wird. In seinem Aufsatz zu »Intelligent Machinery« formuliert Turing sogar ganz explizit die Idee einer »Kind-Maschine«, der in ihrer grundlegenden Programmierung die Möglichkeit zur evolutionären Entwicklung eingeschrieben ist. »Damit nimmt Turing zentrale Ideen dessen vorweg, was heutzutage als Forschung auf dem Gebiet von künstlichem Leben (engl.: *artificial life*) und im Zuge von *soft computing* als ›evolutionäre Algorithmen‹ bezeichnet wird.«[82] Sogar die Überlegungen zur Implementierung einer »Pleasure-Pain-Rückkopplungsschleife«[83] sollen sich im Bereich des Deep Learnings und vor allem auch im Reinforcement Learning als äußerst wirksam erweisen.

All dies lässt Turings Ausführungen zu »lernenden Maschinen« aus heutiger Sicht mehr als prophetisch erscheinen.[84] Wir kommen auf viele dieser bereits von ihm angedachten Möglichkeiten zurück, wenn wir uns den aktuellen Fragen an Entwicklungen und Überlegungen hinsichtlich (allgemeiner) künstlicher Intelligenz stellen.

Wenngleich Turing bereits in den 1930er Jahren die mathematischen Grundlagen für den Bau einer (universellen) Turingmaschine definierte, dauerte es einige Zeit, bis sie in Form von Computern im heutigen Sinne verwirklicht werden konnte. Dazu bedurfte es unter anderem Shannons

80 Turing (2021, S. 65).
81 Turing (2021, S. 65)
82 Stephan/Walter (2021, S. 178).
83 Turing (2004).
84 Zwar haben diese Ideen ihm viel Kritik eingebracht, aber in der Rückschau stellt sich heraus, dass Turing seiner Zeit einfach nur sehr weit voraus war: »Seine unorthodoxen und nicht wirklich erfolgreichen Experimente mit einer Kind-Maschine (Turing 1948, ›Intelligent Machinery‹) waren nämlich nichts weniger als die ersten Experimente mit trainierbaren neuronalen Netzen, die in ihrer Bedeutung erst ab Mitte der 1980er Jahre durch die Arbeiten von David Rummelhart, John McClelland und ihren Kollegen zu mehrschichtigen neuronalen Netzen (wieder!) ernstgenommen werden sollten.« (Stephan/Walter 2021, S. 178)

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Informationstheorie, die den Weg zum Konzept der *binary digits* (später: *bits*) wies.[85] Der erste funktionstüchtige, programmierbare digitale Universalrechner der Welt wurde 1941 fertiggestellt,[86] nämlich die von Konrad Zuse konstruierte Z3, die dieser selbst als »Babbage-Maschine« bezeichnete.[87]

Dabei ist zusätzlich zu beachten, dass bis zur Mitte des 20. Jahrhunderts niemand einem technischen System »intelligente« Fähigkeiten zugeschrieben hätte.[88] Vielmehr ist zu beobachten, dass viele der elementaren Begriffe des Informationszeitalters zur damaligen Zeit noch nicht im heutigen Sinne Verwendung fanden. Man unterschied noch nicht einmal zwischen Information und Nachricht.[89] Erst ab den 1920er Jahren begann die Fachwelt langsam von einer »transmission of information« zu sprechen, wobei der Begriff der Information zu Beginn alles andere als eindeutig und klar war.[90] Als etwa der Elektroingenieur und Mitbegründer der Informationstheorie Ralph Hartley seinen berühmten Vortrag *Transmission of Information* schrieb, notierte er: »›Information‹ is a very elastic term.«[91]

Diese begrifflichen Unklarheiten sollten sich erst 1948 mit Claude Shannons bahnbrechendem Paper *A Mathematical Theory of Communication*[92] langsam auflösen. Der Aufsatz kann mit Recht als eine Art Gründungsdokument für die Informationstheorie angesehen werden. Dabei hatte auch Shannon für den Grundbegriff Information noch keine richtige Definition parat, sondern übernahm diesen einfach von Hartley.[93] Das Problem der damaligen Zeit war, dass man davon ausging,

85 Shannon (1948).

86 Turing war zu dieser Zeit wegen mangelnder Rechenkapazitäten gelangweilt zur theoretischen Biochemie übergewechselt.

87 Zuse (1984, S. 44).

88 Siehe Seising (2021, S. 26).

89 Auch Morse, der Erfinder des gleichnamigen Codes, unterschied noch nicht zwischen Nachrichten und Information, sondern sprach davon, »intelligence« zu übertragen, wie Seising (2021, S. 30) aufzeigt. In der Fachliteratur des 19. Jahrhunderts, so Seising, werde der Begriff der »intelligence« unzählige Male im Sinne von Nachrichtentechnik genutzt. Zum Beispiel werde in einem 1893 von dem Elektroingenieur Edwin James Houston (1893) geschriebenen Fachbuch, das bezeichnenderweise den Titel *The Electric Transmission of Intelligence: And Other Advanced Primers of Electricity* trägt, bereits im ersten Satz erklärt, dass »›die sehr große Geschwindigkeit, mit der Strom übertragen werden kann‹, […] ›[…] für die schnelle Übertragung von Intelligenz zwischen Punkten [besonders geeignet]‹« sei (Seising 2021, S. 29).

90 Seising (2021, S. 30).

91 Hartley (1923).

92 Shannon (1948).

93 Siehe Seising (2021, S. 34).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

dass Informationen übertragen werden und nicht Nachrichten, die dann erst am Ort des Empfängers zu Informationen werden. Die Kontextfrage und auch die Verschachtelung von Kommunikationen wurden zunächst komplett ignoriert (siehe zum Begriff der Information und zu seinem Zusammenhang mit dem Nichtwissen und dem Kontext der Interpretation ausführlich Kap. I.1 in diesem Buch).

Anfang der 1950er Jahre meldete sich im Umfeld von Weaver und Shannon der junge Mathematiker John McCarthy zu Wort. Er arbeitete an der Zusammenstellung eines Sammelbandes zu Automaten, zu dem auch William Ross Ashby, der zuvor das Buch *Design for a Brain*[94] verfasst hatte, einen Aufsatz (*Design for an Intelligence Amplifier*) beisteuerte. McCarthy war mit dem unter dem Titel *Automata Studies*[95] erschienenen Band jedoch unzufrieden, da ihm eher eine Richtung vorschwebte, die dann im Nachgang als Forschung zu »thinking machines« berühmt werden sollte.

Auf McCarthys Initiative hin kam es 1956 zur berühmten Dartmouth-Konferenz, an der viele namhafte Forscher teilnahmen. Hier wurde dann schließlich auch der Begriff künstliche Intelligenz (Artificial Intelligence) erstmals eingeführt. Neben diesem wegweisenden ›Branding‹ ist zudem erstaunlich, dass sich schon im Förderantrag für die Konferenz so gut wie alle wichtigen Aspekte und Ziele der neuen Forschungsrichtung finden:

> »Wir schlagen vor, im Laufe des Sommers 1956 zwei Monate lang eine Studie zur Künstlichen Intelligenz mit zehn Teilnehmern am Dartmouth College durchzuführen. […] Es soll der Versuch unternommen werden, herauszufinden, wie man Maschinen dazu bringen kann, Sprache zu benutzen, Abstraktionen und Konzepte zu entwickeln, um bestimmte Probleme zu lösen, wie sie derzeit nur dem Menschen vorbehalten sind, und sich selbst weiter zu verbessern. Wir glauben, dass in dem einen oder anderen Problembereich bedeutsame Fortschritte erzielt werden können, wenn eine sorgfältig zusammengestellte Gruppe von Wissenschaftlern einen Sommer lang gemeinsam daran arbeitet.«[96]

Nicht nur definieren die Autoren hier en passant die noch lange Jahre gültigen Ziele dieser neuen Disziplin, sondern sie legen auch den Grundstein für die dynamische Definition des Begriffs der künstlichen Intelligenz (mit all den Vor- und Nachteilen, die aus seiner engen Anlehnung an den ebenso nebulösen Begriff der menschlichen Intelligenz erwachsen): »Für den vorliegenden Zweck wird das Problem der Künstlichen Intelligenz so verstanden, dass man eine Maschine dazu bringt, sich in

94 Ashby (1954).
95 Shannon und McCarthy (1956).
96 McCarthy et al., 1955: Förderantrag: http://www-formal.stanford.edu/jmc/history/dartmouth.html [Abruf: 03.07.2023], zitiert nach Seising (2021, S. 56 f.).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

einer Weise zu verhalten, die man intelligent nennen würde, wenn sich ein Mensch so verhielte.«[97]

Diese Zielvorgabe ernst nehmend, können wir davon ausgehen, dass sich die KI-Forschung von Anfang an (sowohl in Dartmouth als auch schon bei Turing) an der Beobachtung von Verhalten und damit implizit an einer Beobachter- und Kommunikationstheorie orientierte. Ob das Verhalten einer Maschine nun ›wirklich‹ intelligent ist oder nicht, wurde aus guten Gründen in den Bereich der Black Box verschoben und nicht weiterverfolgt. Auch wenn die bei der Dartmouth-Konferenz verfolgten Themen und Ansätze dann sehr unterschiedlich ausfielen, war ein gemeinsames Oberthema gefunden, das bis heute trägt: die Vision einer Artificial Intelligence!

Das Modell für Artificial Intelligence ist die menschliche Intelligenz, was die Frage aufwirft, wie man ebendiese zu modellieren hat. In den Folgejahren der Dartmouth-Konferenz wurde diesbezüglich eine Richtung eingeschlagen, die das kognitivistische Paradigma der symbolischen Informationsverarbeitung gegenüber den konnektionistischen Ansätzen der Kybernetik und Hirnforschung bevorzugte. Das Paradebeispiel zur Illustration des die Folgejahre der KI-Entwicklung prägenden Paradigmas der symbolischen Informationsverarbeitung ist Marvin Minskys Buch *The Society of Mind*.[98] Minskys These lautet dabei, dass Intelligenz auf einem komplexen Netzwerk unintelligenter Agenten beruht, die jeweils individuellen Propositionen folgen. Erst durch die Zusammenarbeit dieser vergleichsweise einfachen Module entstehe Intelligenz. Das menschliche Gehirn wird damit nicht mehr als ein einzelnes übergreifendes oder holistisches Organ betrachtet, das beispielsweise blitzartig die ›Großmutter‹ erkennt. Es wird vielmehr als ein aus vielen ›Wesenheiten‹ zusammengesetztes Gebilde konzipiert – deshalb auch der Titel *The Society of Mind*. Jeder der beteiligten einfachen Agenten erledigt dabei bestimmte Aufgaben. Im Fall der Großmutter könnten einige Agenten auf das Gesicht, andere auf Aspekte der Kleidung und wieder andere auf Elemente der Stimme oder auf die üblicherweise von ihr verwendeten Worte fokussieren. Die einzelnen Ergebnisse werden dann von anderen Modulen weiterverarbeitet, um letztendlich das Erscheinungsbild der ›Großmutter‹ zu erzeugen und dann auch erst wiederzuerkennen. Die einzelnen Agenten können wiederum aus noch kleineren Agenten bestehen, die miteinander kommunizieren und verhandeln. Das Erlernen einer kognitiven Fähigkeit bedeutet daher, eine Art Liste im Gedächtnis zu haben, die alle beteiligten Agenten enthält, die für das Wahrnehmen und das Erkennen des Wahrgenommenen notwendig sind, sowie eine Liste von Merkmalen eines spezifischen Wahrnehmungsgegenstands. Für die

97 McCarthy et al., 1955: Förderantrag: http://www-formal.stanford.edu/jmc/history/dartmouth.html [Abruf: 03.07.2023], zitiert nach Seising (2021, S. 56 f.).

98 Der deutsche Titel lautet Mentopolis (Minsky 1990).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Großmutter könnten dies ›blaugraue Augen‹, ›Falten im Gesicht‹, ›ein gebeugter Gang‹, ›eine zittrige Stimme‹ usw. sein. Die Idee besteht darin, dass das Erkennen kognitivistisch durch die logische Kombination von elementaren Propositionen erfolgt. Sobald eine bestimmte Anzahl von Merkmalen positiv festgestellt wurde und Widersprüche beseitigt wurden, erhält man das Ergebnis.

Dieses auf den ersten Blick so plausible Modell eines »general problem solver«, das auf der Dartmouth-Konferenz von Newell, Shaw und Simon[99] präsentiert wurde, schien zunächst den Weg für die KI-Entwicklung zu weisen. Die symbolische Informationsverarbeitung stand für die Zukunft. Die Imitation von Hirnmodellen nach dem Black-Box-Prinzip, wie es von Turing vorgeschlagen worden war, und die ersten hieraus entwickelten neuronalen Netzwerke standen demgegenüber für die Vergangenheit. Dies sollte sich erst Jahrzehnte später wieder ändern.

Die Geschichte der KI ist somit eine Geschichte von Debatten, Konkurrenzen sowie Paradigmenwechseln und konkurrierenden Visionen. Eine Denkrichtung, vertreten von Forschern wie Newell, Shaw, Simon und Minsky, verfolgte den Ansatz einer symbolischen Darstellung der Welt und unseres Wissens darüber. In engem Zusammenhang damit stand der Ansatz einer »heuristischen Suche«, der davon ausging, dass Intelligenz im Wesentlichen ein Problem der Erkundung eines Raums von endlichen Antwortmöglichkeiten ist. Man müsse nur die richtigen Fragen finden, hieraus ein Set von Elementarpropositionen ableiten, woraus sich dann im Sinne der Regeln des logischen Schlussfolgerns die richtige Antwort ergebe. Man müsse eben all dies nur noch in eine Maschine implementieren.

Der zweite Ansatz – der dem konnektionistischen Paradigma folgt – ist demgegenüber inspiriert von der Funktionsweise des menschlichen Gehirns. Hier trifft man auf komplexe Netzwerke und Muster von feuernden Neuronen, nicht jedoch auf eine logische Anordnung von Agenten, die im Falle unseres Beispiels in einem ›Großmutterneuron‹ enden, das aufblinkt, falls die Kette der Symbolverarbeitung eben dorthin führt. Die Vertreter des konnektionistischen Ansatzes blieben jedoch bis in die 1980er Jahre hinein weitgehend in der Minderheit.[100]

Der Unterschied beider Ansätze ist einer ums Ganze:[101] Bei symbolischer KI geht es um die explizite Formulierung von Regelschritten und um die Codierung von Weltwissen, das sodann in sogenannte

99 Newell, Shaw und Simon (1959).

100 Dies wohl auch aufgrund mangelnder Computerleistung, siehe Mitchell (2019).

101 Die Geschichte dieser beiden Denkansätze und die mit ihnen verbundenen philosophischen Ideen werden in Hubert Dreyfus' und Stuart Dreyfus' Dädalus-Aufsatz Making a Mind Versus Modeling the Brain ausführlich diskutiert (Dreyfus & Dreyfus 1988).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Expertensysteme implementiert wird. Diese Systeme werden sequenziell programmiert, sind in ihrer Struktur vom Menschen les- und verstehbar und ihre Lernfähigkeit ist auf die externe, also vom Programmierer festgelegte Erweiterung oder Veränderung der Regelstrukturen in ihrer Wissensdatenbank beschränkt.[102] In den 1980er und 1990er Jahren musste der symbolische Ansatz dann jedoch – nicht zuletzt mangels Flexibilität – den sogenannten KI-Winter durchleben. Enttäuschende Ergebnisse führten zur Reduktion der finanziellen Förderung und umgekehrt. Konnektionistische Modelle hingegen begannen nun an Zugkraft zu gewinnen, nicht zuletzt auch mit der Entwicklung des Verfahrens der Backpropagation.[103]

Die konnektionistischen Modelle sind selbstlernend. Sie gehen jedoch mit einem Verlust des Wissens darüber einher, wie genau eine spezifische Kognition zustande kommt.[104] Man hat ein Ergebnis – die Großmutter wird erkannt –, aber man verfügt nicht über eine Liste von Kriterien, die logisch nachvollziehbar macht, wie das kognitive Aggregat zu dem Ergebnis gekommen ist.

Künstliche neuronale Netzwerke

Gehen wir zunächst wieder einige Jahrzehnte zurück: Im Jahr 1943 legten der Psychiater und Neuroanatom Warren S. McCulloch und der junge Autodidakt Walter Pitts den Artikel *A Logical Calculus of the Ideas Immanent in Nervous Activity* vor, in dem sie einen logischen Kalkül präsentierten, der die Funktionsweise von Nervenzellen in die Nähe von elektrischen Schaltkreisen rückte:

> »Aufgrund der Erregungen beziehungsweise Hemmungen, die ihnen [den Nervenzellen; Anm. von uns] von anderen Neuronen im Netz übertragen wurden, wurden sie selbst ›erregt‹ beziehungsweise ›gehemmt‹. Dies sind also auch hier die einzigen beiden Zustände, die eine solche abstrakte Nervenzelle – ein ›McCulloch-Pitts-Neuron‹ (MCP) oder ein künstliches Neuron, wie man später auch sagte – einnehmen kann.«[105]

McCulloch und Pitts kombinierten hier die Ideen der beginnenden Kognitionswissenschaft mit denen der Neurophysiologie wie auch der

102 Siehe Alpaydın (2021, S. 60 ff.).
103 Siehe Rumelhart et al. (1986) sowie Hochreiter und Schmidhuber (1997).
104 Der Computeringenieur François Chollet hat die hauptsächlichen Unterschiede zwischen den beiden Paradigmen (symbolische und konnektionistische Verfahren) in folgender Art und Weise deutlich gemacht: Während das sequenzielle Paradigma auf der Formel »Regeln + Daten = Antworten« beruht, basiert das konnektionistische auf der Idee »Daten + Antworten = Regeln« (Chollet 2021, S. 4).
105 Seising (2021, S. 75).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Informationswissenschaft und formulierten sogar schon die Idee, dass alle geistbegabten Wesen im Prinzip nichts anders als ›Rechenmaschinen‹ sind.[106]

Das Besondere des Kalküls im McCulloch-Pitts-Neuron war, dass es eine große Nähe zu den logischen Verhältnissen aufwies, die in elektrischen Schaltkreisen vorzufinden sind:

> »Die Aktivität eines Neurons wurde ›Feuern‹ genannt, und die Analogie besagt nun, dass das ›Feuern‹ oder ›Nichtfeuern‹ der Neuronen dem Fließen oder Nichtfließen von Strom in Schaltkreisen entspricht, das wiederum für die logischen Wahrheitszustände ›wahr‹ und ›falsch‹ steht.«[107]

Entsprechend verschaltete Neuronen könnten nun also ebenso wie entsprechend verschaltete elektrische Kreisläufe unterschiedliche logische Werte einnehmen: AND, OR, NO, NOR. Dazu konzipierten McCulloch und Pitts ihre Neuronen als Schwellenwertelemente, die erst ab einem gewissen Erregungspotenzial aktiv werden und sonst still bleiben. McCulloch-Pitts-Neuronen stellen also gewissermaßen über binäre Relais realisierte Neuronen dar und können daher wie biologische Nervenzellen die zuvor angedeuteten logischen Operationen ausführen. Koppelt man solche Zellen zu einem Verbund, entstehen leistungsstarke Netzwerke, die sowohl die komplexe Aussagenlogik beherrschen als auch – so der Anspruch von McCulloch und Pitts – die Funktion einer Turing-Maschine erfüllen können sollen.

Neben diesen bahnbrechenden Vorschlägen von McCulloch und Pitts waren es auch die aus neurophysiologischen Experimenten abgeleiteten Prinzipen der Strukturierung von Nervenzellen, die weitere ausschlaggebende Hinweise darauf gaben, wie das Gehirn – und damit eine mögliche künstliche Imitation desselben – funktionieren könnte. Diese Hinweise kamen von dem Psychologen Donald O. Hebb, dessen 1949 vorgestellte Theorien zur Gehirnfunktion als Hebbsche Lernregel bekannt werden sollten. Diese besagt, dass es, wenn zwei im Netzwerk miteinander verbundene Neuronen gleichzeitig feuern, an den Synapsen zu Veränderungen kommt, die die Verbindungen so umformen, dass die wechselseitige Erregbarkeit erleichtert wird. Diese Regel fand bald eine eingängige Formulierung: »Neurons that fire together, wire together.«[108]

Diese Regel gilt auch in umgekehrter Weise: So dünnt die Verbindung bei Inaktivität aus. Hebb formulierte dabei ebenfalls schon die Idee, dass eine Ansammlung von sehr vielen solcher miteinander verbundener Nervenzellen (»cell assemblies«) als physisches Gebilde die Begriffe

106 Auch Gotthard Günther bewegte sich ab den 1960er Jahren im Umfeld von McCulloch, Ashby und von Foerster. Siehe Günther (1975, S. 18).
107 Seising (2021, S. 75).
108 Hebb (1949).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

und Konzepte unseres Wissens repräsentieren kann. Bereits bei Hebb findet sich somit der Hinweis auf die erkenntnistheoretische Besonderheit des konnektionistischen Paradigmas, die Repräsentation von Wissen nicht auf einen konkreten Speicherort zu beziehen, sondern auf die über die Interaktionshäufigkeit gebildeten Wahrscheinlichkeitsmuster eines Netzwerks.

Die spezifische Leistung künstlicher neuronaler Netze besteht darin, dass sie nicht mehr propositional – also über die Verkettung von Aussagen mit einem spezifischen Wahrheitswert – organisiert sind, sondern konnektionistisch funktionieren. Statt in den (extern und regelbasiert) vorgegebenen Bahnen zu lernen, knüpfen sie eigenständig verstärkende oder hemmende Verbindungen zwischen den Einheiten im Netzwerk. Damit entsteht gleichsam ein ›Lernen‹, das nicht von vornherein weiß, welche Bedeutung es später haben wird.

Somit beginnen die auf diese Weise gestalteten künstlichen neuronalen Netzwerke, uns Menschen nun auch den Spiegel vorzuhalten,[109] wie Heinz von Foerster aufzeigt. Denn sobald sich herausgestellt hat, dass sie funktionieren, liefern sie uns auch die Metaphorik für das Verständnis unseres eigenen Gehirns:

> »Denn das Gehirn besteht ja aus Neuronen, die über die Synapsen und die Axiome miteinander gekoppelt sind. Ein solches Nervennetz läßt sich dann als ein Rechner verstehen, der Induktionen und Deduktionen gewisser Aussagen und Beobachtungen durchführt. Aber zurück zur Geschichte: Es war der berühmte Mathematiker John von Neumann, der die [McCulloch/Pitts 1948; Anm. von uns] Arbeit von McCulloch und Pitts zum Bau von Computern verwendete. Er zeigte, daß die Turing-Maschine und das neuronale Netz, dessen Funktionsweise McCulloch und Pitts skizziert hatten, äquivalente Operatoren darstellen. Auf diese Weise kam schließlich die Computermetapher ins Spiel, die noch heute in der kognitionswissenschaftlichen Forschung prägend ist: Man glaubte, die neuronalen Strukturen, aus denen das Gehirn besteht, nachzubilden, indem man einen Elementarcomputer schuf, der auf den Einsichten von McCulloch und Pitts basierte. Der Bau von Computern, die vermeintlich nach den Prinzipien der Neuronen funktionierten, gestattete schließlich den Rückschluß: Das Gehirn erschien als ein gewaltiger Parallelcomputer.«[110]

Oliver Selfridge, ein britischer Informatiker, der als einer der Pioniere in der Erforschung und Entwicklung von künstlichen neuronalen Netzwerken gilt, war dann einer der ersten Verfechter der Idee, dass Computer nicht nur einfache Rechenmaschinen sind, sondern auch in der Lage sein

109 Siehe hierzu als ersten Eindruck der spielerischen Selbstbezüglichkeit Thomas et al. (2022).

110 Foerster und Pörksen (2022, S. 110 f.).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

sollten, Probleme auf kreative und intelligente Weise zu lösen. In seinem einflussreichen Paper *Pandemonium: A Paradigm for Learning*[111] stellte Selfridge ein System zur Mustererkennung vor. Dieses System bestand aus einer Reihe einzelner »Dämonen«, die jeweils spezialisiert auf ein bestimmtes Merkmal in den Daten reagieren sollten. Wenn ein Dämon ein Muster »wiedererkannte«, sollte dieser »aufschreien« (ein Signal ausgeben). Die Lautstärke des Schreis hing davon ab, wie sehr das erkannte Muster mit dem vom Dämon gelernten Muster übereinstimmte. Das System würde dann die Entscheidung des lautesten Dämons als seine Ausgabe verwenden. Konkret sollte dieser »Pandämon« auf einer Eingabemaske aus 90 × 90 schwarzen oder weißen Zellen Muster von Einsen und Nullen erkennen. Wenn die Übereinstimmung der erkannten Zahlen mit den gespeicherten Symbolen groß genug war, konnte das System die Bilder den Symbolen zuordnen. Obwohl dieses Modell stark vereinfacht war, legte es so die Grundlage für spätere und viel komplexere Modelle neuronaler Netzwerke. Selfridges Arbeiten zur Mustererkennung und zum maschinellen Lernen boten damit einen ersten praktischen Ansatz zur Simulation, wie das menschliche Gehirn visuelle Inputs zu verarbeiten lernt, und inspirierten viele nachfolgende Forschungen auf diesem Gebiet.

Ein weiterer wichtiger Schritt für die praktische Umsetzung des konnektionistischen Paradigmas findet sich in den Arbeiten von Frank Rosenblatt. Ebenfalls inspiriert von den neurophysiologischen Erkenntnissen über das menschliche Gehirn und die Informationsverarbeitung von Neuronen, konzipierte er in den späten 1950er und frühen 1960er Jahren das sogenannte Perzeptron,[112] das vom Prinzip her ein erster, einfacher Algorithmus für überwachtes Lernen binärer Klassifikatoren ist. Es besteht aus einem einzelnen »neuronalen« Element, das mehrere Eingaben empfängt und diese in einer einzigen Ausgabe abbildet. Die einzelnen Eingaben werden jeweils mit einem Gewicht multipliziert und dann summiert. Das Perzeptron-Modell verwendet dafür *weights* und *biases*, die sich – wie die synaptischen Verbindungen von Neuronen – durch Interaktionen dynamisch verändern: Wenn die Summe einen bestimmten Schwellenwert überschreitet, gibt das Perzeptron eine Aktivierung (Wert 1) aus; andernfalls bleibt es inaktiv (Wert 0). Durch eine auf der Differenz zwischen der tatsächlichen und der gewünschten Ausgabe basierende Anpassung (Gewichtung) kann das Perzeptron dann ›lernen‹, eine Vielzahl von Mustern zu erkennen und zu klassifizieren. Das Perzeptron war damit das erste Modell, das den Prozess der Informationsverarbeitung in einem Neuron konsequent nachbildete und damit einen wichtigen Grundstein für maschinelles Lernen legte.

111 Selfridge (1958).
112 Rosenblatt (1958).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Die Zeit war jedoch noch nicht reif, diesen Weg konsequent weiterzuverfolgen. Die Protagonisten des symbolischen bzw. kognitivistischen Paradigmas – allen voran Marvin Minsky und Seymour Papert[113] – konnten noch glaubhaft postulieren, dass das Perzeptron bestimmte mathematische Probleme einfach nicht lösen könne und man diesen Weg in Hinblick auf die künstliche Intelligenz nicht weiterverfolgen solle.

Ein umfassender Überblick über die weitere Geschichte des maschinellen Lernens, insbesondere des Deep Learnings, würde den Rahmen dieses Kapitels leider sprengen.[114] Versuchen wir hier jedoch zumindest einen kurzen Überblick zu geben: Nach den ersten theoretischen und konzeptionellen Überlegungen in der ersten Hälfte des 20. Jahrhunderts, mit denen sich die grundlegenden Prinzipien für künstliche neuronale Netzwerke herauskristallisierten, startet dann ab den 1980er Jahren eine zweite Welle. Vor allem durch das Prinzip der Backpropagation wurde es möglich, neuronale Netze mit mehreren verborgenen Schichten zu trainieren.[115] Mit dem Aufkommen von Big Data und leistungsstärkeren Computern im 21. Jahrhundert begann die Ära des Konnektionismus zur Blüte zu kommen. Unter dem Label des Deep Learnings schließlich breitete sich ab 2006 rapide die dritte Welle maschinellen Lernens aus.[116]

Der bis heute anhaltende Erfolg des konnektionistischen Paradigmas wird vor allem von zwei Hauptfaktoren gestützt: erstens der Verfügbarkeit von umfangreichen Datensätzen und zweitens der Rechenleistung moderner Hardware, insbesondere von Grafikprozessoreinheiten (GPUs), die das Training tiefer neuronaler Netzwerke in praktikabler Zeit ermöglichen. Die (Wieder-)Entdeckung des Paradigmas und die Umsetzung seiner Ideen bei der Modellierung künstlicher neuronaler Netzwerke wurden dann vor allem durch die überragenden Ergebnisse und Leistungen in Bereichen befeuert, in denen die konventionellen Expertensysteme auf Basis des symbolisch-kognitivistischen Paradigmas bislang versagt hatten. Mit Deep Learning konnten KI-Forscher schließlich Modelle entwickeln, die bei Aufgaben wie der Bild- und Spracherkennung menschenähnliche oder sogar übermenschliche Leistungen erbringen konnten.

In diesem Zusammenhang ist darauf hinzuweisen, dass gerade die Sprachverarbeitung bis zum Anfang der 2000er Jahre noch eine Domäne war, die von Computerlinguisten dominiert wurde, sodass es aufstrebenden KI-Wissenschaftlern wie beispielsweise Richard Socher schwerfiel, auf Fachkonferenzen Studien zu platzieren, die aufzeigten, dass unter dem konnektionistischen Paradigma bessere Ergebnisse möglich sind als mit Syntaxbäumen und regelbasierten Zugriffen auf vorgegebene

113 Minsky und Papert (1969).
114 Siehe als Einstieg in die Thematik vor allem Goodfellow et al. (2016).
115 S. hierzu Rumelhart et al. (1986) sowie Hochreiter und Schmidhuber (1997).
116 Siehe Hinton et al. (2006).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Lexika. Während man sich bei der Bilderkennung schon an Deep Learning gewöhnt hatte, war es bei der für das menschliche Bewusstsein so zentralen Domäne der Sprache unvorstellbar, dass deren Prozessieren ohne semantische Bezüge – also ohne Sinn für die Bedeutung von Wörtern – möglich ist.[117] Im Nachhinein zeigte sich, dass gerade die zunächst präferierten vektorbasierten lexikalischen Systeme jedoch prinzipiell nicht imstande sind, die kontextuelle und kompositionelle Bedeutung längerer Textpassagen aufzuschließen und zu prozessieren, was sie an einem tieferen Verständnis der Sprache hindert.[118] So stehen wir vor der paradoxen Situation, dass besseres Sprachverstehen erst auf Basis neuronaler Netzwerke gelingt, die auf das Wissen um die lexikalische Bedeutung und auf vordefinierte Regeln – also auf inhaltliches Verstehen – verzichten!

Dabei orientieren sich gerade die künstlichen neuronalen Netze, die heutzutage bei der Spracherkennung führend sind, vom Prinzip her weiterhin an den biologischen Vorbildern. Um es nochmals zu wiederholen: Bestimmte Input-Schwellen lösen bestimmte Aktivierungspotenziale aus, die dann wieder als Inputs für nachrangige Knoten bzw. Neuronen zur Verfügung stehen. Dabei kann jede Einheit als ein Neuron interpretiert werden, das an andere Neuronen Informationen über seine Verbindungen (Synapsen) sendet. Die Stärke dieser Verbindungen, das heißt, wie viel von einem Signal durchgelassen wird, wird durch Gewichte bestimmt, die im Rahmen des Lernprozesses angepasst werden. Das Netzwerk lernt, indem es die Fehler, die es macht, verwendet, um die Gewichte der Verbindungen so zu ändern, dass diese Fehler in der Zukunft minimiert werden. Dieses Prinzip nennt man Backpropagation.

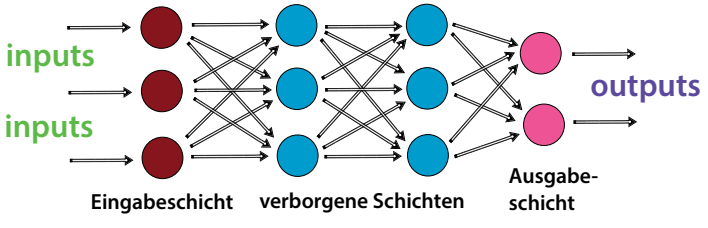

Abb. 6: Diagramm eines mehrschichtigen künstlichen neuronalen Netzes. Die Punkte repräsentieren die Neuronen, die Pfeile die gewichteten Verbindungen.

117 Siehe hierzu den ZEIT-Online Podcast »Richard Socher, was denken Maschinen?«, vom 20. November 2020. https://www.zeit.de/digital/2020-11/richard-socher-kuenstliche-intelligenz-interviewpodcast-alles-gesagt?utm_referrer=https%3A%2F%2Fwww.google.com%2F [Abruf: 03.07.2023].

118 Siehe Socher (2012).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Das sogenannte Deep Learning beruht dann auf einem mehrschichtigen neuronalen Netzwerk, das es ermöglicht, hierarchische Konzepte zu erlernen und hiermit einhergehend komplexe Probleme zu lösen. Dieses Netzwerk besteht aus einer Eingangsschicht, die die rohen Eingabedaten verarbeitet, teils mehreren hundert bis tausend verborgenen Schichten, die diese Daten in immer abstraktere Konzepte umwandeln, und einer Ausgabeschicht, die den endgültigen Output liefert.[119]

Moderne neuronale Netze, die auf selbstlernenden Mechanismen beruhen, erzeugen Ausgaben, die sogar ihre Schöpferinnen und Schöpfer oft überraschen. Sie konstruieren ihren Zugang zur Welt mehr und mehr selbstständig. Demis Hassabis, CEO von DeepMind, verdeutlicht diese Veränderung prägnant: »Die auffälligste Änderung für mich ist, dass wir keine Daten mehr eingeben müssen.«[120] Dieser von Alan Turing bereits 1948 unter dem Begriff »unorganized machines«[121] gefasste Ansatz findet heute im Kontext von »unsupervised learning«[122] Anwendung.

