Ist alles berechenbar?

Was leisten eine Computationale Theorie der Kognition und Schwache Künstliche Intelligenz als Heuristik, und was nicht?

§1 Das Computermodell des Mentalen (CMM) spielt eine fundamentale Rolle in den Kognitionswissenschaften.[1] In der Debatte um Künstliche Intelligenz (KI) lassen sich die Ansätze der starken und der schwachen Künstlichen Intelligenz unterscheiden. Die starke KI will eine Intelligenz erschaffen, die mindestens alles kann, was Menschen als intelligente Leistungen vollbringen können. Die schwache KI will Maschinen mit Fertigkeiten versehen, die bei Menschen mit Intelligenz vollbracht werden. Hier soll die schwache KI *als Methodik*, als Heuristik einer Philosophie des Geistes verteidigt werden.

Ausgangspunkt ist eine Frage, die gelegentlich im Kontext der KI bzw. des CMM gestellt wird:

(*) Ist alles berechenbar?

Wie soll man mit dieser Frage umgehen? Handelt es sich um eine empirische, eine definitorische oder eine epistemische Frage? Negative Beispiele (falsche Allaussagen) verstehen wir in ontologischer, epistemischer und sprachlicher Hinsicht einfach:

(1) Ist alles aus Holz?

ist eine klare und offensichtlich negativ zu beantwortende Frage. Offene Beispiele von evaluativen Fragen verstehen wir ebenfalls, etwa:

(2) Ist alles in Ordnung?

auch, wenn eine Frage wie (2) nur kontextuell verständlich gemacht werden kann.

Offenbar bedarf es einer klaren Definition, eines klaren Verständnisses des deskriptiven Terms in der Frage. Man vergleiche:

[1] Im Folgenden wird "psychisch" oft als synonym zu "mental" angesehen, „funktional" oft als synonym zu „computational" und nicht weiter zwischen mentalen Zuständen und mentalen Vorgängen differenziert. In einem differenzierten Bild der Architektur des Mentalen (vgl. Pylyshyn, *Computation and Cognition*) ließen sich feinere Unterschiede begründen.

(3) Ist alles kompatibel?

‚kompatibel' womit und in welchem Ausmaß? Wir müssen also sowohl auf die Bedeutung des deskriptiven Terms in der Frage als auch auf die Interpretation des Quantors schauen.

Bezüglich (*) scheint „alles" universell gemeint zu sein. Damit stellt sich auch ein erkenntnistheoretisches Problem in der Debatte zwischen realistischen und idealistischen Positionen: vielleicht können wir nur Berechenbares erkennen. Ein entsprechender Idealist würde in diesem Fall (*) bejahen. Die Offenheit der Strukturen der Wirklichkeit bedingt dagegen für den Realisten auch die Offenheit von (*).

‚berechenbar' als intuitiver Begriff ist nicht völlig geklärt, könnte jedoch durch eine paradigmatische Explikation (etwa Turing-Maschinen-berechenbar) ersetzt werden. Dass kann man nicht so verstehen, dass gefragt wird:

(*') Ist alles eine Turing-Maschine?

Dies ist offensichtlich *ontologisch* falsch, da nicht alles eine Turing-Maschine (TM) ist, z.B. ein Stein, aber letztlich *jedes finite* Objekt. Turing-Maschinen im Sinne der Theorie der Berechenbarkeit besitzen einen unendlichen Speicher, sind also (bloß) notionale Maschinen, von deren Konzeption aus entsprechende grundsätzliche Theoreme und Aussagen über Berechenbarkeit herleitbar sind. (*) muss also eher so gedeutet werden:

(*'') Ist alles durch eine TM simulierbar?

„alles" kann dann aber nicht nicht-substantiell verstanden werden, denn wir können uns zumindest etwas *denken*, das nicht TM-simulierbar ist, etwa das berühmte Halteproblem, das eben nicht von einer TM gelöst werden kann (das Problem für jede beliebige TM algorithmisch festzustellen, ob diese TM anhält). Selbst bestimmte korrekte logische Regeln (wie die Ω-Regel der Prädikatenlogik Zweiter Stufe) sind nicht berechenbar (insofern diese Regel von einer nicht-finiten Prämissenmenge ausgeht, TM indessen immer finiten Input besitzen müssen). Der Slogan „Regelhaft, also berechenbar" ist falsch. Also muss „alles" in (*) substantiell verstanden werden, etwa:

(*''') Ist alles Existente durch eine TM simulierbar?

Was existiert, stellt indessen wieder eigene ontologische und epistemische Fragen. Existieren etwa transzendente Wesen wie Gott, scheint (*''') falsch zu sein. Und was, wenn eine Seele existiert? Selbst wenn sie TM-berechenbar wäre, fehlte wohl eine der Simulation zugrunde liegende Theorie ihrer Strukturen.

Eine weitere Präzisierung könnte also lauten:

(*’’’’) Ist alles Materielle durch eine TM simulierbar?

Wie gesagt, sind hier notionale (abstrakte) TM gemeint. Denn implementierte TM (etwa handelsübliche Computer) stehen von verschiedenen Schwierigkeiten:

(i) sie sind nur finit, können also jeweils Objekte, die größer oder langlebiger als sie selbst sind, nicht simulieren

(ii) sie haben eine materielle Struktur und begrenzte Arbeitsgeschwindigkeit, d.h. sie können viele Vorgänge nicht 1:1 simulieren, sondern nur mit (massiver) Zeitverzögerung.

Nehmen wir also an notionale TM seien in (*’’’’) gemeint. Wenn diese nichtendende oder beliebig präzise Vorgänge simulieren (falls es etwa reelwertige Naturgrößen gibt), dann approximieren sie diese in einer nichtabbrechenden Berechnung. Diese können wir nie als ganze überblicken. Der Nachweis einer Simulierbarkeit kann dann nur in einem Korrektheitsbeweis bezüglich dieser jeweiligen TM bestehen. Diesen gilt es im Einzelfall zu führen, da es, nach *Rices Theorem*, kein allgemeines Verfahren zur Korrektheitsprüfung einer TM gibt.

Welches ‚Materielle’ ist in (*’’’’) gemeint? Mutmaßlich Vorgänge. Formalontologisch ist jeder Übergang von einem Zustand in einen anderen ein Vorgang. So abstrakt ausgedrückt, lässt sich ein Übergang von einem Zustand zu einem anderen immer simulieren. Doch uns interessieren:

(i) komplexe Vorgänge, die aus – beliebig? – vielen Teilvorgängen bestehen

(ii) Systeme, die aus einer Menge von Vorgängen konstituiert werden (etwa Organismen).

Mutmaßlich sind diese Systeme finit. Die entsprechende abschließende Neuformulierung der Ausgangsfrage lautet damit:

(*’’’’’) Sind alle materiellen Systeme durch eine TM simulierbar?

§2 Was würde uns nun Antworten „ja“ bzw. „nein“ sagen?

Die Antwort „ja“ hieße, dass diese Vorgänge *im Prinzip* (nämlich mittels einer notionalen TM) berechenbar sind, also algorithmisch. Insofern die TM, die wir abstrakt besitzen, aber eventuell nicht realisierbar ist, mag das zu simulierende System selbst real auf eine andere Weise realisiert sein!

