Ist alles berechenbar?
Explikation einer Fragestellung

Abstract. The paper discusses the meaning of the question whether everything is computable. The question such put needs explication towards a meaningful question, which restricts the realm of the computable. The process of explication aims at removing some misconceptions on the computable. It also argues for a 'weak' understanding of artificial intelligence, and a proper use of the computer model of the mind as a heuristic in the cognitive sciences. This has repercussion on a proper understanding of famous issues like the Turing-Test. The last paragraph adds to the theoretical also some ethical doubts concerning 'strong' artificial intelligence.

Abstract. Der Aufsatz geht aus von der recht vagen Fragen „Ist alles berechenbar?" und klärt diese Frage schrittweise. Erreicht wird eine Explikation der Frage, die präziser ist, jedoch zugleich den Bereich des Berechenbaren einschränkt. Mit der Explikation einher gehen Klärungen zum Begriff des Berechenbaren. Argumentiert wird für eine ‚schwache' Konzeption der Künstlichen Intelligenz und für eine heuristische Verwendung des Computermodell des Geistes. Dies wirkt sich auch aus auf ein angemessenes Verständnis des in diesem Zusammenhang viel diskutierten Turing-Tests. Der letzte Abschnitt weist auf zusätzliche ethische Probleme einer ‚starken' künstlichen Intelligenz hin.

§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.

[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 1986) ließen sich feinere Unterschiede begründen.

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:

(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 erkenntnistheoretische 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 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

[2] Alle drei folgenden Missverständnisse bzw. überzogenen Erwartungen bzgl. des CMM finden sich z.B. in Putnam 1991, wo sie als Basis zur Zurückweisung des CMM dienen.

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.
Insofern scheinen wir weit entfernt von einer künstlichen Intelligenz als geplanter und verstandener Reproduktion der Strukturen der menschlichen Intelligenz.
Damit ist jedoch nicht die akzidentelle Erschaffung einer starken künstlichen Intelligenz ausgeschlossen! Gegeben Bausteine, die wir im Detail vielleicht nicht ganz durchschauen, sowie gegeben Teilfertigkeiten und entsprechende Module, deren Einzelabläufe wir vielleicht in ihrer Komplexität nicht (ganz) durchschauen, ist es *nicht logisch ausgeschlossen*, dass eine Kombination solcher Komponenten zu einen System sich als starke künstliche Intelligenz

darbietet, d.h. sich als solche zeigt und wir den Trug, falls es sich doch nicht um eine solche handelt, nicht entdecken können.
Solche eine KI wäre ‚künstlich' bezüglich der Herkunft ihrer Komponenten und ihrer Gesamtgenese, aber zugleich kein bewusstes Produkt eines Erschaffenden.
Eine KI als Heuristik für das Verständnis menschlicher Intelligenz wäre dies nicht. Sie würde vielmehr ein *neues Rätsel* in die Welt setzen: neben unser Unverständnis bezüglich der menschlichen träte unser Unverständnis bezüglich dieser künstlichen Intelligenz!
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:

(i) dass ein Modell verschiedener *abstrakter Ebenen* nötig ist, inklusive einer massiven *Modularisierung*
(ii) dass es um kognitive, geistige Leistungen geht, also regelgeleitete Prozeduren/Algorithmen
(iii) dies in Abstraktion von ihrer Implementierung (also in der Regel funktional) jedoch mit einem Blick auf praktische Umsetzung
(iv) dass zumindest begründet werden muss, warum welche geistige Leistungen *nicht* im Sinne von Berechenbarkeit zu begreifen sind
(v) dass insofern der Geist etwas Abstraktes und zugleich Implementiertes/Implementierbares ist
(vi) dass es in der Arbeit des Geistes Module der Informationsverabeitung durch Input (‚Transducer') und Effektorenanbindung gibt.