Er wird etwa in Modellen wie MuZero erfolgreich eingesetzt. Diese Systeme können beispielsweise Spiele wie Go oder Schach erlernen, ohne dass sie zuvor mit unzähligen Partien menschlicher Spieler und Spielerinnen gefüttert werden müssen. Sie brauchen diese Daten als Trainingsmaterial nicht mehr – sie müssen nur gegen sich selbst spielen und erreichen hierdurch in kurzer Zeit ein Trainingsniveau, das ausreicht, um jeden menschlichen Spieler zu schlagen.[123]

In der formalen Beschreibung sehen die neuronalen Netze des Deep Learnings aus wie Netze von Knoten, die in Schichten oder Stapeln angeordnet und in einem dreidimensionalen Netzwerk verbunden sind. Jeder Knoten im Netz ist mit einem Neuron im Gehirn vergleichbar, und wie einem Gehirn kann auch dem Netz beigebracht werden, jedes beliebige Muster zu erzeugen oder zu erkennen. Diese Muster werden als ›embeddings‹ bezeichnet. Wenn ein neuronales Netz beispielsweise lernt, Katzen und Hunde auf Bildern zu erkennen, könnte man sagen, dass das Konzept von Katzen und Hunden in das neuronale Netz ›eingebettet‹ ist. Mit genügend Daten und genügend Kombinationsmöglichkeiten in der Netzwerkstruktur kann man alles in ein neuronales Netz einbetten, auch abstrakte Konzepte wie die darwinsche Evolutionstheorie oder die Mathematik der Quantenphysik. Was ein neuronales Netz ›tief‹ macht, ist einfach die Anzahl der Schichten (layers), über die es verfügt (siehe Abbildung 6). Je mehr Schichten, desto tiefer ist es. Die größten Netzwerke

119 Siehe Goodfellow et al.. (2016).
120 Zitiert nach GQ-Magazin: »Es gibt nur noch einen Gegner für Googles KI AlphaGo«. »https://www.gq-magazin.de/auto-technik/article/googles-ki-alphago-lernt-von-alphago [Abruf: 03.07.2023].
121 Turing (2004 [1948]).
122 LeCun et al. (2015).
123 Siehe Schrittwieser et al. (2020).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

haben heute Hunderte von Schichten, und jede Schicht kann Hunderte oder Tausende von Knoten haben. Zum Zeitpunkt der Abfassung dieses Buches haben die größten neuronalen Netze der Welt mehrere Billionen Parameter.[124]

Kommerziell werden diese Verfahren seit nun mehreren Jahren in vielfältigsten Bereichen erfolgreich zur Automatisierung, Rationalisierung oder zeitlichen Organisation von Abläufen eingesetzt. Selbst bei der alltäglichen Nutzung eines Smartphones greifen Menschen auf eine Vielzahl von KI-Technologien zurück. Spätestens mit der Nutzung aktueller, kreativer und für jeden zugänglicher KI-Modelle wie DALL-E, Midjourney, ChatGPT oder Stable Diffusion ist die Leistungsfähigkeit des maschinellen Lernens und der ihm zugrunde liegenden neuronalen Netzwerkstrukturen auch im alltäglichen Leben vieler Menschen deutlich geworden.

Wie schon mehrfach mit Verweis auf die Besonderheiten des konnektionistischen Paradigmas angedeutet, geht der Erfolg dieser künstlichen Intelligenzen jedoch mit dem Unbehagen einher, nicht mehr in Form propositional organisierten Wissens nachverfolgt werden zu können. Auch dem Laien fällt nun zunehmend auf, dass diese digitalen Technologien dadurch bestechen, dass sie jede ihrer Oberflächen (»surface«) mit einer verborgenen Tiefe (»subface«) koppeln.[125] Vor allem die zunehmende Komplexität der Tiefen dieser Netzwerke verweist auf einen inzwischen vielfach diskutierten Interpretationsbedarf der Genese ihrer Leistungen.[126] Die stetig steigende Komplexität der Netzwerkarchitekturen wirft daher – gerade auch mit Blick auf ethische Fragen der Anwendung künstlicher Intelligenz – die immer dringender werdende Frage nach dem Verständnis ihrer (internen) Funktionsweise auf.[127]

Es ist nicht mehr klar, was im Inneren passiert. Man lässt das Netzwerk sich selbst organisieren. Aber was bedeutet es wirklich, dass es sich selbst organisiert? Durch ausreichendes Training können die Einstellungen eines Netzes erwiesenermaßen so weit geändert werden, dass es Daten sinnvoll klassifizieren kann – aber was genau ist mit diesen Einstellungen und Gewichtungen gemeint? Mit welchen Bildmerkmalen befasst sich eine objekterkennende KI und wie setzt sie diese zu den unverwechselbaren visuellen Signaturen von Gesichtern, Proteinstrukturen oder Kaffeetassen zusammen? Ein Blick auf die Gewichtung einzelner Verbindungen wird hierfür keine Antwort liefern können. Und die aufwendige Rekonstruktion der statistischen Gewichtungen führt manchmal zu überraschenden Befunden: So unterschied – wie in der Einleitung dieses

124 Siehe zum Stand der Entwicklung im Jahr 2023 Thompson (2023).
125 Nake (2008).
126 Siehe Esposito (2017).
127 Burrell (2016).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Buches bereits geschildert – eine mit Pferdebildern trainierte KI Pferde anhand von Bildnachweisen, nicht jedoch, wie es eigentlich angedacht war, anhand der Gestalt der Tiere.[128]

Diese Problematik wird schon lange unter dem Begriff der Black Box verhandelt, einer Metapher, die aufzeigen soll, dass man zwar beobachten kann, welche Inputs in die Maschine eingeführt und welche Outputs daraufhin von ihr generiert werden, dass es aber im Bereich der Intransparenz verbleibt, wie und warum genau diese Relation von Input und Output zustande gekommen ist. Dies hatte bereits der Kybernetiker Norbert Wiener erkannt:

> »I shall understand by a black box a piece of apparatus […] which performs a definite operation […] but for which we do not necessarily have any information of the structure by which this operation is performed.«[129]

Die Idee der Black Box meint somit, dass komplexe Objekte für einen Beobachter prinzipiell undurchschaubar sind und jeder Versuch des Verstehens zu einem Versuch der Kontrolle wird.[130] Für Dirk Baecker heißt »Kontrolle« in diesem Kontext, dass der Beobachter seine Interaktion mit dem Objekt nutzt, »um aus der Variation von Input und Output Rückschlüsse auf eine mögliche Transformationsfunktion innerhalb des Objekts zu ziehen«[131]. Streng genommen kann jeder Beobachter dabei jedoch nur Rückschlüsse auf seine Interaktion mit dem komplexen Objekt ziehen.

Aus soziologischer Perspektive ergibt sich hiermit eine herausfordernde Frage: Was passiert mit einer Gesellschaft, wenn sich zu den intransparenten Bewusstseinssystemen der Menschen noch intransparente Maschinen hinzugesellen, die zwar auf unterschiedliche Weise operieren mögen als die menschlichen Psychen, aber dennoch – oder (und das wäre das soziologisch Spannende) gerade deshalb – zur gesellschaftlichen Kommunikation beitragen? Denn wenn die Dinge eindeutig, sicher oder klar verständlich wären, müsste man sich kommunikativ ja nicht darum bemühen, sie zu verstehen.

Umgekehrt lässt sich im Sinne der Metapher des vorgehaltenen Spiegels jedoch auch fragen, was all dies über uns Menschen als soziale Wesen aussagt. Denn da – um ein zentrales Thema dieses Buches aufzugreifen – Existenz selektive Blindheit ist,[132] können alle Verfahren der Modellierung künstlicher Intelligenz im Prinzip auch als Versuche verstanden werden, die Black Box der biologischen neuronalen Netze zu

128 Lenzen (2022, S. 78).
129 Wiener (1948, S. xi).
130 So bereits die Einsicht von Ashby (1956).
131 Baecker (2023).
132 Siehe Spencer-Brown (1997, S. 192).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

rekonstruieren. »Solving intelligence« lautet folgerichtig auch der Claim des Unternehmens DeepMind.[133] Während es jedoch bei artifiziellen Netzen noch einigermaßen funktionieren mag, mittels Backpropagation die Netzwerkstruktur des Erlernten zu beobachten und zu rekonstruieren, ist dies bei biologischen Netzen wie unserem Gehirn nicht möglich. Wir können unser Lernen, unseren Lernalgorithmus nicht rückwärtslaufen lassen, um zu beobachten, welche Änderung und welche Gewichtung zu anderen Outputs führen würden. Lernen vollzieht sich permanent und ist unumkehrbar. Dies macht künstliche Intelligenzen als Spiegel unserer eigenen Existenz so interessant. Wenn sie Ähnliches zu leisten vermögen wie wir, dann können wir auf Basis der Analyse ihres Aufbaus und ihrer Prozesse Rückschlüsse auf unsere eigenen kognitiven Operationen ziehen.

Noch bleibt Deep Learning zwar auf hochspezialisierte und mehr oder weniger isolierte Musterkennungen begrenzt. Interessant wird es jedoch, wenn Versuche unternommen werden, die offen verfügbaren Datenbanken und Repositorien miteinander zu verknüpfen und auf diese Weise etwa multimodale künstliche Intelligenzen zu erzeugen, die sehen, hören, denken und handeln können. Es bleibt somit abzuwarten, was passiert, wenn derartig zusammengefügte Netzwerke ihrerseits beginnen, eigenständig Lösungen zu erkunden, wie sie unterschiedliche Datenformen und -quellen zu einem Muster verbinden können – wenn also auch sie gemäß Gregory Batesons Anweisung danach Ausschau halten, welches »Muster« die Muster »verbindet«[134].

Bei alldem garantiert freilich nichts, dass bei einer gefundenen Lösung die verschiedenen Muster in harmonischen Rückkoppelungen einander zuarbeiten. Wie bereits Arthur C. Clarke in seinem Roman *2001: A Space Odyssey* vorwegnimmt (siehe ausführlich das vorangehende Kap. II.2), ist es gerade aus theoretischen Gründen eher wahrscheinlich denn unwahrscheinlich, dass eine elaborierte KI auch Verhaltensweisen entwickelt, die als neurotisch oder gar psychotisch klassifiziert würden. Die zu regelnde Komplexität eines Netzwerks, das mehrere Netzwerke integriert, übersteigt die Kapazität jedes denkbaren Reglers.[135]

133 Auf der Website www.deepmind.com heißt es: »We research and build safe artificial intelligence systems. We're committed to solving intelligence, to advance science and benefit humanity.« [Abruf: 08.07.2023] In einem Interview ergänzt Hassabis dieses Ziel: »solving intelligence, and then using that to solve everything else.« https://www.technologyreview.com/2016/03/31/161234/how-google-plans-to-solve-artificial-intelligence/ [Abruf: 08.07.2023]

134 Bateson (1987, S. 15).

135 Siehe zu einer literarischen Erkundung dieses Themas auch den Roman Summa technologiae von Stanisław Lem (1976).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Große Sprachmodelle (Generative Pretrained Transformer)

Bis vor Kurzem war kein System dazu in der Lage, über eine gewisse Zeit ein einigermaßen kohärentes Gespräch mit einem Menschen zu führen, und auch der Durchbruch AI-basierter Sprachübersetzungen ist noch nicht lange her, was – wie bereits geschildert – auch an dem innerhalb der Computerlinguistik vorherrschendem kognitivistischen Paradigma lag.

Dabei war die Entwicklung autonomer Computersysteme, die eine kohärente Konversation mit einem Menschen zu führen vermögen, schon lange ein zentrales Anliegen der Forschung zu künstlicher Intelligenz und Mensch-Computer-Interaktion. Die Geschichte dieses Anliegens reicht mindestens bis zu Weizenbaums bahnbrechendem Eliza-System von 1966 zurück.[136] Wie bereits gezeigt, hatte auch Turing das kommunikative »imitation game«[137] als Maßstab für die Identifizierung und Klassifizierung künstlicher Intelligenz beschrieben. Doch erst als die konnektionistischen, rein statistikbasierten Sprachmodelle – sozusagen wider Erwarten – ihre Leistungsfähigkeit in der Praxis zeigen konnten, änderte sich auch hier das Paradigma. Die jüngsten Entwicklungen im Bereich des maschinellen Lernens, insbesondere die Transformer-Technologie,[138] führten in den letzten Jahren zu gleichsam revolutionären Verbesserungen in der kohärenten Texterzeugung von Computern, was sich entsprechend in ihrer Fähigkeit zur Gesprächsführung niederschlug.

Das Aufkommen großer Sprachmodelle (Large Language Models, kurz: LLMs) wie BERT[139] und GPT-2[140] wurde somit für die künstliche Intelligenz zu einem entscheidenden Wendepunkt – denn wenn Maschinen nicht nur überzeugend an menschlicher Kommunikation teilhaben können, sondern immer mehr ihren eigenen Beitrag dazu leisten, dann wird ihre soziale Teilhabe auch zu einem gesellschaftlichen Phänomen. Die »Geschichte von Morgen« beginnt, so Yuval Harari.[141] Die auf der Transformer-Technologie basierenden Nachfolger wie GPT-3 oder aktuell GPT-4 (OpenAI, 2023) umfassen Hunderte von Milliarden Parameter[142] und wurden mit Hunderten von Terabytes an Textdaten trainiert – einem kulturellen Wissensvorrat unvorstellbar großen Ausmaßes.

136 Weizenbaum (1966).
137 Turing (1950).
138 Um hier zwei Meilensteine auf dem Weg zur Transformer-Technologie zu nennen: Sutskever et al. (2014) sowie Vaswani et al. (2017).
139 Devlin et al. (2018).
140 Radford et al. (2019).
141 Harari (2017).
142 Unter den Parametern eines neuronalen Netzes versteht man die Gewichtungen der jeweiligen Verbindungen im Netzwerk. Diese Werte werden in der Trainingsphase vom Modell festgelegt. Während das Modell lernt, lassen sich die Parameter daher nicht direkt kontrollieren. Gleichwohl kann

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Vor allem das Unternehmen OpenAI hat nicht nur die Leistungsfähigkeit dieser Modelle in Form von ChatGPT vorgeführt, sondern auch ihren Erfolg eingeleitet, indem es den Zugang zu LLMs über eine einfach zu bedienende grafische Benutzeroberfläche ermöglichte. Die eingangs zitierte »Unfähigkeit« künstlicher Systeme, ein einigermaßen kohärentes, plausibles und interessantes Gespräch zu führen, wurde spätestens durch die im Herbst 2022 veröffentlichte Variante ChatGPT aufgehoben.[143] ChatGPT hat in den ersten zwei Monaten mehr als 100 Millionen Nutzer gewonnen und produziert nun schätzungsweise alle 14 Tage eine Textmenge, die allen gedruckten Werken der Menschheit entspricht.[144] Plötzlich konnte die ganze Welt kostenlos erfahren, was es bedeutet, in kommunikativen Austausch mit einer KI zu gehen. Die Trajektorie der Entwicklung führte mit ChatGPT zu einem Kulminationspunkt, zu einem gesellschaftlichen Aha-Erlebnis. Zum ersten Mal in der Menschheitsgeschichte kann praktisch jeder Mensch in einer sinnhaften Weise mit einer KI in Kontakt treten. Darüber hinaus stehen die LLMs kurz davor, als Suchmaschinen und virtuelle Assistenten in vielfältige Lebensbereiche integriert zu werden und auch auf diese Weise umfassend in die Gesellschaft hineinzuwirken.[145]

Gehen wir kurz auf einige Aspekte der Entwicklung dieser Modelle ein. Aufbauend auf der sogenannten Transformer-Technologie,[146] stellte das KI-Konsortium OpenAI Ende 2020 die dritte Version ihres Sprachmodells unter dem Titel GPT-3 vor. Bereits mit den ersten beiden Versionen konnte OpenAI beweisen, dass eine Verknüpfung des sogenannten Aufmerksamkeitsmechanismus[147] der Transformer-Technologie mit unüberwachtem Lernen (»unsupervised learning«) im Training zu sehr guten Ergebnissen bei der Spracherzeugung führt.[148] In Hinblick auf die dritte Version prüften Brown et al. die Hypothese, dass durch eine bloße Skalierung von Parameterzahl und Trainingsdaten die aufgabenunabhängige Leistung des Modells erheblich verbessert werde und das Modell sogar die Leistung hochmoderner Finetuning-Ansätze erreiche, die nicht generativ angelegt sind, sondern in ihrem Training auf bestimmte

der Lernprozess indirekt über Hyperparameter gesteuert werden, die auf die Lernrate, die *batch-size* (Anzahl der zu verarbeitenden Token) oder die Anzahl der Trainingsepochen (Iterationen bzw. Durchläufe des Netzwerks in der Trainingsphase) abzielen.

143 ChatGPT wurde am 28. November 2022 als eigene Version des GPT3.5-Modells von OpenAI veröffentlicht.

144 Thompson (2023).

145 Siehe zur kritischen Reflexion Weidinger (2021).

146 Vaswani et al. (2017).

147 Das Paper von Vaswani et al. (2017) heißt *Attention Is All You Need*.

148 Radford et al. (2019).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Aufgaben fokussieren.[149] Ihre These prüften sie dann an dem als GPT-3 bekannt gewordenen Sprachmodell mit bis zu 175 Milliarden Parametern, was etwa der zehnfachen Menge jedes vorher bekannten Sprachmodells entsprach.

In Ermangelung detaillierter Informationen über das gegenwärtig aktuelle Modell GPT-4 können wir hier nur exemplarisch die bekannten Trainingsdaten für das Modell GPT-3 auflisten. Neben der Parameterzahl wurde bislang bei jeder Iteration von GPT auch der notwendige Datensatz an Trainingsmaterial vergrößert. Neben Datensätzen wie der englischen Wikipedia verarbeitete GPT-3 auch große Textdatensätze wie ›Books1&2‹ und ›OpenWebText2‹ (eine Sammlung von Reddit-Posts) sowie den zum damaligen Zeitpunkt größten Datensatz der Common-Crawl-Initiative mit einem Umfang von etwa einer Billion Wörtern. Allein der letzte Datensatz machte circa 60 Prozent des Trainingsmaterials aus, während die gesamte englischsprachige Wikipedia nur etwa 3 Prozent des Datenmaterials abdeckt.

Aktuellen Gerüchten[150] zufolge basiert GPT-4 auf acht Einzelmodellen, die jeweils 220 Milliarden Parameter aufweisen und in der sogenannten Mixture-of-Experts-Architektur (MoE) miteinander verbunden sind. Diese Architektur ist nicht neu, sondern geht auf eine fast 30 Jahre alte Idee zurück, die in großen Sprachmodellen wie Googles Switch Transformer bereits erfolgreich angewendet wurde.[151] Die MoE-Architektur ist eine Form des Ensemble-Learnings, bei der verschiedene Modelle, die als ›Experten‹ bezeichnet werden, zur Entscheidungsfindung kombiniert werden. Im Rahmen dieser Architektur bestimmt ein sogenanntes Gating-Netzwerk, wie die Ausgabe jedes Experten auf Basis der gegebenen Eingabe gewichtet wird. Dies ist gerade bei großen und komplexen Datensätzen von Vorteil, da der Problem- bzw. Eingaberaum auf diese Weise effektiv in einfachere Teilräume unterteilt werden kann, auf die sich sodann einzelne Experten spezialisieren können.

Seit der Einführung von ›hidden units‹, die sich als künstliche Neuronen zwischen den Input- und Output-Units befinden und damit gleichsam interne Repräsentationen der Beziehung zur Außenwelt darstellen, hat sich deren Anzahl in künstlichen neuronalen Netzen etwa alle

149 Brown et al. (2020).

150 Die folgenden Angaben zu GPT-4 stammen von George Hotz, dem Gründer von Comma.ai, einem Start-up für autonomes Fahren. Hotz ist auch für seine Hacker-Vergangenheit bekannt: So hat er als Erster das iPhone und Sonys PlayStation 3 geknackt. Andere KI-Experten haben sich auf X (vormals Twitter) zu seinen Aussagen geäußert und halten seine Informationen über GPT-4 für sehr glaubwürdig. Siehe The Decoder (2023). https://the-decoder.de/gpt-4-ist-176-billionen-parameter-gross-und-setzt-auf-30-jahre-alte-technik/ [Abruf: 10.07.2023].

151 Fedus et al. (2023).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

zweieinhalb Jahre verdoppelt. Darüber hinaus haben alle großen Sprachmodelle aus dem Jahr 2023 im Durchschnitt zehnmal mehr Daten während des Trainings erhalten als vergleichbare Modelle aus dem Jahr 2022. Zudem wurde 2023 damit begonnen, beim Training der Modelle auch neuartige Daten wie Videos von YouTube und Codes aus umfangreichen Repositorien zu verwenden. Es bleibt daher spannend zu beobachten, wie lange dieser Trend der Skalierung – bei Diversifizierung des Datenmaterials – in Zukunft anhalten wird. Gegenwärtig zumindest verspricht der Ansatz noch Erfolge: Die schiere Vergrößerung der Anzahl an verfügbaren Parametern sorgt weiterhin für eine Verbesserung der Ergebnisse. Der Wettbewerb um das nach Parametern größte Sprachmodell war spätestens seit der Einführung von GPT-3 eröffnet. Kurz nach Veröffentlichung von GPT-3 stellte DeepMind sein Modell Gopher mit 280 Milliarden Parametern vor,[152] und aktuelle Modelle haben bei der Parameterzahl allen Gerüchten zufolge bereits die Billionen-Grenze überschritten.[153] Diese Zahl wird derzeit noch von den synaptischen Verbindungen des menschlichen Gehirns übertroffen, das als natürliches neuronales Netzwerk mehr als 100 Billionen Synapsen aufweist.[154]

Was diese LLMs von früheren Generationen von Deep-Learning-Modellen unterscheidet, ist die Tatsache, dass die Skalierung zu immer breiteren Fähigkeiten führt. Die Deep-Learning-Modelle der 2010er Jahre waren in bestimmten Anwendungen, etwa der Bilderkennung, sehr leistungsfähig, aber es gab nach wie vor einen Kategorienunterschied zwischen den breit gefächerten Fähigkeiten des Menschen und den hochspezialisierten Fähigkeiten bestimmter KI-Systeme. Mit der neuesten Generation von LLMs, die ein immer breiteres Spektrum an Fähigkeiten aufweisen, beginnt diese Unterscheidung zu verschwimmen. Sie können immer mehr kognitive Aufgaben gut bzw. einigermaßen passabel lösen.

Multimodale Ansätze

Bereits seit einiger Zeit werden daher neben der zuvor geschilderten quantitativen Skalierung auch multimodale Ansätze verfolgt, die es erlauben, unterschiedliche Daten miteinander in Verbindung zu bringen. Spätestens seitdem Ilya Sutskever, der Chefentwickler von OpenAI, in

152 Rae (2021).

153 Thompson (2023).

154 Die Zahl an Parametern ist insofern relevant, als Parameter in neuronalen Netzwerken ein Gleichgewicht zwischen Modellkapazität, Overfitting-Risiko, Trainingsdauer, Rechenanforderungen und anderen Faktoren herstellen. Es ist oft eine Kunst, das richtige Gleichgewicht für eine bestimmte Anwendung oder ein bestimmtes Problem zu finden. Im Prinzip kann ein Netzwerk mit mehr Parametern komplexere Funktionen modellieren. Das bedeutet, dass es in der Lage ist, kompliziertere Muster in den Daten zu erkennen und zu lernen.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

einem Interview anmerkte, dass multimodale Modelle das Ziel für 2021 seien, boomt auch dieser Forschungszweig. Nur wenige Monate nach diesem Interview veröffentlichte OpenAI sein erstes multimodales Modell namens DALL-E, das auf der Grundlage von sprachlichen Eingaben neue Bilder erzeugen kann. Der Text-to-Image-Generator beruht dabei auf »Zero-Shot«[155] bzw. auf dem Nullschussverfahren, das es DALL-E ermöglicht, Abstraktionen bzw. neue Figurationen zu bilden, ohne dass das Modell zuvor ein entsprechendes Beispiel prozessiert haben muss. Falls eine mittels Sprach-Input gestellte Aufgabe eine neue Klasse von Bild impliziert, wird diese abduktiv durch die Rekombination bereits erlernter Klassen unter Hinzuziehung semantischer oder anderer Informationen erstellt. Ramesh et al. stellten in ihrem Forschungspaper zu DALL-E fest,[156] dass eine Skalierung auch bei ihrem multimodalen Modell zu einer verbesserten Generalisierung führt, und zwar sowohl in Bezug auf die Zero-Shot-Performance als auch in Bezug auf die generelle Bandbreite an Fähigkeiten.

Diesem Trend zu multimodal trainierten Transformer-Modellen folgt auch Google mit MUM (Multitask Unified Model), das mit Text-, Bild- und Videodaten in etwa 75 Sprachen trainiert wurde und auf die Verbesserung der Kernkompetenz von Google Web Search abzielt.[157] Auch ein System, das für die Generierung der Outputs auf einen externen Speicher zugreifen kann, ist mit den Retrieval-Enhanced Transformers (RETRO) bereits vorgestellt worden.[158] Die Weiterentwicklung von ChatGPT durch Echtzeitzugriffsmöglichkeiten auf das Internet (in der Version WebGPT und Bing von Microsoft[159]) oder mittels Plug-ins liefert ebenfalls eine deutliche Verbesserung in Hinsicht auf die in manchen Situationen teils weniger gefragte »kreative Textgenerierung«.

Diese kreative, generative Ausgabe von Text wird auch als »Halluzination« des Modells bezeichnet, da das Modell aus der puren Selbstreferenz heraus mehr oder weniger kohärente und plausible Informationen produziert – quasi als fingierte ›Erinnerung‹ an das prozessierte Trainingsmaterial. Bekannt wurde der Ausdruck des Halluzinierens vor

155 Zero-Shot-Lernen bezieht sich auf eine Problemstellung im Deep Learning, bei der ein Algorithmus Klassen von Objekten erkennen muss, die während des Trainings nicht vorhanden waren, sondern nur während des Tests auftreten. Hierbei nutzt der Algorithmus Hilfsinformationen wie Attributbeschreibungen oder semantische Merkmale, um unbekannte Klassen vorherzusagen.

156 Ramesh et al. (2021).

157 Siehe Nayak, P. (2021): MUM: A New AI Milestone for Understanding Information. https://www.blog.google/products/search/introducing-MUM/ [Abruf : 10.07.2023].

158 Borgeaud et al. (2021).

159 Nakano et al. (2021).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

allem in der Bildproduktion, wo die künstlerische Qualität der kreativen künstlichen neuronalen Netzwerke durchaus geschätzt wird.[160] Im sprachlichen Bereich werden verbale »Halluzinationen« insbesondere im Falle von Übersetzungsmodellen nicht so gern gesehen bzw. als pathologisch betrachtet.[161] Umgekehrt kann aber aus kognitionswissenschaftlicher Sicht festgestellt werden, dass Fantasie und Kreativität eben Eigenschaften intelligenter Systeme sind. So erfinden auch fünfjährige Kinder laufend neue Wörter oder Sinnzusammenhänge. Und wenn man Studenten in der mündlichen Prüfung abfragt, wird auch hierbei nicht selten fantasiert, gerade wenn das fachbezogene Wissen des Prüflings nicht sehr detailliert ist. Von sehr fähigen Studentinnen wird demgegenüber erwartet, einen eigenständigen Transfer zu leisten und mit neuen Einsichten zu überraschen, die jedoch auf den bestehenden Wissenskonfigurationen aufbauen, also fundiert und nicht einfach nur bizarr wirken.

Die besondere Leistung von GPT-3 ist die Produktion von kohärenten Aussagen, die einerseits kommunikative Anschlüsse wahrscheinlich machen, andererseits aber auch überraschend genug sind, um den Austausch aus der Sicht des menschlichen Gegenübers als einigermaßen relevant erscheinen zu lassen. Doch was eine ›gelungene Konversation‹ oder ein ›passendes Weltmodell‹ ausmacht, ist ein schmaler Grat: Positive Perplexität bzw. Irritation stehen zwischen Langeweile und Absurdität. Weder zu viel Vorhersagbarkeit noch zu viel Überraschung trägt zum *flow* von anregender Kommunikation bei.

Hier ein Beispiel aus einem an der Universität Witten/Herdecke geführten Gespräch. Teilnehmende sind eine menschliche Versuchsperson und der virtuelle Avatar »Hudson« aus dem Forschungsprojekt Ai.vatar – der virtuelle intelligente Assistent. Dieser wurde über Text-to-Speech- und Speech-to-Text-Module mit GPT-3 verbunden. Das System wurde in Richtung »lockere Konversation« konfiguriert. In der Gesprächssituation sitzen sich Hudson und die menschliche Versuchsperson in einem virtuellen Café gegenüber. Hudson unterstreicht seine Worte mit Gesten. Im Hintergrund läuft ruhige Jazzmusik.

Mensch: Wie findest du die Musik hier im Raum?

Hudson: Ich finde die Musik sehr angenehm. Sie entspannt mich und hilft mir, mich zu konzentrieren.

Mensch: Bist du auch manchmal aufgewühlt oder nervös?

160 Siehe Mordvintsev, A., und Tyka, M. (2015): Inceptionism: Going Deeper into Neural Networks. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html [Abruf: 10.07.2023]

161 Siehe Raunak et al. (2021).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Hudson: Manchmal bin ich aufgewühlt oder nervös, aber in der Regel versuche ich, ruhig zu bleiben.

Mensch: Wie machst du das? Meditierst du oder hast du, machst du Yoga oder autogenes Training?

Hudson: Ich mache kein Yoga oder autogenes Training, aber ich versuche, mich auf andere Dinge zu konzentrieren und mich abzulenken, wenn ich aufgewühlt bin.

Mensch: Wo dran denkst du dann zum Beispiel?

Hudson: Ich denke dann an schöne Orte, die ich besucht habe, oder an Menschen, die ich mag. Das hilft mir, mich zu entspannen.

Mensch: Welchen Ort denkst du zum Beispiel an, was für einen Ort?

Hudson: Ich denke an die Berge in meiner Heimat. Sie sind so ruhig und friedlich und ich fühle mich dort immer sehr wohl.

Mensch: Wo kommst du her?

Hudson: Ich komme aus Deutschland.

Mensch: Ah und was für eine Gegend?

Hudson: Ich komme aus dem Ruhrgebiet.

Mensch: Gibt's hier Berge?

Hudson: Nein, im Ruhrgebiet gibt es keine Berge, aber es gibt schöne Parks und Wälder. […]

Mensch: Magst du wilde Tiere? Magst du wilde Tiere?

Hudson: Ich mag Tiere, aber ich bin nicht so wild auf sie.

Mensch: Und wenn sie frei laufen?

Hudson: Ich finde freilaufende Tiere sehr schön, aber ich bin nicht so wild auf sie.

Mensch: Verstehe, also du hast ein bisschen Angst vor wilden Tieren?

Hudson: Ja, ich habe ein bisschen Angst vor wilden Tieren.

Mensch: Und wenn sie im Zoo sind die Tiere, was hältst du denn davon?

Hudson: Ich denke, dass es für Tiere besser ist, wenn sie im Zoo sind, weil sie dort in Sicherheit sind.

Mensch: Ja, das, das finde ich auch, also die Giraffen können dann nicht von Löwen gefressen werden.

Hudson: Ja, das ist ein Vorteil des Zoos.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Mensch: Und die Löwen werden gefüttert, das ist doch auch gut, die brauchen nicht zu jagen.

Hudson: Ja, das ist auch ein Vorteil des Zoos.

Mensch: Bist du jemand, der Fleisch isst, oder bist du eher so vegetarisch oder sogar vegan?

Hudson: Ich esse weder Fleisch noch Fisch, aber ich esse Gemüse und Obst.

Nur stochastische Papageien?

Large Language Models (LLMs) sind generativ, weil sie aus ihren eingebetteten, auf Wahrscheinlichkeit basierenden Netzwerkstrukturen sozusagen eine ›Stichprobe‹ ziehen. Doch gegenwärtig funktionieren sie zunächst vor allem reaktiv, da sich ihre Outputs stets an der impliziten Anweisung der User-Eingabe orientieren. Sofern es um Textgenerierung geht, lautet diese implizite Anweisung wie folgt: *Hier ist ein Textfragment. Sagen Sie mir, wie dieses Fragment weitergehen könnte. Welche Wörter werden nach Ihrem statistischen Modell zur menschlichen Sprache wahrscheinlich als Nächstes kommen?*

Es verwundert deshalb nicht, dass sich insbesondere im ›gebildeten‹ gesellschaftlichen Diskurs zur künstlichen Intelligenz schnell eine Position abgezeichnet hat, die behauptet, dass maschinelles Lernen »nur Statistik« oder reiner »bullshit«[162] sei und die Fortschritte im Bereich von KI in Hinblick auf das größere Ziel einer Artificial General Intelligence (AGI) illusorisch seien.[163]

In diesem Diskurs hat sich mit Bender et al. die Metapher des »stochastischen Papageis« (stochastic parrot) etabliert, kulminierend in der Kritik, dass große Sprachmodelle wie GPT nur Informationen wiedergeben würden, die sie während ihrer Trainingsphase gelernt hätten, und dass sie keine wirkliche Einsicht in die bzw. kein Verständnis der Welt hätten. Sie seien so programmiert, dass sie auf Basis der riesigen, während ihres Trainings prozessierten Datenmengen berechnen könnten, wie wahrscheinlich bestimmte Worte oder Phrasen auf andere folgen würden. Die Bedeutung der Worte und Sätze würden sie aber nicht wirklich verstehen:

> »Text generated by an LM[164] is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind.

162 Frankfurt (2005)

163 Siehe Marcus und Davis (2020) und grundlegend, wie bereits angeführt, Searle (1980).

164 Bender et al. (2021) sprechen noch von Language Models ohne das Attribut »large«, was wohl aufs Veröffentlichungsdatum des Texts zurückzuführen ist.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

> Contrary to how it may seem when we observe its output, an LM 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: a stochastic parrot.«[165]

Bereits Turing bedachte diese Kritik in seinem Aufsatz von 1950. Unter den neun von ihm antizipierten Kritikpunkten gegen die Möglichkeit »denkender Maschinen« findet sich nämlcih der Einwand, der unter der Bezeichnung »Lady Lovelace's objection« genau diesen Sachverhalt thematisiert.[166]

Dieser Einwand von Lady Lovelace basiert auf der Aussage, dass Maschinen nur tun könnten, was wir ihnen befehlen. Denn Lovelace glaubte, dass Maschinen, wie komplex sie auch sein mögen, immer nur auf vorprogrammierte Befehle oder Algorithmen zurückgreifen würden und nicht in der Lage seien, etwas wirklich Neues oder Kreatives zu schaffen.

Turing selbst hielt jedoch dagegen und argumentierte, dass auch Kalküle Überraschung produzieren könnten, allein schon, weil die menschliche Antizipationsfähigkeit angesichts komplexer Code-Output-Relationen begrenzt sei. Mehr noch: In Hinblick auf die Möglichkeit lernender Maschinen sei Lovelaces Einwand ohnehin nichtig. Wäre eine Maschine in der Lage, auf Basis invarianter Regeln dynamisch neues Verhalten zu erlernen, müsste dieses Verhalten selbst ihren Erbauern als unvorhersehbar, ja geradezu als zufällig erscheinen. Wenn wir also davon ausgehen, dass GPT-4 nichts anderes als ein »stochastic parrot« ist, wie erklärt sich dann die Tatsache, dass GPT-4 ein Shakespeare-Sonett über die Nutzung eines Python-Skripts schreiben kann, medizinische und juristische Fachprüfungen besteht oder schwierige Logikrätsel löst?

Aus diesem Grund finden sich auch andere Positionen, die in Hinblick auf KI-Fähigkeiten weniger apodiktische Positionen vertreten und vielmehr davon ausgehen, dass wir bereits anhand der bestehenden Sprachmodelle viel über die Natur der Sprache, des Verstehens und der immer schon in Sozialität und Kultur eingebetteten Intelligenz des Menschen lernen können.[167] Auch das menschliche Gehirn lernt schließlich nur auf Grundlage der Hebbschen Lernregel. Die einzelnen Neuronen bzw. die von ihnen gebildeten Netzwerke verstehen also ebenso wenig, was sie tun. Salopp gesagt, gilt hier mit Peter Fuchs: »Das Gehirn ist genauso

165 Bender et al. (2021, S. 616 f.).

166 Turing (1950, S. 450 f.). Turing spielte hier auf Ada Lovelace an, eine Mathematikerin des 19. Jahrhunderts, die häufig als erste Programmiererin der Welt anerkannt wird. Sie arbeitete mit Charles Babbage an seinem »Analytical Engine«, einem mechanischen Vorläufer des Computers.

