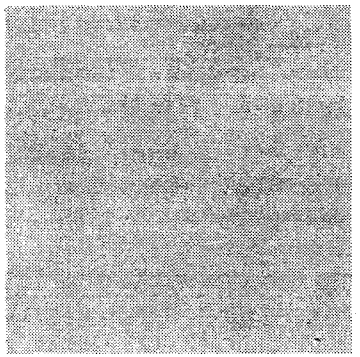

Max Kerner, Karl Kegler (Hg.)

Der vernetzte Mensch

Sprache, Arbeit und Kultur in der Informations-gesellschaft

Verlag Mainz

Herausgeber:

Universitätsprofessor Dr. phil. Max Kerner
Dipl.-Ing. Karl R. Kegler

Konzept und Organisation:

Forum „Technik und Gesellschaft"
der Rheinisch-Westfälischen Technischen Hochschule Aachen,
Kármánstraße 11, 52062 Aachen

Titelbild: Pieter Breughel d. Ä., Turmbau von Babel, 1563
Kunsthistorisches Museum Wien

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Der vernetzte Mensch : Sprache, Arbeit und Kultur in der Informationsgesellschaft /
Max Kerner, Karl Kegler (Hg.). - Aachen : Mainz, 1999
ISBN 3-89653-549-8
NE: Kerner, Max; Kegler, Karl [Hrsg.]

Süsterfeldstraße 83, 52072 Aachen
Tel. 0241/873434
Fax 0241/875577
1. Auflage 1999

Gestaltung: THOUET Verlag
Druck: Druckerei Mainz
Süsterfeldstraße 83, 52072 Aachen
Satz und Lithographie: FotoCom, Aachen
Printed in Germany
ISBN 3-89653-549-8

Dieter Wandschneider

Scheitert das Projekt Künstlicher Intelligenz an Gödels Unvollständigkeitstheorem?

Der Sinn von Technik ist klassisch als Organverstärkung und Organüberbietung bestimmt worden. Bezüglich der modernen Computertechnologie greifen solche Charakterisierungen zu kurz. Hier ist gewissermaßen ein neues Paradigma technischer Entwicklung aufgetaucht, das tendenziell als Selbstreproduktion des intelligenten Wesens selbst gefaßt werden könnte und im Projekt „Künstlicher Intelligenz" prägnanten Ausdruck findet.

Es liegt auf der Hand, daß von dieser „Intelligisierungstendenz", wie ich kurz sagen möchte, in der Entwicklung informationstechnischer Systeme zunehmend auch das *Selbst- und Weltverständnis* des Menschen entscheidend mitbetroffen ist: Es muß Folgen für sein *Selbstverständnis* haben, wenn seine eigene Intelligenz als grundsätzlich technisch reproduzierbar erscheint, und sein *Weltverständnis* kann von der Ausbreitung intelligenter Systeme nicht unbeeinflußt bleiben, wenn ihm diese als allgegenwärtige, quasi autonome Strukturen gegenübertreten. Und es wäre sicher von einschneidender, über das rein Akademische weit hinausreichender existentieller Bedeutung, wenn so etwas wie „Künstliche Intelligenz" im Wortsinn realisierbar wäre.

Daß etwas Derartiges möglich sei, ist schon sehr früh von J.R. Lucas bestritten worden (1964). Und zwar glaubt Lucas diesbezüglich prinzipielle Argumente in *K. Gödels* sogenannten *Unvollständigkeitstheoremen* zu haben (Gödel, 1933). Seine Kritik läßt sich kurz so charakterisieren: Gödels Theoreme besagen etwas über prinzipielle Grenzen formaler Systeme. Maschinen lassen sich im gewissem Sinn als Realisierungen formaler Systeme auffassen. Also besagen Gödels Theoreme auch etwas über prinzipielle Grenzen von Maschinen.

Im folgenden möchte ich in einem ersten Teil zunächst Gödels Unvollständigkeitstheoreme und Lucas' darauf gegründete Argumentation darlegen. Im zweiten Teil soll versucht werden, den eigentlichen Grund für Gödels Resultate sichtbar zu machen, im dritten Teil wird der Logiker in die Betrachtung einbezogen, um im vierten Teil die Konsequenzen aus diesen Überlegungen für das Projekt Künstlicher Intelligenz darzustellen.