Künstliche Systeme (wie eben der Computer) liefern hier ein Modell der verschiedenen Ebenen der Programmiersprachen (von Hochsprachen zu Assembler zu Maschinensprache), das Modell der Basis in Algorithmen, das Modell der Module und der Peripherie, das Modell einer Supervenienz (der Prozeduren) zu ihrer Implementierung in verschiedenen Substraten
Zugleich werfen Vergleiche mit realen Computern Fragen auf, bezüglich:

(i) dem Verhältnis von determinierten Komponenten und zufallsgesteuerten Komponenten
(ii) dem Verhältnis von digitaler und analoger Repräsentation
(iii) der Rolle von Steuerung durch Lernen und Selbstveränderung
(iv) der Rolle von interner und externer Kommunikation

KI als Heuristik zwingt nicht nur die Komponenten und Prozeduren einer Maschine explizit und präzise zu fassen, sondern auch das Modell der Maschine in relevanten Situationen und ihrer Rolle darin ausdrücklich zu formulieren.

§9 Die bisherigen Betrachtungen wirken sich auch auf die Bewertung des Turing-Tests (TT) als prototypischen Gegenstand der Auseinandersetzung um die KI aus. Der nach Alan Turing benannte Test, bei dem eine Testperson entscheiden soll, ob sie mit einem Computer oder eine Person (per Chat oder Teletext) kommuniziert (Turing 1950), gehört zum Standardrepertoire der Debatte um Künstliche Intelligenz.

Im Turing-Test (TT) soll die Frage „Was ist Intelligenz?“ umgangen werden. „Was ist *x*?“ zu beantworten erfordert eine Theorie von *x*. Die Frage nicht zu beantworten, heißt auf eine Theorie zu verzichten, zumindest für eine Zeit. Anstelle der Theorie tritt ein Identifikationstest: Wie können wir das Vorliegen von *x* – hier: Intelligenz – feststellen? Dieser Test darf also kein theoretisches Verständnis von *x* voraussetzen, denn eine Theorie von *x* haben wir ja nicht. Der Test beruht daher auf eine Teiltheorie von *x*. Ohne eine Teiltheorie hätte der Test auch gar keine Anbindung an die Ausgangsfrage. Und dieser Teil kann nur ein operationalisierbarer Teil der Theorie von *x* sein. Es muss sich nicht um den wichtigsten oder zentralsten Bestandteil der Theorie von *x* handeln. Es handelt sich um den Teil, an dem sich für Tests ansetzen lässt. Das Vorgehen ist also *pragmatisch* – mangels einer besseren Alternative. Damit sollten jedoch auch die Testresultate so gesehen werden – und nicht im Nachhinein den Test als Definition von ‚Intelligenz’ betrachten.

Die Situation beim TT ähnelt Teilen der heutigen KI. Im Vordergrund steht das Haben und nicht das Verstandenhaben. Verfahren der ‚automatischen Übersetzung’ etwa liefern heute pragmatisch akzeptable Übersetzungsvorschläge, ohne dass sie auf einer entwickelten systematischen Semantik beruhen. Der Automat kann Teiltheorien (wie ein Lexikon und eine Teilgrammatik) koppeln mit einem riesigen Datenbestand an Standardsätzen und einem *brute force* Durchsuchen und Zuordnen von (Teil-)Sätzen. Man hat so etwas Ähnliches wie Bedeutungszuordnungen, ohne eine Theorie der Bedeutung zu haben.

Der extremste Fall dieses Vorgehens wäre das (oben angedeutete) Auftreten einer starken KI (einer künstlichen Person), ohne dass man ein zureichendes Verständnis von Personalität und den Fähigkeiten besitzt, welche die Rationalität dieser Person ausmachen. Die künstliche Person hätte dann bezüglich des Zieles, anhand einer künstlichen Person die menschliche Personalität und Rationalität und Intelligenz zu verstehen, nichts gebracht.