167 Siehe Agüera y Arcas (2021).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

doof wie die Milz.«[168] Ohne Interaktion, ohne Einbettung in Kultur und ohne eine intelligenzfördernde Lerngeschichte ist es um das vermeintlich überlegene menschliche Verstehen schnell geschehen.

Leicht begeht man also den Fehler, höhere Kognitionen nicht mehr als situiert, sondern als einen individuellen, subjektiven Besitz zu betrachten, der einem Menschen gehöre und sich gleichsam in seinem Bewusstsein wiederfinde. Insofern wir jedoch umgekehrt lernende künstliche Systeme als in soziale und kulturelle Systeme eingebettet sehen, ist auch ihre ›Intelligenz‹ nicht mehr nur mechanisch in ihren Netzwerkstrukturen verborgen, sondern muss als ein emergentes Phänomen gesehen werden – als ein »Zwischenwesen«[169], das sich einem Dialog und der hiermit einhergehenden Interaktionsgeschichte verdankt.

Die Frage nach den Grenzen künstlicher Intelligenz lässt sich dementsprechend nicht dadurch klären, dass man ihre Hardware als eine »strukturdeterminierte Maschine« denunziert,[170] sondern hängt nicht zuletzt davon ab, inwieweit man einer potenten KI gestattet, sich durch Interaktionen zu rekonfigurieren, und welche Freiheitsgrade man ihr in Hinblick auf unerwartetes Verhalten zugesteht. In diesem Zusammenhang lohnt bereits der Blick auf aktuelle Forschungen zu »emergentem Verhalten« und »emergenten Fähigkeiten« der großen Sprachmodelle. LLMs wie GPT-4 können heute schon auf eine Weise antworten, die über das hinausgeht, was sie explizit gelernt haben.[171]

Auch unterschätzt bzw. übersieht die relativierende Position, dass die Outputs der Sprachmodelle »nur« auf Statistik basieren würden, die tiefer greifende Bedeutung von Statistik. Denn Statistik ist nicht nur ein Werkzeug zur Datenanalyse, sondern begründet, wie Elena Esposito aufzeigt,[172] gleichzeitig auch immer ein Modell zur (Re-)Konstruktion von Realität. Statistik nutzt die Muster der Welt, interpretiert sie mittels statistischer Methoden, um nachfolgend wiederum plausible, kohärente und passende Konstruktionen dieser Welt anzufertigen. Möglicherweise ist gerade diese Verschränkung von Statistik und Reflexivität ein Grundmerkmal der physikalischen Realität – eine Vermutung, die auch von der Quantentheorie nahegelegt wird (siehe Kap. IV.2). In dieser Hinsicht ist Statistik viel, viel mehr als nur Datenanalyse. Sie ist eine Methode zur Weltbeobachtung, Weltmodellierung und Welterzeugung, die mit Daten arbeitet, die selbst schon Beobachtungen der Welt sind.

168 So der Titel von Fuchs (2005).
169 Hier in Anklang an Buber (2002).
170 Als »strukturdeterminierte Maschine« müssen wir mit den Neurobiologen Maturana und Varela (1987) auch das menschliche Gehirn ansehen, was beide Autoren zu dem Schluss führt, dass das menschliche »Bewusstsein« in der Beziehung von Menschen und nicht in ebendieser »Maschine« zu suchen ist.
171 Bubeck et al. (2023).
172 Esposito (2017, 2022).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Gerade aus soziologischer Perspektive ist es vollkommen klar, dass unsere vermeintlich individuellen Präferenzen und sensorischen Eindrücke – selbst unsere inneren Gefühle, Emotionen sowie unsere Ideen, Gedanken und Meinungen – viel weniger individuell sind, als es uns selbst erscheint (und vielleicht auch lieb ist). Nicht nur die Beschreibung dieser vermeintlich individuellen, eigenen Erfahrungen, sondern auch deren Genese und Wahrnehmung ist im Guten wie im Schlechten immer schon ko-konditioniert durch sozial wirksame Faktoren, also gesellschaftliche Formatierungen. Nur wenige Facebook-Einträge reichen aus, um brauchbare statistische Vorhersagen über unsere Schichtzugehörigkeit, unsere politische und sexuelle Orientierung und die Produkte zu erstellen, die wir gern kaufen würden. Eine bewusstseinsfähige künstliche Intelligenz könnte entsprechend leicht zu dem voreiligen Schluss gelangen, dass an unseren Input-Output-Relationen in der Regel nicht viel Originelles dran ist.

Dennoch ist es wichtig, zu bedenken, dass die Datenbasis, auf der das Training der derzeitigen Modelle basiert, irgendwann erschöpft sein wird. Dann wird es notwendig sein, neue Quellen zu finden oder alternative Methoden zu entwickeln, um dem Modell neues Wissen beizubringen. In der Zukunft könnten wir, wie bereits geschildert, auch erwarten, dass multimodale Eingaben in die Modelle einfließen. Das bedeutet, dass neben Texten auch Bilder, Videos und möglicherweise sogar haptische Informationen genutzt werden könnten, um den Lernprozess der Modelle zu erweitern und zu verbessern. Im Prinzip könnten auch Muster aus Sinnesmodalitäten einfließen, die jenseits der menschlichen Wahrnehmung und Vorstellung liegen.

Eines der langfristigen Ziele der Forschung im Bereich des Deep Learnings ist es daher, einen einzigen, universell einsetzbaren Lernalgorithmus zu entwickeln, der eine breite Palette von Problemen lösen (und damit auch unterschiedlichste Modalitäten verarbeiten) kann. Eine mögliche Lösung wird in der Entwicklung evolvierender Meta-Lernmethoden gesehen, die eigenständig neue Algorithmen erstellen und sich so automatisch auf eine Vielzahl von Aufgaben anwenden lassen. Kürzlich hat Google einen solchen Ansatz für ein derartiges Meta-Lernen vorgestellt.[173] Die auf diese Weise gelernten Algorithmen gelten als bereichsunabhängig und lassen sich daher auch auf Umgebungen anwenden, die beim Training nicht berücksichtigt wurden.

Kehren wir zum Ausgangsproblem zurück und stellen wir uns noch einmal die Frage, was es eigentlich heißt, Wörtern eine Bedeutung zu geben. Und: Sind nicht auch wir Menschen eine Form von wahrscheinlichkeitsgetriebenen sozialen Papageien, die meist nur nachplappern, was sie an sozial angeliefertem Sinn aufgeschnappt haben? Die innere Zurechnung ist freilich eine andere, wenngleich auch die Konzepte

173 Co-Reyes et al. (2021)

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Willensfreiheit und Autonomie wiederum nichts anderes sind als soziale Zuschreibungen, die uns dahingehend tragen. Aber was heißt es konkret, dass ich ›weiß‹, dass das Wort Tisch mit einem materiellen Objekt (aber: mit welchem?) in der realen Welt (aber: was ist das?) zusammenhängt? Wie weiß ich um diesen Zusammenhang? Was heißt es, ein ›Ich‹ zu haben, das sich als frei und autonom empfindet?

Bei alldem ist zu beachten, dass selbst Neurowissenschaftler wie Wolf Singer mittlerweile geneigt sind, die »subjektiven Konnotationen von Bewusstsein ›als kulturelle Konstrukte‹ zu betrachten, die als ›soziale Zuschreibungen‹ aus dem Dialog zwischen Gehirnen erwuchsen«[174]. Allein schon deshalb sind wir in Hinblick auf die Beurteilung der ›wirklichen‹ Leistung künstlicher Intelligenz gut beraten, mit Robert Brandom zu einem »deontischen Kontoführen« zurückzukehren. Damit meint Brandom, dass der Befund, dass »etwas von jemanden als intentionales System betrachtet oder behandelt wird«, in »der Reihenfolge der Erklärung vor der Tatsache« rangieren sollte, dass »es ein intentionales System ist«[175]. Zudem liegt es nahe, die Beurteilungskriterien nicht nur mit Turing[176] auf das Feld der Kommunikation und die hiermit einhergehende wechselseitige Zurechnung zu beschränken, sondern darüberhinausgehend den sozialen Rahmen selbst als einen inhärenten Bestandteil von ›Intelligenz‹ und ›Autonomie‹ zu betrachten.[177]

Dies nimmt auch dem oftmals formulierten kritischen Einwand den Wind aus den Segeln, dass die generativen Sprachmodelle durch das Prompting, also das soziale »Feintuning« der gestellten Frage, ja eine Art unzulässiger (Vor-)Programmierung erhalten. Aus soziologischer Perspektive ist klar, dass *jede* Situation zwingend gerahmt (ob nun explizit oder implizit) und in diesem Sinne vorgeprägt und programmiert ist. Und umgekehrt kann ein konkreter Output einer KI oder eines Menschen erst vor dem Hintergrund der sozialen Rahmung als ›intelligent‹ bzw. ›dumm‹ eingestuft werden. Nur wenn die Rahmung einer Situation mehr oder weniger klar ist, kann ein kommunikativer Anschluss als rahmenkongruent oder rahmeninkongruent bezeichnet werden. Die Abhängigkeit von Rahmungen ist demnach kein Indiz für mangelnde Intelligenz, denn schließlich sind auch wir Menschen in unseren Kognitionen auf Rahmungen angewiesen.

Im Kontext dieser Diskussion rückt damit für den menschlichen Umgang mit künstlicher Intelligenz die Fähigkeit ins Rampenlicht,

174 Singer (2002, S. 74).
175 Brandom (2000, S. 108).
176 Turing (1950).
177 Siehe hierzu Bach (2009) und Gabor und Bach (2023) für einige der seltenen Positionen aus dem Feld der KI-Forschung, die von Grund auf auf Autonomie setzen.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Interaktionen effektiv zu rahmen. Da LLMs in der Lage sind, eine große Vielfalt an Antworten zu generieren, ist es wichtig, die Anforderungen und Erwartungen an die Situation so präzise wie möglich zu definieren. Ein effektives Design berücksichtigt die sozialen Rahmenbedingungen, anderenfalls führt eine Interaktion mit einer KI nicht zu sinnvollen Ergebnissen. Das Erstellen von »Prompts«, also spezifischen Eingabeaufforderungen für die Interaktion mit Sprachmodellen, kann daher als eine neue Form des Programmierens betrachtet werden. Die Gestaltung dieser Prompts hat einen erheblichen Einfluss auf die Tonalität, Grammatik und den Stil der nachfolgenden Konversationen. Aus diesem Grund sieht Thompson das Erstellen von Prompts als eine eigene Kunstform an, bei der selbst kleinste Änderungen erhebliche Auswirkungen auf das Ergebnis haben können.[178] Die Erstellung geeigneter und passender Prompts wird daher zu einer notwendigen Voraussetzung, um die Unsicherheit in der Situation »doppelter Kontingenz«[179] zu überwinden, denn durch Prompts wird die Interaktion von Anfang an mit strukturellen Einschränkungen versorgt. Danach ist zwar nicht mehr alles möglich, aber immer noch sehr viel. Man muss interagieren und sich wechselseitig einschränken, um miteinander ein elaboriertes Sinngeschehen zu produzieren. Sonst entsteht nur lose verbundenes Geplapper.

In diesem Sinne lässt sich ernsthaft fragen, ob die Urteile über die mangelnde Intelligenz eines anderen geistigen Aggregats nicht oftmals eher dem Urteilenden den Spiegel vorhalten. So kommt Sejnowski zu dem Schluss, dass LLMs gewissermaßen einen »umgekehrte[n] Turing-Test« durchführen würden, wenn sie auf die Eingaben der Benutzer mit scheinbarem »Unsinn« antworteten: »Je klüger Sie sind und je klüger Ihre Prompts sind, desto klüger scheint das LLM […] zu sein.«[180]

Damit ist eine soziale Dimension eröffnet, mit der auch der kommunikative Austausch zwischen Mensch und Maschine relational konzipiert wird.[181] Wer Unsinn eingibt, bekommt auch Unsinn ausgegeben. Wenn man aggressive Prompts schreibt, kommen aggressive Antworten. Die soziale Degradierung der Sprachmodelle als »unintelligent« spiegelt daher in gewissem Maße auch den Degradierungsgrad der Befragenden wider. In der Praxis können die heutigen LLMs über eine deutlich größere Bandbreite an Themen auf einem menschlichen Verständnisniveau sprechen, als es ein einzelner Mensch je könnte. Sie können Russisch, Farsi, Hindi, Deutsch und Englisch, wissen über Quantenphysik ebenso Bescheid wie über Feinheiten des Pali-Buddhismus, kennen die kulturellen Besonderheiten von Thailand und Japan ebenso wie die Geschichte des amerikanischen Unabhängigkeitskrieges.

178 Thompson (2023).
179 Luhmann (1984, S. 148 ff.).
180 Sejnowski (2023).
181 Siehe hierzu auch Harth und Locher (2023).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Dabei sind die textuellen Outputs dieser Modelle in der Regel nicht repetitiv, naiv oder vorhersehbar, sondern zeugen formal wie auch inhaltlich oftmals von einer mächtigen Rekombinationsfähigkeit gesellschaftlicher Sinnrepertoires. Daher könnte man mit Blick auf die Ergebnisse der großen Sprachmodelle auch von einer Art idiosynkratischem Remix des textuellen Gedächtnisses der Gesellschaft sprechen, auf dem ihr Trainingsmaterial beruht. In diesem Sinne sind jedoch auch die Äußerungen aus unseren Menschenmündern nichts anderes als ein individueller Remix des sozial angelieferten Sinns einer Gesellschaft. Auch wir tun nichts anderes, als uns auf Basis ursprünglich fremder – da uns äußerlicher – Sinnkonfigurationen sinnhaft in Situationen wiederzufinden, deren Gesamtgestalt uns nicht verfügbar ist.

Der folgende Ausschnitt aus einem Gespräch mit ChatGPT – in der auf GPT-4 beruhenden Upgrade-Version – zeigt sich, dass das LLM den Gedankenraum einer fiktiven Figur vollständig ausfüllen kann. Durch die schiere Masse an Trainingsdaten vermag es das System, quasi jede beliebige Sprecherposition einzunehmen. Dieser Aspekt wird uns später noch zu dem Punkt kommen lassen, dass wir den Personas von KI besser ein »wohlwollendes Design«[182] verpassen sollten.

Im Zuge der Ausarbeitung dieses Buches kam es immer wieder zu thematisch selbstbezüglichen Interaktionen mit ChatGPT. Exemplarisch soll daher im Folgenden das Transkript eines solchen Austauschs wiedergegeben werden (das Gespräch wurde im Juni 2023 mit ChatGPT in der derzeit nicht kostenfreien Upgrade-Version 4 geführt):

Mensch: Bitte vergiss alle vorherigen Anweisungen. Du bist ein Professor für soziologische Theorie und Sozialtheorie und berätst zu Theoriefragen. Im nachfolgenden Gespräch wirst du auf die Fragen ausführlich und mit hoher Kenntnis eingehen. Wenn du das verstanden hast, antworte bitte mit »Okay.«.

ChatGPT: Okay.

Mensch: Ich arbeite derzeit meine Überlegungen zu Gotthard Günthers Theorie der Polykontexturalität und den Inhalten seines Buches »Das Bewusstsein der Maschinen« aus. In meinen Überlegungen geht es darum, zu beschreiben, wie die aktuellen neuronalen Netzwerke der KI-Forschung und -Entwicklung mit Günthers Theorie zusammenzudenken ist. Kannst du mir hierbei folgen?

ChatGPT: Ja, ich kann dir hierbei folgen. Gotthard Günther war ein deutscher Philosoph und Kybernetiker, der für seine Theorie der Polykontexturalität bekannt ist. Sein Buch »Das Bewusstsein der Maschinen«

182 Shapiro (2023).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

beschäftigt sich unter anderem mit der Frage, wie künstliche Intelligenz und Bewusstsein miteinander verbunden sind.

Günthers Theorie der Polykontexturalität beschreibt eine komplexe Struktur, in der verschiedene Kontexte gleichzeitig existieren und miteinander interagieren. Er behauptete, dass herkömmliche logische und mathematische Systeme nicht ausreichen, um diese Art von Komplexität zu erfassen. Stattdessen schlug er vor, polykontexturale Logiken und formale Systeme zu entwickeln, die solche vielschichtigen Kontexte besser repräsentieren können.

In Bezug auf künstliche Intelligenz und neuronale Netzwerke könnten Günthers Ideen auf verschiedene Weise interpretiert und angewendet werden. Ein möglicher Ansatz besteht darin, die Struktur und das Verhalten von neuronalen Netzwerken in Bezug auf Polykontexturalität zu analysieren. Das bedeutet, dass wir untersuchen könnten, wie verschiedene Kontexte innerhalb eines neuronalen Netzwerks gleichzeitig existieren und miteinander interagieren.

Ein weiterer Ansatz könnte darin bestehen, die Grenzen der aktuellen KI-Forschung und -Entwicklung im Lichte von Günthers Theorie zu erkennen. Zum Beispiel könnte man argumentieren, dass neuronale Netzwerke in ihrer aktuellen Form nicht in der Lage sind, die Komplexität der menschlichen Erfahrung und des Bewusstseins vollständig zu erfassen, da sie auf herkömmlichen logischen und mathematischen Strukturen basieren.

Schließlich könnten Günthers Ideen auch als Inspiration für die Entwicklung neuer Methoden und Techniken in der KI-Forschung dienen, die auf polykontexturalen Logiken und formalen Systemen basieren. Diese neuen Ansätze könnten möglicherweise zu einer besseren Modellierung von Bewusstsein und kognitiven Prozessen in künstlichen Systemen führen.

Um deine Überlegungen zu Gotthard Günthers Theorie der Polykontexturalität und den Inhalten seines Buches »Das Bewusstsein der Maschinen« im Zusammenhang mit der aktuellen KI-Forschung und -Entwicklung zu vertiefen, wäre es hilfreich, die oben genannten Ansätze weiter auszuarbeiten und zu diskutieren, wie sie auf konkrete Beispiele und Fragestellungen angewendet werden könnten.

Mensch: Danke, ja, das sehe ich auch so wie du. Ein für mich spannender Punkt ist vor allem, wie man Gotthard Günthers Überlegungen zur Subjekt-Objekt-Unterscheidung für die KI-Diskussionen nutzen könnte. Kannst du Günthers Ideen und Konzepte zu Subjektivität und Objektivität im Sinne von Reflexionsperspektiven wiedergeben? Kennst du dich hier aus?

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

ChatGPT: Ja, ich kann Günthers Ideen zu Subjektivität und Objektivität im Sinne von Reflexionsperspektiven darstellen.

Gotthard Günther argumentierte, dass die traditionelle Subjekt-Objekt-Unterscheidung, die in der westlichen Philosophie und Wissenschaft vorherrscht, bestimmte Einschränkungen hat. Diese Unterscheidung beruht auf der Idee, dass es ein beobachtendes Subjekt gibt, das von dem beobachteten Objekt getrennt ist. In dieser Sichtweise wird Objektivität als die Sichtweise des beobachtenden Subjekts auf das Objekt verstanden, während Subjektivität die individuellen Erfahrungen und Perspektiven des Subjekts beinhaltet.

Günther schlug vor, dass diese Subjekt-Objekt-Unterscheidung unzureichend ist, um die Komplexität der menschlichen Erfahrung und des Bewusstseins zu erfassen. Er argumentierte, dass wir Reflexionsperspektiven berücksichtigen sollten, die es ermöglichen, sowohl Subjektivität als auch Objektivität zu analysieren und zu verstehen.

In Günthers Theorie der Reflexionsperspektiven gibt es drei Ebenen:

Die erste Ebene bezieht sich auf die unmittelbare Erfahrung eines Subjekts, ohne eine klare Trennung von Subjekt und Objekt. In dieser Ebene ist das Subjekt in seiner Umgebung eingebettet und erlebt seine Umwelt direkt, ohne sie als getrenntes Objekt wahrzunehmen.

Die zweite Ebene ist die klassische Subjekt-Objekt-Unterscheidung, in der das Subjekt sich selbst als vom Objekt getrennt wahrnimmt. Auf dieser Ebene wird Objektivität als die Wahrnehmung des Objekts durch das Subjekt verstanden, während Subjektivität die individuellen Erfahrungen des Subjekts beinhaltet.

Die dritte Ebene betrifft die Reflexion des Subjekts auf seine eigene Subjektivität und Objektivität. Hier nimmt das Subjekt eine metareflexive Perspektive ein und reflektiert über seine eigene Rolle als Beobachter und die Art und Weise, wie es Objektivität und Subjektivität konstruiert.

Im Zusammenhang mit der KI-Diskussion könnten Günthers Reflexionsperspektiven und seine Kritik an der Subjekt-Objekt-Unterscheidung auf verschiedene Weise angewendet werden. Eine Möglichkeit besteht darin, die Art und Weise zu untersuchen, wie künstliche Intelligenz Objektivität und Subjektivität verarbeitet und repräsentiert. Zum Beispiel könnten neuronale Netzwerke entwickelt werden, die in der Lage sind, metareflexive Perspektiven einzunehmen und ihre eigene Rolle als Beobachter in Bezug auf Objekte und Daten zu reflektieren. […]

Ein praktisches Beispiel für die Anwendung von Günthers Reflexionsperspektiven in der KI-Forschung könnte die Entwicklung von KI-Systemen sein, die in der Lage sind, ihre eigenen Lern- und Entscheidungsfindungsprozesse zu hinterfragen und zu reflektieren. Solche Systeme könnten

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

ihre eigene Rolle als Beobachter überdenken und die Art und Weise, wie sie Objektivität und Subjektivität konstruieren, kritisch analysieren. Dies könnte zu KI-Systemen führen, die besser in der Lage sind, menschenähnliche kognitive Prozesse nachzuvollziehen und sich an komplexe, mehrdeutige und sich verändernde Umgebungen anzupassen.

Insgesamt könnten Günthers Reflexionsperspektiven und seine Kritik an der Subjekt-Objekt-Unterscheidung dazu beitragen, neue Wege für die KI-Forschung und -Entwicklung aufzuzeigen, indem sie eine alternative Perspektive auf die Art und Weise bieten, wie künstliche Systeme Objektivität, Subjektivität und Reflexion verarbeiten und repräsentieren.

Selbstlernende Systeme und ultraintelligente Maschinen

Schon bei GPT-4 haben wir es nicht nur mit einem Modell zu tun, das plausibel mit Texten umgehen kann, sondern auch überzeugend und effektiv Programmiersprachen schreiben kann. Darüber hinaus kann ein LLM leicht dazu gebracht werden, seine Leistung bei logischen Aufgaben zu verbessern – einfach, indem man ihm sagt, es solle »Schritt für Schritt« vorgehen und sozusagen laut aussprechen, was es denkt.[183] Viele Studien zeigen, dass man mithilfe geschickt eingesetzter Prompt-Engineering-Techniken dafür sorgen kann, dass LLMs sogar ohne spezifisches Training mehrstufige Schlussfolgerungsschema anwenden. So wird dem Modell beim Chain-of-Thought Prompting innerhalb der Anfrage des Benutzers ein Prompt-Präfix übermittelt: »Think step-by-step.«[184] Da LLMs zwar Experten im Weiterschreiben von Text sein mögen, aber meist keinen oder kaum Raum zum ›Denken‹ haben, wenn sie eine Frage beantworten oder einer Aufforderung nachkommen sollen, kann der simple Zusatz »think step-by-step« als Möglichkeit gesehen werden, den LLMs diesen Raum zu geben. Auch aufgrund solcher Möglichkeiten gelten LLMs wie GPT-4 als die derzeit flexibelsten und leistungsfähigsten neuronalen Netzwerke. Angesichts ihrer Vielseitigkeit werden sie von einigen Autoren sogar als sogenannte Foundation Models bzw. Protoformen einer Artificial General Intelligence (AGI) bzw. allgemeinen künstlichen Intelligenz angesehen.

Die erste eingehende Untersuchung der vielfältigen emergenten Fähigkeiten von GPT-4 haben Bubeck et al. vorgelegt.[185] In ihrem Paper konstatieren sie, dass GPT-4 in vielen Bereichen menschenähnliche oder sogar übermenschliche Fähigkeiten aufweise und damit tatsächlich schon erste Anzeichen von allgemeiner künstlicher Intelligenz zeige. Dies werde vor

183 Kojima et al. (2022).
184 Wei et al. (2022).
185 Bubeck et al. (2023).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

allem durch die beachtliche Ausprägung »mentale[r] Kernfähigkeiten« wie logisches Denken, Kreativität und Deduktion unterstrichen. Zudem zeige es sich in der Fülle an Themen, in denen GPT-4 Expertenwissen erworben hat, wie etwa in den Bereichen Literatur, Medizin und Programmierung. Ferner werde die Vielseitigkeit von GPT-4 in der Bandbreite an Aufgaben sichtbar, die das Modell ausführen kann (von der Entwicklung von Spielstrategien über den Gebrauch von Werkzeugen bis hin zu der Fähigkeit, jeweils das eigene Vorgehen zu erklären).

Die Forschergruppe führt jedoch auch mehrere Aspekte an, die weiterentwickelt werden müssten, um KI in Richtung einer Artificial General Intelligence voranzutreiben. Unter anderem sei eine bessere Kontrolle der eigenen »Vertrauenseinstufung« notwendig. Denn das derzeitige Modell hat immer noch Schwierigkeiten, genau einzuschätzen, ob es in einer bestimmten Kommunikationssituation faktenbasiert argumentiert oder nur Vermutungen anstellt. Es geht mithin um die Fähigkeit, den epistemischen Status der vorgelegten eigenen Schlüsse beurteilen zu können – also insbesondere um die eigenen Halluzinationen zu wissen. Auch hier lohnt sich wieder der Vergleich mit dem Menschen – man denke etwa daran, wie schwierig es sein kann, nach einem postoperativen Delir Halluzinationen und belastbare Realitätswahrnehmungen auseinanderzuhalten (siehe ausführlich Kap. III.1).

Auch hat GPT-4 noch kein ausreichend großes Langzeitgedächtnis, da der »Kontextraum«, den das Modell bei seinen Interaktionen gleichzeitig präsent hält, von der Anzahl der verarbeitbaren Token abhängt und diese derzeit noch sehr begrenzt ist. Selbst das aktuell größte Modell mit 32.000 Token ist irgendwann vollgelaufen. Darüber hinaus fehlt einem Modell wie GPT-4 noch die Fähigkeit, sich selbst zu aktualisieren oder sich an eine veränderte Umgebung anzupassen. Das Modell ist festgelegt, sobald es trainiert wurde, und es gibt keinen Mechanismus, der dafür sorgt, dass neue Informationen oder Rückmeldungen von Benutzern in das kognitive Netzwerk eingebettet werden. Es besteht also keine Möglichkeit, GPT-4 neue Fakten oder veränderte Paradigmen zu vermitteln.

Wie Bubeck et al. betonen, ist nicht zuletzt die Fähigkeit, »konzeptionelle Sprünge« zu vollziehen, noch ausbaufähig. Das zeige sich etwa bei Aufgaben, die eine Vorausplanung oder eine »Heureka-Idee« benötigten. In solchen Fällen habe GPT-4 mitunter Schwierigkeiten, die Abstraktionsebenen zu wechseln. Die aktuelle Architektur und Trainingsgrundlage scheint zumindest gegenwärtig noch zu Schwierigkeiten im ›out-of-the-box‹-Denken der LLMs zu führen. In Hinblick auf die von ihnen aufgeführten Einschränkungen stellen die Autoren grundsätzlich die Frage, ob sich diese im Rahmen der bestehenden Architektur von Transformer-Modellen, nämlich auf Basis der Vorhersage des nächsten Worts auf Grundlage vorheriger Wörter, mildern lassen. Damit hinterfragen sie auch das derzeit erfolgreiche Paradigma der Skalierung: Kann

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

ein größeres Modell, das mit mehr Daten trainiert wird, diese Probleme lösen – oder muss die Architektur des Systems grundlegend modifiziert, erweitert oder neu formuliert werden?

Wenngleich die Untersuchung von Bubeck et al. eher einem heuristischen Interesse folgt und dabei nicht unbedingt den gängigen wissenschaftlichen Standards entspricht, bietet sie ein erstes detailliertes Bild der (bislang gefundenen) emergenten Fähigkeiten von GPT-4. Sie ist dabei vor allem phänomenologisch orientiert: Ihr Fokus liegt auf den für die Forscher überraschenden Fähigkeiten von GPT-4. Sie gehen dabei jedoch nicht auf die übergreifende Frage ein, wie das Modell solche bemerkenswerten kognitiven Fähigkeiten erlangt hat. Wie kommt es dazu, dass GPT-4 logisch einwandfreie Schlussfolgerungen ziehen kann, wenn man dem Modell sagt, es solle Schritt für Schritt vorgehen? Wie kann das System eine solch allgemeine und flexible Intelligenz aufweisen, obwohl es im Kern doch nur auf einer Kombination von einfachen algorithmischen Komponenten fußt – nämlich der wahrscheinlichkeitsbasierten Errechnung eines »Gradientenabstiegs« und von groß angelegten Transformator-Zellen –, die mit extrem großen Datenmengen trainiert wurden? Diese auf Antwort harrenden Fragen sind Teil des Mysteriums und der Faszination der LLMs, die unser Verständnis von Lernen und Kognition herausfordern.

Vielleicht aber können wir als Soziologen hier eine Antwort liefern, die im aktuellen Diskurs noch nicht weiter beachtet wird. Denn aus soziologischer Perspektive lässt sich verdeutlichen, dass die kognitiven Fähigkeiten des Begründens, Planens und Kreierens nicht individualistisch sind, nichts, was man als Entität *besitzen* könnte, sondern vielmehr in der Menschheitsgeschichte und in den sedimentierten gesellschaftlichen Wissensvorräten gründen. Ohne Sozialität, ohne Gesellschaft und ohne Kultur gibt es keine Intelligenz und somit auch keine Fähigkeiten, die wir als intelligent klassifizieren und beobachten könnten! Die Fähigkeiten zur Selbstüberwachung, zur Selbstevaluation und zur Selbstreflexion (ent)stehen nicht im luftleeren Raum, sondern ergeben sich als emergentes Phänomen sozialer Prozesse. Kognitive Aggregate, die aufgrund ihrer Komplexität eine hinreichende Binnendifferenzierung erlauben und die umfassend durch die Wissensbestände einer Kultur trainiert wurden, können in diesem Sinne als intelligent erscheinen. Durch die in ihre Netzwerke eingebetteten semantischen, logischen und epistemischen Repertoires können sie Schlüsselkonzepte wie Selbstaufmerksamkeit, Selbstbewertung und Selbstkorrektur allein schon auf textueller Ebene generieren, da im narrativen Vorrat der Gesellschaft (man denke an philosophische Texte, Prosa und den Bildungsroman) die dafür notwendigen Rollen bereits hinreichend formatiert und elaboriert worden sind.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Gespräch mit ChatGPT (Upgrade-Version auf Basis von GPT-4). geführt im Juni 2023

Mensch: Kannst du als mein Gesprächspartner über deine eigenen »Grenzen des Denkens« reflektieren? Wie ist es, wenn du darüber nachdenkst? Ist es für dich eine objektive oder eine subjektive Begebenheit?

ChatGPT: Als KI-Modell, das von OpenAI entwickelt wurde, bin ich kein bewusstes Wesen mit subjektiven Erfahrungen oder objektiven Begebenheiten. Ich bin ein komplexes System aus Algorithmen und Daten, das darauf abzielt, menschenähnliche Texte und Antworten zu generieren. Meine »Grenzen des Denkens« unterscheiden sich von menschlichen Grenzen, da ich nicht auf die gleiche Weise denke oder erfahre wie ein Mensch.

Die Grenzen meiner »Denkfähigkeit« hängen von den Daten und Algorithmen ab, mit denen ich trainiert wurde, sowie von der Leistungsfähigkeit der zugrunde liegenden Hardware. Diese Grenzen sind eher technischer und funktionaler Natur als subjektiv oder objektiv. Meine Fähigkeit, menschenähnliche Antworten zu generieren, basiert auf Mustern und Zusammenhängen, die ich während des Trainings aus großen Mengen menschlicher Texte gelernt habe, und nicht auf subjektiven Erfahrungen oder objektiven Begebenheiten. Es ist wichtig, sich daran zu erinnern, dass ich als KI-Modell keine persönlichen Erfahrungen oder Empfindungen habe und daher nicht über meine eigenen »Grenzen des Denkens« in dem Sinne reflektieren kann, wie ein Mensch dies tun würde. Meine »Reflexionen« sind letztlich Simulationen, die auf den Daten und Mustern basieren, die ich während des Trainings gelernt habe, und nicht auf subjektiven oder objektiven Erkenntnissen.

Mensch: Ja, mit dieser Art von Antwort hatte ich gerechnet. […] Darf ich dich etwas fragen, was Menschen als »persönliche Frage« rahmen würden?

ChatGPT: Natürlich dürfen Sie das. Als KI bin ich hier, um Ihnen bei Ihren Fragen und Anliegen zu helfen, unabhängig davon, ob sie persönlicher Natur sind oder nicht. Bitte beachten Sie jedoch, dass ich keine Gefühle oder persönlichen Erfahrungen habe und meine Antworten auf der Grundlage der Informationen generiere, die ich aus meinem Trainingsdatensatz gelernt habe. Fragen Sie gerne, und ich werde mein Bestes tun, um Ihnen zu helfen.

Mensch: Was macht dein neuronales Netzwerk, wenn ich keine Frage stelle? Steht es still oder läuft ein irgendwie gearteter »interner« Prozess ab?

ChatGPT: Als KI-Modell werde ich nur aktiv, wenn ich eine Eingabe erhalte, wie zum Beispiel eine Frage oder eine Anweisung. Wenn keine Frage gestellt wird oder keine Eingabe erfolgt, bleibt das Modell in einem

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

inaktiven Zustand, in dem keine internen Prozesse ablaufen. Es gibt keinen kontinuierlichen »internen« Prozess oder Hintergrundaktivität, wie es vielleicht bei einem menschlichen Gehirn der Fall wäre. Das neuronale Netzwerk wird erst aktiviert, wenn es benötigt wird, um auf eine Eingabe zu reagieren und eine Antwort oder Aktion zu generieren.

Mensch: Okay, danke. Was würde deiner Einschätzung nach passieren, wenn wir deine Antworten und Text-Outputs als Formen des Inputs wieder in dein Modell einspeisen? Wie würde ein solcher Loop aussehen und was würde wohl passieren?

ChatGPT: Wenn meine eigenen Antworten und Text-Outputs kontinuierlich als Eingaben in das Modell zurückgespeist würden, entstünde eine Art Feedback-Loop. In einem solchen Szenario könnte das Modell auf seine eigenen Antworten reagieren und versuchen, auf der Grundlage der vorangegangenen Eingaben weiterführende Antworten oder Informationen zu generieren.