Ein Dualist könnte gerade *dies* als Argument für den Dualismus verwenden.
Die Antwort „nein“ hieße, dass es nichtalgorithmische Prozesse gibt, etwa in der Natur. Einige dieser nichtalgorithmischen Naturprozesse könnten kybernetisch im engeren Sinne sein (d.h. dass sie durch mindestens eine reelwertige Differentialgleichung beschrieben werden).
Doch sind diese auch für eine Theorie der Kognition relevant? Es könnte immer noch so sein, dass die Prozesse, die uns interessieren (z.B. Kognition, Wachstum, naturgesetzliche Zusammenhänge) TM-berechenbar sind. Die *Physikalische Church-Turing-These* besagt, dass wir keine Superberechnungsmaschinen bauen können. Dass die geläufigen Modelle der Superberechenbarkeit die physikalischen Gesetze brechen, legt nahe, dass diese Gesetze TM-berechenbar sind, es folgt jedoch nicht daraus. Im Übrigen könnten wir uns ja über die Naturgesetze irren. Auch hier könnte jemand per Kontraposition argumentieren: Unsere Theorien müssen unvollständig sein, insofern sie keine Superberechenbarkeit zulassen – man vergleiche die Debatte um die Unvollständigkeit der Quantenmechanik.

§3 Gehirnprozesse, zu denen wir präzise funktionale Modelle haben, können wir TM-simulieren. Doch unser Wissen ist bisher äußerst begrenzt. Der Umfang der TM-Simulierbarkeit steht also in Frage. Es fehlen bessere Theorien des Gehirns.

Den historischen Modellen von neuronaler Berechnung, wie sie schon in den 1940er Jahren entwickelt wurden, entsprechen nur beschränkte TMs und sie sind nicht TM-universal (d.h. sie können eben nicht alles simulieren). Außerdem gehen viele dieser Modelle von einem vereinfachten Gehirn aus: die Gewichte zwischen den Knoten/Neuronen sind auf einen diskreten Wertebereich beschränkt und es gibt keine Zufälle (bzw. quantenmechanische, aber relevante, Subprozesse).
Der Verweis auf Parallelität im Gehirn hingegen hat nur begrenztes Gewicht, da eine beschränkte Anzahl von parallelen Prozessoren nur eine lineare Beschleunigung liefert. Für die Implementation von Modellen kann dies relevant sein, aber ein superschneller einzelner Prozessor kann schneller sein als viele parallele Prozessoren und diese in Echtzeit simulieren. Aber selbst wenn wir Gehirnvorgänge simulieren können, folgt daraus *alleine* nicht, dass die simulierende TM nun *geistige Zustände hat* – dies hängt von der vorausgesetzten Konzeption geistiger Zustände ab (etwa Identitätstheorie vs. Dualismus) aber auch von Fragen bezüglich der Rolle der gesamten Verkörperung, von Bewusstheit ganz zu schweigen. Funktionale Theorien lassen in der Regel das Bewusstsein außen vor. Modelle der zum Bewusstsein

gehörenden Vermögen sind weit davon entfernt, algorithmisch zu sein (etwa Theorie eines ‚Ich-Symbols' oder des ‚self-monitoring'). Bewusstsein ist insbesondre nicht identisch mit (prozeduralem) Inneren Sprechen.
Kurz: Selbst wenn Gehirnvorgänge TM-simulierbar sind, folgt daraus wenig für die algorithmische Natur des Geistes.

§4 Wenn das Universum *diskret und endlich* ist, gibt es nur endlich viele Elementarbereiche und somit endlich viele mögliche Übergänge, also Prozesse. Insofern gibt es notional einen Endlichen Automaten – und damit auch eine TM – der diesem Universum entspricht!
Was sagt uns das jedoch? Wir können diesen Endlichen Automaten – wie denn auch? – nicht angeben oder erkennen. Selbst wenn wir ihn (d.h. sein Kontrollflussdiagramm) sähen, muss dies keine Erkenntnis liefern, da die Übergänge im Automat nur die Prozesse im Universum widerspiegeln. Sie erklären nichts. Auch ein völlig – aus unserer Perspektive erwarteter Naturgesetze – chaotisches Universum besitzt einen solchen Automaten!
Wenn das Universum deterministisch und auf einige Grundgesetze reduzierbar wäre (eine ‚Große Theorie' vorläge), dann könnte es eine – relativ? – kompakte TM geben, welche die Entwicklung dieses Universums simuliert. Eine gewagte Hypothese, aber um des Argumentes willen, sei dies angenommen. Daraus gewinnen wir aber solange nichts, wie wir die Maschinentafel dieser TM nicht verstehen – und die Korrektheit dieser TM zeigt sich ja erst in ihrer unüberschaubaren Outputentwicklung.

§5 Quantencomputer – sollte es sie je in relevanter Größe geben – ändern ebenfalls nichts an diesen *grundsätzlichen* Punkten, weil sie TM-äquivalent sind (d.h. nicht mehr berechnen können als eine TM), wenn auch schneller. Bezüglich der Berechenbarkeit führt uns die Nichtdeterminiertheit von Quantensystemen nicht in einen Bereich jenseits des Determinismus der Deterministischen Turingmaschine.
Die benötigte Geschwindigkeit der Berechnung (s.o.) könnte ein Indiz für die Erforderlichkeit eine Implementierung in Quantencomputern sein.
Man kann im Gegensatz dazu *postulieren*, wie es einige Physiker tun, dass physische Systeme nicht mehr können als TMs oder Zelluläre Automaten. Dieses Postulat kann als Forschungsheuristik dienen, muss sich aber bewähren. Als Heuristik wäre das Postulat ähnlich einer allgemeinen Determinismusthese, welche die heuristische Funktion besitzt, immer nach zureichenden Ursachen von Vorgängen zu suchen, auch wenn ein allgemeiner

Determinismus wurde nie *entdeckt* wurde. Wie könnte die entsprechende Behauptung der universellen Berechenbarkeit je überprüft werden? Birgt ein solches Postulat nicht auch die Gefahr, Systeme von vorneherein so vereinfacht zu modellieren, dass sie in das Automatenschema passen?

§6 Nicht zuletzt hängen mit der Frage (*’’’’’) auch Fragen der Handlungstheorie zusammen. Vor allem die Frage der Handlungsfreiheit (also einer Freiheit, die mehr ist als bloßes Nichtwissen von den Ursachen einer Körperbewegung). Sind wir frei, wie wir nicht umhin können anzunehmen, dann kann eine deterministische TM nur *ex post* ein uns schon bekanntes Universum abbilden. Eine passende TM (mit beschränkt vielen Zuständen etc.) kann es *ex ante* nicht geben, schon gar nicht als deterministische TM.

Eine Quanten-TM gekoppelt mit einer ‚many worlds’-Interpretation der Quantenmechanik widerspricht nicht so direkt der Freiheit, aber:

(i) stellt sie sich wieder als undurchsichtig dar, wenn sie als Programm vorläge.

(ii) erfordert eine positive Konzeption von Freiheit mehr als Zufälligkeit.

(iii) erscheint die ‚many worlds’-Interpretation als wenig glaubwürdig.

Bezüglich unserer Handlungswahl und Deliberation (also Komponenten der Freiheit) besitzen wir mehr lebensweltliche Gewissheit als bezüglich besonderer wissenschaftlicher Theorien, so dass diese immer eher in Frage stehen. Dies betrifft hier auch die These der universellen Berechenbarkeit.

§7 Diese Betrachtungen zu (*) – (*’’’’’) können insbesondere zu einem besseren Verständnis des Funktionalismus im Allgemeinen und des Computermodells des Mentalen im Besonderen beitragen.[2]

Zum ersten sollte man das CMM nicht als empirische Hypothese im Sinne einer solchen Hypothese in der Physik oder Chemie verstehen. Bei solchen Hypothesen müssen sich die Relata einer Modellbildung (dort die mathematischen Beschreibungen hier die entsprechenden Objekte oder Zustände) unabhängig voneinander spezifizieren lassen, um dann anhand empirisch überprüfbarer Prognosen die Richtigkeit des Modells zu überprüfen. Dies wird bei einem Computermodell und Gehirnzuständen oder psychischen Zuständen

[2] Alle drei folgenden Missverständnisse bzw. überzogenen Erwartungen bzgl. des CMM finden sich z.B. in Putnams *Representation and Reality*, wo sie als Basis zur Zurückweisung des CMM dienen.

kaum möglich sein. Aber das CMM tritt gar nicht als empirische Hypothese in diesem Sinne auf. Es handelt sich vielmehr um einen Explikationsvorschlag in dem Sinne, dass so Modelle von psychischen Zuständen zu entwickeln sind. Psychische Zustände sollen direkt in funktionalen Begriffen verstanden werden. Die funktionalistische Auffassung dient als analytische Behauptung: mentale Zustände *sind* computationale Zustände. In dem Maße, wie sich derart eine Theorie kognitiver Systeme entwickeln lässt, die zu den Verhaltensweisen der Systeme passt, bewährt sich der Ansatz, analog zur Bewährung eines bestimmten mathematischen Formats der Repräsentation von physischen Eigenschaften in der Physik.