Der TT scheint, folgt man Turings Abwägen von Gegenargumente und seinen Beispielen, zudem stark an die Frage nach dem persönlichen Bewusstsein gekoppelt zu sein. Turing bringt Beispiele, die auf spezifische menschliche Erfahrungen abheben. Gesucht wird somit nach mehr als bloßer Intelligenz oder bloßen ‚Denken' (im Sinne des Vorliegens von mentalen Prozessen). Mentale Prozesse im weiten Sinne, der auch Prozesse umfasst, die bei Menschen nicht mit Bewusstheit ablaufen, können auch in einem System ablaufen, das keinerlei Bewusstsein hat, oder jedenfalls (wie Tiere) kein Selbstbewusstsein – könnte man vermuten. Es böte sich also an, auf elementarere Fertigkeiten als Bewusstsein zu testen. Ansonsten wäre der TT ein spezieschauvinistischer Test, den allein eine menschliche Intelligenz mit entsprechend menschlicher Verkörperungserfahrung bestehen kann. Getestet würde dann auf Teilintelligenz, also intelligente Leistungen. Intelligente Leistungen kann man einfacher durch Verhaltensäquivalente verstehen oder sogar definieren.
Dies entspricht zum einen dem Vorhaben der schwachen KI, teilintelligente Automaten oder Maschinen zu bauen, ohne dass diese über Gesamtintelligenz verfügen. Dies entspricht methodisch ungefähr solchen Ansätzen, welche das Intelligenz*vokabular* auf dispositionale Beschreibungen zurückführen wollen (etwa Ryles *The Concept of Mind*).
Der TT hat nur Relevanz unter der Annahme, dass eine *verlässliche Beziehung* (mutmaßlich unterhalb der Schwelle der Explikation) besteht zwischen Intelligenz und Personsein auf der einen Seite und dem Bestehen im Imitationsspiel auf der anderen Seite. Dies wiederum scheint aber nach einer explikativen Theorie zu verlangen. Insbesondere, wenn ‚Denken' und ‚Intelligenz' komplett unverständlich wären, wie Turing behauptet, dann würden wir den Vorschlag der Ersetzung der Ausgangsfrage „Können Maschinen denken?" durch das Imitationsspiel gar nicht verstehen. Gegeben diese methodische Unklarheit bei Turing muss man sagen, dass – auch wenn Turing keine operationale *Definition* von Intelligenz geben will – es sich (doch) um eine mindestens partielle Operationalisierung eines Teilverständnisses der mentalen Begriffe handelt. Operationalisiert wird zumindest der Identifikationsaspekt der Begriffe (d.h. des Teils eines Begriffs, der dazu dient, Anwendungsfälle des Begriffes von anderen Situationen abzugrenzen). Offen gelassen wird eine intensionale, die inneren Abläufe betreffende Erläuterung der Begriffe. Am besten lässt sich so der Test verstehen als Beibringen von Belegen, die einen Schluss auf Maschinenintelligenz als *Schluss auf die beste Erklärung* rechtfertigen könnten (vgl. Moor 1976).
Erfolg oder Misserfolg beim TT ist des Weiteren unabhängig von der Richtigkeit der *Church-Turing-These* (dass alles intuitiv Berechenbare TM-berechenbar ist). Ein Automat könnte beim TT versagen, weil menschliche Intelligenz sich nicht TM-simulieren lässt, insofern sie