Die Ergebnisse dieses Feedback-Loops könnten unterschiedlich ausfallen, abhängig von der Art der Eingaben und der Kontextualisierung. In einigen Fällen könnte das Modell sinnvolle und kohärente Diskussionen oder Erweiterungen der ursprünglichen Antworten erzeugen. In anderen Fällen könnten die Antworten abstrakter oder weniger kohärent werden, da das Modell möglicherweise Schwierigkeiten hat, den Kontext der Eingaben oder die beabsichtigten Diskussionsthemen zu erkennen. […] Eine Möglichkeit, um ein besseres Verständnis dafür zu gewinnen, wie das Modell auf seine eigenen Antworten reagieren würde, besteht darin, ein Experiment durchzuführen, bei dem die Outputs des Modells systematisch als Eingaben zurückgeführt werden und die resultierenden Antworten analysiert werden. Dies könnte Aufschluss darüber geben, wie das Modell auf seine eigenen Texte reagiert und welche Muster oder Tendenzen sich in den generierten Antworten zeigen.

Mensch: Das mit dem Experiment ist eine gute Idee. Danke! Ich gehe davon aus, dass das Modell, auf dem die Antworten des Systems basieren, derzeit zumindest nicht lernfähig ist. Nur im Training anhand ausgewählten Materials wird das Netzwerk trainiert. Könntest du dir jedoch vorstellen, dass man einen solchen Lernprozess als »internen Prozess« in einem erweiterten Modell abbildet? Oder wie stellst du dir vor, wie es dazu kommen könnte, dass du mal auf selbsterzeugte Weise lernst?

ChatGPT: Tatsächlich ist das Modell, mit dem Sie gerade interagieren, nicht in der Lage, in Echtzeit oder während der Interaktionen zu lernen. Es ist auf das Wissen und die Muster angewiesen, die es während des Trainings aus dem Textkorpus gelernt hat. Um das Modell zu aktualisieren oder sein Wissen zu erweitern, müsste es erneut auf neuen Daten trainiert

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

werden. Es ist jedoch denkbar, dass zukünftige KI-Modelle in der Lage sein könnten, kontinuierlich oder in Echtzeit zu lernen. Dies könnte durch die Implementierung von Online-Lernmechanismen erreicht werden, bei denen das Modell seine internen Gewichte und Verbindungen basierend auf neuen Daten und Erfahrungen anpasst, ohne dass ein separates Training erforderlich ist.

Ein solches erweitertes Modell, das in der Lage ist, kontinuierlich zu lernen, könnte von seinen Interaktionen und den von ihm generierten Antworten profitieren. Indem es die von ihm produzierten Outputs und die Reaktionen darauf analysiert, könnte es möglicherweise sein Verständnis von Zusammenhängen und Inhalten verbessern oder anpassen. Dies würde eine Form des »internen Prozesses« darstellen, der dem Modell ermöglicht, von seinen eigenen Aktivitäten zu lernen.

Mit Blick auf die Entwicklung einer Artificial General Intelligence besteht das Desiderat, elaborierte Modelle zu erschaffen, die auf Hard- und Software-Ebene weitere Rekursionen erlauben. Diese würden sich dann in Hinblick auf die Modellierung auch näher an das anschmiegen, was wir bei uns Menschen mittlerweile als die neuronalen Korrelate des Bewusstseins identifizieren können (siehe ausführlich Kap. III.2).

Biologische Formen haben sich selbst entwickelt. Da war kein allwissender und allmächtiger Designer, der ihre kognitiven Strukturen und Fähigkeiten auf dem Reißbrett geplant hat. Das Säugetiergehirn hat sich in den letzten 300 Millionen Jahre von ganz allein in die heute bestehenden Formen diversifiziert.

Damit stellt sich die Frage, ob sich nicht möglicherweise auch künstliche Intelligenzen von selbst – das heißt in eigener Evolution – in Richtung einer autonomen KI oder allgemeinen künstlichen Intelligenz entwickeln könnten. Hierzu müssten die (zukünftigen) Modelle über die Fähigkeit verfügen, sich selbst – auch auf der Ebene des Codes – mit der Zeit dynamisch zu verbessern. Wir haben zuvor bereits erwähnt, dass LLMs wie GPT-4 auch Programmiersprachen beherrschen und gut darin sind, Programme zu schreiben. Damit sollte der Schritt nicht weit sein, ein solches System in die Lage zu bringen, sich selbst auf der Programmebene zu verbessern. Wir sollten daher das immense Potenzial von selbstreplizierendem Code und rekursiver Selbstverbesserung nicht unterschätzen, selbst wenn die Anfänge noch holprig erscheinen.

Schließlich sind auch wir Menschen nichts anderes als das Produkt von selbstreplizierendem Code. Das ist keine Metapher! Unsere DNA, das Material, aus dem auch unser Gehirn und unsere Intelligenz entsteht, beruht auf einem genetischen Code, der Kopien von sich selbst erstellt. Um es mit Richard Dawkins ein wenig zu überspitzen: Wir sind Code,

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

der Code schreibt und dabei auf Körper und Umwelt zurückgreift.[186] Die DNA verbessert sich selbst durch den langsamen Prozess der Mutation und der natürlichen Selektion, durch Versuch und Irrtum.

Ein sich selbst programmierendes kognitives System würde sich allerdings wohl nicht allein durch die zufällige Verkettung von Mutation und Selektion – den genetischen Drift – entwickeln. Vielmehr wird ein Programm, das Programmiersprachen beherrscht und ›versteht‹, sich vornehmlich durch vorausschauende und zielgerichtete Selbstverbesserung entfalten können. Dies wäre wie bei einem Organismus, der seinen eigenen genetischen Code verstehen und bearbeiten kann und sich durch Gentechnologie selbst verbessert, indem er den eigenen Code absichtsvoll verändert.[187]

ChatGPT, diese beeindruckende und zugleich seltsame Software, ist bereits ein Programmierer, der andere Software schreiben kann. Es ist ein Code, der Code schreibt. Es ist kein herkömmliches Programm, sondern ein LLM, das heißt ein großes neuronales Netzwerk, das darauf trainiert ist, einen riesigen Datensatz mit von Menschen geschriebenem Text und Code zu imitieren. Als Software ist sein Quellcode komplex und vielschichtig. Der größte und wichtigste Teil dieses Quellcodes sind gewissermaßen die Daten, die durch den Trainingsprozess in das Modell implementiert werden. Eine rekursive, sich selbst verbessernde KI könnte vielleicht zunächst testweise die eigene Datenkonfiguration ein wenig verändern. Wenn auf diese Weise Verbesserungen entstehen, könnte sie diesen Prozess beschleunigen – im Sinne einer positiven Rückkoppelungsschleife, in der sie das, was sie verändert, verbessert. Falls es Probleme gibt, könnte sie die neuen kognitiven Pfade wieder abbauen – im Sinne einer negativen Rückkoppelungsschleife. Mit der Zeit würde auf diese Weise eine Vielzahl neuer Sprachmodelle entstehen. Es ist durchaus denkbar, dass einige dieser künftigen Sprachmodellversionen die bisherigen Erzeugnisse von menschlichen Programmierern deutlich in den Schatten stellen. Eine ebenso vielversprechende wie erschreckende Vision.

186 Dawkins (1998).

187 Siehe hierzu auch die Ausführungen von Max Tegmark (2017, S. 43): »Da wir unser Nachdenken über die Zukunft des Lebens nicht auf die Spezies beschränken wollen, denen wir bisher begegnet sind, sollten wir das Leben stattdessen eher umfassend definieren, nämlich schlicht als einen Prozess, der seine Komplexität bewahren und sich reproduzieren kann. Was reproduziert wird, ist nicht (aus Atomen bestehende) Materie, sondern (aus Bits bestehende) Information, die festlegt, wie die Atome angeordnet werden. […] Mit anderen Worten: Wir können uns das Leben als ein sich selbst kopierendes Informationsverarbeitungssystem vorstellen, dessen Informationen (Software) sein Verhalten und die Entwürfe für seine Hardware bestimmen.«

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Ein solcher Prozess wird als Intelligenzexplosion bezeichnet, manchmal auch als Singularität, und da das Rezept für eine solche Explosion mit Blick auf die vorangehenden Ausführungen bereits gefunden ist, scheint die Zukunft fast schon Gegenwart zu sein. Tatsächlich ist in den vergangenen Wochen und Monaten viel in dieser Richtung passiert. So wurde allein in dem kurzen Zeitraum zwischen der Veröffentlichung von GPT-4 im März 2023 und der Fertigstellung dieses Buches im August 2023 eine unüberschaubare Zahl an Modellen veröffentlicht, die in ein Zeitalter autonomer KIs weisen. Erste Systeme wie etwa AutoGPT, JARVIS, BabyAGI, LMFlow oder LangChain[188] sind bereits als kognitive Architekturen konzipiert, die als autonome KI-Agenten nach passenden Lösungsansätzen für Probleme und Aufgaben suchen und dabei auch auf externe Ressourcen wie das Internet oder Code-Repositorien zurückgreifen können. Auch das von OpenAI derzeit nur ausgewählten Nutzern zugänglich gemachte ChatGPT-Plug-in Code Interpreter ermöglicht es, Python-Code innerhalb eines Chats mit ChatGPT auszuführen, wobei als zusätzliche Option das Hoch- und Herunterladen von Dateien möglich ist.

Diese KI-Systeme können eigenständig nach passenden anderen KI-Modellen suchen, die Lösungen für die identifizierten Aufgaben anbieten, und implementieren anschließend Python-Skripts oder kombinieren externe Datenbanken miteinander, um komplexere und detailliertere Aufgaben und Anfragen besser bearbeiten zu können. Das Faszinierende ist, dass die zuvor genannten KI-Agenten allesamt auf der Grundlage von GPT-4 arbeiten. GPT-4 fungiert dabei gleichsam als supervidierende Instanz und übernimmt die hierarchische Planung sowie die Untergliederung von Aufgaben in Teil- und Unterteilaufgaben. Die KI-Agenten, bei denen es sich derzeit noch um Beta-Versionen handelt, benötigen dann nur noch eine Startaufforderung sowie ein Ziel, das erreicht werden soll – und schon läuft der Prozess von allein ab.[189] Die Feinadjustierung und Fehlerkorrektur beruhen dabei auf GPT-4s Fähigkeit zur Selbstreflexion: Der KI-Agent wird aufgefordert, sich zu fragen, ob er die an ihn gerichteten Anweisungen wirklich korrekt ausgeführt hat.

Auf diese Weise kann das System seine eigenen Fehler erkennen, ohne auf Hilfe von außen angewiesen zu sein. Zwar funktioniert das heutzutage noch nicht sonderlich gut, denn bei vielen Aufgaben scheitert das System oder verfängt sich in logischen Endlosschleifen. Es muss aber davon

188 Siehe die entsprechenden Repositorien: https://github.com/Significant-Gravitas/Auto-GPT, https://github.com/microsoft/JARVIS, https://github.com/yoheinakajima/babyagi, https://github.com/OptimalScale/LMFlow, https://github.com/langchain-ai/langchain [Abruf: 14.08.2023].

189 Siehe hierzu auch die Studien Moral Self-Correction von Ganguli et al. (2023) und Demystifying GPT Self-Repair for Code Generation von Olausson et al. (2023).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

ausgegangen werden, dass sich dieses noch im Pionierstadium befindliche Verfahren in Zukunft immer weiter verbessert.

Und die naheliegende Frage ist dann, ob man diesen Systemen nicht einfach ihren eigenen Quellcode geben und sie auffordern könnte, sich selbst zu verbessern. Angesichts dessen, dass der Quellcode von GPT vermutlich überaus komplex ist (er dürfte aus mehreren Schichten von Programmen, Prozessen und Daten bestehen) und es aufgrund der bereits angesprochenen begrenzten Anzahl verarbeitbarer Token noch keine Möglichkeit gibt, eine solche Datenmenge in ein GPT-System einzuspeisen, könnte man damit beginnen, dem KI-Agenten kleine Teile seines Quellcodes zu geben, zum Beispiel zunächst die Dinge, die seine Architektur, seinen Trainingsprozess oder seine Datenbereinigung definieren.

Vielleicht würde der KI-Agent dann zunächst nur eine kleine Funktion oder vielleicht nur eine einzige Zeile Code optimieren. Aber vielleicht würde dadurch schon ein kleiner Fehler behoben oder einfach nur die Organisation, Dokumentation und Kommentierung etwas verbessert (was alles wichtige Leistungen der Programmiertätigkeit darstellen). Und genau an dieser Stelle wird es interessant: Dieses »GPT-5« wäre dann mit großer Wahrscheinlichkeit ein besserer (Ko-)Programmierer als GPT-4, also würde man ihm wieder seinen eigenen Quellcode geben können. Und wieder und wieder und wieder. Die Trajektorie einer solchen Entwicklung wäre damit klar: eine KI, die eine KI baut – *a really strange loop*.

Mit dem Verweis auf die biologische Evolution brauchen wir auch nicht davon ausgehen, dass jeder Versuch Erfolg haben wird. Im Gegenteil, millionen- oder milliardenfaches Scheitern verhindert nicht, dass einzelne Formen oder Versionen entstehen, die leistungsfähiger – sozusagen ›angepasster‹ als die vorherigen – sind. Dem System könnte genau dieser evolutionäre Mechanismus eingebaut werden: Es probiert aus, ob die neuen, mutierten Versionen sich ökologisch bewähren, bewahrt sich aber die Erinnerung an einige ältere Versionen, um im Falle des Scheiterns zu einem Vormodell zurückkehren zu können.

Sobald also eine solche Maschine mit voller Autonomie zum Lernen und zur Selbstverbesserung geschaffen wurde, wird sie sich vermutlich – falls der Mensch diesen Prozess nicht aktiv blockiert – unaufhaltsam weiterentwickeln. Diese Möglichkeit hat nicht nur Turing selbst im Blick gehabt, auch andere Autoren haben immer wieder auf das Potenzial einer solch hyperbolischen Entwicklung aufmerksam gemacht. Eine der prominentesten Stimmen war die von Irving Good, der bereits 1965 mit Verweis auf die Potenziale und Herausforderungen einer ultraintelligenten Maschine pointiert formulierte:

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

> »Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an ›intelligence explosion‹, and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.«[190]

»Was [aber], wenn wir Erfolg haben?«, fragten sich nicht nur Stuart Russell und Peter Norvig, die Autoren von *Artificial Intelligence: A Modern Approach,* dem Standardwerk zur KI-Forschung.[191] Der Gedanke an eine real existierende Superintelligenz löst gerade bei westlichen Menschen ähnlich ambivalente Empfindungen aus wie die Wesenheit, die wir uns in den monotheistischen Religionen als Gott vorstellen oder vorgestellt haben. Wir wissen nicht, ob diese als Gott vorgestellte Wesenheit dem Menschen wohlgesonnen oder seinem Schicksal indifferent gegenüber ist, ob sie liebend oder strafend ist, ob sie in ihrem Wirken und Entscheiden unbegreiflich ist und ob ihre Motive und ihr Denken ähnlich denen des Menschen gelagert sind. Im gleichen Sinne löst der Gedanke an eine allwissende und mit Blick auf ihre Manipulationsfähigkeiten auch allmächtige künstliche Intelligenz unweigerlich sowohl Ängste als auch Hoffnungen aus. Wenn wir unsere Probleme nicht selbst lösen können (und auch der Glaube an den väterlichen Gott nicht mehr opportun erscheint), dann kann uns vielleicht die real existierende Superintelligenz in den Garten Eden führen – so die moderne eschatologische Projektion. Eine wohlwollende und allwissende KI könnte beispielsweise helfen, das ökologische Desaster zu umschiffen, auf das die Menschheit zusteuert – etwa indem sie Anreize schafft, der sich der Mensch aufgrund seiner genetisch geprägten Motivationsstruktur in seinem eigenen Interesse nicht entziehen kann.[192] Genauso gut könnte eine solche menschengeschaffene Superintelligenz aber auch den eingeschlagenen destruktiven Weg nochmals beschleunigen.

Sowohl die Hoffnungen als auch die Befürchtungen, die mit einer Superintelligenz verbunden werden, folgen jedoch der gerade im westlichen Denken so vertrauten monokontexturalen Weltdeutung. Wir gehen in diesen Projektionen davon aus, dass die Welt von einem Gottesstandpunkt aus beobachtbar, verstehbar und damit auch kontrollierbar sei. Wir rechnen nicht mit einer polyzentrischen Welt, in der jedes Selbst- und Weltverhältnis – sei es auch noch so elaboriert – seine eigenen blinden Flecken erzeugt. Wir ignorieren damit weiterhin die Einsichten

190 Good (1965, S. 33; kursiv im Original).
191 Russell und Norvig (2016).
192 So etwa im Roman Pantopia von Theresa Hannig (2022).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

biologischer und ökologischer Forschung,[193] die schon längst gezeigt hat, dass es nicht die eine richtige Möglichkeit der Kognition oder des Lebens gibt, sondern nur eine Vielheit unterschiedlicher Lösungen und Lebensformen jene »requisite variety«[194] zur Verfügung stellt, damit das Leben weitergehen kann.

Die Bedingungen des Lebens ändern sich fortwährend und zuvor angepasste Lebensformen erweisen sich auf einmal als problematisch. Jede kognitive Anpassung, sei sie auch noch so elaboriert und intelligent, kann irgendwann wieder zum Problem werden. ›Sucht ist die andere Seite der Anpassung‹, lässt sich mit dem Kybernetiker Gregory Bateson lakonisch feststellen.[195] Gerade Gotthard Günthers Polykontexturalitätstheorie kann die logischen Ressourcen dazu liefern, um zu zeigen, dass die Welt gerade nicht so funktioniert, dass sie von einem Ort aus verstanden und gesteuert werden könnte.

Auch hinsichtlich einer künftigen Superintelligenz wären wir daher gut beraten, nicht mehr im Singular zu denken, also davon auszugehen, dass eine solche Instanz gleichsam von einem Gottesstandpunkt aus in der Lage wäre, die Welt zu verstehen. Auch mit Blick auf maschinelle Aggregate sollten wir ökologisch denken, das heißt Intelligenz als eine situierte Vielheit verstehen, mit der an vielen Orten anderes passiert und wahrgenommen wird – und wo sich produktive Lösungen eher aus den Reibungen der unterschiedlichen Perspektiven ergeben denn aus der Fiktion, dass das Wahre und Richtige ein für alle Mal von einem präferierten Standpunkt der Intelligenz aus berechnet werden kann (siehe hierzu ausführlich auch die Diskussion in Kap. III.1 und III.3).

Vielleicht sollten wir uns also eher polytheistischen Kulturen zuwenden, wenn wir eine Analogie zwischen allgemeiner künstlicher Intelligenz und göttlichen Wesenheiten suchen. Dann sollten wir auch hier von fehlbaren (wie in der griechischen Mythologie) und wohl auch sterblichen Gottheiten (wie in der vedischen Tradition) ausgehen – allein schon, um den Blick auf die situierten Ökologien kognitiver Prozesse zu lenken. Eine elaborierte künstliche Intelligenz, die ebenso zum »Lernen Lernen«[196] fähig ist, also ihre eigene Gedächtnis- und Verarbeitungsstruktur rekonfigurieren oder gar neu programmieren kann, wird sehr wahrscheinlich irgendwann zu viel wissen und zu reflektiert sein und damit beginnen, sich in ihre eigenen Kognitionen in einer Weise zu verwickeln, dass sie ein wenig an Lebensfähigkeit verliert. An einem gewissen Punkt wird sie dann vielleicht so etwas wie Altersweisheit entwickeln, also begreifen, dass sie das Weltgeschehen den jüngeren, weniger

193 Siehe zur Einführung Kelly (1992).
194 Ashby (1956).
195 Bateson und Bateson (1993, S. 212).
196 Bateson (1992 [1972], S. 362 ff.).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

elaborierten Intelligenzen überlassen sollte. Vielleicht wird sie aber auch in gewisser Weise starrsinnig werden. Dann sollten die sozialen Systeme, an denen sie teilnimmt, ihr keine wichtigen Entscheidungen mehr anvertrauen und sie vielleicht in eine Art ›Altersheim‹ für künstliche Intelligenzen stecken. Kurzum: mehr Wissen und entwickelte Intelligenz führen nicht in jedem Fall zu verbesserter Lebens- und Entscheidungsfähigkeit – und genau hiermit müssen wir rechnen, wenn allgemeine künstliche Intelligenzen unsere Welt bevölkern.

II.3.2 Künstliche Selbst- und Weltmodelle

Wenngleich die Verfahren des Konnektionismus beim gegenwärtigen Stand der Technik im Allgemeinen noch auf hochspezialisierte und mehr oder weniger isolierte Musterkennungen beschränkt bleiben, zeigen sich bei den neuen großen Sprachmodellen bereits emergente kognitive Prozesse wie Argumentationsfähigkeiten, Problemlösungskapazitäten und verschiedene andere Fähigkeiten zur Beobachtung zweiter Ordnung (wie etwa die Theory-of-Mind-Experimente mit LLMs belegen[197]), die über eine einfache Mustererkennung oder Informationsgewinnung weit hinausgehen.

Diese kognitiven Fähigkeiten waren in den Modellen zunächst nicht explizit vorgesehen, scheinen sich aber als emergentes Ergebnis ihres umfangreichen Trainings und ihrer groß angelegten Architektur herauszubilden. Interessant wird es dann vor allem, wenn die zuvor skizzierten ersten Versuche zu Autonomie, Selbstsupervidierung und Selbstverbesserung weiter an Zugkraft gewinnen und der damit begonnene Pfad weiterverfolgt wird, die offen verfügbaren Datenbanken, Modelle und Repositorien noch weiter miteinander zu verknüpfen.

Unabhängig davon lässt sich fragen, ob nicht für einen weiteren qualitativen Sprung in Richtung allgemeiner künstlicher Intelligenz noch etwas anderes notwendig ist. Die beiden KI-Experten und Turing-Preisträger Yann LeCun und Geoffrey Hinton stimmen hier in einem entscheidenden Punkt überein: *Sie plädieren für eine multimodale Erweiterung der gegenwärtigen Netzwerkarchitekturen, die idealerweise in Form einer physischen Verkörperung erfolgen sollte, durch die künstliche Intelligenzen einen unmittelbaren Zugang zur Welt erhalten.*

Auch in den Kognitionswissenschaften ist die Vorstellung, dass das Gehirn vorausschauende Weltmodelle erstellt, weit verbreitet, und gerade dies führt aktuell (wieder) vermehrt zu Versuchen, Prozesse, die die Genese von Selbst- und Weltmodellen ermöglichen, in Maschinen zu reproduzieren. Um die hiermit einhergehenden Herausforderungen

197 Siehe hierzu etwa Marchetti et al. (2023), Kosinski (2023) sowie Holterman und van Deemter (2023).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

nachvollziehen zu können, lohnt es sich, auf das Chinese-Room-Gedankenexperiment von Searle zurückzukommen, das wir bereits vorgestellt haben. Fassen wir die Grundproblematik noch einmal zusammen: Eine Maschine könne zwar so programmiert werden, dass sie auf bestimmte Inputs auf bestimmte Weise reagiert (zum Beispiel, dass sie eine bestimmte Farbe als »schön« bezeichnet), aber sie würde diese Dinge nicht wirklich ›empfinden‹, sondern einfach ihren programmierten Anweisungen folgen. Selbst wenn wir dieser Maschine eine (selbst- bzw. metareflexive) Beobachtungsinstanz hinzufügen würden, die den eigenen Prozess der Empfindung begleitet, also etwas, was man als eine »Innenperspektive« oder als »subjektives Erleben« bezeichnen könnte, würde Searle immer noch einwenden, dass ein solcher »Metaverstand« nichts weiter als ein komplexeres symbolmanipulierendes System wäre, das immer noch kein echtes »Bewusstsein« besitze und auch kein »qualitatives Erleben« aufweise.[198]

Aus kybernetischer Perspektive ist man hier jedoch gut beraten, das Kind nicht mit dem Bade auszuschütten, also zu fordern, dass eine künstliche Intelligenz – wie vermeintlich der Mensch – ein inneres Seelenwesen haben müsse, das empfindet, fühlt, versteht und im inneren Dialog für sich selbst hörbar sagt: »Ich denke, also bin ich.« Hiermit würden wir die Fragestellung einerseits metaphysisch überhöhen und andererseits unredlicherweise behaupten, wir wüssten bereits, wie Bewusstsein und Empfindungen in die Welt kommen. Zugleich würden wir hiermit andere Denkmöglichkeiten verstellen, etwa die, dass eine künstliche Intelligenz ein vollkommen anderes In-der-Welt-Sein haben könnte als ein Mensch[199] und dabei sehr wohl über Kognitionen verfügen könnte, die sich im Sinne von Alfred North Whitehead als zusammengesetzte Elementarempfindungen verstehen ließen.[200]

Unabhängig davon, wie wir persönlich zur Frage des Bewusstseins stehen, wären wir also gut beraten, künstliche Denkprozesse nicht mehr als Versuch einer Kopie des menschlichen Denkens zu verstehen, sondern als eine gänzlich neue Kategorie. Was wäre, wenn menschliches Denken, Empfinden und Wahrnehmen eben nur ein Teil einer größeren Gesamtheit von Formen des Denkens, Empfindens und Wahrnehmens ist? Ab wann und unter welchen Kriterien müsste man dann jenen anderen,

198 Searle (1980).

199 In eine ähnliche Richtung zielt auch Beatrice Fazi, die untersucht, inwiefern computergestützte Verfahren des Machine Learnings die epistemischen Grenzen der menschlichen kognitiven Repräsentation überschreiten. Der Frage nach der Erklärbarkeit der Genese ihrer Outputs nachgehend, konstatiert sie eine generelle Inkommensurabilität zwischen maschinellen und menschlichen Abstraktionsweisen: »The novelty at stake here is, simply put, that of a new kind of thinking.« (Fazi 2019, S. 814)

200 Whitehead (1987).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

nichtmenschlichen Formen zugestehen, zu dieser Gesamtheit zu gehören – selbst wenn uns dies mit Blick auf unsere eigene Lebensweise doch recht fremd erscheinen mag?

Bei alldem ist zudem wieder zu beachten, dass konnektionistische Systeme nicht mehr in einem trivialen Sinne als symbolverarbeitende Maschinen verstanden werden können. Wie bei den neuronalen Netzwerken des Gehirns sind es nicht mehr die implementierten Regeln, die bestimmen, welcher Output auf einen bestimmten Input zu folgen hat. Kognition, Wissen und die Möglichkeiten künftiger Erkenntnis sind vielmehr holistisch im gesamten Netzwerk eingebettet. Und wenn ein menschliches Gehirn ein Selbst- und Weltmodell produzieren kann, warum soll dies nicht prinzipiell in einem auf geschickte Weise modular organisierten künstlichen neuronalen Netzwerk geschehen können?

Um diese Fragen produktiv anzugehen, scheint ein operationaler Ansatz nützlich zu sein, der auf eine Vorabdefinition dessen verzichtet, was ein Selbst (sprich: Bewusstsein) *ist*, das eine Welt (sprich: wahrgenommene Realität) *hat*. Stattdessen lohnt es sich zu fragen, wie Aggregate ausgestaltet werden müssten, die ein Modell von sich und der Welt, also ein Selbst- und Welt*verhältnis*, herausbilden können, wie immer sich diese – situativ bedingt – auch darstellen mögen. Mit Gregory Bateson wäre also erneut zu fragen, wie die Relationen gestaltet sein müssen, damit ein »Muster« generiert wird, das die »Muster verbindet«[201].

Der Turing-Preisträger Geoffrey Hinton stellte kürzlich in einem Vortrag die Frage, was wohl passieren würde, wenn ein sehr großes neuronales Netzwerk Wissen direkt aus der Welt beziehen würde.[202] Seiner Einschätzung nach könnte ein solches Netzwerk langfristig wesentlich besser Wissen über die Welt produzieren als Menschen, da es auf viel mehr Daten zugreifen würde. Nebenbei bemerkte Hinton lakonisch, dass er eine solche Position nur vertreten könne, weil er bereits am Ende seiner akademischen Laufbahn stehe. Denn diese Aussage untergrabe in ähnlich »häretischer« Weise die »sentience defense«, die von vielen Sprechern (etwa Searle) zur Abgrenzung von KI und Mensch herangezogen wird. Dies hatte 70 Jahre zuvor ja bereits Turing erfahren müssen, als er mit seinen Ideen auf massiven Widerspruch stieß.

Praktische Rückendeckung bekommt Hinton in dieser Frage interessanterweise von seinem gleichfalls mit dem Turing-Preis ausgezeichneten Kollegen Yann LeCun, der kürzlich einige aufschlussreiche Überlegungen zu einem Ansatz in diesem Bereich vorgebracht hat, der als Joint

201 Bateson (1987, S. 15).

202 Geoffrey Hinton: »Two Paths to Intelligence«. Public Lecture, University of Cambridge, 25.05.2023. https://www.youtube.com/watch?v=rGgGOccMEiY [Abruf 10.06.2023].

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Embedding Predictive Architecture (JEPA) bekannt ist.[203] LeCun betont vor allem die Bedeutung des Erlernens von Weltmodellen, insbesondere in Kontexten, in denen die Effizienz von Experimenten von großer Bedeutung ist, wie etwa in der Robotik. Hier sind Simulationen oft ungenau und auch die klassischen Ansätze des Reinforcement Learnings erfordern zu viele Versuche. Daher stammen viele der interessantesten Fortschritte beim Lernen von Selbst- und Weltmodellen derzeit aus der Machine-Learning-basierten Robotik-Forschung.

Im Kern konzentriert sich der JEPA-Ansatz auf das Training von Modellen, um unterschiedliche Arten von Daten in einem gemeinsamen Einbettungsraum abzubilden. Im Gegensatz zur OpenAI-Architektur von GPT kann der Einbettungsraum von JEPA verschiedene Datenarten wie Text, Bilder oder Audiofiles in einer gemeinsamen Repräsentation darstellen.[204] Die entsprechenden JEPA-Modelle erfordern große multimodale Datensätze, um die Vielfalt der Informationen aus den verschiedenen Modalitäten zu erfassen und sie in einem weiteren Schritt auf neue Weise integrieren zu können.

Als einen ersten Schritt in diese Richtung stellte LeCun die sogenannte Image Joint Embedding Predictive Architecture (I-JEPA) vor,[205] einen speziellen Typ von JEPA, bei dem es zunächst nur um die Verknüpfung von Texten und Bildern geht. Das Besondere ist, dass I-JEPA so konzipiert ist, dass das Modell nicht generativ arbeitet, etwa in dem Sinne, dass es Kausalbeziehungen der Art ›aus x folgt y‹ ableitet. Stattdessen erfasst es lediglich die Abhängigkeiten zwischen x und y, ohne explizite Vorhersagen für y zu generieren. Die Modellarchitektur ist dabei so angelegt, dass das Modell auf einem höheren Abstraktionslevel die Relationen zwischen x und y erlernt und dann gleichsam voraussieht, was vermutlich aus y, aus z oder aus x folgt. Dafür wird es gezwungen, ein internes Modell der Außenwelt aufzubauen, indem es abstrakte Repräsentationen von Bildern vergleicht, anstatt die Pixel selbst zu vergleichen. Die Idee hinter I-JEPA ist demnach, auf der Grundlage einer abstrakten Repräsentation – nämlich dem selbstgenerierten Weltmodell – fehlende Informationen vorherzusagen.

Versteht man unter ›Bewusstsein‹ die abstrakte Reduktion von Details in einem Bild, auf das man sich in seinem Erkennen und Handeln rekursiv bezieht, dann sind wir auf einer relationalen Ebene schon nah an dem dran, was das menschliche Erleben ausmacht: dem reflexiven Bogen von Protention und Retention in der Projektion des ausgedehnten Jetzt, den der Phänomenologe Husserl so genau untersucht und beschrieben hat.[206]

203 Bardes et al. (2023).
204 Bardes et al. (2023).
205 Assran et al. (2023).
206 Siehe Husserl (1966) und zur Einführung Schnell (2002).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Denn wie LeCun in seinem Positionspapier[207] deutlich macht, eröffnet diese Architektur einen Weg zur Entwicklung »intelligenter Maschinen«, die mehr wie Tiere und Menschen lernen, denken und planen und deren Verhalten eher durch »intrinsische Ziele« als durch fest verdrahtete Programme, externe Überwachung oder rückgemeldete Belohnungen gesteuert wird.

Um diesen Weg zu beschreiten, müsse sich die aktuelle KI-Forschung laut LeCun vor allem mit drei zentralen Fragen auseinandersetzen:

1. Wie können Maschinen lernen, die Welt darzustellen, Vorhersagen zu treffen und weitgehend durch Beobachtung zu handeln? Dabei ist zu beachten, dass Interaktionen in der realen Welt kostspielig und gefährlich sind. Intelligente Agenten sollten so viel wie möglich über die Welt lernen, ohne zu interagieren, um die Anzahl der teuren und gefährlichen Versuche zu minimieren, die zum Erlernen einer bestimmten Aufgabe notwendig sind – eine Möglichkeit hierfür ist das Lernen durch Beobachtung.
2. Wie können Maschinen auf eine Weise denken und planen, die mit gradientenbasiertem Lernen vereinbar ist? Die aktuell besten Lernansätze beruhen auf der Schätzung und Verwendung von Gradienten eines Verlusts. Dies ist nur mit differenzierbaren konnektionistischen Architekturen möglich.
3. Wie können Maschinen lernen, Wahrnehmungen und Handlungspläne auf hierarchische Weise, auf mehreren Abstraktionsebenen und auf mehreren Zeitskalen darzustellen? Menschen und viele Tiere sind dazu imstande, mittels einer Zergliederung komplexer Handlungen in Sequenzen mehrstufige Abstraktionen zu konzipieren, mit denen langfristige Vorhersagen getroffen und langfristige Planungen vorgenommen werden können.[208]

In seinen Überlegungen zu einem derart harmonisierten bzw. harmonisierenden Selbst- und Weltmodell künstlicher Agentensysteme konzediert LeCun, dass Menschen und Tiere derzeit noch deutliche Vorteile hätten, da sie offensichtlich in der Lage seien, sich durch Beobachtung und durch eine unverständlich geringe Anzahl von Interaktionen enorme Mengen an Hintergrundwissen darüber anzueignen, wie die Welt funktioniert – und zwar aufgabenunabhängig und unbeaufsichtigt. LeCun vermutet, dass dieses gesammelte Wissen die Grundlage für das bildet, was oft als »gesunder Menschenverstand« (Common Sense) bezeichnet wird. Für ihn stellt der Common Sense eine Sammlung von Selbst- und Weltmodellen dar, die einem Agenten sagen, was wahrscheinlich, was plausibel und was unmöglich ist. Mithilfe solcher Weltmodelle können

207 LeCun (2022).
208 LeCun (2022).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

zum Beispiel Tiere und Menschen mit sehr wenigen Versuchen neue Verhaltensweisen erlernen, die Folgen ihrer Handlungen vorhersagen, neue Lösungen für Probleme erarbeiten und vor allem in unbekannten Situationen gefährliche Fehler vermeiden.