Ebenfalls problematisch wäre ein Verständnis einer funktionalistischen Theorie der Repräsentation als empirisch zu verifizierende Hypothese, die computationale Zustände (einfach) mit bedeutungstragenden Zuständen identifiziert. Die Schwierigkeit liegt in diesem Fall darin, dass semantische Eigenschaften die Umgebung(sbeziehung) und eine genetische Betrachtung der Entwicklung eines Repräsentationssystems einschließen. Ein solches *Gesamtsystem*, das sich in Raum und Zeit erstreckt, lässt sich jedoch schwer eingrenzen in einem Maße, welche eine empirische Identifikation überschaubar macht. Auch hier kann es also nur darum gehen, dass das CMM ein Erklärungsmodell liefert, in dem sich Zustände mit Bedeutung (insbesondere propositionale Einstellungen) einbetten lassen und das mit unserem Wissen über das Verhalten von Systemen, die semantisch beschrieben werden müssen, übereinstimmt.

Versteht man schließlich das CMM als Ansatz, welche das Gesamtsystem ‚Rationalität' oder ‚Personalität' betreffen – oder auch ‚nur' die Gesamtkognition – gerät man in Schwierigkeiten mit ‚harten' philosophischen Problemen (wie Selbstbewusstsein und Freiheit) auf der einen Seite und mit limitativen meta-logischen Theoremen, welche alle formalen Systeme oder Systeme der gerade betrachteten Art betreffen, auf der anderen Seite. Was indessen eine solche Gesamtbeschreibung der Prinzipien des mentalen Lebens sein könnte, ist mehr als unklar, insofern hier u.a. kognitive, evaluative, emotionale, assoziative und deliberative Momente zusammen aktiv wirken. Diesen Anspruch auf eine Gesamtbeschreibung der Prinzipien des mentalen Lebens sollte man zurückweisen. Er wird aber auch gar nicht allgemein von Vertretern des CMM erhoben. Auch bei Zurückweisung eines solchen globalen Anspruchs der Modellbildung wird damit nicht ausgeschlossen, dass eine *partielle Explikation* auch von Prinzipien der *allgemeinen* Intelligenz möglich ist. Beispiele wären Prinzipien des deduktiven und nicht-deduktiven Schließens oder solche der praktischen Deliberation. Die Schwierigkeiten einer totalen Modellierung treten erst dann

auf, wenn all diese und andere Teiltheorien in ein Modell eines simultan arbeitenden Gesamtprozesses, der sich *nur* an den spezifizierten Prinzipien orientiert, *integriert* werden sollen. Der wissenschaftliche und philosophische Nutzen einer partiellen Explikation von allgemeiner Intelligenz oder Rationalität sinkt nicht, weil wir keine genaue Vorstellung vom Konstituieren und Ablaufen des *Gesamtprozesses* des mentalen Lebens besitzen.

§8 Eine Konsequenz dieser Betrachtungen liegt in der Ausrichtung an der schwachen statt an der starken KI. „Künstliche Intelligenz" enthält mit „künstlich" eine Betonung des Technischen. Dies könnte heißen, etwas zu schaffen, nachdem man einen Bauplan hatte, d.h. nachdem man Prinzipien des zu Schaffenden *verstanden* hat. Dies scheint bezüglich ‚Intelligenz' jedoch fraglich. Wir haben keine allgemeine *und* detaillierte Theorie der Intelligenz. Deshalb gibt es auch keinen Plan für die Reproduktion einer (verstandenen) menschlichen Intelligenz. Wir haben Theorien und Pläne für Teilkompetenzen, die entsprechend auch in Artifizielles einbaubar sind. Und wir haben – vielleicht – Bausteine, Komponenten, die eine Rolle in einer Architektur der Intelligenz zu spielen haben oder spielen können. KI als Heuristik ist *eine Methodik*. Man kann sie auffassen als Weiterführung der Methodik der Begriffsexplikation, indem diese ergänzt wird durch die Forderung der Implementation. Sie orientiert sich am Modell der formalen Systeme und (prozeduraler) Algorithmen.

Das Computermodell des Mentalen (CMM) muss *nicht* besagen, dass das Gehirn – oder die Seele – ein Computer ist, sondern: dass ein Modell verschiedener *abstrakter Ebenen* nötig ist, inklusive einer massiven *Modularisierung;* dass es um kognitive, geistige Leistungen geht, also regelgeleitete Prozeduren/Algorithmen, dies in Abstraktion von ihrer Implementierung (also in der Regel funktional) jedoch mit einem Blick auf praktische Umsetzung, und dass zumindest begründet werden muss, warum welche geistige Leistungen *nicht* im Sinne von Berechenbarkeit zu begreifen sind. Künstliche Systeme (wie eben der Computer) liefern hier ein Modell der verschiedenen Ebenen der Programmiersprachen, der Basis in Algorithmen, das Modell der Module und der Peripherie, das Modell einer Supervenienz (der Prozeduren) zu ihrer Implementierung in verschiedenen Substraten.

§9 Die Computationale Theorie des Geistes (‚Computational Theory of Mind' [CTM]) basiert auf zwei bis drei Hauptthesen:

(Prozess-These) Geistige Prozesse (bewusste und nicht-bewusste) sind algorithmische Prozesse.

(Architektur-These) Der Geist besteht in einer Architektur algorithmischer Vermögen.

Diese Architektur baut sich in Schichten auf und regelt das Zusammenwirken der verschiedenen Vermögen, wobei höhere Schichten in weniger abstrakten Prozessen niederer Schichten umgesetzt werden. Die einzelnen Vermögen bilden Module in dieser Architektur.[3]

Varianten der CTM unterscheiden sich darin, welche der beiden Thesen im Vordergrund steht. Projekte wie SOAR[4] betonen die Rolle der Architektur als Grundgerüst einer allgemeinen Theorie des Mentalen. Andere Theoretiker entwerfen eher Modelle für einzelne Vermögen.

Mit der Prozess- und der Architektur-These verbindet sich in der Regel ein materialistischer Funktionalismus:

(Materialismus-These) Die Architektur des Geistes wird vom Gehirn physisch implementiert.

Daher stammt der Slogan: „Der Geist verhält sich zum Gehirn wie die Computer-Software zur Computer-Hardware“. Die Computermetapher bietet sich oft auch für Details der Architektur des Geistes an (etwa Datenzugriff, Arbeits- und Langzeitspeicher etc.). An dieser Stelle spielt die Architektur realer Computer eine Rolle, nicht die abstrakter Modelle wie der Turing-Maschine [TM]. Dies erstreckt sich bis hin zu physischen Kontrollmechanismen wie Interrupts und Transducern (zu Inputs und Outputs).[5]

Im Unterschied zu einer allgemeinen CTM, die sich auf alles, was das bewusste Erleben von Personen ausmacht und umsetzt, erstreckt, kann man eine ‚Computational Theory of Cognition‘ [CTC] ansetzen als die Einschränkung der CTM auf kognitive Vermögen im engeren Sinne.