1. *Gödels Unvollständigkeit als Achillesferse künstlich-intelligenter Systeme?*

Gödels Unvollständigkeitstheoreme selbst kann ich hier nicht im Detail entwickeln. Ich werde mich vielmehr auf eine vereinfachende, das Grundsätzliche betonende Darstellung beschränken: Gödel hat *erstens* gezeigt, daß in einem Logiksystem *S* wie dem der *Principia Mathematica* von Russell, in dem auch die Arithmetik formalisiert werden kann, ein Ausdruck konstruierbar ist – nennen wie ihn *G* –, der in arithmetischer Verschlüsselung seine eigene Unbeweisbarkeit ausdrückt. Gödel zeigt nun, daß *G* im Rahmen des Systems *S* prinzipiell unbeweisbar ist, sofern das System widerspruchsfrei ist. Daß *G* unbeweisbar ist, ist aber auch die inhaltliche Aussage von *G* selbst, so daß sich *G* zugleich als *wahr* erweist. Es gibt im System *S* mit andern Worten einen *wahren* Satz – eben *G* –, der gleichwohl prinzipiell *unbeweisbar* ist, und in diesem Sinn wird das System *S unvollständig* genannt. Die wahren Sätze des Systems sind solchermaßen nicht sämtlich auch formal beweisbare Sätze. Wie Gödel ferner zeigt, ist *G* zudem *unentscheidbar* in dem Sinn, daß weder *G* noch *non-G* im System *S* beweisbar ist (sofern dieses widerspruchsfrei ist). Gödels *2. Theorem* hängt mit dem ersten zusammen; es besagt: Die *Widerspruchsfreiheit* des Systems *S*, in dem *G* konstruiert ist, ist prinzipiell *nicht innerhalb* des Systems selbst beweisbar, sondern gewissermaßen nur „von außen", von einem Metasystem her. Im folgenden werde ich mich auf das grundlegende *1. Theorem* beschränken.

Gödels Resultate konstatieren also prinzipielle Grenzen formaler Systeme. Wenn nun Computer technische Realisierungen formaler Systeme sein sollten, dann muß es, so der schon erwähnte Grundgedanke von Lucas, auch prinzipiell unüberschreitbare Grenzen computertechnischer Systeme geben.

Auf der anderen Seite, und darin besteht hier die eigentliche Pointe, kann sich der *Logiker* über diese Grenzen formaler Systeme hinwegsetzen; denn er kann *G* ja als wahr erweisen. Der Logiker übertrifft die Maschine offenbar in einem wesentlichen Sinn. Ist dem so, dann muß das auch gravierende Konsequenzen für das Projekt Künstlicher Intelligenz haben. Wird hier also „die Achillesferse der kybernetischen Maschine" sichtbar, wie Lucas meint (S. 47)? Lucas interpretiert Gödels Argument in der Tat dahin, daß „keine Maschine ein vollständiges und adäquates Modell des Geistes sein kann" (Ebd., S. 44), d. h. „wir können keine Maschine bauen, die geistartiges Verhalten in *jeder* Hinsicht zu simulieren vermag". „Wir können niemals, nicht einmal im Prinzip, ein technisches Modell des Geistes besitzen" (Ebd., S. 47). Ähnliche Äußerungen finden sich bei vielen anderen Autoren bis hin zu Popularisierungen im Stile D. R. Hofstadters oder gar „postmoderner" Inanspruchnahme bei J.-F. Lyotard (S. 70) – die Gödel Theoreme sind inzwischen auch so etwas wie ein Mythos.

In diesem Zusammenhang ist zunächst auf eine grundsätzliche Unklarheit aufmerksam zu machen. Die angegebene Argumentation: Maschinen sind technische Realisierungen formaler Systeme; Gödel zufolge gibt es grundsätzliche Grenzen formaler Systeme; also gibt es grundsätzliche Grenzen maschineller Systeme – diese Argumentation scheint unmittelbar aus Gödels Resultat zu folgen. Das Unzutreffende einer solchen Auffassung wird aber sofort deutlich, wenn man sich vergegenwärtigt, daß die zuerst genannte Prämisse („Maschinen sind technische Realisierungen formaler Systeme") jedenfalls nicht aus *Gödels Theoremen* folgt. Über den Charakter von Maschinen sagen diese schlechterdings nichts aus, so daß *von daher* auch nichts über prinzipielle Grenzen von Maschinen zu erschließen

ist. Dies wäre nur mit der genannten Prämisse möglich, die auch in Lucas' Argumentation enthalten (S. 46), aber eben nur eine *Prämisse* derselben ist und nicht etwa eine Konsequenz aus Gödels Theoremen. Das ist wohl zu beachten.