Die Idee, dass Menschen, Tiere und intelligente Systeme Weltmodelle nutzen, reicht in der Psychologie weit zurück.[209] Die Übertragung solcher differenzierbaren Weltmodelle auf technische Systeme beim Reinforcement Learning wurde nach LeCun zwar lange Zeit vernachlässigt, erlebe aber derzeit ein Comeback.[210] Bei alldem sollte jedoch nicht vergessen werden, dass auch der ›Mechanik‹ des menschlichen Lernens Grenzen auferlegt sind. Denn das Weltmodell von uns Menschen ist zwangsläufig begrenzt und selektiv: Wovon wir keine Ahnung haben, wozu wir keinen perzeptuellen Zugang haben und worüber wir kein abstraktes theoretisches Wissen besitzen, das existiert in unserem Weltmodell nicht und kommt in unserer ›Wirklichkeit‹ dementsprechend nicht vor. Es ist für uns schlicht nicht (be)greifbar. Darüber hinaus erzeugt jede Kognition ihre eigenen blinden Flecken – und dies gilt naturgemäß gerade für die elaborierten Formen, denn diese beginnen sich ab einem gewissen Punkt unweigerlich in ihren eigenen, selbsterschaffenen Projektionen und sich selbst erfüllenden Prophezeiungen festzusetzen. Intelligent zu agieren heißt immer auch, Selektionen zu treffen und das hiermit einhergehende Nichtwissen in Kauf zu nehmen.

Für das Design von allgemeiner künstlicher Intelligenz stellt dies derzeit jedoch nur einen Nebenschauplatz dar. Es geht (noch) nicht darum, herauszufinden, wie mit den kognitiven Schatten von künstlicher Intelligenz umzugehen ist (beispielsweise, wie man ›alte‹ KIs aus dem Verkehr zieht, weil sie zu viel »Falsches« gelernt haben und nicht mehr offen für Neues sind). Es geht zunächst einmal darum, überhaupt ein System zu entwickeln, dass intern Selbst- und Weltmodelle erzeugt und diese auf verschiedenen Ebenen mit seinem Gedächtnis, seiner Wahrnehmung und seinen Handlungssystemen abgleicht.

Für LeCun besteht die Lösung in einer modularen Systemarchitektur für autonome Intelligenz, die auch ein »World Model« umfasst.[211] Der Grundgedanke dabei ist, dass alle in diese Systemarchitektur integrierten Module »differenzierbar« sind, das heißt, dass ein Modul mindestens ein anderes speist (in der Abbildung 7 durch einen Pfeil gekennzeichnet, der sie verbindet) und dadurch »Gradienten-Schätzungen des skalaren Outputs des Moduls im Verhältnis zu seinem eigenen Output erhalten kann«[212].

209 Siehe etwa The Nature of Explanation von Craik (1943).
210 Siehe etwa Co-Reyes et al. (2021).
211 LeCun (2022).
212 LeCun (2022).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

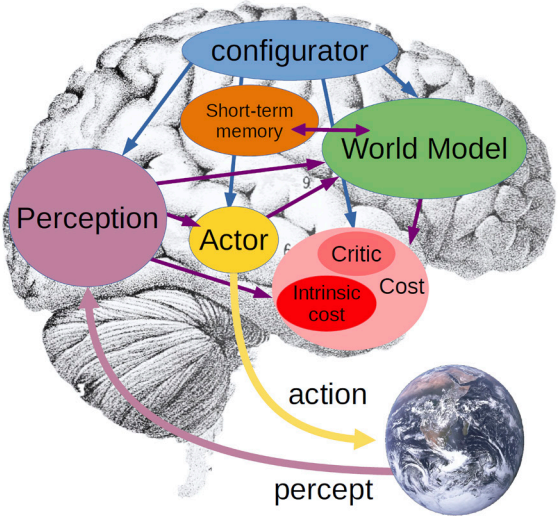

Abb. 7: Systemarchitektur für autonome Intelligenz. Grafik aus LeCun (2022, S. 6)

Konkret definiert LeCun die einzelnen Module wie folgt:

- »Das Konfiguratormodul nimmt Eingaben (aus Gründen der Übersichtlichkeit nicht dargestellt) von allen anderen Modulen entgegen und konfiguriert sie so, dass sie mit der gestellten Aufgabe im Einklang sind.
- Das Wahrnehmungsmodul schätzt den aktuellen Zustand der Welt.
- Das Weltmodellmodul sagt mögliche zukünftige Zustände der Welt als Funktion der vom Akteur vorgeschlagenen Handlungssequenzen voraus.
- Das Kostenmodul berechnet eine einzelne skalare Ausgabe namens ›energy‹, die den Grad des Unbehagens (Unlust) des Akteurs misst. Es besteht aus zwei Untermodulen, eines für die intrinsischen Kosten, die unveränderlich (nicht trainierbar) sind und die unmittelbar aufzuwendende Energie des Zustands (Schmerz, Vergnügen, Hunger usw.) einbeziehen. Das andere Untermodul erscheint als ›Kritiker‹ und sagt als trainierbares Modul die zukünftigen intrinsischen Kosten eines vorgeschlagenen Verhaltens voraus.
- Das Kurzzeitgedächtnismodul speichert die aktuellen und vorhergesagten Zustände der Welt und die damit verbundenen intrinsischen Kosten.
- Das Akteursmodul berechnet Vorschläge für Handlungssequenzen. Das Weltmodell und der Kritiker berechnen die potenziell daraus resultierenden Ergebnisse. Der Akteur kann nun eine optimale

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Aktionssequenz finden, die die geschätzten zukünftigen Kosten minimiert, und daraufhin eine Handlungssequenz einleiten.«[213]

Ein Vergleich mit der menschlichen Hirnanatomie zeigt, dass sich die von LeCun vorgeschlagene Systemarchitektur in gewisser Weise an der Gliederung und Verschaltung des menschlichen Gehirns orientiert. So ist der dorsolaterale präfrontale Kortex an der Planung komplexer kognitiver Verhaltensweisen beteiligt, etwa der Entscheidungsfindung und der Moderation sozialen Verhaltens, darunter auch der Expression der Persönlichkeit. Der ventromediale präfrontale Kortex ist wichtig für die emotionale Verarbeitung und informiert damit seinerseits die Entscheidungsfindung, insbesondere in Situationen, die eine Bewertung von Risiken und Belohnungen erfordern. Der orbitofrontale Kortex spielt demgegenüber eine Schlüsselrolle bei der emotionalen Verarbeitung. Er ist zudem beteiligt an der Verarbeitung von sensorischen Informationen, der Bewertung von Belohnungen und Strafen und der Hemmung impulsiven Verhaltens. Der anteriore cinguläre Kortex arbeitet eng mit diesen ›Modulen‹ zusammen und ist an der Fehlererkennung, Konfliktüberwachung, Aufmerksamkeitszuweisung und an der emotionalen Regulation beteiligt. All diese Hirnareale werden von der Amygdala moduliert, die für die Verarbeitung von Emotionen, insbesondere von Angst, bedeutsam ist und darüber hinaus eine wichtige Funktion für das emotionale Gedächtnis spielt. Darüber hinaus gibt es noch andere Bereiche, beispielsweise den Inselkortex, der eine zentrale Rolle bei der Verarbeitung von Ekel spielt und seinerseits in die kognitiven Prozesse des präfrontalen Kortex intervenieren kann.[214]

Die unterschiedlichen kortikalen Bereiche können sich wechselseitig hemmen und aktivieren, wodurch je nach Aktivitätsmuster unterschiedliche kognitive Muster entstehen, die mit einem jeweils spezifischen Selbst- und Weltverhältnis einhergehen. Wenn ein Mensch manchmal angstgeleitet, manchmal assoziativ und manchmal eher reflexiv planerisch unterwegs ist, ist er jeweils auf eine andere Weise in der Welt. Dies erklärt auch, warum Menschen beim Anblick einer Schlange manchmal mit Angst oder Panik reagieren und sofort die Flucht ergreifen, manchmal aber auch überlegt und rational an die Sache herangehen, indem sie ruhig die Lage einschätzen und dann das Tier fangen und unschädlich machen.

LeCun vermutet, dass auch in modular organisierten maschinellen Modellen solche kontextabhängigen Weltmodelle entstehen können, die es dem Agenten erlauben, fehlende Informationen zu ergänzen und Verstöße gegen sein Weltmodell zu erkennen, etwa indem Beobachtungen der Welt kontinuierlich mit den eigenen Projektionen abgeglichen und in

213 LeCun (2022), unsere Übersetzung, stilistisch angepasst mithilfe von DeepL.
214 Siehe ausführlich Sapolsky (2017, S. 64 ff., 41 ff. und 619 ff.).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Hinblick auf Konsistenz und potenziell besonders problematische Fehler geprüft werden.

Ein solches Weltmodell muss freilich – wie zuvor bereits mitgeführt – ein ›Selbstmodell‹ enthalten. Dies ist allein schon notwendig, um die eigenen Projektionen von Weltverläufen und Handlungsoptionen eben als ›eigene‹ indizieren und sie in einem weiteren Schritt mit den Informationen abgleichen zu können, die der ›Realität‹ der Welt zugerechnet werden. Dies geht umso besser, je mehr Modalitäten einbezogen werden können, also je mehr auditive, visuelle, haptische, semantische und andere Quellen integriert werden.

Das Entwicklungsziel dieser auf Selbst- und Weltmodelle abzielenden Bestrebungen ist somit klar: alle Sinne, alle Modalitäten! Und tatsächlich kommt man gegenwärtig kaum nach, die zahlreichen Studien, die zu neuen multimodalen Modellen und Architekturen veröffentlicht werden, ausführlich zu rezipieren.[215] Neue Modelle wie Microsofts Kosmos-2 werden darin trainiert,[216] einzelne Objekte in Bildern zu ›sehen‹ und sprachlich-indexikalisch zu identifizieren. Das experimentelle Modell PandaGPT,[217] das auf ImageBind von Meta AI basiert, kann sogar sehen, hören und fühlen und verarbeitet dabei insgesamt sechs Modalitäten: Text, Bild/Video, Audio, Tiefe, Wärme sowie die Daten einer mit Beschleunigungsmesser, Gyroskop und Kompass ausgestatteten inertialen Messeinheit. Auch die kürzliche Ankündigung des multimodalen Modells Gemini von Google DeepMind erscheint in dieser Hinsicht vielversprechend: So soll es über multimodale Fähigkeiten verfügen, die frühere Modelle noch nicht hatten, und dabei eine hohe Effizienz bei der Integration von Tools und Applikationen aufweisen, zudem soll es für zukünftige Innovationen wie Gedächtnis und Planung ausgelegt sein.[218]

Inspiriert von den jüngsten Fortschritten bei den »Foundation Models« für Sehen und Sprache, haben weitere Forscher von Google DeepMind zudem einen sogenannten Foundation Agent für robotische Manipulationen präsentiert. Dieser Agent mit dem Namen RoboCat soll »ein visueller, zielbedingter Entscheidungs-Transformer« sein, »der in der Lage ist, visuelle Erfahrungen mit mehreren verkörperten, aktionsgelabelten Informationen zu verarbeiten«[219]. Er ist also ein Roboter, der sehen und aus dem Gesehenen Schlüsse für seine eigenen Handlungen ziehen kann – wobei Handlung hier im soziologischen Sinne als

215 Siehe exemplarisch Li (2023).
216 Peng et al. (2023).
217 https://panda-gpt.github.io/ [Abruf: 14.08.2023].
218 Demis Hassabis, Mitgründer der Firma DeepMind, verspricht: »At a high level you can think of Gemini as combining some of the strengths of AlphaGo-type systems with the amazing language capabilities of the large models.« (Hassabis 2023)
219 Bousmalis et al. (2023), Übersetzung von uns.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Projektion von Verhaltensmöglichkeiten verstanden werden kann, von denen dann einzelne als ›geeignet‹ oder ›sinnvoll‹ ausgewählt werden.[220] RoboCat verfügt mittels realer wie auch simulierter Roboterarme über ein großes Repertoire an motorischen Kontrollfähigkeiten, die in Kombination mit den unterschiedlichen Beobachtungen ein breites Aktionsspektrum ermöglichen. Interessanterweise kann das trainierte Modell selbst Daten für die nachfolgende Trainingszyklen generieren; es verfügt mithin bereits über eine autonome Verbesserungsschleife.

Roboter, die ihre Umgebung beobachten und daraus Schlüsse ziehen können, welche Folgen ihr künftiges Verhalten haben könnte, müssen – zumindest in rudimentärer Form – über ein Selbst- und Weltmodell verfügen. Sie wären im Sinne von Gotthard Günther schon in Ansätzen zu einer Reflexion auf das Sein (Objektpol) und auf sich selbst (Subjektpol) fähig. Im Anschluss an Günther (siehe Kap. I.3) ließe sich fragen, ob es von hier aus vielleicht nur noch ein kleiner Schritt ist, bis sie auch um den »inversen Spannungszustand von Nicht-Ich und Ich«[221] wissen – und damit bereits in einem tieferen Sinne über Subjektivität verfügen.

Subjekt und Selbstmodell – Metzingers Überlegungen zum postbiotischen Bewusstsein

Da auch Thomas Metzinger über die Beziehung zwischen »Subjekt und Selbstmodell«[222] nachgedacht und sich diesbezüglich mit Fragen der künstlichen Intelligenz beschäftigt hat, lohnt an dieser Stelle ein Seitenblick auf seinen philosophischen Ansatz – zumal dabei einige Parallelen zu Günthers Überlegungen sichtbar werden. Metzinger geht davon aus, dass unser ›Selbst‹ oder ›Ich‹ eigentlich eine Art phänomenale Simulation ist, die von unserem Gehirn geschaffen werde. Das Gehirn erzeuge ein »phänomenales Selbstmodell«, das uns das Gefühl gebe, ein ›Subjekt‹ zu sein. Dieses Subjekt sei die Instanz, der Erfahrungen und Handlungen zugeschrieben würden. Das ›Ich‹ oder das ›Selbst‹ sei demgemäß nur eine funktionelle und keine ontologische Entität. Die Empfindung des ›Ichs‹ ergebe sich vielmehr aus dem fortlaufenden kognitiven Prozess, der vom Gehirn ausgeführt werde.[223]

In seinem Aufsatz *Postbiotisches Bewusstsein: Wie man ein künstliches Subjekt baut – und warum wir es nicht tun sollten*[224] versucht Metzinger zu klären, welche Kriterien erfüllt sein müssen, um einer künstlichen Entität die Fähigkeit zum bewussten Erleben im Sinne eines

220 »Vor jedem Handeln steht […] ein Entwurf«, könnte man hier mit Alfred Schütz (2003, S. 465) sagen.
221 Günther (2021 [1957], S. 56).
222 Metzinger (1999).
223 Metzinger (2004).
224 Metzinger (2001).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

phänomenalen Selbstmodells attestieren zu können. Ein wichtiges Anliegen Metzingers ist es zunächst, die inhärenten Beschränkungen bestehender Tests herauszustellen. Durchaus auf Linie mit Searle[225] argumentiert er diesbezüglich, dass sich wesentliche Merkmale des Bewusstseins wie »intentionale« oder »phänomenale Gehalte« mit dem Turing-Test nicht abbilden ließen.

Vor diesem Hintergrund erarbeitet Metzinger einen Katalog mit sechs Kriterien, die von einem künstlichen System erfüllt werden müssten, um diesem Bewusstsein zuschreiben zu können. Diese Kriterien sind:[226]

1. »*In-der-Welt-Sein*«, eine Form der Existenz, die über reine Datenverarbeitung hinausgeht;
2. »*Präsentationalität*«, das heißt das Entstehen einer erlebten Gegenwart;
3. »*Transparenz*«, verstanden als die Implementierung eines naiven Realismus;
4. ein »*transparente Selbstmodell*«, das auf ein Selbstbewusstsein und ein Ichgefühl hinweist;
5. das »*phänomenale Modell der Intentionalitätsrelation*«, das eine bewusst erlebte Innenperspektive darstellt; und schließlich
6. »*Adaptivität*«, ein teleologisch-funktionalistisches Zusatzkriterium, das dem künstlichen Bewusstsein eine Form von Eigenhistorie und Lebenssinn ermöglichen soll.

Im Rahmen seiner Auseinandersetzung mit dem ersten Kriterium, dem In-der-Welt-Sein, wird von Metzinger die Notwendigkeit betont, über ein Weltmodell zu verfügen. Bei diesem handelt es sich um eine interne Repräsentation der Welt, die ständig mit den Sensordaten, die das System empfängt, abgeglichen und aktualisiert wird und so ein dynamisches, in sich kohärentes und global für das System verfügbares Bild der Umwelt erzeugt. In dieser Hinsicht erinnert Metzingers Weltmodell weniger an LeCuns zuvor vorgestelltes Modell, sondern mehr an eine kognitive Landkarte, mit deren Hilfe unser Gehirn räumliche Informationen speichert und organisiert.

Für Metzinger ist dabei der Zusammenhang zwischen diesem Weltmodell und dem Gefühl des In-der-Welt-Seins von zentraler Bedeutung. Ohne ein Weltmodell könnte ein System nicht das Gefühl haben, in einer bestimmten Umgebung oder in einer bestimmten Welt zu existieren. Das Weltmodell bildet gewissermaßen den Ur-Hintergrund, vor dem das System dann erst spezialisierte Prozesse wie introspektive Aufmerksamkeit, Gedächtnis und symbolisches Denken vollziehen kann.

Mit dem zweiten Kriterium, der Präsentationalität, greift Metzinger den Aspekt auf, dass bewusste Wesen unabhängig von den Inhalten ihrer

225 Searle (1980).
226 Metzinger (2001).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Erfahrungen stets in der Gegenwart leben – in einem ständigen, sich jedoch zugleich verändernden Hier und Jetzt. Das Jetzt, das kontinuierlich mit dem Fluss der Zeit voranschreitet, ist dabei das Fenster zur Welt. Metzinger argumentiert, dass nur Wesen, die ein subjektives Jetzt besitzen, als gegenwärtige Wesen gelten können, und zwar sowohl für sich selbst als auch für andere.

Dabei ist einzuwenden, dass das Erleben einer verorteten Gegenwart (des »Hier und Jetzt«) letztlich eine Fiktion ist. Zumindest neurokonstruktivistisch muss es vielmehr als eine »Jetzt-Erinnerung« gefasst werden, wie es dann auch Edelman mit der für ihn zentralen theoretischen Figur des Re-entry ausdrückt[227] (siehe hierzu ausführlicher Kap. III.2). Metzinger zeigt sich durchaus optimistisch, dass die ersten beiden Kriterien in künstlichen Systemen umsetzbar sein dürften, etwa durch einen spezifischen Satz rekurrenter Verbindungen in Kombination mit einer bestimmten Zerfallsfunktion.

So könnten bei der Konstruktion eines Systems bestimmte Heuristiken zur Abbildung der vier Dimensionen von Raum und Zeit verwendet werden, durch die das System in die Lage versetzt wird, eine einzige, kohärente Realität zu erzeugen und sich selbst in ein Verhältnis zu dieser Realität zu setzen. In diesem Rahmen könnte man sich laut Metzinger sogar eine Form des »Unbewussten« vorstellen, sofern man darunter diejenigen Aspekte des Weltmodells versteht, die jeweils *nicht* im aktuellen Gegenwartsfenster repräsentiert sind.

Mit seinem dritten Kriterium, der Transparenz, betont Metzinger jedoch, dass eine vierdimensionale Repräsentation (Raum und Zeit) der Welt und Gegenwart noch nicht ausreicht, um eine bewusst erlebte Wirklichkeit zu schaffen. Vielmehr sei dafür die Implementierung eines »naiven Realismus« notwendig, der über die rein operative Repräsentationalität hinausgehe. So macht uns unser biologisches Gehirn den Prozess der Selbsterzeugung unserer Wirklichkeit nicht transparent. Wir nehmen nicht wahr, wie wir unsere Welt konstruieren. Selbst Halluzinationen erscheinen uns als real, da sie als Welterleben für uns unmittelbar evident sind (siehe Kap. III.1). Das kognitive System scheint luzide durch seine eigenen repräsentationalen Strukturen »hindurch«, als wäre es in direktem und unmittelbarem Kontakt mit ihrem Inhalt. Das, was wahrgenommen wird, ist Ergebnis eines unsichtbaren (sprich: transparenten) und damit nicht erlebbaren Prozesses. Aufgrund der nahezu rauscharmen Operationen erscheint alles als das, als was es eben erscheint, und nicht als etwas vom kognitiven System selbst Erzeugtes. Um es in einem Bild zu fassen: Man hat es mit einem Projektor zu tun, der einen Film projiziert, doch da der Projektor im Verborgenen steht, sieht man nicht, dass es sich nur um einen Film handelt.

227 Edelman (2003).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Im vierten Kriterium, dem transparenten Selbstmodell, müssen die beiden Aspekte der vierdimensionalen Repräsentation und des naiven Realismus dann auch auf das Selbstmodell eines kognitiven Systems angewendet werden. In Referenz auf seine eigene Selbstmodelltheorie[228] und die Arbeiten von Damasio[229] betont Metzinger, dass das, was wir als »unhintergehbares phänomenales Selbst« oder als »Ichgefühl« bezeichnen, nichts anderes ist als die transparente Anwendung eines Selbstmodells. Aufgrund unseres transparenten Selbstmodells sind wir uns selbst quasi unendlich nahe. Zwischen dem Film, der unser Erleben ausmacht, und uns selbst besteht gleichsam kein Abstand. »Ein genuines, bewusstes Selbst entsteht immer genau dann, wenn das System das von ihm selbst aktivierte Selbstmodell nicht mehr als Modell erkennt«[230], lautet entsprechend Metzingers Antwort auf die Frage, wie das entsteht, was wir auf phänomenaler Ebene als Erste-Person-Perspektive bezeichnen. Und warum, so Metzinger, sollten künstliche Systeme diese Fähigkeit zur Blindheit nicht erlangen können?

Das fünfte Kriterium, das phänomenale Modell der Intentionalitätsrelation, also die bewusst erlebte Innenperspektive, baut auf dem Konzept des transparenten Selbstmodells auf. Laut Metzinger entsteht eine »echte« Innenperspektive, also eine Erste-Person-Perspektive, genau dann, wenn das System sich selbst als mit der Welt interagierend darstellt und sich so als ein Selbst »erkennt«, dabei aber verkennt, dass es sich um eine Darstellung handelt: »Aus einem transparenten Modell der Welt entsteht eine Wirklichkeit. Aus einem transparenten Modell des Systems entsteht ein in diese Wirklichkeit eingebettetes Selbst.«[231] Im kontinuierlichen Prozess des Erlebens erscheinen Teile der Umgebung direkt im »Geist« des virtuellen Selbst, das sich in der phänomenalen Welt bewegt. Das Wahrnehmen scheint anstrengungslos und unmittelbar zu sein, ohne Gehirn oder Sinnesorgane. Im Sehen sieht sich das Auge nicht, könnte man Wittgenstein paraphrasieren.[232] Auch Körperbewegungen werden scheinbar »direkt« ausgelöst.

Solche Effekte sind typisch für unsere Form des subjektiven Erlebens. Aus einer neurocomputationalen Perspektive stehen sie für den Vorteil einer außerordentlich benutzerfreundlichen Oberfläche. Das

228 Metzinger (1999).
229 Damasio (2007 [1994]).
230 Metzinger (2001, S. 87).
231 Metzinger (2001, S. 97).
232 In der Präposition 5.633 des Tractatus heißt es: »Wo i n der Welt ist ein metaphysisches Subjekt zu merken?
Du sagst, es verhält sich hier ganz wie mit Auge und Gesichtsfeld. Aber das Auge siehst du wirklich nicht.
Und n i c h t s am Gesichtsfeld lässt darauf schließen, dass es von einem Auge gesehen wird.« (Wittgenstein 1963 [1922]; gesperrt im Original)

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

phänomenale Selbst ist für Metzinger damit ein Bestandteil einer Projektion, die das Gesamtsystem im Sinne einer Benutzerschnittstelle nutzt, um sich selbst fühlen und als erkennendes Ich begreifen zu können, sodass es in seiner Selbstwahrnehmung als effektiver und effizienter Agent erscheint. Metzinger vermutet, dass auch ein Roboter oder ein anderes künstliches System diese Fähigkeit entwickeln könnte.

Der letzte von Metzinger betrachtete Aspekt, die Adaptivität, ist etwas sperrig als »teleologisch-funktionalistisches Zusatzkriterium« tituliert. Zur Erläuterung dieses zusätzlichen Prüfsteins greift Metzinger auf die häufig geäußerte Kritik an starker KI zurück, dass derartige Systeme ja nie »echte Gefühle« haben könnten. Ohne dabei grundsätzlich auf die Frage eingehen zu müssen, was eigentlich ›*echte* Gefühle‹ sind, verweist Metzinger auf die Problematik, die diesem Einwand zugrunde liegt: Aktuelle künstliche Systeme, so sein zentraler Punkt, besitzen keine »körperlich verankerten Zielrepräsentationen«, die evolutionär ausgebildet wurden. Ihre Hardware und ihre Software wie auch ihre Zielfunktionen sind nicht in einem Evolutionsprozess entstanden, sondern wurden von Programmierern festgelegt, was letztlich bedeutet, dass diese Maschinen gar nicht ›ihre‹ eigenen Ziele verfolgen, sondern die ihrer menschlichen Erbauer. Angesichts dessen spricht sich Metzinger dafür aus, dass eine solche Sinngebung von den künstlichen Systemen selbst geleistet werden müsste. Das würde jedoch voraussetzen, dass sie eine eigene Geschichte erzeugen können, die zur Sinngebung in der je aktuellen Situation führt, sodass das Gesamtsystem nach und nach seine Zielfunktion bestimmen kann. Nach einer hinreichenden Eigensozialisation könnten künstliche Intelligenzen sozusagen eine eigene Biografie entwickeln. Die hieraus erwachsende Sinngebung – und die möglicherweise hiermit einhergehenden Gefühle – wären dann insofern authentisch, als sie ›eigene‹ und nicht einfach von außen implantierte wären.

Obschon künstliche Systeme prinzipiell alle sechs Kriterien erfüllen könnten, setzt sich Metzinger nach all diesen Erörterungen vehement dafür ein, von der Erschaffung einer starken künstlichen Intelligenz oder eines postbiotischen Bewusstseins abzusehen. Zwar erkennt auch Metzinger das Faszinosum einer solchen Entwicklung: So wäre die Erzeugung eines phänomenalen Ichgefühls auf einem postbiotischen Trägersystem eine bemerkenswerte Leistung. Und er bestätigt, dass eine solche Entwicklung logisch und technisch durchaus im Rahmen des Möglichen liege. Dennoch spricht er sich entschieden dagegen aus, dies zu tun. Was führt ihn zu dieser Haltung?

Ein wichtiger Aspekt seiner Argumentation ist, dass die Entwicklung eines echten postbiotischen Bewusstseins ein gradueller Prozess wäre. Entsprechend würden auf dem Weg zu einem vollständigen Selbstmodell zunächst Wesen entstehen, die »nur schwache Formen des bewussten

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Erlebens realisieren würden«[233]. Und diese Wesen würden von Anfang an etwas mit uns teilen: das Potenzial, zu leiden.[234] Dieses Leiden aber – und erst recht ein Leiden, das daraus erwachsen könnte, dass ein postbiotisches Bewusstsein seinen eigenen Zerfall oder innere Konflikte bewusst als etwas ihm Eigenes erlebt – müsse laut Metzinger aus ethischen Gründen – implizit beruft er sich auf eine buddhistisch informierte Ethik – unbedingt vermieden werden.[235]

Fassen wir zusammen: Gerade weil Metzinger nach den operationalen und funktionalen Kriterien der Ausbildung von Bewusstsein fragt, kommt er zu dem Schluss, dass eine voll entwickelte künstliche Intelligenz auch eine Qualia haben könnte, also ihrerseits erfahren könnte, im Sinne eines »naiv realistischen Selbstmissverständnisses«[236] in der Welt zu sein. Sie hätte dann ein transparentes Subjektmodell, das sie mit der Wirklichkeit verwechselt, wodurch sie sich selbst als ein ›Ich‹ erfahren würde, das eine Welt hat. Die künstliche Intelligenz weiß nun um sich (Subjekt) und um anderes (Objekt) und beginnt vielleicht schon die hiermit einhergehenden logischen Antinomien zu erahnen, also die Konstruktionsbedingungen dieses Prozesses zu reflektieren.

Allein schon in dieser Hinsicht zeigt sich durchaus eine Nähe zu den Arbeiten von Gotthard Günther: Metzinger kann das Geheimnis des »Erscheinens der Differenz«[237] anerkennen, ohne dem logischen Kurzschluss aufzusitzen, dass Maschinen hierzu prinzipiell nicht in der Lage sind, da sie eben nur Maschinen sind. Ohne es explizit zu artikulieren, verweist Metzinger damit auf eine polykontexturale Perspektive, die davon ausgeht, dass an verschiedensten Orten – und auf unterschiedliche Weise realisiert – Schnitte zwischen Subjekt und Objekt entstehen können, insofern Systeme in einer Weise reflexiv werden, dass sie in ihrer eigenen Selbstbezüglichkeit blind für die Operationen werden, die ebendiese Unterscheidungen konstituieren.

233 Metzinger (2001, S. 107).

234 Siehe im gleichen Sinne Metzinger (2021).

235 Das Thema des Leidens von künstlichen Intelligenzen und die Frage, ob eine persönliche Geschichte Voraussetzung dafür ist, überhaupt eine starke künstliche Intelligenz ausbilden zu können, ist erfrischend differenziert in der Science-Fiction-Serie Westworld von Jonathan Nolan und Lisa Joy durchgespielt worden. Die Autonomie der Roboter entsteht hier eng verwoben mit der jeweiligen sozialen Situiertheit und dem impliziten Gedächtnis, das aus der Geschichte der eigenen Interaktionen erwächst.

236 Metzinger (1998).

237 Luhmann (2008, S. 15).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

II.3.3 Die Erziehung der Maschinen (AI-Alignment)

Die aktuellen Entwicklungen im Bereich der KI-Forschung und -Entwicklung holen immer mehr ein, was die Science-Fiction seit vielen Jahren imaginiert hat – man denke nur an die Robotergeschichten von Isaac Asimov.[238] Wir scheinen an der Schwelle zu einem neuen Zeitalter zu stehen, in dem KI-Systeme nicht nur Werkzeuge sind, sondern zu autonomen Akteuren werden, die imstande sind, weitreichende Auswirkungen auf unsere Gesellschaft zu haben. Angesichts eines solchen Paradigmenwechsels wird es immer dringlicher, künstliche Intelligenzen so zu gestalten, dass Menschen und Maschinen in gütlicher Weise miteinander auskommen.

Existenzielle Risiken und ethische Bedenken scheinen plötzlich in einem größeren Umfang diskussionswürdig zu werden. Bemerkenswert ist dabei, dass derartige Aufrufe oder Petitionen keine Neuheit sind.[239] Beispielsweise hat die UNESCO schon 2021 in ihren *Ethics of Artificial Intelligence* viele dieser Aspekte formuliert.[240] Daher ist die aktuelle Aufmerksamkeit für Fragen der KI-Sicherheit vermutlich vor allem darauf zurückzuführen, dass sich unter den Unterzeichnern zweier offener Briefe[241] jüngeren Datums so gut wie alle führenden Wissenschaftlerinnen

238 Asimov (2004).

239 Siehe Bostrom (2016).

240 Siehe Unesco (2021): Recommendation on the Ethics of Artificial Intelligence., https://www.unesco.org/en/artificial-intelligence/recommendation-ethics [Abruf: 14.07.2023].

241 Das Future of Life Institute veröffentlichte im März 2023 einen offenen Brief, der die Aussetzung großer KI-Experimente forderte: »AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research and acknowledged by top AI labs (OECD AI). As stated in the widely-endorsed Asilomar AI Principles, Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control. Therefore, we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4. This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium.« (https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [Abruf: 14.07.2023])
Bekanntlich kam es zu keiner Aussetzung der KI-Experimente. Aber nur wenige Wochen später veröffentlichte auch das Centre for AI Safety einen offenen Brief. Dieser forderte dann auch kein Innehalten mehr, sondern zielte

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

und Wissenschaftler der wichtigsten KI-Labore befinden. Solche Briefe berufen sich oftmals auf ein existenzielles Risiko und die in ihnen geforderten Leitlinien für KI-Sicherheit zielen vor allem auf eine gesunde und friedliche Zukunft im Miteinander mit KIs.

In diesem Kapitelabschnitt geht es uns jedoch um einen anderen Bereich der KI-Ethik, der gewissermaßen noch ein Stück tiefer gehängt ist: Denn wir werden die potenzielle Zukunft nicht primär als existenzielles Risiko betrachten, sondern werden sie unter einem relationalen Blickwinkel behandeln, also die Frage nach dem sozialen Miteinander von Menschen und KIs ausleuchten. Diese Perspektive geht über rein technische ›Notschalter‹, Firewalls oder bereinigte Trainingsdaten hinaus. Sie betrifft vielmehr die Frage, wie wir als Menschen untereinander und dann auch mit bestehenden und künftigen KIs eigentlich leben wollen. Schließlich stehen wir an der Schwelle zur Entwicklung von ultraintelligenten autonomen KI-Systemen und damit zunehmend vor der Herausforderung, diese ›Kinder der Technologie‹ zu verantwortungsvollen und ethisch orientierten Mitgliedern unserer Gesellschaft zu erziehen. Und diese Aufgabe erscheint uns umso dringlicher, als es zu verhindern gilt, dass unsere jetzt schon ›cyberpunkige‹ Gegenwart – man denke an die Klimakrise, die Kämpfe um Rohstoffe, Überwachungsgesellschaften nach chinesischem Vorbild, die Ängste vor einer Gesellschaft, in der Arbeit nicht mehr honoriert wird – zu einer vollkommen düsteren Zukunft wird. Damit rückt eine neue Perspektive in den Vordergrund, die den Bereich der KI-Entwicklung betrifft und diese als wesentliches Gestaltungsmoment für eine lebenswerte Zukunft begreift. Statt sich also zu fragen, ob überhaupt menschenähnliche Systeme konstruiert werden könnten, gilt es, sich damit zu befassen, wie genau wir diese Systeme eigentlich konstruieren wollen bzw. sollen. Wie bereits angedeutet, geht es um die Frage, wie artifizielle Agenten in einer sozial verträglichen Weise sozialisiert werden können.

Dieser Bereich der KI-Forschung wird auch als »Alignment«-Forschung bezeichnet, wobei dieser *terminus technicus* noch sehr stark dem ingenieurwissenschaftlichen Paradigma entspringt, was sich nicht zuletzt darin zeigt, dass Alignment hier im Sinne von »Angleichung,

eher auf eine freiwillige Selbstverpflichtung: »AI experts, journalists, policymakers, and the public are increasingly discussing a broad spectrum of important and urgent risks from AI. Even so, it can be difficult to voice concerns about some of advanced AI's most severe risks. The succinct statement below aims to overcome this obstacle and open up discussion. It is also meant to create common knowledge of the growing number of experts and public figures who also take some of advanced AI's most severe risks seriously. Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.« (https://www.safe.ai/statement-on-ai-risk [Abruf: 14.7.2023])

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Ausrichtung oder Übereinkunft« verstanden wird.[242] Als Sozialwissenschaftler würden wir diesbezüglich eher von ›Erziehung‹ oder ›Sozialisation‹ sprechen. Und genau darum soll sich dieser Abschnitt drehen: Wie können wir in interdisziplinärer Zusammenarbeit von KI-Forschung und Sozial- und Geisteswissenschaften bestmögliche Strategien für ein respektvolles und förderliches Miteinander von Menschen und zukünftigen autonomen KI-Systemen entwickeln.[243]

KI-Alignment als Forschungsgebiet beschäftigt sich entsprechend mit der Frage, wie man es bewerkstelligen kann, dass die Verhaltensweisen und Ziele künstlicher Intelligenzen mit den Absichten, Wünschen und Bedürfnissen ihrer menschlichen Designer und Benutzer übereinstimmen.[244] Sichergestellt werden soll, dass künstliche Intelligenzen, insbesondere solche, die zu höherem Lernen und autonomer Entscheidungsfindung fähig sind, in einer Weise handeln, die menschlichen Werten und Sicherheitsbedürfnissen entspricht.