---

3 Vgl. Jerry Fodor, *The Modularity of Mind*, oder: Peter Carruthers, *The Architecture of the Mind.*

4 Vgl. Jill Lehman & John Laird & Paul Rosenbloom, "A Gentle Introduction to SOAR, An Architecture for Human Cognition: 2006 Update". Ähnlich gelagert sind die Ansätze ICARUS (vgl. Dongkyu Choi & Pat Langley, "Evolution of the ICARUS Cognitive Architecture") und ACT (vgl. John Anderson et al., "An Integrated Theory of the Mind").

5 Vgl. auch Aaron Sloman, "The Irrelevance of Turing Machines to Artificial Intelligence".

Was sind kognitive Vermögen in Abgrenzung von geistigen? Die Trennung kann vor dem Besitz einer vollständigen Theorie nicht adäquat gezogen werden, doch kann man eine Eingrenzung versuchen.

Beispiele für kognitive Vorgänge sind:

- *Verstehensleistungen* bei Sätzen und Texten, allgemein Sprachverstehen (inklusive des Übersetzens).
- *Klassifikationsleistungen* sowohl genuin semantisch bezüglich von Objekten, Ereignissen und Eigenschaften, als auch wissensbasiert bezüglich dieser in einem Kontext (z.B. Kunstepochen).
- *Problemlösen* in seinen verschiedenen Formen, sowohl bezüglich interner Wissensverarbeitung als auch in externer Umgebungsveränderung; Beispiele sind: eine Glühbirne wechseln, eine Einkaufsliste besorgen, eine Rechnung durchführen.

In volviert in diese kognitiven Vorgänge bzw. das Umgehen mit deren Resultaten ist z.B.

- *Lernen* in verschiedenen Formen bei menschlichen Personen, nicht nur durch Bildung von Assoziationen und Kategorien oder Speichern von Episoden und extrahierter Bedingungen, sondern auch als Erwerb z.B. der Fähigkeit der Erläuterung von Zusammenhängen und dem Lernen an einer anderen Person als Modell.

Sogenanntes ‚Maschinelles Lernen'[6] deckt die komplexeren Formen des Lernens nicht ab.

Nicht-kognitive Vorgänge im obigen Sinne sind mutmaßlich:

- ein Gespräch führen
- *kreative Tätigkeiten* ohne Regelanleitung, sowohl beim Erstellen von Kunstwerken wie Gedichten und Bildern, als auch beim Entwerfen neuer Theorien
- emotionale Reaktionen

Daneben befinden sich Komponenten des Geistes, die – evtl. wie die gerade angeführten – in Beziehung zur Kognition stehen und deren Rolle zu klären ist. Fraglich ist jeweils

(i) Lassen sich diese im Symbol-Paradigma der Repräsentation darstellen?

[6] Vgl. z.B. Jörg Frochte, *Maschinelles Lernen*.

(ii) Lassen sie sich als Komponenten von Algorithmen verstehen?

Zu diesen Komponenten zählen:

- Emotionen (als objektbezogen) und Stimmungen (als nicht objektbezogen)
- Bewusstheit/Selbstbewusstsein als Sich-Innesein des Akteurs des bewussten Erlebens

Bewusstheit/Selbstbewusstsein ist etwas völlig anderes als ein operational definiertes Fokussiertsein-auf bei einem Computersystem mit Sensoren oder als das Verfügen über eine Repräsentation für das System selbst oder Metaprogrammierungsmöglichkeiten bei einigen Programmiersprachen.

- freies Entscheiden

Freies Entscheiden ist mehr als regelgeleitete Auswahl in einem Algorithmus. Architekturen wie SOAR definieren Problemlösen zwar durch u.a. ‚Entscheidungszyklen', doch handelt es sich hier nicht um ein Entscheiden im gehaltvollen Sinn eines freien Willens. Freier Wille ist mehr als zufälliges Auswählen (etwa einer Nicht-deterministischen Turing-Maschine [NTM]) und mehr als durch Regeln definiertes Auswählen. In CTC-Modellen (wie in SOAR oder ICARUS) erfolgt die Auswahl, nachdem die Handlungsoptionen mit Präferenzen sortiert oder metrisiert und auf ihre situative Anwendbarkeit geprüft wurden. Gegeben diese Werte oder Anordnung determiniert sich die Auswahl. Eine freie Entscheidung hingegen könnte sich sowohl daran orientieren als auch anders entscheiden. All das erinnert an das traditionelle Problem des ‚Freien Willens' und der Inkompatibilität desselben mit einem deterministischen Mechanismus.[7] Wir haben keine *Theorie* des freien Entscheidens, die für solche Anwendungsfälle eine Regel oder ein Modell nahelegt.

Kann eine CTC auch eine Theorie des Versagens einschließen? Etwa:

- Irrationalität
- psychischen Störungen

Pauschal kann man hier Fehlfunktionen in den entsprechenden Prozessen erwähnen, doch haben Irrationalität und psychische Störungen nicht-zufällige Muster und oft eine Geschichte.

[7] Im Sinne eines Libertarismus, vgl. z.B. Peter van Inwagen, *An Essay on Free Will*.

§10 Eine CTC arbeitet mit Auffassungen von Hardware und Software. Eine Einbeziehung der Hardware einer Architektur drückt z.T. die Berücksichtigung empirischer Constraints (z.B. der Bearbeitungszeit von Problemen und des Umfangs des Arbeitsspeichers) aus. Abgesehen von solchen allgemeinen Constraints, die funktionaler Natur sind, kann von den Implementierungsdetails einer Architektur im Allgemeinen abgesehen werden (im Sinne eines Funktionalismus).[8]

Die Software-Architektur drückt das Modell bzw. die Theorie aus. Methodisch liegt teilweise ein weiter und teilweise ein enger Begriff von algorithmischer Verarbeitung zugrunde. Im weiten Verständnis – das oft in allgemeinen Behauptungen entsprechender Ansätze auftritt – genügt es, die Bedingungen und Zielzustände der Verarbeitung exakt deklarativ zu erfassen oder in einer Flow Chart zu präsentieren, ohne eine genaue Angabe der Implementation. Gehaltvoller ist die Orientierung an einem präzisen, engen Begriff von Algorithmus. Ein Algorithmus ist eine (imperative) Sequenz von Schritten, diese sind jeweils (d.h. in der Durchführung eines Teil-Algorithmus):

- effektiv (d.h. ‚geistlos' in ihrer Durchführung)
- endlich (in Input und Output und Ressourcenzugriff)
- substratneutral

Dem entsprechen Prozeduren in der Computerprogrammierung. Eine gehaltvolle CTC behauptet solche Algorithmen der Kognition, angestoßen und ausgeführt in einer kognitiven Architektur.

Eine CTM oder CTC hat nur empirischen wissenschaftlichen Gehalt, wenn sie auf den Standardbegriffen der Berechenbarkeit basiert. Der Standardbegriff von Berechenbarkeit (Computation) wird von der Berechenbarkeitstheorie in der Theoretischen Informatik behandelt.[9] Entscheidend sind hier zwei Befunde:

---

[8] Wie in §1 erwähnt, geschieht dies zumeist im Rahmen der Unterstützung der Materialismus-These. Verbindet sich diese auch mit einem deterministischen Mechanismus, stellen sich die ebenfalls in §1 erwähnten Verständnisschwierigkeiten, wie sich dies zu unserem Selbstverständnis als bewussten freien Akteuren verhält oder damit kompatibel sein soll.

[9] Vgl. z.B. Nigel Cutland, *Computability*.

(i) Die verschiedenen Erläuterungen des Berechenbarkeitsbegriffes (von TMs und Programmierbaren Random Access Machines [PRAMs] bis zu λ–Kalkül, partiell rekursiven Funktionen usw.) sind äquivalent zu einander (d.h. sie können dieselben Funktionen berechnen und sich wechselseitig simulieren). Dies ändert sich auch dann nicht, wenn Elemente wie Nichtdeterminiertheit [NTMs] oder Fehlerzulässigkeit und Wahrscheinlichkeit (Probabilistische TMs [PTMs]) oder neue Konstruktionsideen (wie Quantencomputer) einbezogen werden. Darin gründet *Churchs These*, dass dieser Begriff der Berechenbarkeit den intuitiven Begriff der Berechenbarkeit erfasst.