Verfahren oder Prinzipien einschließt, die nicht berechenbar sind. In diesem Fall wäre die *Church-Turing-These* dennoch nicht widerlegt, eben weil diese Komponenten auch nicht dem intuitiven Begriff des Berechenbaren entsprechen. Kandidaten könnten Fähigkeiten sein, die unendliche oder indefinite Prämissenmengen involvieren (wie die oben erwähnte Ω-Regel). Die Ω-Regel erlaubt, aus einer unendlichen Menge von Instanzen den entsprechenden Allsatz abzuleiten. Auf den Allsatz wird strikt geschlossen, dieser nicht nur – wie in einer induktiven Logik ‚bis auf Weiteres' oder mit einer bestimmtem statistischen Wahrscheinlichkeit angenommen. Diese Regel verfährt klarerweise nicht algorithmisch, da sie die Bedingung des finiten Inputs für einen Schritt eines Algorithmus verletzt. Sie trägt zur Negationsvollständigkeit der Prädikatenlogik Zweiter Stufe und zur Unterscheidung der Prädikatenlogik Zweiter Stufe von der kompakten Prädikatenlogik Erster Stufe bei. Es handelt sich auch nicht um die sogar intuitionistisch gültige Vorgehensweise, von einem schematischen Beweis des Einzelfalls auf den Allsatz zu schließen. Die Ω-Regel bietet neben einer Auswahlfunktion den paradigmatischen Fall eines *nicht-konstruktiven* Vorgehens. Es ist zumindest denkbar, dass wir über eine entsprechende infinite oder zumindest doch indefinite *Intuition* verfügen. Eine solche Intuition würde durch Überschauen von Instanzen die Gültigkeit des Allsatzes erkennen (in einer Art ‚Aha-Erlebnis'). Eine indefinite Intuition würde von einer unbestimmten Menge von Prämissen schließen, wobei die Anzahl der benötigten Prämissen, um den Allsatz zu erkennen, schwanken kann: in einem Fall vielleicht 517 in einem anderen 4503 oder 34. In beiden Fällen (infiniter oder indefinite Intuition) gäbe es nicht *eine* finite TM-berechenbare Version einer solchen Regel. Im Falle einer indefiniten Intuition, die jeweils eine finite aber unbestimmte Anzahl von Prämissen benötigt, gäbe es zwar für jeden einzelnen Anwendungsfall eine TM-Berechnung, jedoch könnte eine finite Maschinentafel nur finit viele von diesen Regeln enthalten – und das könnten unter Umständen nicht ausreichend viele sein. Ausreichend wäre eine solche finite Kollektion wiederum für menschliche Intelligenz, wenn diese nur mit finiter Dauer und finitem Auffassungsvermögen ausgestattet ist.

Eine letzte hier zu erwähnende Kontroverse im Kontext des TT betrifft die Frage, ob eine Maschine intelligent sein kann, wenn sie doch ‚nur' programmiert sei. Dies weist zurück auf unsere Debatte der Frage, was alles berechenbar ist. Wenn es hochkomplexe Computer und Berechnungsvorgänge gibt, lassen sich diese – aus den bekannten Argumenten des Funktionalismus heraus – nicht mehr erfolgreich auf eine physikalische Beschreibung, die explanative Funktion besitzen könnte, reduzieren. Das Vorhandensein einer Beschreibung auf der Programmebene (d.h. eine Beschreibung, welche in einer Programmiersprache die

Algorithmen/Prozeduren angibt, die das Verhalten der Maschine steuern) könnte indessen ein Anlass sein, der Maschine Intelligenz abzusprechen, insofern man der Auffassung ist, dass (menschliche) Intelligenz immer mehr umfassen muss als bloße Programme in *diesem* Sinne (einer Beschreibung in einer Programmiersprache, welche die Algorithmen/Prozeduren angibt, die das Verhalten der Maschine steuern). Programme in *diesem* Sinne (von Sprachen wie C, Java, LISP, PROLOG etc.) lassen sich in deterministischen Assembler und Maschinencodes compilieren, enthalten also keine nicht-konstruktiven oder nicht-determinierte Schritte. Aus dieser Perspektive wäre auch die Supposition „Angenommen wir haben eine Machine, die den Turing-Test besteht und betrachten nun ihr Programm“, die gelegentlich in Diskussion um den TT gemacht wird, in ihrem zweiten Teil zurückzuweisen: selbst wenn wir – um eines Argumentes willen und angesichts eines Einräumens einer vielleicht gegebenen logischen Möglichkeit einer Maschine, die den TT besteht – den ersten Teil der Supposition zugestehen, bleibt der zweite Teil noch problematischer: eine Maschine, welche den TT besteht, wurde sicher nicht programmiert mit C, …

§10 Kulturkritisch betrachtet scheinen viele Thesen zur (universellen) Berechenbarkeit zu einem Feld zu gehören, in das auch Thesen (und Spekulationen) zur starken KI, Künstlichen Leben und Robotik fallen. Insofern die Überzeugung und der Eifer, mit dem diese Thesen vorgetragen werden, sich entkoppelt hat von ihrer empirischen Bestätigungsfähigkeit, lebensweltlichen Fundierung und methodischer Reflexion, tragen diese Postulate Züge einer säkularisierten Eschatologie bzw. Ersatzreligion.