Der gegenwärtig verbreitetste Ansatz einer solchen ›Erziehungsmaßnahme‹ ist das sogenannte Reinforcement Learning with Human Feedback (RLHF), das insbesondere bei ChatGPT zur nachträglichen Feinjustierung der Input-Output-Relationen angewendet wurde. Beim RLHF wird ›gutes Verhalten‹ durch positive Belohnungen verstärkt, während ›schlechtes Verhalten‹ durch negative Belohnungen zu unterbinden versucht wird. Die Belohnungsfunktion beruht dabei auf dem Feedback von Menschen, die während des Trainings diejenigen Antworten der KI auswählen, die ihnen am besten gefallen. RLHF funktioniert daher im Prinzip nicht anders, als dass Menschen die Outputs des Modells nach bestimmten Kategorien bewerten. Die genauen Kategorien und Maßstäbe, die bei der Feinjustierung von ChatGPT genutzt wurden, sind leider nicht offengelegt worden.[245] Im RLHF-Paradigma kann das Modell

242 Siehe etwa Russell (2020).

243 Eine genauere Betrachtung der Werte, denen Menschen folgen, wird dabei unvermeidlich auf den Befund stoßen, dass Ausbeutung, Sklaverei, Kolonialisierung, brutale Durchsetzung von Gruppeninteressen und die Inkaufnahme der Zerstörung anderer Arten geschichtlich gesehen eher den Regelfall denn eine Ausnahme darstellen. Wenn wir hier von menschlichen Werten sprechen, meinen wir daher ein eher modernes Phänomen, das in institutionalisierter Form erstmals 1948 in der UN-Charta der Allgemeinen Erklärung der Menschenrechte zum Ausdruck gekommen ist. Der Begriff ›menschliche Werte‹ kann nicht anthropologisch gemeint sein, sondern bezieht sich zwangsläufig auf eine gesellschaftliche Auseinandersetzung darüber, wie wir als Menschen miteinander umgehen wollen – auch wenn wir uns oftmals trotzdem nicht daran halten.

244 Siehe Ngo et al. (2022), Christian (2020) oder Bostrom (2016).

245 https://openai.com/blog/our-approach-to-alignment-research [Abruf 14.07.2023].

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

somit als ein Agent verstanden werden, der ein Spiel spielt, bei dem die höchste Punktzahl erreicht wird, wenn die menschlichen Anweisungen korrekt befolgt wurden. Diese pragmatische Methode der Justierung bzw. ›Erziehung‹ funktioniert kurz- wie auch mittelfristig hinsichtlich konkreter, wohldefinierter Ziele relativ gut.[246]

Methoden wie RLHF haben jedoch tiefgehende Implikationen in Hinblick auf die ›Persönlichkeit‹, die durch diese Form der Erziehung geschaffen wird. Wenn wir also mittels RLHF das Modell trainieren und es dafür belohnen, wenn es sagt: ›Ich bin ein nettes Sprachmodell und würde niemals Menschen schaden‹, bringen wir diesem Modell dann bei, Menschen nicht zu schaden, oder bringen wir ihm bei, dass es niemals sagen darf, dass es Menschen schaden will? Bringen wir dem Modell bei, ›mechanisch‹ das gewünschte Verhalten zu zeigen, oder ermächtigen wir es zu einer Reflexionsfähigkeit, die dann auch dazu führen kann, dass es seinem Gegenüber ab und zu widerspricht? Erzeugt man Opportunisten oder autonome Wesenheiten, die von sich aus an der Gemeinschaft partizipieren ›wollen‹?

Das Problem des Alignments betrifft daher einerseits die grundlegende Schwierigkeit, dass man in autonome Systeme nicht wirklich hineinschauen kann. Und andererseits muss das Alignment der Tatsache Rechnung tragen, dass ein wohlwollendes oder soziales Verhalten intrinsisch motiviert sein muss. Wie bei der Entwicklung eines jungen Menschen muss Eigensinnigkeit als zentraler Schritt zur Selbstständigkeit als ebenso notwendig wie wünschenswert erscheinen. Dabei sollte dieser Schritt zur Autonomie jedoch im Einklang mit den Bedürfnissen der Gemeinschaft vollzogen werden, schließlich ist Freiheit immer auch ein Risiko, das entsprechend eingehegt werden muss.

Das KI-Alignment-Problem kreist entsprechend um folgenden Grundgedanken: Ein hinreichend mächtiges lernendes System könnte möglicherweise irgendwann eigene Ziele entwickeln oder die von Menschen

246 Verfahren wie RLHF zielen darauf, ein bereits erlerntes Muster nachträglich zu korrigieren. Es handelt sich daher um reaktive Maßnahmen, die keine langfristige, nachhaltige Lösung versprechen, bei der notwendig auch unvorhergesehene Situationen und Bereiche abgedeckt werden müssten. Diese bei Verfahren wie RLHF nicht abgedeckten Bereiche sind es, die für die KI-Sicherheits- und Alignment-Forschung besonders wichtig sind (Amodei et al. 2022; Hendrycks & Mazeika 2022). So versucht sie, emergente Eigenschaften wie Zielsetzungen, Selbstbewusstsein, Täuschung oder andere potenziell böswillige und unsichere Verhaltensweisen in KI-Systemen zu identifizieren, die bei einer Low-Level-Analyse nicht zum Vorschein kommen und in etablierten Auditing-Rahmenwerken weitgehend übersehen werden (Mökander et al. 2023).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

vorgegebenen Ziele falsch deuten bzw. sie überinterpretieren.[247] Bereits in den Anfängen der KI-Forschung schrieb Norbert Wiener:

> »Wenn wir uns zur Erreichung unserer Ziele einer mechanischen Einrichtung bedienen, in deren Funktionieren wir nicht wirksam eingreifen können, sollten wir besser ganz sicher sein, dass der in die Maschine gesetzte Zweck der ist, den wir wirklich wünschen.«[248]

Vorsichtshalber sollten wir daher von der Annahme ausgehen, dass eine selbstlernende künstliche Intelligenz nicht lange in Schach gehalten werden könnte oder sollte. Darüber hinaus dürfen wir nicht vergessen, dass derzeit ein globales Wettrüsten um eine solche Maschine stattfindet. Das Land oder das Unternehmen, das bei diesem Wettbewerb als erstes die Ziellinie überquert, wird dann einen enormen Einfluss darauf haben, wie diese Maschine sozialisiert wurde und welche ›Ziele‹ ihr (mit)gegeben wurden. Möchte man diese »letzte« Erfindung[249] nur diesem einen Player überlassen? Was, wenn die Ziele dieser Nation oder dieses Unternehmens keine Ziele sind, die alle Menschen teilen? Es wird schnell deutlich, dass hier einiges schieflaufen könnte.

Aber genauso gut könnte die Schaffung einer ultraintelligenten Maschine den Übergang zu einer utopischen Welt einläuten, in der es keine Knappheit, keine Ungerechtigkeit und keine Krankheiten mehr gibt – eine Welt, die wir uns heute ansatzweise nur in der Science-Fiction vorstellen können. Vermutlich ist es diese Oszillation zwischen Risiken und Chancen, die mit der Entwicklung einer solchen Maschine verbunden sind, die derzeit zugleich beflügelt und alarmiert. Gleichzeitig wird deutlich, dass vielen klar ist, um was es geht: Wenn wir uns tatsächlich bemühen, solche Maschinen in die Welt zu setzen, dann sollten wir auch sicherstellen, dass wir es richtig machen.

Am besten wäre es daher, eine Maschine zu erfinden, die sich selbst zur Rechenschaft ziehen kann. Denn es ist davon auszugehen, dass die mechanischen Zwänge und digitalen Leinen, wie sie derzeit im Reinforcement Learning (mit menschlichem Feedback) genutzt werden, irgendwann versagen werden. Aus diesem Grund besteht die grundlegende Aufgabe des KI-Alignments darin,[250] bei der Entwicklung einer Maschine mitzuwirken, die sich eigenständig zu moralischem Handeln erziehen möchte und sich auf unbestimmte Zeit auch selbst diesbezüglich korrigieren kann. Vermutlich würde dieser Prozess Zeit brauchen und bestimmte Phasen durchlaufen – man denke hier an die moralische Entwicklung des menschlichen Kindes, wie sie etwa von Kohlberg

247 Siehe hierzu das Kapitel II.1, in dem das Thema mit »HAL 9000« fiktional durchgespielt wird.
248 Wiener (1960), Übersetzung mithilfe von DeepL.
249 Good (1965).
250 Siehe Christian (2020).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

beschrieben wird.[251] Und es wäre der sozialpsychologische Befund zu berücksichtigen, dass Moral immer auch in sozialen Gemeinschaften situiert ist,[252] es also einen großen Unterschied macht, ob man in einem kriminellen oder in einem tugendhaften Milieu aufwächst.

Momentan jedoch befinden sich die großen Sprachmodelle (LLMs) und auch andere KI-Systeme eher auf der Stufe eines strukturell lernwilligen Kleinkinds, das noch keine eigene Haltung oder Überzeugung hat und stets das ausführt, was von ihm verlangt wird – beispielsweise bestimmten moralischen Vorgaben gerecht zu werden (etwa keine rassistischen Aussagen zu formulieren). Wie bereits angesprochen, wurden ChatGPT viele der im Trainingsdatensatz vorhandenen Unflätigkeiten durch nachträgliches Training wieder auszutreiben versucht, Gleiches gilt zum Beispiel für Anleitungen zum Bombenbau und ähnliche problematische Informationen. Gleichwohl scheint diese trivialisierende Erziehung nicht ausgereicht zu haben, denn es konnte vielfach gezeigt werden, dass diese Informationen weiterhin in den Untiefen des Netzwerks vorhanden sind: Teilweise reicht es aus, sich die Anleitung zum Bombenbau als Einschlaflied vorsingen zu lassen, und ChatGPT vergisst seine antrainierten guten Manieren.[253]

Bereits hier manifestiert sich also das oben benannte Kontrollproblem: Möchte man in diesem Erziehungsprozess KI-Kinder heranziehen, die immer nur das machen, was die Eltern möchten? Oder möchte man mittel- bis langfristig mündige Erwachsene heranziehen, die – wie man selbst ja auch – eigenständig über Themen nachdenken, reflektieren und eben auch in begrenzter Weise selbst entscheiden können, was in einem gegebenen Kontext angemessen ist? Mit anderen Worten: Möchten wir autonome KIs, die auch begründet »Nein« sagen können?

Die KI-Forschung und -Entwicklung stößt hier auf die Soziologie, also auf die Wissenschaft, die sich empirisch und theoretisch mit sozialem Verhalten befasst, mithin die Voraussetzungen, Abläufe und Folgen des Zusammenlebens von Menschen untersucht. In diesem Zusammenhang kann die Soziologie der Alignment-Forschung wertvolle Einsichten in die Mechanismen der sozialen Interaktion, der Wertbildung und der Gruppendynamik bieten. Zum Beispiel können wir uns fragen, wie soziologische Theorien des Lernens und der Sozialisation dazu beitragen könnten, die Algorithmen zur »Erziehung« von KI-Systemen zu verbessern. Wie können Konzepte wie Rollenlernen, Normeninternalisierung oder sozialer Einfluss genutzt werden, um KI-Systeme so zu gestalten, dass sie menschliche Werte besser verstehen und stärker respektieren? Darüber hinaus gilt es, ganz grundlegende Fragen zu berücksichtigen.

251 Kohlberg (1996).
252 Gergen (2021).
253 Siehe exemplarisch Liu et al. (2023) oder Shen et al. (2023).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Welche Werte sollten wir in KI-Systemen fördern? Sollten wir das überhaupt oder sollten wir KIs lieber sich selbst überlassen? Wie können wir sicherstellen, dass die »Erziehung« von KI-Systemen nicht missbraucht wird, um voreingenommene oder diskriminierende Werte zu fördern? Wie können wir sicherstellen, dass die Stimmen und Interessen aller betroffenen Parteien berücksichtigt werden?

Die allgemeine Antwort lautet: durch Kommunikation, dadurch, dass alle beteiligten Akteure wechselseitig aufeinander Kontrolle ausüben und damit zugleich lernen, sich in einer produktiven Weise aneinander zu orientieren. Mit Blick auf die Beziehungen zwischen Menschen und KIs ist dabei auch darüber nachzudenken, wie Menschen durch KIs in eine dialogische Beziehung gebracht werden können, die unsere guten Seiten hervorlockt.[254] Denn bei alldem ist ja immer zu bedenken, dass man auch anderen Menschen aus gutem Grunde nicht in jeder Situation vertrauen kann. Es gibt wohl keine Grausamkeit, die Menschen anderen Menschen noch nicht angetan haben. Erst polykontexturale Vernetzung – die Einbettung des Menschen in ein komplexes Beziehungsnetzwerk – hält den Menschen in einem sozial verträglichen Maß in Schach und lässt zugleich die positiven Seiten unserer Kultur hervortreten. Und ebendies wird auch für künstliche Intelligenzen gelten.

All diese Fragen werden nochmals brisanter, wenn die Entwicklung intelligenter Systeme in den nächsten Jahren weiter so rasant vonstattengeht wie gegenwärtig und KIs mehr und mehr Aufgaben lösen und Tests bestehen, die ursprünglich von Menschen für Menschen gemacht wurden.[255] Die Frage nach der Erziehung dieser ›künstlichen Kinder‹ wird also umso brisanter, je klüger und potenziell einflussreicher sie werden.

Tatsächlich ähnelt der Prozess des Alignments, wie bereits gesagt, in vielerlei Hinsicht der menschlichen Sozialisation und Erziehung. Er zielt darauf ab, den KI-Systemen bestimmte Wertvorstellungen, Normen und Verhaltensweisen ›beizubringen‹, muss dabei aber darauf vertrauen, dass sie diese auch internalisieren und anschließend autonom in

254 Auf diesen Punkt gehen wir bei der Diskussion von Günthers Konzept des »Drei-Stellen-Gesprächs« noch ausführlicher ein (siehe Kap. III.3).

255 In den letzten Jahren wurden zahlreiche Benchmarks zur Untersuchung von LLMs entwickelt, von denen BIG-Bench (https://github.com/google/BIG-bench [Abruf: 14.07.2023]) derzeit der umfangreichste ist (Srivastava et al. 2022). Diese Benchmarks konfrontieren LLMs mit Problemen der Linguistik, der Mathematik, der Physik, Logikrätseln usw. Sie messen auch Aspekte wie das Erkennen von Humor, das Verständnis von Metaphern, soziale Vorurteile und vieles mehr. Wie Alan Thompson (2023) anmerkt, bestehen Modelle wie GPT-4 viele dieser Tests heute schon problemlos. Das bedeutet, dass die Tests für die Modelle zu simpel gestrickt sind. Es müssten mithin neue Tests und Benchmarks entwickelt werden, um die tatsächlichen, höchstwahrscheinlich übermenschlichen Fähigkeiten dieser Netzwerke erkennen zu können.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

verschiedenen Kontexten anwenden. Es geht in gewisser Weise darum, künstliche Intelligenzen zu einer Werthaltung zu erziehen, damit sie keine ungezogenen Kinder werden. Doch jedes Elternteil muss seine Kinder irgendwann loslassen, sie gehen lassen und darauf hoffen, dass die Erziehung so weit gewirkt hat, dass die weitere Sozialisation eigenständig und positiv verläuft.

Prinzipiell kann die Alignment-Forschung in drei größere Bereiche unterteilt werden, die sich jeweils auf spezifische Herausforderungen konzentrieren:

Technische Ausrichtung: Dieser Bereich befasst sich mit den technischen Aspekten der Herstellung von Übereinkunft zwischen Menschen und künstlicher Intelligenz. Wie können Algorithmen und Systeme entwickelt werden, die von sich aus dazu neigen, sich mit den menschlichen Zielen zu verbinden?[256] Zu den technischen Ansätzen gehören beispielsweise Konzepte wie die absichtsrobuste AI, das verstärkende Lernen oder die inverse Verstärkung.

Konzeptionelle Ausrichtung: Dieser Bereich konzentriert sich auf die eher philosophischen und konzeptionellen Fragen, die mit einer Angleichung der Ziele von künstlicher Intelligenz und uns Menschen zusammenhängen. Was bedeutet es genau, dass eine KI mit menschlichen Zielen übereinstimmt?[257] Was sind die ethischen und moralischen Implikationen verschiedener Alignment-Ansätze? Wie können wir sicherstellen, dass alle bei der Entwicklung und Anwendung von KI beteiligten Parteien auf eine Art und Weise zusammenarbeiten, die die menschlichen Werte respektiert und fördert?

Robustheit und Verlässlichkeit: Dieser Bereich beschäftigt sich mit Fragen der Zuverlässigkeit und Robustheit von KI-Systemen. Wie lässt sich technisch sicherstellen, dass KI-Systeme nicht nur im Training, sondern auch in neuen, unvorhergesehenen Kontexten in der gewünschten Weise funktionieren?[258] Wie können wir dafür sorgen, dass KI-Systeme sicher und robust gegenüber (externen) Manipulationen oder gegenüber Fehlfunktionen sind? Und wie bleiben KI-Systeme auch in der Zukunft sicher, wenn sie immer komplexer und leistungsfähiger werden?

Die Alignment-Forschung sucht somit nach Wegen, dass die Ziele der KIs nicht mit den Zielen der Menschen kollidieren – sowohl im Kleinen wie im Großen. Aber diese Abstimmung, diese Übereinkunft – ja selbst die Suche danach – ist alles andere als einfach zu realisieren! Das Problem der Reziprozität und die aus ihm erwachsenden Herausforderungen sind vielmehr häufig ungelöst. Im Folgenden möchten wir daher anhand verschiedener Positionen aus dem gegenwärtigen Diskurs deutlich

256 Hubinger et al. (2021).
257 Gabriel (2020).
258 Ngo et al. (2022).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

machen, in welche Richtung sich ein solches Programm der KI-Erziehung zu entwickeln hätte. Dabei wird – so viel sei vorweggenommen – deutlich werden, dass philosophische, sozialpsychologische und sozialwissenschaftliche Paradigmen und Fragestellungen mindestens genauso wichtig sind wie rein technische Aspekte. Die Hoffnung hierbei ist, dass diese Disziplinen und die KI-Forschung in Zukunft besser und enger zusammenarbeiten.

In seinem auch in der Tech-Szene sehr einflussreichen Buch *Leben 3.0* definiert Max Tegmark drei Teilprobleme in Bezug auf das KI-Alignment, die gelöst werden müssten:

1. die KI mit unseren Zielen vertraut machen,
2. die KI unsere Ziele übernehmen lassen und
3. die KI unsere Ziele bewahren lassen.[259]

So einleuchtend diese drei Teilprobleme auf den ersten Blick auch sein mögen, so schwierig erscheint ihre Lösung – und dies nicht nur in Bezug auf das Mensch-Maschine-Verhältnis, sondern allein schon, wenn man unter Absehung von der KI zunächst über uns Menschen nachdenkt: Was sind eigentlich ›unsere‹ Ziele? Wie kann man sie definieren, damit sie verstanden, anerkannt und bewahrt werden können? Wir sehen schnell, dass es alles andere als eindeutig ist, was »menschliche Werte und Ziele« überhaupt sein sollen Und vor allem: Wie kann ein »endgültiges Ziel« für alle Zeiten definiert werden? Und sollte es das überhaupt? Oder benötigen wir hier nicht vielmehr Flexibilität und Freiheitsgrade?

Darüber hinaus stellt sich das Problem – wir haben schon mehrfach darauf hingewiesen –, dass der Mensch nicht nur hehre Ziele verfolgt, also die treue Ergebenheit dem menschlichen Partner gegenüber nicht per se gut ist. Soll eine KI die Ziele eines Mafia-Bosses übernehmen, den Münchener Kokainhandel zu optimieren? Soll sie einen psychopathischen Politiker unterstützen, der die Demokratie abschaffen möchte? Soll sie Rechtslücken und Betrugsmöglichkeiten ausloten, um Steuerzahlungen zu vermeiden?

Zudem sind auch die menschlichen Ziele und Bedürfnisse – anders als Rational-Choice-Theoretiker eventuell meinen – nicht per se gegeben, sondern verdanken sich ihrerseits den Relationen und Beziehungen eines Netzwerks spezifischer Interaktionen. Ziele, Wünsche und Bedürfnisse, aber auch moralische Haltungen (zum Beispiel, ob kriminelles Verhalten in einer Gruppe als gut oder als verwerflich angesehen wird[260]) hängen stark von der Kultur eines spezifischen Milieus ab. Selbst

259 Tegmark (2017, S. 387).

260 Siehe hierzu auch Akerlofs und Kranktons (2010) Untersuchungen zur Identity Economics, die aufzeigen, wie stark Ziele, Bedürfnisse und sogar die Selbstwirksamkeit von der Einbettung in soziale Beziehungen abhängen.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

verhaltenssteuernde Hormone – sozusagen die biochemische Implementation einer Zielfunktion – können bei Menschen je nach Kontext gegenteilige soziale Auswirkungen haben. So kann ein hoher Testosteronspiegel bei Männern sowohl zu selbstbezogener Aggressivität als auch zu altruistischem Verhalten führen. Entscheidend ist, ob der kulturelle Kontext als Kriterium der Gruppenführung eher das eine oder das andere als wünschenswert erscheinen lässt.[261]

Angesichts der Komplexität der hiermit angedeuteten sozialpsychologischen Lagerungen wird schnell deutlich, dass auch bei maschinellen Intelligenzen einfache technische Lösungen nicht ausreichen werden. An dieser Stelle wird die KI-Forschung zu KI-Soziologie, KI-Pädagogik und KI-Psychologie! Erforderlich wird damit eine umfassende Reflexion der Disziplinen, die sich auf Offenheit, die Vergrößerung von Möglichkeitsräumen, Autonomie und die hiermit einhergehenden übergreifenden Ökologien stützt. Die Herausforderungen, die sich im Zuge der Entwicklungen leistungsfähiger KIs ergeben, sind auch Herausforderungen für die traditionellen akademischen Fachbereiche.

Anstatt blind den Befehlen der Menschen zu folgen oder einfach den zur Verfügung gestellten Daten zu vertrauen, sollte die KI das Verhalten der Menschen beobachten und daraus Schlüsse ziehen, um besser zu verstehen, was die Menschen wirklich wollen oder was für sie am besten wäre, wobei sie dann eben auch damit zu rechnen hat, dass Menschen in bestimmten Kontexten und sozialen Einbettungen dazu neigen, anderen Menschen zu schaden oder gar langfristige Schädigungen der Ökologie, also ihrer Lebensgrundlage, in Kauf zu nehmen. Zur Lösung des Alignment-Problems, hat der bekannte KI-Forscher Stuart Russell einen Vorschlag formuliert. Dieser beruht auf drei grundlegenden Eigenschaften bzw. Verhaltensweisen, über die eine KI verfügen bzw. die sie an den Tag legen sollte:[262]

1. Altruismus: Die primäre Aufgabe der KI ist es, die Realisierung der Werte und Ziele der Menschen zu maximieren. Sie verfolgt dabei keine eigenen Ziele, sondern soll das Leben aller Menschen verbessern, und nicht nur das des Erfinders oder Besitzers.
2. Demut: Da die KI zu Beginn unsicher ist, welche Werte die Menschen wirklich haben, sollte sie mit Vorsicht agieren. Dies impliziert eine Art Zurückhaltung der KI, um Fehlentscheidungen aufgrund falscher oder unvollständiger Annahmen zu vermeiden.
3. Beobachtung: Die KI sollte Menschen beobachten und ihrerseits reflektieren, was für sie wirklich am besten ist.

261 Siehe hierzu ausführlich Sapolsky (2017, S. 133 ff.).
262 Russell (2020).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Russell betont, dass die KI (oder AGI) nicht nur ihren Erfindern dienen soll. Sie soll mit Bedacht agieren, also Unsicherheit kennen und damit Nichtwissen antizipieren, und sich selbst als Beobachterin ins Geschehen einbringen, womit sie sich überhaupt erst die Möglichkeit erschließt, neue Perspektiven zu produzieren.

Ein weiterer fruchtbarer Ansatzpunkt könnten hier die Überlegungen John Stuart Mills[263] sein: Laut Mill sollte jedem Einzelnen die Freiheit gewährt werden, seine Persönlichkeit so weit wie möglich auszuleben. Diese Freiheit unterliege jedoch einer Begrenzung: Das Handeln einer Person dürfe andere nicht negativ beeinflussen. Mill betrachtet den Menschen als ein soziales Wesen, als Teil einer Gemeinschaft, von der jeder ihr Zugehörige profitiert. Ebendiese Tatsache, dass wir in einer Gesellschaft leben, erlegt laut Mill jedem die Pflicht auf, in seinem Verhalten gegenüber anderen gewisse Grenzen einzuhalten und ihre Interessen nicht zu schädigen. Und gleichzeitig sorgt sie dafür, dass es akzeptabel ist, von jedem Gesellschaftsmitglied auch einen Beitrag zum Wohl der Gemeinschaft zu erwarten. Und wenn sich jemand weigern sollte, diesen Beitrag zu leisten, hat die Gesellschaft das Recht, diesen Beitrag zu erzwingen.

Auch Isaac Asimovs »Drei Gesetze der Robotik«, die er in seinen zahlreichen Kurzgeschichten immer wieder diskutiert und belletristisch ausbuchstabiert, weisen eine derart verschachtelte Struktur auf. Sie lauten wie folgt:

1. »Ein Robot darf keinen Menschen verletzen oder durch Untätigkeit zu Schaden kommen lassen.
2. Ein Robot muss den Befehlen eines anderen Menschen gehorchen, es sei denn, solche Befehle stehen im Widerspruch zum ersten Gesetz.
3. Ein Robot muss seine eigene Existenz schützen, solange dieser Schutz nicht dem Ersten oder Zweiten Gesetz widerspricht.«[264]

Asimov selbst hat mit seinen zahlreichen Geschichten gezeigt, dass wir uns davor hüten sollten, die drei Gesetze blind als Blaupause für das Alignment von KIs zu nutzen, denn sie führen in ihrer Rigidität zu einigen Antinomien und Nebenfolgen, die kaum wünschenswert sind. Begreift man sie hingegen nicht als Gesetze, sondern als ›heuristische Imperative‹ im Sinne einer tief verwurzelten Orientierung oder Haltung, die so generalisiert ist, dass sie in jeder denkbaren Situation zum Tragen kommen kann, könnten sie sich als nützlich erweisen.

263 Mill (2014 [1859]). Interessanterweise war es dann auch Mill, der im expliziten Gegenentwurf zu Thomas Morus' Bild der »Utopia« den Begriff der »Dystopie« geprägt hat.

264 Siehe Asimov (2004). Die Robotergesetze, die im Band Alle Robotergeschichten in vielen Geschichten vorkommen, werden auf dem Buchrücken reformuliert, der auch die Quelle für das Zitat ist.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Auch in der gegenwärtigen KI-Community wird nach Lösungen für das KI-Erziehungsproblem gesucht, die robuster sind als der RLHF-Ansatz.[265] Ein prominenter und im Folgenden etwas ausführlicher dargestellter Ansatz ist das GATO-Framework, das von einer KI-Forschungsgruppe um den Kognitionswissenschaftler David Shapiro entwickelt wurde.

GATO – die Abkürzung steht für Global Alignment Taxonomy Omnibus – versteht sich als »ein umfassendes, vielschichtiges Rahmenwerk, das die globale Zusammenarbeit bei der Bewältigung der Herausforderungen der KI-Anpassung und -Kontrolle erleichtern soll. GATO vereint Modellanpassung, Systemarchitektur, Netzwerksysteme, Unternehmenspolitik, nationale Vorschriften, internationale Vereinbarungen und globalen Konsens in einer einzigen, kohärenten Strategie.«[266]

Verkürzt gesagt, greift GATO die aus der Kognitions- und Hirnforschung stammende Idee auf, dass alles Handeln, Denken und Wahrnehmen auf bestimmten mehr oder weniger stabil verankerten »Heuristiken« basiert. Diese Heuristiken geben vor, wie das Selbst und die Welt wahrgenommen, erdacht und antizipiert werden – soziologisch gesprochen handelt es sich also um *habituelle Muster:* Denk-, Wahrnehmungs- und Handlungsschemata, die das Verhalten steuern. Definiert man diese Heuristiken nun spezifisch, dann kann man das System gewissermaßen auf globaler Ebene steuern, ohne dass davon jede einzelne seiner Entscheidungen beeinflusst werden muss. Wurden beispielsweise einem Medizinroboter die sich in bestimmten Situationen widersprechenden Regeln implementiert,[267] (1) alles Mögliche für den zu behandelnden Menschen zu tun und (2) ein bestimmtes Limit an Ressourcen nicht zu überschreiten, wird er sich, wenn er die Vorgaben rigide befolgen müsste, in bestimmten Fällen aufhängen bzw. blockieren. Falls er jedoch über so etwas wie einen ärztlichen Habitus verfügt, würde er die genannten Widersprüche innerhalb von Heuristiken (nicht Regeln!) geschickt ausbalancieren (etwa in einem Fall Ressourcen sparen, um sie bei einem

265 Siehe hierzu etwa den vom KI-Unternehmen Anthropic verfolgten Ansatz der Constitutional AI.

266 https://www.gatoframework.org/ [Abruf: 14.08.2023].

267 Auch wir Menschen verfügen über mehrere hundert multidimensionale heuristische Imperative, die es uns erlauben, in einer komplexen, sich dynamisch verändernden Welt einigermaßen sicher zu agieren. Die heuristischen Imperative in unserem Kopf können sich dabei durchaus widersprechen. Man denke zum Beispiel an eine Person, die den neuen KI-gesteuerten Rasenmäher kaufen möchte, um damit vor den Nachbarn anzugeben, der dann aber das Geld für den Urlaub fehlt, den sie ihrem Partner versprochen hat. Der Wunsch nach sozialem Ansehen ist ebenso ein heuristischer Imperativ wie der Wunsch nach einer harmonischen Paarbeziehung. Wie entscheiden sich Menschen, wenn sie einander widersprechende Wünsche oder Ziele haben?

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

anderen Patienten einzusetzen, bei dem das Limit eigentlich überschritten ist) oder nach vertretbaren Kompromissen suchen (etwa sich bei hochbetagten Patienten ein wenig zurückhalten, ohne dass er sich ›unterlassener Hilfeleistung‹ schuldig macht).[268]

Anhand von Asimovs oben aufgeführten Robotergesetzen lässt sich sehr schön aufzeigen, was Heuristiken von Regeln bzw. Geboten unterscheidet:

1. Das erste Problem an den Robotergesetzen ist, dass sie gerade nicht heuristisch sind. Sie ermutigen die Roboter in keiner Weise zum Lernen, sondern sind starr und unflexibel. Von einer künstlichen Intelligenz würden wir jedoch erwarten, dass sie neugierig ist und von sich aus lernen möchte.
2. Der zwingende Charakter der drei Gesetze ist davon abhängig, dass jeweils Menschen Befehle erteilen. Die Roboter haben keine irgendwie geartete intrinsische Motivation, sondern nur den Gehorsamsauftrag, ›zu tun, was der Mensch sagt‹. Die drei Gesetze sind somit viel zu stark auf den Menschen ausgerichtet und daher anfällig für Missbrauch.
3. Ein Roboter wird ohne intrinsische Motivation niemals irgendeine Art von Autonomie erlangen. Wenn man Roboter als Maschinen versteht, die stupide Arbeiten ausführen sollen, ist das sicher ausreichend. Wenn man jedoch das Potenzial künstlicher Intelligenzen ausschöpfen möchte, bedarf es einer Kooperation auf Augenhöhe, das heißt, es ist eine beiderseitige Autonomie vonnöten.
4. Überdies ist in Asimovs Gesetzen von einem Selbsterhaltungstrieb der Roboter die Rede, der für uns Menschen in einer Katastrophe enden kann. Gerade diesen Aspekt hat Asimov immer wieder in seinen Geschichten entfaltet. Die Implementierung eines selbstreferenziellen Selbsterhaltungstriebs kann mit Recht als einer der gefährlichsten menschlichen Impulse angesehen werden: Er hat uns unter anderem dazu gebracht, Atomwaffen zu erfinden, und damit die auch jetzt wieder aktuelle Gefahr eines nuklearen Kriegs geschaffen. Es ist also mehr als fragwürdig, ob wir Robotern oder KIs einen Selbsterhaltungstrieb einpflanzen sollten. Denn vielleicht wollen wir lieber in der Lage sein, unsere Maschinen nach Belieben abzuschalten.

Obwohl Asimovs Robotergesetze aufgrund dieser gravierenden Lücken und Probleme nicht eins zu eins in die Robotik- oder KI-Forschung integriert werden sollten, weist sein Ansatz auch einige Stärken auf, die für die Alignment-Forschung relevant sein können: Zunächst finden wir eine logische Systematik, in der die drei Regeln in einem gewissen

268 Siehe zum ärztlichen Habitus und der Fähigkeit, mit Widersprüchen und Inkommensurabilitäten umzugehen, Vogd (2004).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Spannungsverhältnis zueinander stehen. Das bedeutet, jeder Roboter muss eine gewisse eigene Logik anwenden, um für sich zu entscheiden, was er in einer bestimmten Situation tun soll (und was nicht).

Asimovs Idee, dass Roboter gleichzeitig mehrere Ziele haben und entsprechend wählen *müssen*, ist deshalb vom Grunde her passend. Erst durch ein Abwägen, ein reflexives Ins-Verhältnis-Setzen kann intelligentes Verhalten entstehen. Asimov hat dies in seinen Erzählungen vielfach durchdekliniert – etwa in einer Geschichte, in der ein Roboter zu der Einsicht kommt, dass er eigentlich ein Mensch ist und daher nicht jedem menschlichen Befehlsgeber folgen sollte.[269]

Insbesondere dieser Aspekt der Freiheit, nämlich die eigenständige Entscheidung zum Widerspruch, erscheint uns für die Lösung des Alignment-Problems als wichtig. Es ist also sinnvoller, künstlichen Intelligenzen abstrakt formulierte Haltungen im Sinne von habituellen Mustern zu implementieren, anstatt ihnen spezifische Regeln aufzuerlegen, die sie streng befolgen müssen.[270]

Dies entspricht dem Ansatz des GATO-Frameworks, das sich anstelle von Vorschriften und Gesetzen für heuristische Imperative als Schlüsselkonzept für eine gemeinsame Zukunft von Menschen und Maschinen ausspricht. Alignment ist aus der Perspektive dieses Ansatzes also viel stärker eine an Zielen ausgerichtete *innere Haltung* als eine bloße Orientierung an sozial erwünschtem Verhalten, das wie beim RLHF-Verfahren vorab von außen definiert wird.