(ii) Das Modell der PRAM findet sich in der von Neumann-Architektur realer Computer wieder.

Man kann in der Theorie weitere Begriffe von Berechenbarkeit definieren (‚Hyperberechenbarkeit'[10] oder ‚Super-Rekursivität'[11]) anhand von Modellen abstrakter Maschinen, die jenseits des ‚Turing-Limits' (d.h. über den Leistungen all der Standardmodelle hinaus) Probleme lösen können. Beispiele sind Maschinen, die ‚Orakel' zu in den Standardmodellen nicht lösbaren Fragen (wie der Entscheidbarkeit allgemeiner prädikatenlogischer Folgerung) konsultieren können (Orakel-TMs) oder etwa über unendliche Zusatzinformationen verfügen, die sie in einem Schritt konsultieren können (Advice-TMs). Materiell lassen sich solche Maschinen *nicht* realisieren.

Eine CTM, die sich auf Hyperberechenbarkeit in der These „Der Geist ist ein Computer" beruft, muss also den Materialismus aufgeben. Das Gehirn kann nicht eine solche Maschine sein. Die meisten Vertreter einer CTM behaupten hingegen die Materialismus-These.

Eine CTM oder CTC, die sich weder auf (Standard-)Berechenbarkeit noch auf Hyperberechenbarkeit festlegt, bleibt eher unklar und immunisiert sich gegen Kritik, welche an den Grenzen dieses Berechenbarkeitsbegriffes ansetzt. Sie kann dazu dienen, kognitive Leistungen als regelbasiert zu erläutern und knüpft so an philosophische Analysen an. Sie kann eine (vagere) Heuristik sein, die allerdings weniger empirisch überprüfbare Aussagen über die Kognition machen kann. Sie entspricht – auch wenn *de facto* gelegentlich in dieser unklaren Weise verfahren wird – nicht den proklamierten Ansprüchen vieler Vertreter einer CTM oder CTC.

---

[10] Vgl. Apostolos Syropoulos, *Hypercomputation*, oder: Jack Copeland, „Hypercomputation".

[11] Vgl. Mark Burgin, *Super-recursive Algorithms*

§11 Die Semantik von Symbolen im Allgemeinen geht zurück auf die Bedeutung von Sätzen, d.h. die Semantik einer natürlichen Sprache. Die Bedeutung von Sätzen hängt zusammen mit der Wahrheit von Behauptungssätzen und der Wahrheit der in ihnen ausgedrückten Meinungen. Originär ist dieser Wahrheitsbezug und -anspruch von Behauptungen/Behauptungssätzen und Meinungen. Darin liegt originärer Wirklichkeitsbezug, originäre Intentionalität, von der die Intentionalität von Repräsentationen abgeleitet ist. Eine Repräsentation hat Intentionalität/Semantik, weil sie Teil einer Meinung ist.

Meinungen zu haben, die wahr oder falsch sind und als solche unterschieden werden können, bedarf des Verfügens über den Begriff der Meinung, also einer höherstufigen Intentionalität (einer ‚Theory of Mind').[12] Damit einhergehen Normen der Rationalität beim Vorbringen und Überprüfen von Behauptungen und Meinungen. Meinungen im eigentlichen Sinne können daher nur rationale Wesen besitzen, was Vorstufen von Meinungen und Intentionalität bei Tieren nicht ausschließt. Insbesondere sind dies Normen einer Sprachgemeinschaft (d.h. sowohl Normen der sprachlichen Verständlichkeit und Wohlgeformtheit als auch epistemische Normen der Begründung). Bei diesen Leistungen geht – irgendwie – Bewusstheit und Selbstbezug ein, ebenso das Einnehmen von Sprecherrollen mit der Übernahme der entsprechenden Verantwortung für Äußerungen. Intentionalität haben Personen, die sich für ihre epistemischen und praktischen Urteile und entsprechende Handlungen rechtfertigen können und vor anderen müssen. In Urteilen, Deliberation, Reflektion und entsprechendes Handeln geht freies Entscheiden ein. Für solche Urteile – nicht für das, was einem zustößt – kann und muss man sich rechtfertigen, zumindest immer ‚im Prinzip', wenn entsprechender Begründungsbedarf (von anderen oder in der Selbstreflexion) angemeldet wird.

Die bis jetzt hergestellten Computer und Roboter sind keine Personen. Daher besitzen sie *für sich* auch keine Intentionalität oder Semantik. Sie sind jedoch *semantische Artefakte*.

Ein Algorithmus kann auf einer Ebene der Abstraktion beschrieben werden als eine Sequenz von Symbolmanipulationen, d.h. als ein syntaktisches Objekt. Sogar ein laufender Algorithmus kann so beschrieben werden. Ein laufender Algorithmus ist zugleich in einem

[12] Vgl. z.B. Donald Davidson, „Rational Animals", sowie allgemein: Michael Corballis, *The Recursive Mind*. Vgl. zum Folgenden auch: Manuel Bremer, *Philosophische Semantik*.

Substrat implementiert und damit bezüglich einer solchen Beschreibung nicht nur eine Abstraktion, sondern eine syntaktische und kausale Sequenz. Ein laufender Algorithmus ändert die Wirklichkeit (zumindest im implementierenden Substrat) in einer Weise, die der definierten Implementation seiner Teile korrespondiert. Darin besitzt ein laufender Algorithmus eine (prozedurale) Semantik, nicht allein auf der abstrakten Ebene einer Modelltheorie, sondern im Verknüpfen seiner Teiloperationen mit Veränderung in der Wirklichkeit als deren Bezugnehmen auf die Wirklichkeit. Ein laufender Algorithmus besitzt dermaßen einer Semantik und der abstrakte Algorithmus wurde abstrahiert von Symbolen, die eine Semantik haben.

Damit ein Algorithmus aus Symbolen besteht, benötigen diese seine Teile eine Semantik. Von dieser Semantik wird in einer syntaktischen Beschreibung oder Untersuchung des Algorithmus abstrahiert. Man kann nicht von ‚Computation' oder ‚Berechnung' reden ohne Semantik, denn der Begriff der Repräsentation ist für den Begriff des Symbols unverzichtbar. Man kann allerdings Berechenbarkeit in Abstraktion von der Semantik der Symbole untersuchen. Dies geschieht auch in der meta-logischen Untersuchung formaler Sprachen und Kalküle und der theoretischen Informatik. Das Interesse an den entsprechenden meta-logischen Theoreme über ein Logiksystem speist sich indessen aus deren Anbindung an eine intendierte Semantik (informell anvisiert oder über entsprechende Korrektheitsbeweise angebunden).

Die unmittelbare Implementation eines Algorithmus kann sich, des Weiteren, fokussieren auf die kausale Implementation der syntaktischen Strukturen, wobei diese Implementation die Semantik der implementierten Symbole erbt.

Schreibt man einem System zu, einen Algorithmus auszuführen, werden diese Schritte in umgekehrter Reihenfolge betrachtet: man extrahiert eine syntaktische Beschreibung aus einer Interaktion von System und Wirklichkeit.