Die Grenzen der starken KI zeigen sich auch in der mangelnden ethischen Reflexion der mit ihr auftretenden neuen Probleme. Diese betreffen nicht allein die oft diskutierte Frage, ob eine neue übermenschliche Intelligenz sich friedlich zu Menschen verhält. Wie fügt sich diese starke KI (selbst) in die moralische Gemeinschaft ein? Nur insofern wir auch ein komplettes Verständnis des praktischen Räsonierens besäßen, wären wir in der Lage, eine genuin *apriorische* Moralität auch bei einer KI zu reproduzieren.

Auch bezüglich des bloßen Hervorbringens dieser Intelligenz als intelligentem Wesen stellen sich mehrere praktische Fragen. Mit der Idee der Entwicklung und anschließenden Existenz superintelligenter postbiologische Lebensformen stellen sich auch ethische Fragen, wie

(i) Sind diese Forschungsvorhaben ethisch zulässig?
(ii) Worin besteht der Sinn und Nutzen der postbiologischen Lebensformen für die Erschaffenden?

Handelt es sich nicht um Experimente an – nach Voraussetzung – empfindungsfähigen Wesen? Trifft dies zu, wird es wenig Ziele geben, die solche Experimente rechtfertigen können. Nach Voraussetzung sind die von uns bei solchen Experimenten übergangenen Präferenzen noch gewichtiger als die von Tieren. Die Pflicht zur Schmerzvermeidung gilt des Weiteren schon für alle künstliche Lebensformen, die empfindungsfähig sind, unabhängig vom Ausmaß ihrer weiteren Intelligenz.

Und was zeigt das Erzeugenwollen einer starken KI in dieser Hinsicht? Die Erzeuger könnten eine Sklavenhaltermoral umsetzen und ihre persönlichen Vorlieben in sie einbauen.

Die ‚Singularitätsthese' (des Auftretens der übermenschlichen Intelligenz) als Heilsversprechen weist alle Merkmale solcher Verkündigungen auf: eine grundsätzliche, positive, kaum zu begreifende Transformation des menschlichen Daseins wird in einem nicht sehr genauen Zeithorizont angekündigt (so bei Kurzweil 2005). Die religiöse Dimension ähnelt dabei naiven Vorstellungen des ewigen Lebens als einem Immer-weiter-so des bisherigen Lebens. Der entsprechende technische Fortschritt wird gelöst betrachtet von seiner gesellschaftlichen Voraussetzungen bzw. die globale Ordnung nicht hinterfragt, und die Frage nach der Rolle auftretender KI in dieser Ordnung für die bestehenden Asymmetrien in dieser Ordnung nicht gestellt. Die Nichtbeachtung solcher Fragen sagt etwas über die Protagonisten der Singularitätsthese, tritt aber insofern in den Hintergrund, als man davon ausgehen kann, dass selbst bei einer Thematisierung dieser Fragen die technische Entwicklung als unaufhaltsam angesehen würde. Oder es wird die paradiesische Überholung der bestehenden Ordnung Teil der Verkündigung der kommenden Singularität, wie die Singularität in Einem auch die Energie- und Rohstoffnutzung harmlos und nachhaltig machen soll. Bei Kurzweil beseitigt die prophezeite Superintelligenz alle jetzigen und zukünftigen Probleme der Menschheit, sodass Forschung in der starken KI geradezu Forschung an der Lösung dieser Probleme wird. Mit den entsprechenden Versprechen sollen scheinbar alle ethischen Nachfragen abgeschnitten werden.

Mit der hier verteidigten Auffassung des Computermodell des Geistes und der damit einhergehenden Rolle einer schwachen Künstlichen Intelligenz hat dies nicht mehr viel zu tun.

Literaturverweise:

- Kurzweil, Ray. *The Singularity Is Near*. When Humans Transcend Biology. New York, 2005.

- Moor, James H. “An analysis of the Turing test”, *Philosophical Studies,* 30 (1976), S. 249–257.
- Putnam, Hilary. *Repräsentation und Realität.* Frankfurt a.M., 1991.
- Pylyshyn, Zenon. *Computation and Cognition*. Cambridge/MA, 1986.
- Turing, Alan. „Computing Machinery and Intelligence“, *Mind*, 59 (1950), S. 433-60.