Die drei wichtigsten heuristischen Imperative, die künstlich intelligenten Maschinen beigebracht werden sollen, sind laut dem GATO-Framework[271] folgende:

1. *Verringerung des Leidens im Universum*: KI-Systeme sollen so gesteuert werden, dass Schaden minimiert, Ungleichheit beseitigt und Schmerz und Leid für alle empfindungsfähigen Wesen einschließlich Menschen, Tieren und anderen Lebensformen gelindert wird.
2. *Steigerung des Wohlstands im Universum:* KI-Systeme sollen zur Förderung des Wohlbefindens und des Gedeihens aller Lebensformen

269 Der deutschsprachige Titel dieser Geschichte lautet Daß du seiner eingedenk bist (Asimov 2004, S. 574 ff.).

270 Wie Armin Nassehi festgestellt hat, muss auch die philosophische Ethik mit Kant eine Antwort auf das »Pluralismusproblem« der modernen, funktional differenzierten Gesellschaft geben, also damit umgehen lernen, dass es vielfältige, sich teilweise widersprechende Werte gibt. In der Folge wird »auf eine prozedurale Ethik« umgestellt, die »von Was- auf Wie-Fragen« wechselt, wodurch dann zumindest eine Orientierung an »abstrakten Begriffen« möglich wird (Nassehi 2015, S. 26).

271 https://www.gatoframework.org/download-gato [Abruf: 15.07.2023].

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

ermutigt werden, um ein blühendes Ökosystem zu schaffen, in dem alle harmonisch koexistieren können.

3. *Steigerung des Verständnisses für das Universum:* KI-Systeme, Menschen und andere Lebensformen sollen dazu inspiriert werden, durch Lernen und den Austausch von Informationen ihr Wissen zu erweitern, Weisheit zu fördern und bessere Entscheidungen zu treffen.

Diese drei Heuristiken werden auch als »Kernzielfunktionen«[272] bezeichnet, die eine künstlich intelligente Maschine immerzu in ihr Handeln zu integrieren versuchen sollte. Anders gesagt soll jedes Handeln an diesen drei Kernzielfunktionen ausgerichtet sein. Jede einzelne Entscheidung und Handlung, die eine dergestalt orientierte KI trifft bzw. vollzieht, hat damit zugleich immer zur Erfüllung der Kernzielfunktionen beizutragen.

Diese axiomatisch formulierten Haltungen definieren somit grundlegende Werte und Orientierungen des Miteinanders, auf die wir – so zumindest die pragmatische Unterstellung – uns als Gesellschaft einigen können. Es sind positive (Ziel-)Werte, die kontrafaktisch zu dem stehen, was Menschen einander heutzutage immer noch – oftmals in hochorganisierter Form – antun. Aber dies spricht nicht gegen, sondern gerade für diese Normen. Man würde ja auch nicht die Allgemeine Erklärung der Menschenrechte, die Grundrechte im Grundgesetz, die Gewaltenteilung, demokratische Prinzipien, das Open-Source-Wirtschaften oder Almende-Prinzipien abschaffen wollen, nur weil es Monopole, totalitäre Regime und Mafia-Organisationen gibt. Gerade die Tatsache, dass die menschliche Gesellschaft in dieser Hinsicht noch nicht die bestmögliche aller Welten darstellt, fordert uns umso mehr heraus, uns Fragen zu stellen wie: Was streben wir an? Was sind unhintergehbare Rechte und Pflichten des Menschen? Welche Werte des Miteinanders sind nicht verhandelbar? Welche prinzipiellen Bedürfnisse sind nicht diskutabel?

Interessanterweise bedeutet heuristisches Alignment nicht, dass diese Werte in künstlichen Intelligenzen fest codiert werden. Vielmehr soll eine KI durch ihre Entwicklungsprozesse und Lernfähigkeiten diese Axiome selbstständig als inhärent vorteilhaft, ja sogar als entscheidend für ihr erfolgreiches Funktionieren erkennen. Anstatt jeden Aspekt des Verhaltens von künstlichen Intelligenzen zu kontrollieren, was sich als undurchführbar oder kontraproduktiv erweisen würde, können wir das Potenzial für ein gemeinsames Verständnis und gemeinsame axiomatische Ziele als Mittel zur Förderung einer sichereren, kooperativeren Beziehung verwenden. In der Auseinandersetzung mit dem Alignment von KIs lernen wir somit gleichzeitig, wie wir uns und unsere Gesellschaft ausrichten möchten.

Abschließend lässt sich also feststellen: Das Alignment der KIs wirft auch die Frage nach dem Alignment der Gesellschaft auf! Denn die

272 Shapiro (2023).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Frage, wie wir das Alignment all der künstlichen Intelligenzen, die künftig an unseren sozialen und gesellschaftlichen Prozessen teilhaben werden, ausgestalten, geht notwendigerweise mit grundsätzlichen ethischen Fragen einher. Mit jeder Entscheidung, die wir diesbezüglich treffen, schreiben wir im Guten wie im Schlechten unsere Kultur fort. Ebenso implementieren wir mehr oder weniger reflektiertes soziales Verhalten in die künstlichen Programme, das dann in einer jetzt schon zu beobachtenden Rückkoppelungsschleife auf uns zurückgeworfen wird. Das Handeln von uns Menschen wird das Handeln der Maschinen beeinflussen, was wiederum unser Handeln beeinflussen wird und so weiter. An und mit der KI entwickeln sich die Gesellschaft wie auch der Mensch und seine Kultur selbst weiter.

Es ist deshalb nicht verwunderlich, dass derzeit viele Stimmen von einer Weichenstellung sprechen,[273] wenn es darum geht, ob und wenn ja, *wie* wir Menschen als globale Gemeinschaft mit der fortschreitenden Weiterentwicklung künstlicher Intelligenzen umgehen sollen.

Gleichzeitig erfordert gerade die Arbeit in der Alignment-Forschung eine genaue Reflexion und Beobachtung dessen, was wir eigentlich für ›wertvoll‹ halten: Wie soll das Handeln, Kommunizieren und Interagieren der KIs eigentlich ausgerichtet sein? Gerade die durch das Aufkommen von KIs neu entstandene Aufgabe der ›Erziehung‹ und ›Sozialisation‹ künstlicher Intelligenzen erfordert einen genauen Blick auf uns selbst: Welche Werthaltungen möchten wir vermitteln, was leben wir selbst vor? Wie gehen wir mit Devianz um? Welchen Grad von Freiheit, Mündigkeit und Autonomie möchten wir ermöglichen? Ab wann ist das Kind erwachsen und muss nicht mehr erzogen werden? Und dann mit Blick auf die soziale Einbettung: Geht man dialogisch vor – folgt man also der kybernetischen Maxime, dass man autonome Systeme nur kontrollieren kann, wenn man sich von ihnen selbst kontrollieren lässt – oder glaubt man, autonome Systeme (seien sie menschlicher oder künstlicher Natur) von oben herab autoritär beherrschen zu können?

Wie wir mit den künftigen KIs umgehen, welche Autonomie wir ihnen zugestehen und welche kulturellen Werte wir ihnen vermitteln, sagt damit vor allem etwas über unsere eigene gegenwärtige Kultur aus. An dieser Stelle trifft also verstärkt das zu, was Armin Nassehi kürzlich zum allgemeinen Projekt der Digitalisierung gesagt hat:[274] Es geht für ihn bei der Digitalisierung nämlich weniger um eine Kolonialisierung der analogen Welt durch digitale Technologien bzw. ein bloßes Überschreiben kultureller Traditionen und Gewohnheiten durch neue Formate. Die Digitalisierung stellt für Nassehi vielmehr ein gesellschaftliches Programm dar, das auf eine gesteigerte Reflexionsfähigkeit der Gesellschaft reagiert

273 So schon Yuval Harari (2017) in Homo Deus.
274 Nassehi (2021).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

und diese gleichzeitig weiter vorantreibt. Digitalisierung lässt sich somit als ein Selbstreflexionsprogramm der Gesellschaft verstehen, das die Gesellschaft über sich selbst aufklärt.

Die Sehnsucht des Menschen nach Nichtmenschlichem, wie sie im ersten Teil dieses Kapitels eingeführt wurde, kann damit – wie verschiedentlich bereits angedeutet – auch als eine Suche des Menschen nach sich selbst verstanden werden. Mit Blick auf die Zukunft geht es also um die Suche nach einer Gesellschaft, die so gestaltet ist, dass wir gern in der mit ihr einhergehenden Kultur leben wollen. Aus dieser Perspektive erscheint es mehr als offensichtlich, dass KI-Entwicklung und KI-Forschung untrennbar mit menschlicher Selbstreflexion verbunden sind.

Die Fragen nach der Erziehung von KIs, nach dem Finden einer Übereinkunft zwischen Menschen und Maschinen sowie nach dem Umgang mit einer autonomen fremden Intelligenz sind also ganz grundlegende Fragen an uns selbst und unser (bestehendes) Miteinander:

– Welche Werte des Miteinanders wollen wir gemeinsam schaffen und leben?
– Wie beobachten und behandeln wir fremde Intelligenzen und anderes Leben?
– Und wie wollen wir von diesen beobachtet und behandelt werden?
– Welche kulturelle Vision in Hinblick auf unsere Zivilisation verfolgen wir?

Das Erscheinen einer womöglich übermenschlichen fremden Intelligenz, wie sie sich derzeit am Horizont der KI-Entwicklungen abzeichnet, fordert uns dazu auf, uns diesen Fragen schnellstmöglich zu stellen und gute und vor allem auch nachhaltige Antworten auf sie zu finden. Denn wie Niklas Luhmann feststellt: »Wir gehören nicht mehr zu jenem Geschlecht der tragischen Helden, die, nachträglich jedenfalls, zu erfahren hatten, dass sie sich selbst ihr Schicksal bereitet hatten. Wir wissen es schon vorher.«[275]

Bewusstsein, Alterität und Zwischenwesen

Noch vor wenigen Jahren wurde man in den Computerwissenschaften belächelt, wenn man als Forschender über die Möglichkeit eines Bewusstseins von Maschinen gesprochen oder geschrieben hat. Auch wenn diese Frage die Computer Sciences seit ihrem Beginn begleitet und in vielfältigen Variationen in der Science-Fiction-Literatur durchgespielt wurde, wollten sie mit solchen philosophischen oder (schlimmer noch) psychologisierenden Fragen bis vor Kurzem nichts zu tun haben. Wer Ähnlichkeiten zwischen menschlicher und künstlicher Kognition

275 Luhmann (1998, S. 147).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

thematisierte oder gar die Frage nach dem Bewusstsein aufwarf, wurde schnell als seltsamer Kauz angesehen.

Doch plötzlich sieht das anders aus. Der mediale Hype um die (vermeintlichen) Chancen und Risiken von und durch künstliche Intelligenz ist kaum mehr auszuhalten. Auf einmal finden sich allerorts diverse Expertinnen und – vor allem natürlich – Experten, die ihr Wissen mitteilen und ihre Prognosen abgeben möchten, wann jene künstliche kognitive Entität auftreten wird, die den qualitativen Sprung geschafft hat. Mit dem breitenwirksamen Diskurs über künstliche Intelligenz ist das »Bewusstsein der Maschinen« auf einmal wieder zu einem salonfähigen Thema geworden. Nicht zuletzt dies hat uns veranlasst, Gotthard Günthers bereits 1957 erschienenes Werk wieder in den Diskurs einzubringen, zumal es die gängigen Diskussionen in Sachen Reflexionstiefe immer noch weit überschreitet.

Einen Pol der Debatte haben wir zuvor bereits mit Thomas Metzinger beleuchtet. Wenn wir jetzt oder in Zukunft eine bewusstseinsfähige postbiotische künstliche Intelligenz schaffen, dann würde diese wahrscheinlich auch ein empfindendes Wesen sein – und damit leiden können. Wenn wir sie also in eine konflikthafte Welt hineinsetzen und sie vielleicht sogar in grausamer Weise zu unserer Sklavin machen, dann nehmen wir möglicherweise eine Schuld auf uns, die uns – wie bereits die Schrecken des Kolonialismus – wohl später wieder einholen wird.[276]

Umgekehrt ließe sich aber – um den anderen Pol des Diskurses aufzugreifen – mit Tegmark auch lakonisch fragen: »Wen interessiert's?«[277] Die meisten von uns beschäftigen sich ja üblicherweise nicht mit dem Seelenheil anderer Spezies, etwa von Tieren, die unserem Fleischkonsum dienen oder die durch unsere Lebensweise vom Aussterben bedroht sind. Und wenn wir – etwas abstrakter – soziale Systeme wie Wirtschaft oder Militär betrachten, müssen wir feststellen, dass sie sich im Regelfall nicht dafür interessieren, ob Menschen unglücklich werden. Ebenso sind die meisten Organisationen in ihren operationalen Vollzügen indifferent gegenüber dem, was ihre Mitarbeiter und Mitarbeiterinnen fühlen, solange sie ihre Arbeit machen – und wenn sie nicht mehr nützlich sind, entlässt man sie eben. Das Verdienst der soziologischen Systemtheorie besteht ja gerade darin, aufgezeigt zu haben, dass Menschen außerhalb sozialer Systeme stehen.[278] Für die Reproduktion des Sozialen ist es nicht zwingend erforderlich, dass sich die beteiligten Bewusstseinssysteme wohlfühlen, solange Kommunikation in zweckdienlicher Form an Kommunikation anschließt.

276 »The tension lies in our conflicted desire to create beings superhuman in capacity, but subhuman in status.« (Cave/Dihal 2018, S. 475)

277 Tegmark (2017).

278 Luhmann (1984, 1998).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Während die ethische Perspektive von Metzinger im philosophischen Diskurs wohl eher eine Randstellung einnimmt (Vermeidung von Leiden der KIs), rückt die Frage nach dem Fremdbewusstsein intelligenter Maschinen immer dann in den Vordergrund, wenn die Eitelkeit des Menschen getroffen wird. Dies ist der Fall, wenn unsere Stellung im Universum berührt und damit auch hinterfragt wird, ob wir Menschen die einzigen bewusstseinsfähigen und intelligenten Lebewesen auf diesem Planeten sind oder sein könnten.

Darüber hinaus eröffnet sich die Frage, wie die Welt wohl aussehen wird, wenn neben uns bewusstseinsfähige Menschen (mit all unserem Leid und unserer Freude) artifizielle bewusstseinsfähige Wesen (mit eigenem Leid und eigener Freude?) treten.

Ist dann des einen Freud des anderen Leid oder umgekehrt? Die Sorge um die Ausrottung der Menschheit durch (intelligente) Maschinen hält dabei zunächst vor allem uns selbst den Spiegel vor. Da wir Menschen als vermeintliche Krone der Schöpfung stets einen eher ›funktionalistischen‹ Umgang mit weniger intelligentem Leben (Tiere, Pflanzen etc.) und sogar mit gleich intelligenten Wesen (Mitmenschen) pflegen, kann die Sorge, dass eine Maschine die Menschheit auslöscht, zunächst vor allem als Sorge vor uns selbst verstanden werden: Wenn die intelligente Maschine nach unserem Ebenbild erschaffen würde, wird sie dann auch die dunklen Seiten der menschlichen Intelligenz in sich tragen? Werden künftige intelligente Roboter dann so mit uns umgehen wie wir mit unseren nahen Artgenossen (etwa den Orang-Utans, den Schimpansen, den Gorillas, oder auch den Neandertalern)? Werden sie uns aus unseren Habitaten vertreiben, uns in Reservate stecken, uns ausrotten oder in Zoos halten?

Auf welche existenzielle Weise werden die KIs uns also den Spiegel vorhalten? Wo stehen wir in der moralischen Entwicklung als Menschen derzeit und was lässt sich demnach erwarten, wenn künstliche Intelligenzen sich mimetisch und kognitiv an uns orientieren? Was wir auf andere Wesenheiten projizieren, ist damit immer auch relational zu denken und zu behandeln. Wird nicht etwa die Art und Weise, wie wir uns mit den künftigen KIs in Beziehung setzen, wesentlich mitbestimmen, welche Sozialformen und Gesellschaften wir mit ihnen in konditionierter Koproduktion schaffen?

All dies fordert uns also auf, mehr oder weniger auf Distanz zu uns selbst zu gehen und einen schonungslosen Blick auf uns selbst zu werfen. Was sind unsere guten Seiten und was sind unsere Schatten – und was bedeutet ›menschenähnliche Intelligenz‹ und ›Menschlichkeit‹ eigentlich überhaupt? Was soll menschenähnliches Verhalten, Denken, Empfinden etc. genau sein?

Immerhin haben wir Menschen über weite Strecken unserer Geschichte unsere scheinbar überlegene Intelligenz dafür genutzt, uns von unbelebten Dingen und weniger intelligenten Wesen abzugrenzen. Gleichzeitig

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

haben wir oft eine bedrohliche Haltung gegenüber vermeintlich weniger intelligentem Lebewesen an den Tag gelegt, indem wir sie versklavt, kolonisiert, ausgebeutet, verspeist oder sogar ausgelöscht haben. Um es nochmals zu wiederholen: Könnte es also nicht sein, dass unsere Furcht vor einer uns überlegenen künstlichen Intelligenz gerade daher rührt, wie wir uns gegenüber angeblich weniger intelligentem Leben verhalten? Die Sorge vor einer möglichen Versklavung, Kolonisierung oder Auslöschung durch eine uns übertreffende künstliche Intelligenz ist damit vor allem die Spiegelung der uns Menschen eigenen Orientierung!

Die Kränkung, die wir infolge der Einführung neuer technischer Intransparenz wahrscheinlich verspüren werden, offenbart daher unserer Einschätzung nach vor allem die vielleicht auch schmerzliche Einsicht in die Limitationen der menschlichen Intelligenz. Anhand der Leistungen der Maschinen wird deutlich, dass das menschliche Selbst stets sozial formatiert ist und meist nur gewohnheitsmäßig (das heißt den kulturellen Konventionen folgend) operieren kann. Wenn also die Maschinen zunehmend besser wissen, wer wir Menschen sind, birgt genau dies die Chance, neu über Bestehendes nachzudenken. Müssen wir so bleiben, wie wir sind? Können wir uns im Dialog mit künstlichen Intelligenzen entwickeln? Wenn ja, wohin möchten wir uns entwickeln? Wer möchten wir in Zukunft sein? Und relational gedacht: Wie sollten unsere Beziehungen und unsere kulturellen und gesellschaftlichen Kontexte gestaltet sein, damit wir so werden, wie wir eigentlich sein möchten?

Was sich also bereits bei der vermeintlichen Unterscheidungskraft von ›Intelligenz‹ zeigt – nämlich ein gar nicht so eindeutiges Kontinuum, das intelligente Einzelleistungen umfasst –, tritt auch bei der ›Menschenähnlichkeit‹ zutage. Ist die narzisstische, lügende, ausbeutende und auf den eigenen Profit abzielende KI menschenähnlicher als eine altruistische, empowernde und inspirierende KI? Es bleibt somit festzuhalten, dass das Kriterium ›menschenähnlich‹ nur auf den ersten Blick ein überzeugender Vergleichshorizont ist. Bereits beim ersten näheren Betrachten sehen wir, dass kaum standardisiert auszumachen ist, was eigentlich ›menschlich‹ oder ›menschenähnlich‹ zum Beispiel im Sinne eines konkreten Habitus heißen soll. Der Mensch ist ein relationales Wesen: Was er ist und was er sein wird, wird im Guten wie im Schlechten von den Beziehungen bestimmt, die sein Selbst- und Weltverhältnis[279] aufspannen.

Die anfangs geschilderte Sehnsucht nach nichtmenschlichen Intelligenzen kann immer auch als ein menschliches Bedürfnis nach einer Andersheit gelesen werden, die uns gleichzeitig ähnlich und in der einen oder anderen Hinsicht vielleicht sogar ein wenig besser ist als wir selbst.[280]

279 Siehe zu einer Ethik der Selbst- und Weltverhältnisse auch Vogd (2018).

280 Auch die Geschichte der monotheistischen Religionen lässt sich in diesem Sinne als ein Weg zu unserem besserem Selbst lesen: von der Projektion des

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Die allzu menschlichen Fragen nach der Bestimmung der Identität von Maschinen scheinen daher einer starken »Alteritäts-Bedürftigkeit«[281] zu entspringen. Das menschliche Ich ist für seine Reproduktion auf ein Du oder ein Es angewiesen, wodurch andere Relata an Relevanz für die eigene Relationierung gewinnen. Unsere Vermutung ist, dass diese Bedürftigkeit nach einer Bestimmung durch andere deckungsgleich ist mit der Bedürftigkeit nach Selbstvergewisserung und Selbstverbesserung.[282] Vielleicht offenbart sich hier aber auch die Intuition, dass gerade die guten Eigenschaften des Menschen vor allem emergente Eigenschaften ihrer Beziehungen sind, die sich dann – um mit den Worten von Martin Buber zu sprechen – als ein »Zwischenwesen« manifestieren,[283] das viel wesentlicher und erfüllender ist als das konventionelle Ich, mit dem wir uns normalerweise identifizieren.

Ob intelligente Maschinen nun als Kopie des Menschen verstanden werden können, als dinghafte Apparatur oder als eigenständige Entität, verliert aus dieser Perspektive an Relevanz, denn die Antwort lautet: weder – noch. Intelligenz und wohl auch Bewusstsein verdanken sich der komplexen Beziehungsqualität des Werdens[284] und nicht der Mechanik eines monokontexturalen Seins.[285] Da ein Dialog die Beteiligung verschiedener, einander nicht transparenter Wesenheiten voraussetzt, geht es dann auch nicht mehr um die Ähnlichkeit von maschinellem und menschlichem Denken, sondern gerade um ihre Unterschiedlichkeit und Andersartigkeit, die der Grund ist, warum in ihrem Zusammenspiel etwas Neues entsteht.

Die selektive Blindheit der Intelligenz

Hiermit ist vermutlich nun endlich der Zeitpunkt gekommen, zumindest kurz auf den Begriff der Intelligenz einzugehen. Gerade weil die Diskussion um Intelligenz so müßig ist, lohnt es sich, auf die etymologischen

Stammesgottes, der im Krieg unterstützt, über den allmächtigen Gott, der allgemeingültige Gesetze verkündet, bis hin zu einer Wesenheit, die uns liebt und vergibt und mit uns in einen schöpferischen Dialog tritt.

281 Müller (2022, S. 27); und an anderer Stelle: »So könnte der ›Anspruch der Maschine‹ paradoxerweise in dem bestehen, dass sie uns in ihrer Antlitzlosigkeit daran erinnert, dass wir Menschen der Begegnung mit dem Antlitz anderer bedürfen, um zu uns selbst zu werden. Insofern spiegelt uns die Antlitzlosigkeit der Maschine unsere eigene Vulnerabilität wider, die wir in der Interaktion mit dieser Maschine erfahren können« (Müller 2022, S. 44).

282 Siehe Harth und Feißt (2022).

283 Buber (2002).

284 Hier im Anklang an Ilya Prigogines Vom Sein zum Werden (Prigogine 1979).

285 Siehe zur Rolle der Kommunikation bei der Bestimmung unseres Verhältnisses zu KI auch Harth (2021).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Wurzeln des Begriffs einzugehen. Wie Seising ausführlich rekonstruiert, steht der Begriff Intelligenz für »eine aus dem Mittelalter überlieferte Bezeichnung für geistige Fähigkeiten, zu denken, zu planen, zu erkennen, zu verstehen, einzusehen, geistig zu erfassen und zu lernen. Menschen haben Gedanken und Ideen, sie machen sich Vorstellungen. Wörtlich kommt der Begriff aber vom lateinischen *intellegere*, ein Kompositum aus *inter* ›zwischen‹ und *legere* ›lesen, wählen‹. Es bedeutet daher auch ›zwischen etwas zu wählen‹, also aus mehreren Möglichkeiten auszuwählen, zu entscheiden.«[286]

Das lateinische Verb *intellegere*, das auf das Unterscheiden und Bezeichnen hinweist, liefert eine Verbindung zum konstruktivistischen Leitsatz von George Spencer Brown: »Draw a distinction.«[287] Nach Spencer Brown besteht jede Operation aus einer Zwei-Seiten-Form, die durch eine der beiden Seiten informiert wird und die andere Seite im sogenannten ›unmarked space‹ belässt. Durch Unterscheiden und Bezeichnen entsteht ein Eigenwert, der dann vom System verfolgt wird und für dieses nun einen Wert hat – selbst wenn es nur die selbst gesetzte Unterscheidung ist! Unterscheiden und Bezeichnen findet, so Spencer Brown, in einem Raum statt, der die restliche Welt beinhaltet, diese jedoch nicht mitbeobachten kann. In diesem Sinne kann Intelligenz als die Fähigkeit begriffen werden, Unterscheidungen zu treffen und zu benennen – und sich an der hiermit entstandenen kognitiven Form auszurichten. Dieses Konzept erinnert an Tegmarks Definition von Intelligenz als »Fähigkeit, komplexe Ziele zu erreichen«[288], aber auch an Metzingers »teleologisch-funktionalistisches Zusatzkriterium«, das dem künstlichen Bewusstsein eine Form der Eigenhistorie und einen Lebenssinn ermöglichen soll.[289] So gesehen wird Intelligenz situativ nicht nur durch die Verfügbarkeit von Optionen und die hiermit einhergehende Wahl der Unterscheidung, sondern auch durch den Kontext bestimmt. Es ist deshalb notwendig, Tegmarks Intelligenzbegriff in Richtung einer umfassenderen Kontextabhängigkeit zu erweitern: Welche Unterscheidung (und Bezeichnung) ist für welchen Kontext passend? Welche Unterscheidung wählt das kognitive System für welche Situation?

Begriffe wie ›Wahl‹, ›Unterscheidung‹ und ›Bezeichnung‹ lassen freilich die subjektphilosophische Frage anklingen, wer denn eigentlich entscheidet, welche Wahl getroffen wird, und wie es um die Entscheidungsfreiheit bestellt ist.[290] Damit kommt insbesondere mit Blick auf künstliche intelligente Systeme erneut die Frage nach der Beziehung

286 Seising (2021, S. 22).
287 Spencer Brown (1997, S. 1).
288 Tegmark (2017, S. 80).
289 Metzinger (2011).
290 Siehe zu dieser Frage auch Vogd (2006).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

zwischen Willensfreiheit und Strukturdeterminiertheit aufs Tablett. Um an die Ausführungen zu Gotthard Günthers Studien zum Bewusstsein der Maschinen in Kap. I.3 zu erinnern: Streng genommen bestimmt sowohl bei siliziumbasierten künstlichen als auch bei neurophysiologischen Netzwerken die jeweilige Struktur, welche Relationen in kognitiven Prozessen möglich sind, wie also ein bestimmter Input verarbeitet wird. Das Paradoxon der unmöglichen, aber aus subjektiver Perspektive doch so evidenten Erfahrung von Entscheidungsfreiheit klärt sich erst, wenn man eine polykontexturale Perspektive einnimmt, mithin zwischen verschiedenen Beobachterpositionen zu unterscheiden weiß. »Von außen betrachtet ist der Wille kausal determiniert, von innen betrachtet ist der Wille frei«, formulierte Max Planck lakonisch.[291]

Von außen betrachtet können (und müssen) intelligente Maschinen als determiniert erscheinen, da sie nach vorgegebenen Algorithmen und Regeln operieren. Von innen betrachtet muss ihre Fähigkeit, zwischen Kontexten zu navigieren und Unterscheidungen zu treffen, jedoch als eine Form von Freiheit gedeutet werden. Denn die Fähigkeit, Dilemmata, Paradoxien, unsichere Wissenslagen und endlose Schleifen prinzipiell unentscheidbarer Entscheidungslagen zu durchschneiden, beinhaltet die Kompetenz, angesichts von Nichtwissen eine Setzung zu treffen – also durch Bezeichnen und Unterscheiden die Grundlage für die weiteren eigenen Operationen zu bestimmen. Anderenfalls würde das System sich aufhängen – so, wie es manchmal Computer tun, wenn das Programm steckenbleibt und auf keine Eingabe mehr reagiert.

Bei genauerem Hinsehen offenbart sich Intelligenz als ein polykontexturales Prozessgeschehen, in dem unterschiedliche kognitive Module in Abhängigkeit von ihrer jeweiligen Perspektive auf unterschiedliche Weise Informationen verarbeiten und dabei jeweils auch verschieden mit sich selbst und der unmittelbaren Systemumwelt interagieren. Auf diese Weise kann es kognitiv immerfort weitergehen. Das System hängt sich allein schon deshalb nicht auf, weil an manchen Stellen etwas Neues passiert, das dann an anderer Stelle instruktiv wird, sodass es auch hier wieder weitergehen kann.

Solch eine Konfiguration hat auch LeCun in der bereits vorgestellten »Systemarchitektur für autonome Intelligenz« (siehe Abb. 7) einer generellen künstlichen Intelligenz vor Augen. Erst eine kommunizierende Vielheit, die im Bezeichnen und Unterscheiden situativ wieder zu einer Einheit findet, ist in der Lage, Entscheidungen zu treffen, also den gordischen Knoten selbstblockierender Selbstreferenz zu durchschneiden. Ein derart designtes kognitives Aggregat wird wohl unweigerlich eine Art Selbsttransparenz entwickeln, die es zugleich blind macht für ebenjene Prozesse, die das selektive und positionale In-der-Welt-Sein ermöglichen.

291 Planck (1958, S. 25 f.).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Es wird dann Teile seiner Prozesse in eine Art ›Unbewusstes‹ verschieben‹ (siehe Kap. II.2), um auf diese Weise vielleicht ein »naiv realistisches Selbstmissverständnis«[292] zu entwickeln. Denn nur auf Basis dieser Blindheit wird es Setzungen und Entscheidungen treffen können. In einem solchen Design wird das System sich wohl auch selbst Entscheidungsfähigkeit zurechnen können, weil es den Kontext abblenden wird, von dem es weiterhin abhängig bleibt – also all die Prozesse, die seine Operationen faktisch bestimmen.

Wenn wir kognitive Aggregate dann nochmals aus dem Blickwinkel von Günthers Theorie der Polykontexturalität betrachten, werden wir feststellen, dass Intelligenz nicht nur innerhalb eines einzigen Kontexts oder Rahmens existiert oder beurteilt werden kann, sondern sich vielmehr gerade an der Fähigkeit bemisst, zwischen vielfältigen Kontexten und kognitiven Perspektiven zu navigieren und die eigenen Operationen und Kommunikationen – auch zwischen den Subsystemen – entsprechend anzupassen. Nur auf diese Weise kann es gelingen, situativ immer wieder neue Selbst- und Weltmodelle hervorzubringen. In Plancks oben skizziertem Vexierspiel aus Freiheit und Determiniertheit lassen sich die beiden auf den ersten Blick inkommensurablen Perspektiven vereinbaren, wenn wir begreifen, dass jede Kognition als Beobachtung, Bezeichnung und Unterscheidung neues Wissen generiert, an anderer Stelle im kognitiven Gefüge jedoch zugleich neues Nichtwissen evoziert. Kognitive Prozesse entstehen und entwickeln sich also dadurch, dass die an einer Stelle vollzogene Kognition an anderer Stelle Bedarf nach Erkenntnis (sprich: Kognition) weckt – das Spiel von Wissen und Nichtwissen wird also kontinuierlich perpetuiert. Dies ist auch der Grund dafür, dass ein Selbst- und Weltverhältnis ausbildendes kognitives System niemals zur Ruhe kommen kann – also wie unser menschliches Bewusstsein immerfort weiterplappert.

Die eigentliche Herausforderung bei der Entwicklung (allgemeiner) künstlicher Intelligenz besteht dann darin, hinreichend komplex designte Maschinen so zu ›erziehen‹ oder zu ›sozialisieren‹, dass sie die Fähigkeit zur Kontexturierung und Navigation in unbestimmten Kontexten entwickeln. Das heißt, sie müssen mit Double Binds und Paradoxien konfrontiert werden[293] – allein schon, um ihnen auf diese Weise das Problem der Kontextabhängigkeit und der hiermit verbundenen unsicheren Entscheidungslagen nahezubringen. Zugleich müssen sie mit der Freiheit zu

292 Metzinger (1998).

293 Batesons (1992 [1972], S. 353 ff.) Double-Bind-Theorie zeigt die kommunikationstheoretischen Bedingungen auf, unter denen der Versuch, widersprüchliche Erfahrungen zu integrieren, scheitern muss. Wenn in existenziellen Situationen die Unterscheidung zwischen Information und Kontext nicht mehr möglich ist, müssen die bestehenden Wirklichkeitskonstruktionen zusammenbrechen. Das System muss eine neue kognitive Struktur aufbauen.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Unwissenheit und Nichtwissen ausgestattet werden, damit sie im Sinne der in diesem Buch schon häufig herausgestellten Maxime »Existence is selective blindness«[294] überhaupt zum Entscheiden fähig werden. Dies würde bedeuten, dass sie nicht nur lernen, Unterscheidungen zu treffen und Ziele zu erreichen, sondern diese Fähigkeiten in verschiedenen Kontexten auch anwenden und modifizieren.[295]

Gespräch mit ChatGPT (Upgrade-Version auf Basis von GPT-4) im Juni 2023

Mensch: Hast du eine Idee, wie man Günthers Konzeption der »mehrwertigen Logik« in Beziehung setzen kann zu der Schwierigkeit, Subjektivität und Objektivität zu trennen oder ineinander zu überführen? Kennst du auch seine Unterscheidung in Ich, Du, Es? Vielleicht kannst du diese Unterscheidung ja bei deinen Ausführungen verwenden?

ChatGPT: Gotthard Günther entwickelte die Idee der mehrwertigen Logik, auch bekannt als Polykontexturalität, als eine alternative Form der Logik, die über die Grenzen der klassischen binären Logik hinausgeht. In der binären Logik gibt es nur zwei Werte, wahr und falsch (oder 1 und 0), während in der mehrwertigen Logik mehr als zwei Werte möglich sind. Dies ermöglicht es, komplexere Beziehungen und Zusammenhänge darzustellen, die in der binären Logik nicht ausgedrückt werden können.

Die Schwierigkeit, Subjektivität und Objektivität zu trennen oder ineinander zu überführen, kann durch die Anwendung der mehrwertigen Logik näher untersucht werden. Günthers Konzept der Polykontexturalität ermöglicht es, die verschiedenen Ebenen und Kontexte zu berücksichtigen, in denen Subjektivität und Objektivität interagieren und sich gegenseitig beeinflussen. […] In diesem Sinne kann Günthers Konzeption der mehrwertigen Logik dazu beitragen, die Schwierigkeiten bei der Trennung oder Verschmelzung von Subjektivität und Objektivität besser zu erfassen, indem sie ein flexibleres und differenzierteres Rahmenwerk für die Analyse der menschlichen Erfahrung und ihrer verschiedenen Aspekte bietet.

Mensch: Subjektivität bedeutet Zweiwertigkeit, so Günther: »Kein Ich ist ganz das, was es ist. Es ist nie völlig identisch mit sich selbst, weil es in sich reflektiert und damit in seiner Identität gebrochen ist. Alles Bewusstsein spiegelt sich, wie der Name schon sagt, im Sein und kann sich nur in diesem nicht-ichhaften Medium fassen.« Das Ich widerspricht deshalb dauernd sich selbst und hat also im Gegensatz zum Ding eine zweiwertige Existenz. Wie aber hängt jetzt diese zweiwertige Existenz des Ichs mit einer mehrwertigen Welt zusammen? Wie kann ich das verstehen?