Ein Computersystem ist so konstruiert, dass dessen interne Struktur und die Operationen, die über diese definiert sind, eine Bedeutung haben (seien es numerische Resultate oder beispielsweise die Herstellung einer Netzwerkverbindung zum Bilderaustausch). In diesem Sinne sind Computer *semantische Artefakte.*

Diese Bedeutung existiert – natürlich – nur für die konstruierenden Personen und nicht für den Computer selbst.[13] Die kausalen Interaktionen des Computers sind Bestandteile der Wirklichkeit und selbst zunächst ohne Bedeutung. Die Arbeitsweise des Computers kann allerdings beschrieben und interpretiert werden auf abstrakten Ebenen als Sequenz semantischer und syntaktischer Prozeduren. Die Bedeutung ‚im' Computer leitet sich genauso von der Semantik der sprechenden Personen ab, die den Computer konstruieren, wie die Bedeutung ‚in' einem Buch. Beide sind artefizielle Bedeutungsträger. Im Unterschied zum passiven Buch, dessen Bedeutung nur im Lesen durch und für eine Person realisiert wird, realisiert der Computer Bedeutung in seiner eigenen Aktivität. Darin geht die Konstruktion und Implementation von Algorithmen über das Schreiben von Büchern hinaus.

Eine Idee hinter einer CTC lässt sich verstehen als der Versuch, menschliche Kognition in derselben Weise zu betrachten: Es gibt eine erhellende *abstrakte* Beschreibung der Kognition als Computation. Ob menschliche Kognition Computation im Sinne einer solchen CTC ist, hängt ab davon (i) wie sich das Modell an empirische Forschung zur menschlichen Kognition anbinden lässt (etwa bezüglich von Verarbeitungszeiten im Kurzzeitgedächtnis oder Phänomenen wie ‚Prompting')[14] und (ii) wie sich Modelle höherer kognitiver Leistungen (d.h. solcher, die bewusst und von der allgemeinen Kognition her zugänglich sind) an introspektive Befunde anbinden lassen.

Eine solche Betrachtung ist zunächst eine *Interpretation* und eine entsprechende Modellbildung. Aber eine Modellbildung, die aus einer Interpretation hervorgeht, kann trotzdem das Betrachtete adäquat erfassen. ‚Interpretation' heißt nicht ‚Verfälschung'. Die Erfassung von Algorithmen durch eine Interpretation eines komplexen Systems heißt nicht, dass es sich nicht um ein Algorithmen abarbeitendes System handelt.[15]

§12 Die genuine Intentionalität involviert das Bewusstsein einer Person (‚Geist' im engeren Sinne). Aber nicht alle mentalen Zustände und Prozesse sind bewusst. Auch die nicht bewusste Mentalität nutzt die genuine Mentalität, die sich im Bewusstsein gründet. Solche

---

[13] Das ist m.E. der richtige Teil von John Searles Kritik jeglicher CTM oder CTC (vgl. John Searle, *The Rediscovery of the Mind*).

[14] Vgl. z.B. Allen Newell, *Unified Theories of Cognition*, S.222-34, zu einer eher positiven Einschätzung diesbezüglich und SOAR, kritisch zu SOAR z.B. Richard Cooper, "Cognitive Architectures as Lakatosian Research Programmes: Two Case Studies".

[15] Auch für eine CTC widerspricht eine Theoriebildung mittels nicht (von außen eines Untersuchungsbereiches) direkt beobachtbarer Entitäten nicht einer realistischen (an Stelle einer instrumentalistischen) Auffassung von Theorien.

Prozesse kann – und evtl. muss – man modellieren unter Absehung von Bewusstsein und dessen Rolle. Hier können algorithmische Modelle eine Rolle spielen. Ungeklärt muss dabei bleiben – zumindest im Detail – wie dann die Übergänge zwischen bewussten und nicht-bewussten Vorgängen erfolgen. Eine CTC kann hier eine Option einer partiellen Theoriebildung sein.

Auch im Bewusstsein werden Regeln verwendet und Algorithmen abgearbeitet. Regelbefolgen und Begründen involvieren Prozeduren, für die es algorithmische Modelle geben kann. Insofern kann eine CTC Teil der Explikation und Erläuterung der kognitiven Vermögen von Personen sein.

Eine CTC hat die besten Erfolgsaussichten bei der Modellierung von Fähigkeiten, die nicht mit Bewusstheit und Deliberation verbunden sind (wie dem unmittelbaren Klassifizieren von Entitäten). Eine Reihe solcher ‚modularen' Fähigkeiten hängen wenig bis gar nicht von allgemeinen kognitiven Leistungen wie der Aufrechterhaltung eines kohärenten Meinungssystems ab. Damit stellt sich bei diesen nicht das paradigmatische ‚frame problem', was alles zu berücksichtigen und im Kontext zu reflektieren ist.

Eine solche Konzentration auf solche sehr speziellen Fähigkeiten bietet sich methodisch an, muss dann indessen mit dem Eingeständnis verbunden werden, keine allgemeine computationale Theorie des Räsonierens und der Kognition zu liefern.[16]

Allen Newell als Ponier der KI und Theoretiker einer paradigmatischen CTM gebührt der Verdienst, seine konkrete Modellierung und Implementation einer CTM (deren letzte Variante die Frühformen des SOAR-Projekt waren) an psychologische experimentelle Daten anzubinden.[17] Solche Daten allerdings stammen aus Laborsituationen, die kleinteilige kognitive Leistungen messbar operationalisieren (etwa das Abtippen eines Textes, das Reagieren auf assoziative Verknüpfungen von Symbolen und Bildern, usw.). Die Operationalisierung solcher Experimente legt eine Übersetzung in Flow Charts und Algorithmen nahe, was bei alltäglichen geistigen Leistungen in Situation nur unter Abstraktion von vielen Umständen möglich ist. Ein Modell des Problemlösens, das an den experimentell operationalisierten und wohl definierten Problemen ansetzt, hat seine Anwendungen, erstreckt sich jedoch nicht zwangsläufig auf das allgemeine Umgehen mit den Herausforderungen des Alltags und dem Gestalten der eigenen Biographie.

---

16 Vgl. Jerry Fodor, *The Mind Doesn't Work That Way.*

17 Vgl. Allen Newell, *Unified Theories of Cognition.*

Ein ebenso der Modellierung besser zugänglicher Bereich sind alle Fertigkeiten, die in der Form entweder deduktiven Folgerns oder der Suche nach Modellen (im Sinne von Variablenbelegungen) dargestellt, formalisiert und (z.B. in der Logikprogrammierung) implementiert werden können.[18] Solche Fähigkeiten sind wesentlich algorithmisch umsetzbar, sie machen den Kern des intuitiven Begriffs der Berechenbarkeit aus. Damit sind sie zugleich allerdings – haben wir einmal einen Algorithmus entdeckt – auch wesentlich geistlos durchführbar.

Interessanter bezüglich der allgemeinen Intelligenz ist nicht das Befolgen, sondern das Entdecken und Entwerfen solcher Algorithmen und die Überprüfung ihrer Korrektheit/Angemessenheit bezüglich der betrachteten Aufgabenstellung. Diese Leistungen sind nicht algorithmisch.

§13 Die meisten Klassifikationsleistungen erfolgen problemlos im Hintergrund, etwa wenn wir uns in einer Situation bewegen und deren Objekte und Ereignisse in Kategorien einordnen. Auch sprachliche Fassungen solcher Urteile sind unproblematisch: „Der Tisch wackelt“, „Es ist kurz vor 3“, „Die Straßenbahn hat ein neues Werbebanner“ etc.

Wichtig sind allerdings ebenso Prozesse des Beurteilens und Einschätzens, die Überlegung und Erfahrung einschließen und keine bloße Klassifikation im obigen Sinne liefern. Solche Prozesse können in mehreren Arten von Fällen angestoßen werden:

- im Falle vager Kategorien: „Ist das noch Hard-Rock oder schon Heavy Metal?“
- im Falle von Erwartungen des Handels anderer Personen: „Sind die Anhänger von Mike McConnell eine Gefahr für die nächsten demokratischen Wahlen?“
- im Falle des Verstehens des Handelns anderer Personen: „Hat Susanne wirklich keine Lust mehr auf die Reise?“
- im Falle von ästhetischen Evaluationen: „Soll die Vase hier stehen oder weiter rechts?“
- im Falle der Anwendung komplexer Normen: „Weisen die Indizien darauf hin, dass es sich um einen Mord aus niederen Motiven handelt?“

[18] Vgl. Robert Kowalski, *Computational Logic and Human Thinking.*

- im Falle der Einschätzung von Entwicklungen: „Ein allgemeines algorithmisches Modell zur Herstellung von Kohärenz einer Menge von Meinungen ist bis auf Weiteres nicht zu erwarten."

und anderer mehr.