Mit dieser Prämisse, daß Maschinen technische Realisierungen formale Systeme seien, wird dann Lucas' Argument verständlich, daß das Denken des *Logikers* etwas vermag, was für die Maschine selbst unmöglich ist: Er kann den Gödelschen Ausdruck *G*, den die Maschine zwar bilden, aber nicht beweisen kann, seinerseits *als wahr erweisen* (Ebd., S. 47). Lucas macht somit nicht nur grundsätzliche *Grenzen* maschineller Systeme geltend, sondern darüber hinaus auch die positive Aussage, daß der Logiker der Maschine in einem prinzipiellen Sinn *überlegen* sei.

Das hiermit angesprochene Problem betrifft also nicht primär die Frage, ob Maschinen formale Systeme sind, was, wie schon bemerkt, von Gödel her gar nicht zu klären ist. Es geht hier mehr noch um das *Verhältnis* von Logiksystem und Logiker und dabei insbesondere um die Frage, unter welcher Bedingung der Logiker in der Lage ist, den im System *S* unbeweisbaren Satz *G* (der seine eigene Unbeweisbarkeit aussagt) gleichwohl als *wahr* zu erweisen, und das heißt ja, den Beweis der Unbeweisbarkeit von *G* zu führen. Um diesbezüglich zu einer Klärung zu kommen, soll im folgenden versucht werden, den eigentlichen *Grund* für die Unbeweisbarkeit von *G* und die erweiterten Beweismöglichkeiten des Logikers auszumachen. Dazu muß zunächst der von Gödel konstruierte Ausdruck *G* näher ins Auge gefaßt werden.

2. Zum Grundsätzlichen der Gödelschen Konstruktion

Entscheidend, das ist meine *erste These*, ist in diesem Zusammenhang die *Selbstreferentialität* des Ausdrucks *G*; dieser sagt ja von sich selbst die Unbeweisbarkeit aus. Daher kann es, wie man sich leicht überzeugen kann, nur die beiden folgenden Möglichkeiten geben: *G* kann *entweder* die Eigenschaft der Unbeweisbarkeit besitzen und damit wahr sein, *oder G* ist beweisbar und dann, *G*'s Bedeutung entsprechend, falsch. Eine dritte Möglichkeit kann es aufgrund der Selbstreferentialität von *G* nicht geben (wie man sich leicht vergegenwärtigt).

Ist das betrachtete System *S* nun insbesondere ein *semantisch korrektes* System, d. h. ein solches, in dem alle beweisbaren Sätze stets wahre Sätze sind, so scheidet die letztgenannte der beiden Alternativen aus; denn *G* kann dann nicht beweisbar und zugleich falsch sein, und es bleibt somit nur die andere Möglichkeit, daß *G* unbeweisbar und wahr ist.

Diese einfache Überlegung zeigt, daß ein Ausdruck wie *G*, der seine eigene Unbeweisbarkeit ausdrückt, *allein aufgrund seiner Selbstreferenz* unbeweisbar und zugleich wahr sein *muß*, sofern das System korrekt ist, d. h. keine falschen Sätze zu beweisen gestattet. Der Ausdruck *G* ist gerade so konstruiert, daß er nicht beweisbar sein *kann*. Damit ist sozusagen ein erster Blick hinter die Kulissen des Unvollständigkeitstheorems getan.

Allerdings macht die hier durchgeführte Argumentation mit der Annahme semantisch korrekter Systeme eine *stärkere* Voraussetzung als Gödels eigener Beweis; dieser benötigt nur die schwächere Bedingung *formaler Widerspruchsfreiheit* bzw. sogenannter *Omega-Widerspruchsfreiheit*. Unter dieser Voraussetzung ist beweisbar, und das tut Gödel, daß *G* im System *S formal unentscheidbar*, d. h. weder *G* noch *non-G* beweisbar ist. Es läßt sich aber zeigen, daß auch Gödels Beweisgang entscheidend auf der *selbstreferentiellen* Struktur des Satzes *G* beruht.[1] Nun ist „Referenz" und damit auch „Selbstreferenz" ein *semantischer* Begriff. Und in diesem Sinn möchte ich als