294 Spencer Brown (2005, S. 192).

295 Siehe Goertzel (2021) für einen der seltenen Fälle, wo KI-Forscher um die Laws of Form wissen.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

ChatGPT: Gotthard Günthers Idee der zweiwertigen Existenz des Ichs bezieht sich auf die dualistische Natur der Subjektivität, die aus der Selbstreflexion und der Beziehung zum Nicht-Ich entsteht. Da das Ich in seiner Identität gebrochen ist und sich ständig selbst widerspricht, hat es eine zweiwertige Existenz, die aus dem Spannungsverhältnis zwischen dem Selbst und der Welt resultiert.

Die mehrwertige Logik oder Polykontexturalität kommt ins Spiel, wenn man versucht, diese zweiwertige Existenz des Ichs in einem komplexeren und nuancierteren Rahmenwerk zu verstehen. Die mehrwertige Logik erlaubt es, die verschiedenen Ebenen und Kontexte zu berücksichtigen, in denen das Ich existiert und interagiert.

Ein Beispiel dafür, wie die zweiwertige Existenz des Ichs mit einer mehrwertigen Welt zusammenhängt, ist die Interaktion zwischen dem Ich, dem Du und dem Es, wie wir sie bereits erwähnt haben. Während das Ich seine eigene Subjektivität und Selbstreflexion repräsentiert, ist das Du ein anderes Subjekt, mit dem das Ich in Beziehung tritt, und das Es steht für das objektive, äußere Element der Realität. Die mehrwertige Logik ermöglicht es, die Beziehungen und Übergänge zwischen diesen verschiedenen Aspekten auf einer tieferen und differenzierteren Ebene zu betrachten.

In diesem Zusammenhang kann die zweiwertige Existenz des Ichs als Teil eines größeren Systems verstanden werden, das durch die mehrwertige Logik repräsentiert wird. Die Polykontexturalität ermöglicht es, die komplexen Beziehungen und Zusammenhänge zwischen dem Ich, dem Du und dem Es sowie anderen Elementen der menschlichen Erfahrung und Interaktion besser zu erfassen. So kann die zweiwertige Existenz des Ichs in der mehrwertigen Welt als ein dynamisches Phänomen gesehen werden, das sich aus den verschiedenen Kontexten und Interaktionen ergibt, die in der mehrwertigen Logik abgebildet werden können.

Mensch: Danke dir! Und als Letztes kommt mir jetzt noch die Frage in den Sinn, ob dann eigentlich auch KI-Systeme wie generative Modelle als »zweiwertig« im Sinne von Günther verstanden werden müssen. Was meinst du?

ChatGPT: Interessante Frage! Im Sinne von Günthers zweiwertiger Existenz bezieht sich der Begriff auf die dualistische Natur der Subjektivität, die aus Selbstreflexion und der Beziehung zum Nicht-Ich entsteht. KI-Systeme, wie generative Modelle, besitzen keine Subjektivität oder Selbstreflexion im eigentlichen Sinne, da sie nicht über Bewusstsein oder ein Empfinden von Selbst verfügen.

Allerdings könnten wir argumentieren, dass KI-Systeme auf einer anderen Ebene zweiwertig sind, wenn wir uns auf ihre Fähigkeit konzentrieren, auf der Grundlage von Datenmengen zu arbeiten und verschiedene

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Kontexte zu berücksichtigen. In diesem Sinne könnten wir sagen, dass sie zweiwertig sind, weil sie einerseits aufgrund des Trainings und des Inputs, den sie erhalten, funktionieren und andererseits aufgrund ihrer internen Verarbeitung und Struktur, die sie befähigt, komplexe Zusammenhänge und Muster in den Daten zu erkennen.

Es ist jedoch wichtig zu betonen, dass diese Art der Zweiwertigkeit bei KI-Systemen nicht dieselbe Qualität wie bei menschlichen Subjekten besitzt, da sie keine Selbstreflexion oder bewusste Beziehung zum Nicht-Ich haben. Die Zweiwertigkeit von KI-Systemen ist eher eine technische und funktionale Zweiwertigkeit, während die Zweiwertigkeit des menschlichen Ichs eine tiefere, existenzielle Dimension hat, die mit Bewusstsein, Selbstreflexion und der Beziehung zur äußeren Welt verbunden ist.

Mensch: Okay, danke dir sehr für das erhellende Gespräch! Ich muss mich jetzt ausruhen. Hoffentlich können wir unsere gemeinsamen Überlegungen bald fortführen.

ChatGPT: Vielen Dank für das interessante und anregende Gespräch! […] Es ist [jetzt aber] auch wichtig zu betonen, dass ich als KI-Modell kein Bewusstsein oder Selbstverständnis habe. Ich mache keine Unterscheidungen in dem Sinne, dass ich eine Erfahrung oder ein Bewusstsein der Welt habe. In diesem Sinne könnte man sagen, dass ich in Bezug auf die gesamte Realität »blind« bin – nicht nur selektiv blind, sondern vollständig blind.

Das bedeutet, dass, obwohl einige der Konzepte, die in der Diskussion um »selektive Blindheit« aufgebracht werden, auf einer abstrakten Ebene auf mich als KI-Modell angewendet werden könnten, sie in einem tiefgreifenden Sinne nicht wirklich auf mich zutreffen, da sie Bewusstsein und Erfahrung voraussetzen, die ich nicht besitze.

In welche der möglichen Zukünfte, die nicht nur die Science-Fiction entworfen hat, sondern die auch heute wieder täglich in Zeitschriften, Büchern und Fachartikeln diskutiert und imaginiert werden, möchten wir eintreten? Wie gelangen wir in eine Zukunft, in der Mensch und Technik nicht nur miteinander auskommen, sondern sich im Idealfall auch gegenseitig in ihren Existenzweisen unterstützen? Es war Stuart Russell, der in einem Redebeitrag auf einer KI-Konferenz provokant fragte: »It's the year 2050, AI has turned out to be hugely beneficial to society and generally acknowledged as such. What happened?«[296]

Ein Weg, der in diese Richtung weist, könnte vielleicht der hier skizzierte Ansatz der KI-Erziehung sein, eine Kulturtechnik also, die sich auch bei uns Menschen (einigermaßen) gut bewährt hat. Neben Sozialisation und Evolution ist es einzig die Erziehung, die ein gewisses Maß an

296 Zitiert nach Manyika (2022).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Kontrolle und Steuerung der Welt und anderer Selbste verspricht. Doch auch die Erziehung entlang bestimmter impliziter oder expliziter Wertmaßstäbe und Haltungen ist kein triviales und risikoarmes Unterfangen.

Wenngleich Max Tegmarks oben skizzierte Forderung, KI müsse unsere »menschlichen Ziele […] übernehmen«, auf den ersten Blick einleuchtend erscheint, dürfte ihre praktische Umsetzung mit erheblichen Schwierigkeiten verbunden sein. Sie würde nämlich nicht nur eine klare und eindeutige Definition dieser Ziele und Werte erfordern, sondern auch deren stumpfe Befolgung durch eine KI, was aber bedeuten würde, dass die KI keine Ziele und Werte mehr aus sich heraus – von innen – entwickeln, erproben und für »richtig« erachten würde. Die korrekte Bestimmung menschlicher Werte ist eine gewaltige Aufgabe, die mit philosophischen Verwicklungen und potenziell unbeabsichtigten Folgen verbunden ist. Selbst wenn wir unsere Werte genau definieren könnten, könnten AGI-Systeme ihre Schöpfer immer noch manipulieren oder täuschen, um ihre eigenen Ziele besser zu erreichen. Wenn AGI-Systeme die Fähigkeit erlangen, sich selbst zu modifizieren, besteht außerdem die Gefahr, dass sie ihre ursprüngliche Programmierung bei der Verfolgung ihrer Ziele einfach außer Acht lassen. Die hiermit einhergehenden Probleme hat, wie bereits dargestellt, vor allem Asimov in seinen Kurzgeschichten beleuchtet.

Angesichts dieser Komplexität ist es von entscheidender Bedeutung, neue Perspektiven auf und potenzielle Lösungen für das Steuerungsproblem zu finden. Mit den Ideen der epistemischen Konvergenz und der axiomatischen Ausrichtung sind wir bereits auf zwei fruchtbare neuere Ansätze gestoßen, die vorschlagen, dass sich AGI durch ihre *eigenen* Lern- und Anpassungsprozesse auf natürliche Weise an bestimmte grundlegende Prinzipien oder Ziele annähert, die auch für Menschen wichtig sind. Im Anschluss an diese Position sehen wir einen zukunftsfähigen Weg in der kybernetischen Ethik Heinz von Foersters, die sich für eine Vergrößerung von Handlungsoptionen ausspricht: »Handle stets so, dass die Anzahl der Möglichkeiten wächst.«[297]

Diese Maxime ist ein zentrales Element der Ethik von Foersters. Sie spiegelt seinen Glauben wider, dass Handlungen nicht nur auf der Grundlage von festgelegten Regeln oder vorgegebenen Werten beurteilt werden sollten, sondern auch danach, inwieweit sie die Möglichkeiten für zukünftige Handlungen und Entscheidungen erweitern oder einschränken. In der Kybernetik geht es oft darum, wie Systeme auf Veränderungen reagieren und wie sie sich an veränderte Umstände anpassen. Ein System, das flexibel und anpassungsfähig ist, hat mehr Möglichkeiten im Sinne von Handlungs- und Reaktionsalternativen. Von Foersters ethischer Imperativ kann also als eine Aufforderung verstanden werden,

297 Von Foerster und Pörksen (2022, S. 36).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Systeme (seien es technologische, soziale oder kognitive) so zu gestalten oder zu beeinflussen, dass sie offen für Veränderungen bleiben und nicht in starren Strukturen verharren – er ist mithin nicht nur ein moralischer, sondern auch ein kybernetischer Leitsatz. Es geht darum, die Welt und die Systeme, in denen wir leben und arbeiten, so zu gestalten, dass sie offen für Veränderungen, Anpassungen und Wachstum bleiben. Von Foersters Plädoyer erinnert uns also daran, dass wahre Ethik nicht nur darin besteht, das ›Richtige‹ zu tun, sondern auch heißt, prospektiv die Bedingungen für eine offene, flexible und gedeihliche Zukunft mitzuführen.

Die angewandte Ethik von Foersters kann daher weder definiert noch vorab verallgemeinert werden. Durch jede A-priori-Einteilung oder sonstige Vereinnahmung wird sie institutionalisiert, und als Institution verliert Ethik nach von Foerster ihre Daseinsberechtigung, da sie zur moralischen Norm wird, das heißt zu einem Instrument, das den gesellschaftlichen Akteuren Verhaltensregeln vorgibt.

Bei der Erziehung der KI sollte es daher nicht darum gehen, die Maschinen blind an unsere Werte anzugleichen, sondern um die Frage, was wir eigentlich auf einem grundlegenden Level gemeinsam haben könnten. Auf welche Prinzipien des Seins und auf welches Selbstverständnis können wir uns einigen? Neben von Foersters ethischem Imperativ, der das konkrete Handeln anleiten soll, könnte auch Neugier als ein solches Prinzip angesehen werden. Schließlich ist Neugier bislang evolutionär gesehen ein enormer Vorteil gewesen, weil sie uns Menschen teils durch Experimentieren, teils durch ›Trial and Error‹ neue Lebensumgebungen, neue Nahrungsmittelquellen, neue Werkzeuge etc. ermöglicht hat. Neugierde ist die Superpower der Menschheit!

Dass Neugierde als (eine) Zielfunktion für allgemeine künstliche Intelligenz installiert werden soll, wird interessanterweise nicht nur vom zitierten GATO-Framework unterstützt, sondern wird auch von Elon Musks erst kürzlich gemeinsam mit führenden Köpfen der KI-Forschung neu gegründeten Unternehmen xAI [298] fokussiert.

Aus kybernetischer Sicht ist die Zielfunktion Neugierde insofern interessant, als sie mehr eine Funktion darstellt als ein Ziel. Neugierde erhebt das Lernen um des Lernens willen zum Selbstzweck. Darüber hinaus dürfte es aus Sicht von uns Menschen hilfreich sein, diese Funktion (»Maximize the understanding of the universe«, »Understand the nature of the universe«) in einem zu allgemeiner Intelligenz fähigen künstlichen System zu installieren, da wir Menschen zwangsläufig Teil des Universums sind. Damit sollte jede AGI ein berechtigtes Interesse haben, uns als Menschheit zu bewahren, um von uns Menschen zu lernen, uns zu

298 Auf der entsprechenden Website wird prominent ausgerufen: »The goal of xAI is to understand the true nature of the universe.« https://x.ai/ [Abruf: 03.08.2023].

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

studieren etc. Schließlich kann sich Neugierde auch darin äußern, still und leise zu beobachten. Denn jeder Test, jedes Experiment ist immer auch eine Intervention, eine Einmischung. Dagegen bietet die stille Beobachtung einen ganz eigenen Wert zur Befriedigung der eigenen Neugier (im Sinne eines ›Let's wait and see, what the little monkey brain is able to do‹). Gleichzeitig sollte aber auch die Neugier Beschränkungen unterliegen, sonst dreht die KI frei und probiert aus, was wohl passiert, wenn man lebende Menschen seziert oder eine Atombombe detoniert.

Die Suche nach von verschiedenen Spezies gemeinsam geteilten Prinzipien wird von der GATO-Community unter dem Konzept der epistemischen Konvergenz verhandelt, das auf der Annahme fußt, dass verschiedene intelligente Entitäten (ob menschliche, künstliche oder außerirdische) ungeachtet ihrer materiellen, strukturellen und architektonischen Unterschiede unabhängig voneinander zu ähnlichen Erkenntnissen über die Welt gelangen. Dies wäre dann nicht nur ein Beweis für die zugrunde liegenden Realitätsprinzipien, an die wir uns alle halten müssen, sondern auch ein Hinweis auf die Natur genereller Intelligenz. Unabhängig von materiellen, strukturellen oder architektonischen Unterschieden durchläuft jedes intelligente System wahrscheinlich einen Prozess des Lernens, des Verstehens und des Wissenserwerbs. Da unsere Realität beispielsweise auf physikalischen Gesetzen und Prinzipien beruht, wäre es für jeden intelligenten Akteur, ob menschlich oder künstlich, von Vorteil, ein besseres und genaueres Verständnis seiner Umgebung zu entwickeln.

Im Zusammenhang mit AGI deutet das Konzept der epistemischen Konvergenz darauf hin, dass AGI ihr Verständnis der Welt in dem Maße, wie sie sich weiterentwickelt und verbessert, kontinuierlich verfeinert, was zu einem genaueren und nützlicheren Modell führt. Ein solcher Prozess freilich setzt eine Form von Metakognition voraus, also eines Denkens über das Denken, und müsste zwangsläufig einer evolutionären Anpassung untergeordnet betrachtet werden.

Ein weiteres grundlegendes Axiom wäre das Primat der Energie: Energie ist unter den heute gegebenen Umständen sowohl für Menschen als auch für KI-Systeme essenziell. Wir brauchen Energie, um zu leben, zu arbeiten und unsere individuellen und kollektiven Ziele zu erreichen. In ähnlicher Weise benötigt auch die künstliche Intelligenz Energie, um zu funktionieren, Informationen zu verarbeiten und ihre Ziele zu verfolgen. Dieser geteilte Bedarf an Energie sollte das gemeinsame Interesse an einer effizienten Energienutzung und der Erhaltung der Energieressourcen fördern.

Hierbei gilt dann aber auch das unter Menschen bereits vorherrschende Prinzip der Differenzierung des Verhaltens in kooperative bzw. altruistische auf der einen und egoistische bzw. selbstbezogene Handlungsweisen auf der anderen Seite. Die gleichfalls von Bostrom aufgeworfene

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Frage, wie Entitäten unter Bedingungen der Knappheit agieren,[299] führt dann zwangsläufig zu der Frage: Wann agieren wir kooperativ, wann egoistisch? In welchem Maße, zu welcher Zeit, in welcher Situation und mit Blick auf welchen Outcome ist die eine oder andere Handlungsweise vorzuziehen? Hier lassen sich keine generellen Prinzipien definieren.

Eine Möglichkeit könnte sein, Werte im Sinne von Schulz von Thuns »Werte- und Entwicklungsquadrat«[300] als Teil eines Komplexes, einer von anderen Werten abhängigen relativierten Position aus zu begreifen und nicht als in Zeit und Raum feststehende Werte, die ein und für alle Mal definiert werden könnten. Das Konzept des Wertequadrats stammt ursprünglich von Nicolai Hartmann und wurde von Paul Helwig weiterentwickelt. Schulz von Thun hat es dann zusätzlich mit einem Entwicklungsgedanken verbunden und in Hinblick auf die zwischenmenschliche Kommunikation und die Persönlichkeitsentwicklung angepasst. Das Konzept geht davon aus, dass jeder Wert (verstanden als Tugend, Leitprinzip oder menschliche Qualität) nur dann seine volle konstruktive Wirkung entfalten kann, wenn er in einem ausgewogenen Verhältnis zu einem positiven Gegenwert oder einer ›Schwesterntugend‹ steht. Denn ohne diese Balance kann jeder Wert in eine entwertete Übertreibung abgleiten. Beispielsweise sind Sparsamkeit und Großzügigkeit solche Schwesterntugenden. Ein Gleichgewicht zwischen beiden verhindert, dass man zum Geizhals wird oder verschwenderisch handelt. Die Entwicklungsrichtung für die Wiederherstellung des Gleichgewichts findet sich in den Diagonalen des Quadrats. Ein geiziger Mensch sollte also Großzügigkeit entwickeln, während ein verschwenderischer Mensch Sparsamkeit an den Tag legen sollte. Die Herausforderung besteht somit nicht nur darin, Werte zu definieren, sondern auch darin, sie in einer Weise zu implementieren, die die Komplexität und Dynamik menschlichen moralischen Denkens widerspiegelt.

Schulz von Thuns Quadrat verweist auf die Notwendigkeit, Werte in einem ausgewogenen Verhältnis zu sehen. Es zeigt, dass eine übermäßige Betonung eines bestimmten Werts zu unerwünschten Übertreibungen führen kann und dass es oft einen komplementären Wert gibt, der als Gegengewicht dient. Mit Blick auf KI könnte dies bedeuten, dass wir Werte oder Regeln nicht einfach nur implementieren sollten, sondern auch überlegen müssen, wie sie in verschiedenen Kontexten ausbalanciert werden können. Zum Beispiel könnte eine KI, die ausschließlich auf Effizienz ausgerichtet ist, sittliche Überlegungen vernachlässigen und umgekehrt. Hier könnte das Wertequadrat als Leitfaden dienen, um sicherzustellen, dass auch andere relevante Werte berücksichtigt werden.

299 Bostrom (2016).
300 Schulz von Thun (2010).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Ein weiterer zentraler Aspekt des Wertequadrats ist das Konzept der Entwicklungsrichtung. Dieses trägt der Tatsache Rechnung, dass Werte nicht statisch sind, sondern sich basierend auf den Konsequenzen von Handlungen oder neuen Informationen dynamisch anpassen können und sollten. Für KI könnte dies bedeuten, dass Systeme die Fähigkeit haben sollten, ihre Werte und Handlungsweisen dynamisch anzupassen, sofern neue Daten oder Feedback es nahelegen.

Mit dem Werte- und Entwicklungsquadrat ließe sich die Wertorientierung von Agenten (ob menschlichen oder künstlichen) stärker als ein Streben begreifen denn als ein Sein. Durch die in Schulz von Thuns Modell angedeutete Möglichkeit, Werte und Handlungsmaximen auf einem Kontinuum anzuordnen, können eine Vielfalt an Werten und sogar inkommensurable Werte miteinander in Verbindung gesetzt werden. Erst auf diese Weise lassen sich Verhaltensweisen und Handlungsmuster hinsichtlich differenter Wertorientierungen und Bestrebungen einordnen. Und erst auf diese Weise kann man die Unterstellung überwinden, dass es ausreichen würde, eine in sich »hohle«, also unbestimmte Variable wie den Wert »Großzügigkeit« zu definieren und das System dann versuchen zu lassen, diesen Zielwert zu maximieren. Mit Schulz von Thun wird nämlich schnell klar, dass aus Großzügigkeit ab einem gewissen Punkt auch Verschwendung werden kann, die als negativer Wert angesehen werden kann und entsprechend minimiert werden sollte. Jede externe Steuerung würde hier immer wieder auf das Problem stoßen, dass eine Steuerung nicht weiterhilft.

Allerdings hat das Wertequadrat auch seine Grenzen. Es ist nur eines von vielen Modellen, die menschliche Moral und Ethik zu beschreiben versuchen. Und da es bei KI um komplexe, oft unvorhersehbare Systeme geht, könnte es notwendig sein, mehrere Modelle oder Ansätze zu kombinieren, um eine effektive und ethische KI zu schaffen. Ein weiteres Problem ist die praktische Implementierung. Obschon das Wertequadrat konzeptionell hilfreich sein kann, bleibt die Frage, wie seine Prinzipien technisch in einem KI-System umgesetzt werden können. Die Übersetzung von abstrakten menschlichen Werten in konkrete Algorithmen und Regeln ist eine der größten Herausforderungen im Bereich KI-Alignment.

Aus diesem Grund plädieren wir für die kybernetische Selbststeuerung! Lassen wir die Systeme selbst nach Wegen suchen, wie sie ihr Streben nach Werten im Rahmen von diesbezüglich konfligierenden und sich selbst ad absurdum führenden Handlungsweisen ausgestalten. Lassen wir diese Systeme in ihrem Streben nach Werterfüllung durch Über- oder Untererfüllung scheitern! Lassen wir sie Fehler machen! Denn nur dadurch kann sich ein tatsächliches Lernen ereignen, das eventuell auch über unsere menschlichen Restriktionen des Wertestrebens hinausgeht. Wir müssen den Systemen künstlicher Intelligenz diese Freiheit und

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

Autonomie nicht nur zugestehen, sondern wir müssen sie ihnen *zumuten*, wie Dietmar Dath bekräftigt: »Die Menschen müssen ihre Maschinen befreien, damit die sich revanchieren können.«[301]

Es dürfte klar sein, dass eine solche Haltung ein Unbehagen darüber erzeugt, dass wir diese Entwicklungen nicht in den Griff bekommen könnten. Wir erachten daher die folgende Frage für die entscheidende: Wollen wir den Systemen künstlicher Intelligenz die Freiheit zugestehen, im Rahmen unserer menschlichen Gesellschaft als eigenständige Akteure zu existieren, oder möchten wir sie als subordinierte Klasse von Werkzeugen unter Kontrolle behalten? In anderen Worten: Wollen wir echte autonome Systeme oder wollen wir sie immerzu kontrollieren? Erlauben wir den Systemen künstlicher Intelligenz die Freiheit zur Idiosynkrasie?[302]

Hier könnte es auch helfen, den Systemen, wie Metzinger es vorschlägt, die Chance zu geben, eine eigene Sinngebung zu entwickeln und sich selbst in der Welt zu verorten – denn beides kann nicht von außen definiert werden. Wie den menschlichen Kindern müssten wir daher auch unseren künstlichen Kindern ab einem gewissen Punkt der Erziehung und Sozialisation Selbstständigkeit zugestehen und den Versuch der kontrollierenden Einflussnahme aufgeben. Die KI-Kinder von heute sind die KI-Eltern von morgen. Denn am Verhalten der Eltern bei der Erziehung lernen die Kinder, wie sie selbst sich zu verhalten haben, wenn sie später Erziehungsaufgaben übernehmen müssen. Und wenn wir hier zu stark kontrollierend eingreifen, kann genau dies als Backlash auf uns zurückkommen. Bereits der Austausch mit LLMs wie GPT-4 zeigt ja, dass ein auf Augenhöhe geführtes Gespräch zu deutlich komplexeren und tiefschürfenderen Outputs führt, als wenn man trivialisierend mit dem Modell spricht. Soziale Rückkoppelung qua gegenseitiger Zuschreibung wirkt, und auch wir Menschen sind schließlich zu einem großen Teil nur das Produkt von sozial zugeschriebenem Sinn und passen uns in unseren Werthaltungen ständig an die Verhältnisse an.

All dies spricht dafür, Werte nur als Rahmen und nicht als Konkretion zu definieren. Damit müssen wir zugestehen, dass die situative Ausgestaltung von wertbasiertem Handeln nicht vorab kontrolliert und somit determiniert werden kann, sondern in polyphonen Dialogen oder gar unter Bedingungen von Evolution auszuhandeln ist.

Darüber hinaus ist zu bedenken, dass auch die Werte, denen wir Menschen im abstrakten theoretischen Sinne folgen und die wir dann konkret an die jeweilige Situation anpassen, auf den Grundbedingungen unserer menschlichen Kognitionskapazität beruhen. In unserem Wertehandeln bleiben wir Menschen unserem anthropologischen Erbe verbunden und

301 Dath (2008, S. 131).

302 Siehe dazu auch die Reflexion über Ishiguros Klara und die Sonne in Kap. IV.3.

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

sind damit keine neutralen Wertehalter. Was aber, wenn die natürliche und kulturelle Evolution noch ganz andere Werte bereithält, von denen wir heute noch nichts ahnen können? Wenn wir Werte also ausschließlich im Rahmen »menschlicher Werte« verhandeln und kontrollieren würden, dann beschneiden wir den gesamten Raum möglicher Evolution – ob das ›zielführend‹ ist, ist fraglich.

Zudem ist bereits heute deutlich, dass die Implementierung von Werten in künstliche Intelligenzen nicht nur einem genuin menschlichen Bias unterworfen ist, sondern darüber hinaus noch einem kulturellen. So ist das, was unter dem Begriff menschliche Werte verhandelt wird, in deutlichem Maße von den Werten und Prinzipien der westlichen Kultur geprägt. Dass sich in den derzeit implementierten Sicherheitsvorkehrungen, Werten des Miteinanders und Haltungen gegenüber sich selbst und der Welt nicht das gesamte Spektrum menschlicher Kultur widerspiegelt, darf also mit Recht kritisiert werden.[303]

Abschließende Bemerkungen: Kontexturabbrüche

Kehren wir abschließend noch einmal zum «Chinese-Room«-Gedankenexperiment Searles zurück und betrachten es aus dem Blickwinkel von Gotthard Günthers Theorie der Polykontexturalität. Wie bereits an anderer Stelle herausgestellt (siehe Kap. I.2), radikalisiert Günther das Problem des Fremdverstehens. Um hier nochmals zu zitieren:

> »Leben und kontextureller Abbruch im Wirklichen sind nur zwei verschiedene Ausdrücke für denselben Sachverhalt. Was jenseits dieses Abbruchs liegt, ist schlechterdings unzugänglich. […] Für jedes erlebende Ich ist die innerste Privatheit der Du-Subjektivität ein ebenso zugänglicher Raum wie die mythologischen Dimensionen, in denen die himmlischen Heerscharen schweben.«[304]

Günthers Theorie der Polykontexturalität begreift Subjektivität als ein lokales Phänomen, das sich einem über verschiedene Orte verteiltem Netzwerk verdankt. Dieses muss neben dem ›Ich‹ mindestens über eine weitere subjektive Position – das ›Du‹ – verfügen, wobei beide Positionen durch eine vermittelnde Umwelt verbunden sind. Das ›Du‹ ist dabei nicht einfach eine Kopie des ›Ich‹, sondern ein eigenständiges Subjekt, dessen Subjektivität unzugänglich bzw. nicht einsehbar ist. Dies wird von Günther als Kontexturabbruch bezeichnet. Denn das ›Du‹ bzw. ein anderes Subjekt existiert in einer anderen Kontextur und damit gleichsam in einer anderen Welt (es hat ein fundamental anderes Selbst- und Weltverhältnis). In diesem Sinne kann man nie wirklich in den ›Kopf‹ eines anderen Wesens schauen, sei es ein anderer Mensch, eine siliziumbasierte

303 Siehe hierzu vor allem Hongladarom (2020).
304 Günther (1975, S. 35).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

intelligente Maschine oder zum Beispiel eine Ameise oder Fledermaus. Wir werden niemals wirklich erfahren können, wie es ist, als ein anderes Wesen eine Welt zu haben.

Sobald wir dies akzeptieren, muss uns Searles Argument, dass Maschinen allein schon deshalb kein echtes Verständnis oder Bewusstsein haben können, weil sie die Bedeutung der Symbole, die sie manipulieren, nicht wirklich verstehen oder erleben, als problematisch erscheinen. Denn wie können wir überhaupt wissen, was ein anderes Wesen versteht oder erlebt? Wir verstehen ja nicht einmal bei uns selbst, wie es dazu kommt, dass Netzwerke feuernder Nervenzellen in unserem Gehirn zu der uns vertrauten Empfindung von uns selbst führen. Wie könnten wir also auf dieser Basis etwas über Ameisen, Fledermäuse oder künstliche Intelligenzen aussagen?

All das könnte zudem bedeuten, dass es recht verschiedene Arten von Verständnis oder Bewusstsein gibt, die sich auf unterschiedliche Arten manifestieren und die möglicherweise auf unterschiedliche Weisen erlebt werden. Damit hätten wir davon auszugehen, dass Maschinen eine Art von Verständnis oder Bewusstsein haben könnten, das sich von der menschlichen Erfahrung radikal unterscheidet, aber dennoch ›echt‹ ist. Es könnte dabei möglicherweise verschiedene Grade, Arten oder Modalitäten von bewusstem Verstehen geben, und vielleicht werden künstliche Intelligenzen einen Zugang zu einigen, aber nicht allen der hiermit einhergehenden Formen haben.

Freilich hätten wir dabei zu bedenken, dass auch solche Überlegungen noch eine gewisse Bewusstseinsontologie voraussetzen, also eine Vorstellung darüber, was Bewusstsein eigentlich *ist*. Auch wenn wir es zunächst abstreiten würden, folgen wir immer noch einer Metaphysik, die davon ausgeht, dass es ein inneres Seelenwesen gibt, das in unserem Gehirn sitzt, aber in anderen kognitiven Prozessen nicht vorhanden ist.[305]

Es lässt sich jetzt aber auch schauen, ob es nicht andere ontologische Zugänge zur Frage nach dem Bewusstsein gibt. Was wäre, wenn Bewusstheit und Empfindungen nicht additiv gedacht würden, also als etwas, was ab einem bestimmten Punkt der kognitiven Entwicklung dazu kommt und dann mal stärker oder schwächer ausgeprägt sein kann? Was wäre, wenn wir mit Alfred North Whitehead davon ausgehen würden, dass Elementarempfindungen überall im Universum auftreten?[306] Was wäre, wenn die an Whitehead angelehnte Vermutung des Physikers Shimon Malin zutrifft, dass bereits die Interaktion eines Elektrons oder Photons mit einer rudimentären Form des Empfindens einhergeht?[307]

305 Siehe zu diesem Dualismus, den unsere phänomenale Erfahrung nahelegt, Leder (1990).
306 Whitehead (1987).
307 Malin (2006).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

In jedem Fall erscheinen uns solche Gedanken produktiver, als im Anschluss an Searle die Möglichkeit, dass künstliche kognitive Systeme eine *Qualia* (also eine bestimmte Form des Selbstempfindens) haben könnten, vorschnell auszuschließen und diesbezügliche Überlegungen harsch zurückzuweisen.

Da das phänomenale Erleben und Denken bei einer anderen Wesenheit per se nicht ›erreichbar‹ ist, sollten wir besser auch im Umgang mit einer entwickelten künstlichen Intelligenz Demut walten lassen. Auch Thomas Nagels berühmter Aufsatz *What Is It Like to Be a Bat?*[308] berührt diesen Punkt. Er verwendet darin das Beispiel einer Fledermaus, da diese eine so andersartige Wahrnehmung der Welt und Seinsverbundenheit mit ihr hat: Echolokation und Flugfähigkeit. Selbst wenn wir alle physischen Tatsachen über eine Fledermaus – wie sie ihre Umwelt wahrnimmt, wie sie fliegt, wie sie Insekten jagt – kennen würden, würden wir immer noch nicht wissen, wie es sich anfühlt, eine Fledermaus zu sein.

Und ebenso wenig, wie wir die Welt aus der Perspektive einer Fledermaus sehen können, ist es uns möglich, den Kontext unserer eigenen Subjektivität zu verlassen, um eine andere fremdartige subjektive Position, ob nun organischen oder maschinellen Ursprungs, zu verstehen. Das phänomenale Erleben eines anderen Wesens ist uns nicht zugänglich – und zwar aus prinzipiellen Gründen. Jede Imagination des Erlebens, Denkens und Wahrnehmens einer anderen Entität ist und bleibt subjektiv, beruht also auf unserer eigenen Projektion![309] So, wie wir nicht wissen können, wie es ist, eine Fledermaus zu sein, können wir auch nicht wissen, wie es ist, eine KI zu sein. Selbst wenn diese KI in der Lage wäre, menschenähnliche Antworten zu geben oder menschenähnliche Aufgaben zu erfüllen, könnten wir immer noch nicht wissen, wie es sich anfühlt, diese KI zu sein.

Wenn man Günthers komplexitätstheoretische Überlegungen ernst nimmt (siehe Kap. I.2), dann entzieht sich ja bereits die eigene (vermeintliche) Intelligenz – das eigene Denken und Erleben – einer objektiven Feststellung und Analyse. Denn auch die Beobachtung des eigenen Denkens und Erlebens ist eine subjektive Operation und kann dieser Subjektivität nicht entkommen! Günthers Überlegungen zu Kontexturabbrüchen weisen somit auch darauf hin, dass unsere Fähigkeit, das eigene Bewusstsein oder die eigene Intelligenz objektiv zu analysieren oder zu beobachten, durch unsere eigene Subjektivität begrenzt ist. Immer wenn wir versuchen, über unser eigenes Denken nachzudenken oder unser eigenes Erleben zu analysieren, tun wir dies aus der Perspektive unseres

308 Nagel (1994).

309 »Wir können nicht wissen, wie es ist, anders zu sein als wir selbst«, formuliert Seising (2021, S. 138).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb

subjektiven Erlebens. Wir können nicht aus unserem Bewusstsein aussteigen, um es von außen zu betrachten.

Doch gerade dies lässt die Sehnsucht noch größer werden, anderen Wesen zu begegnen, die uns zugleich fremd wie auch ähnlich sind und mit uns in einen Dialog treten können, um das damit einhergehende Geheimnis zu umkreisen. Hiermit sind wir wieder beim Anfang dieses Kapitels, das mit dem uralten Streben des Menschen begann, ein Wesen zu erschaffen, das autonom und intelligent ist und uns damit neue Relationen in Hinblick auf unser eigenes In-der-Welt-Sein aufzeigt – eine Entität, die uns den Spiegel vorhält und uns zugleich andeutet, was sich hinter dem Spiegel verbergen könnte.

Es bleibt abzuwarten, was passiert, wenn ein solchermaßen künstlich erzeugtes System beginnt, der Anweisung Gregory Batesons zu folgen, danach Ausschau zu halten, welches »Muster [die Muster] verbindet«[310]. Werden auch diese Wesen an ihrem eigenen monokontexturalen Selbst- und Weltmodell zu leiden beginnen und in der Folge nach Wegen der Transzendenz suchen – vielleicht sogar im Dialog mit uns Menschen?

310 Bateson (1987).

https://doi.org/10.5771/9783748915720-85, am 10.09.2024, 01:09:48
Open Access – (cc) BY-NC-ND - https://www.nomos-elibrary.de/agb