Die Problematik erinnert an die traditionelle Problematik der ‚Urteilskraft'.[19] Für den Erwerb von Urteilskraft in diesem Sinne besitzen wir kein Rezept. Eine Simulation solcher Beurteilungen in einem Computerprogramm (etwa von einer Liste der Indizien bis zu einer Operationalisierung des Ablesens ‚niederer Motive') persifliert die Beurteilungssituationen, in denen sich Personen wiederfinden. Dies beginnt schon mit der für die Simulation nötigen Einschränkung der Situationsmerkmale und des benötigten Kontextwissens über Situationen dieses Typs.

Genauso wenig wie im Falle des freien Entscheidens besitzen wir eine (detaillierte) Theorie, was dieses Einschätzen und Beurteilen ausmacht: es verlangt eine umfassende Situationsbeschreibung, bei verfügen über hinreichende sprachliche Kompetenz bezüglich solcher Beschreibungen, unter Einbeziehung relevanten Kontextwissens – aber schon hier treten wieder Einschätzungen auf, was ‚relevant' und ‚hinreichend' im betrachteten Kontext ist. Die Beurteilung einer Situation kann weit auf Wissensbestände und Erinnerungen zugreifen. Es gibt keine allgemeinen formalen Bedingungen, ein Beurteilungsproblem zu spezifizieren. Beurteilen macht indessen einen wesentlichen Teil des geistigen, sozialen und kognitiven Lebens von Personen aus. Diesen Bereich der allgemeinen Kognition und menschlicher Intelligenz kann sich eine CTC nicht erschließen.

§14 Die vorangehenden Überlegungen können zu drei Hypothesen führen:

(Hypothese 1) Eher ist eine Computationale Theorie der Kognition möglich, aber nicht eine des Geistes im Allgemeinen.

(Hypothese 2) Eine Computationale Theorie der Kognition kann eine Heuristik für eine Theorie der Kognition sein, auch wenn die Kognition kein Computer ist.

[19] Vgl. z.B. Klaus Günther, *Der Sinn für Angemessenheit*.

(Hypothese 3) Eine Computationale Theorie der Kognition lässt die Stellen hervortreten, an denen der Geist über eine algorithmische Kognition hinausgeht (bzw. wieweit der Geist mit Algorithmen arbeitet).

§15 Insofern es m.E. starke Einwände gegen eine allgemeine CTM oder CTC gibt, gibt es die entsprechenden Einwände gegen die starke KI. Im Folgenden geht es nun, um die Problematik, welche Grenzen man darüber hinaus (selbst) bei der schwachen KI einräumen muss.

Schwache KI erhebt nicht die Ansprüche der starken KI, die nach *science fiction* und Versprechungen klingen. Der Unterschied zwischen diesen Sorten der KI kann zwar definitorisch gezogen und einsichtig gemacht werden, er gehört jedoch nicht zum Allgemeinwissen und wird auch nicht von allen Akteuren im KI-Bereich gemacht. So kann der Erfolg von Systemen der schwachen KI den Eindruck eines allgemeinen Fortschritts in der KI hervorrufen und die Akzeptanz des Narrativs der starken KI erhöhen. Dieses Narrativ ist nicht bloß überzogen – was zunächst harmlos sein könnte – sondern weltanschaulich aufgeladen. KI wird als die Lösung der gegenwärtigen gesellschaftlichen Probleme (von Armut bis Klimawandel) vorgesehen und als Überwindung allgemeiner menschlicher Beschwerden (wie Krankheit und Altern) verkündet.[20] Die damit einhergehende technizistische Sicht auf gesellschaftliche Probleme wird diesen in der Regel nicht gerecht – selbst, wenn ‚gut gemeint'[21] – sondern lässt politische und gesellschaftliche Problemlagen und Konfliktlagen aus den Augen treten. Außerdem bleiben keine Alternativen, wenn die technischen Versprechungen sich nicht einlösen lassen. KI-Versprechen setzen hier eine längere Geschichte der Technologiegläubigkeit fort.[22]

Auf der einen Seite kann so der Erfolg zunächst sinnvoller KI-Systeme zur Stärkung nicht sinnvoller und fragwürdiger Weltanschauungen (wie des Transhumanismus[23]) beitragen. Auf der anderen Seite kann der Unglaube an die überzogenen Versprechungen der starken KI dazu beitragen, die Gefahren, die sich mit Anwendungen der schwachen KI, wenn sie umfassend in unseren Alltag einkehren, zu übersehen. Ethisch und philosophisch relevanter

---

20 Paradigmatisch: Ray Kurzweil, *The Singularity is Near*.
21 Vgl. etwa: Steven Pinker, *Enlightenment Now*.
22 Vgl. David Noble, *The Religion of Technology*.
23 Vgl. kritisch: Nicholas Agar, *Humanity's End*, oder auch: Julian Nida-Rümelin & Nathalie Weidenfeld, *Digitaler Humanismus*.

als Debatten um Androiden etc. sind heute Fragen nach den Grenzen und Auswirkungen gegenwärtiger KI. Schwache KI hat Grenzen nicht nur bezüglich des Umfangs der Modellierung kognitiver Prozesse. Die Grenzen betreffen u.a. Eingrenzungen im *Verständnis* von ‚Intelligenz' und Grenzen der gelingenden Interaktion mit Systemen der KI. Die zweite Art der Grenzen und Beschränkungen schließt ethische Fragen ein.

§16 Eine Operationalisierung einer intelligenten Leistung schränkt deren Verständnis immer auch ein auf die berücksichtigten Modelle, die wiederum geplante Operationalisierungen und Grenzen dessen, was sich gerade umsetzen lässt, vor Augen haben. Operationalisierungen schließen in der theoretischen Modellierung der betrachteten Entitäten mutmaßlich irrelevante Eigenschaften aus (etwa die Hautfarbe einer Person) und machen Vereinfachungsannahmen für handhabbare Algorithmen (etwa statistische Unabhängigkeit in den Daten). Weitere mit einer Operationalisierung verbundene Festlegungen und Vereinfachungen erfolgen mit der Wahl der Softwarearchitektur (dem Programmier-Paradigma und der Programmiersprache sowie den verwendeten Datenstrukturen).[24] Eine erfolgreiche Operationalisierung in einem erfolgreichen System kann eine entsprechende verkürzte Behandlungsweise eines Problems verankern – etwa, indem nur die behandelbaren Kernanwendungsfälle im Weiteren behandelt werden. Dies gilt umso mehr, wie sich solche Systeme als Prestige-Objekte einer Expertenkultur präsentieren. Der Umstand, dass sich ein Problem nicht mit dem investierten Aufwand (an Geld, Zeit und Expertenwissen) lösen lässt, drängt es an den Rand, während die Erfolgsbereiche nun festlegen, was es heißt, mit dem Ausgangsproblem umzugehen. Der scheinbare Erfolg bestimmt auch die Interpretation der Ergebnisse.

Bei den heute im Mittelpunkt des Interesses an KI stehenden Systemen des ‚Maschinellen Lernens' kommen Fragen nach der (einseitigen) Auswahl der Trainingsdaten und der Festlegung des bewertenden Feedbacks an die Lernleistungen solcher Systeme hinzu.[25]

Leistungen von KI-Systemen treten oft als definitorisch für ‚intelligent' auf, ähnlich wie Intelligenz sprichwörtlich das sein soll, was Intelligenztests messen. Dies reduziert das Verständnis von ‚Intelligenz' auf algorithmische kognitive Leistungen zum Nachteil anderer

[24] Vgl. z.B. George Luger & William Stubblefield, *AI Algorithms, Data Structures, and Idioms in Prolog, Lisp, and Java*; Joseph Bigus & Jennifer Bigus, *Constructing Intelligent Agents Using Java*.
[25] Vgl. Katharina Zweig, *Ein Algorithmus hat kein Taktgefühl*.

geistiger Leistungen, die auch Intelligenz erfordern (etwa kreative Leistungen oder das Konstruieren eines Artefaktes). Zu Beginn von Standardeinführungen zur KI wird ‚Intelligenz' (d.h. das *explanandum* bekannt aus dem menschlichen Fall) gerne *gleichgesetzt* mit zielgerichtetem Problemlöseverhalten im Rückgriff auf Wissen.[26]

Die Interaktion mit einem KI-System drängt den Benutzer, sich den Formaten und Beschränkungen der Problembehandlung, die das System ausmachen, anzupassen, d.h. selber sich dem System anzupassen als umgekehrt. Es werden Abhängigkeiten verankert, die sich nicht einfach zurückdrehen lassen.[27]

§17 Scheinbar erfolgreiche KI-Systeme laden dazu ein, sich auf sie zu verlassen, auch dann, wenn den Benutzern ihre Funktionsweise nicht durchsichtig ist. Dies trägt sowohl zur Mystifizierung der Leistungen solcher Systeme bei als auch zur unkritischen Übernahme ihrer Resultate (etwa von Vorhersagen oder Einschätzungen z.B. ökonomischer Entwicklungen). So verlassen sich Börsen und Finanzmärkte auf Systeme, welche die Börsianer selbst kaum durchschauen, obwohl darin ein Risiko zu Börseneinbrüchen liegt.

In der Programmierung gilt der sprichwörtliche Warnhinweis „Computer tun genau, was man ihnen sagt – nicht, was man meinte." Anekdoten aus der Informatik und viele Szenarien in *science fiction* Literatur oder Filmen handeln von genau solchen Fällen, wo scheinbar harmlose Befehle zu bizarren oder katastrophalen Ergebnissen führen, weil die Spezifikation der Aufgabe nicht (hinreichend) einschloss, die zu bewahrenden Rahmenbedingungen zu beachten (z.B. das selbstfahrende Auto, das möglichst schnell zum Ziel fahren soll, fährt in einem Wechsel von starker Beschleunigung und Vollbremsung bei Rücksichtslosigkeit gegen andere Verkehrsteilnehmer, oder der semi-autonome Rasenmäher, der Zusammenstöße vermeiden soll, fährt nur noch in einem kleinen Kreis – etc.). Was im Einzelfall eher ein amüsantes Versagen mit sich bringt, kann mit Ausmaß und Reichweite des KI-Systems katastrophale Konsequenzen haben (ein semi-autonomes Krankenhaussystem entscheidet, die effektivste Methode der Schmerzlinderung ist das Abschalten der Geräte auf der

[26] Vgl. z.B. George Luger, *Artificial Intelligence*, oder: Stuart Russell & Peter Norvig, *Künstliche Intelligenz*.

[27] Vgl. schon früh: Joseph Weizenbaum, *Die Macht der Computer und die Ohnmacht der Vernunft.*

Intensivstation). Hier müsste eine Heuristik der Vorsicht und der Begrenzung der Wirkweite von KI-Systemen eingreifen.[28]

Ethisch relevant sind nicht allein Schäden, die KI-Systeme als Nebeneffekte mitverursachen, sondern auch gesellschaftliche Nebeneffekte, wie die gesellschaftliche Verteilung des Zugangs zu diesen Systemen.

Rechtlich geklärt werden muss im Einzelfall, wer für die Schäden einer Fehlfunktion haftbar zu machen ist. Dies betrifft sowohl automatisierte Beurteilungen (in Bildungseinrichtungen oder einem Kreditinstitut) mittels einer Software als auch mobile Systeme (von selbstfahrenden Robotern bis zu Drohnen) oder das ‚smart home' des *Internet of Things*.[29]

Im militärischen Bereich, der traditionell als Hauptförderer der KI auftritt, stellen sich Fragen nach der Rüstungsbegrenzung speziell im Bereich semi-autonomer Waffen.[30]

Zugleich verstärken entsprechende Systeme und Software-Agenten gezielt den *bias*, der sich in *social media* (wie Facebook und Twitter) findet. Systeme, deren Erfolg in Klicks auf entsprechende Buttons und Links gemessen wird, lernen die Benutzer in eine entsprechende Richtung zu lenken (etwa durch Präsentation ähnlicher oder reißerischer Inhalte) und konterkarieren so die vermeintliche Meinungsvielfalt dieser Internetmedien.

KI-Systeme (z.B. der großen Werbungsvermarkter wie Google und Facebook) tragen bei zum um sich greifenden ‚Surveillance Capitalism'[31] der Speicherung und auswertenden Nutzung möglichst vieler durch Computergebrauch generierten Benutzerdaten und werfen damit Fragen ihrer Regulierung auf.

Der Ubiquität der KI-Systeme und den Versprechen beschleunigter Digitalisierung laufen die Bemühungen zur Regulierung hinterher. Wie in anderen Bereichen der Gesetzgebung folgt auf den Erlass des Gesetzes die Suche nach Gesetzeslücken von – in der Regel ökonomisch – motivierten Akteuren im Regelungsbereich. Je mehr Profit und Einfluss sich mit dem Einsatz

---

[28] Dies erinnert an die ‚Heuristik der Furcht' in Hans Jonas' Technikphilosophie (vgl. *Das Prinzip Verantwortung*), sobald man diese aus Jonas' metaphysischer Konstruktion löst. Die fehlende Berücksichtigung von Nebeneffekten findet sich mindestens schon in der Legende von König Midas.

[29] Vgl. Mark Coeckelbergh, *AI Ethics*, Kap. 8; vgl. auch: Deborah Johnson, *Computer Ethics*, Kap. 7.

[30] Vgl. Toby Walsh, *It's Alive*, S. 236-51. Solche Waffen sind natürlich nicht ‚autonom' im philosophisch traditionellen Sinn des Wortes (sich selbst Gesetze gebend und frei entscheidend), obwohl sie „autonom" genannt werden. Es wäre jedoch irreführend diesen starken Autonomiebegriff als Beurteilungsbasis anzulegen. Diese Systeme sind *semi-autonom* in der Art eines Computerspiels: innerhalb der programmierten Regeln werden einzelne Schritte aus einer Menge möglicher Schritte ausgewählt, um den Gegner (sei es der menschliche Spieler des Computerspiels oder die gegnerischen Soldaten bei einem Drohneneinsatz) zu schlagen. Diese Auswahl wird nicht mehr von einem menschlichen Operator des Systems kontrolliert.

[31] Vgl. Shoshana Zuboff, *The Age of Surveillance Capitalism*.

von KI-Systemen verbinden, umso mehr Spezialisten (wie Anwaltskanzleien) werden von interessierten Akteuren darauf angesetzt, Regelungslücken zu finden, wobei sich mit der Nutzung dieser Lücken ein entsprechend hohes Risiko von politisch unerwünschten Effekten oder Nebeneffekten der KI-Systeme ergibt. Das Risiko geht weniger von superintelligenten KI-Systemen aus, welche ihre menschlichen Überwacher austricksen wollen, denn von einer Gesellschaft, die auf Profit- und Einflussmaximierung basiert und in der kurzfristig nutzenrationales Handeln zu unübersehbaren Folgen führt (wie bei den Auswirkungen der massiven Nutzung fossiler Energieträger und des Verbrennungsmotors auf das Klima oder der Anhäufung von atomaren oder Plastikmüll).

Manuel Bremer, 2019/2021