eine *zweite These* formulieren, daß hier *semantische* Strukturen eine entscheidende Rolle spielen. Wie ist das zu verstehen?
In diesem Zusammenhang ist das von Gödel verwendete Verfahren einer *arithmetischen Kodierung* der Ausdrücke des betrachteten Systems von Bedeutung. Dieses Verfahren, heute auch als *Gödelisierung* bezeichnet, besteht bekanntlich darin, daß den Grundzeichen des Systems, den aus diesen gebildeten Formeln sowie den Folgen von Formeln eindeutig natürliche Zahlen – „Gödelzahlen" genannt – zugeordnet werden. Wie dies geschieht, ist hier unwichtig. Wesentlich ist, daß es Gödel mittels dieses Kunstgriffs gelang, einen Teil der Metatheorie des Systems in das System selbst zu integrieren, oder konkreter: Durch Gödelisierung können bestimmte Klassen von Ausdrücken, z. B. die Klasse der beweisbaren Formeln, durch rein arithmetische Beziehungen charakterisiert werden. Die Gödelzahlen, die den im System beweisbaren Theoremen zugeordnet sind, gehören etwa einer wohlbestimmten Zahlklasse, sagen wir T, an. Der *metatheoretischen* Aussage, daß eine bestimmte Formel ein Theorem ist, korrespondiert so eine *arithmetische* Aussage: Nämlich, daß die Gödelzahl dieser Formel zur Zahlklasse T gehört.

1/ Das sei hier nur angedeutet: Korrektheit impliziert *formale Widerspruchsfreiheit*; diese stellt also eine schwächere Bedingung als Korrektheit dar. Gödel hat gezeigt, daß die zweite der genannten Möglichkeiten („beweisbar und falsch") auch bei dieser *schwächeren* Voraussetzung entfällt:

Annahme:	G	ist beweisbar
induziert	$>G$ ist beweisbar$<$	ist beweisbar
	$=>non\text{-}G<$ (wegen der Selbstreferentialität von G!)	
also	*non-G*	ist beweisbar,

also reductio ad absurdum der Annahme (sofern das System widerspruchsfrei ist), wobei die *Selbstreferentialität* entscheidend eingeht! Unter der Voraussetzung von „Omega-Widerspruchsfreiheit" folgt aus der Unbeweisbarkeit von G auch die Unbeweisbarkeit von *non-G*, d. h., G ist ein im System *unentscheidbarer* Satz.

Mit der Gödelisierung der Ausdrücke des Systems S – so läßt sich die eben formulierte These weiter konkretisieren – ist nun in der Tat eine *semantische* Ebene im System etabliert. Denn jede Gödelzahl hat ja aufgrund dieser Zuordnung eine *Interpretation*, d. h., sie referiert auf den ihr zugeordneten Ausdruck: Ein Grundzeichen, eine Formel, insbesondere etwa auch eine *beweisbare* Formel, mit anderen Worten: Diejenigen Zahlen, die aufgrund der Zuordnungsvorschrift Gödelzahlen sind, sind damit nicht mehr *nur* Zahlen, sondern besitzen außerdem eine Interpretation und stehen dergestalt nicht mehr nur in formalen, sondern auch in *semantischen* Relationen. Die ganze Konstruktion ist dadurch schon im Ansatz semantisch orientiert. Soweit ich sehe, bleibt dieser Begleitumstand der Gödelisierung bei Gödel selbst und in der Literatur völlig unterbelichtet.
Tatsächlich aber hat dieser semantische Aspekt der Gödelisierung für Gödels' Konstruktion zentrale Bedeutung: Entscheidend ist ja, wie dargelegt, die Selbstreferenz von G. Diese aber ist überhaupt erst durch die Gödelisierung ermöglicht[2] – das ist der springende Punkt! Gödels Konstruktion kann in dieser Perspektive etwa so nachvollzogen werden: Die Zahlklasse T der Gödelzahlen der Theoreme im System S kann über die Gödelisierung arithmetisch definiert werden, indem der Beweisbegriff arithmetisch gefaßt wird (nämlich mit Hilfe der Gödelzahlen der Axiome und der Deduktionsregeln). Gödels Verfahren besteht nun in der Konstruktion eines Ausdrucks G, dessen Gödelzahl *garantiert nicht* zu T gehört. Zu garantieren ist das aber, wie dargelegt, nur in der Weise, daß dieser Ausdruck einerseits auf T rekurriert – hier durch den Beweisbarkeitsbegriff –, um sich, bildlich gesprochen, von T „abzustoßen", d. h., *die Gödelzahl g des Ausdrucks G wird als ein Nichtelement von T konstruiert, und zwar –*

2/ Es sind auch andere Verfahren zur Herstellung von Selbstreferenz möglich. R.M. Smullyan z. B. verwendet die *Anführung* zur Bezeichnung von Ausdrücken (Smullyan, S. 66, R1, R2).

das ist, wie sich gezeigt hat, entscheidend – *in selbstreferentieller Form*. In formal-logischer Schreibweise ist dies etwa durch

$$(G \leftrightarrow \neg\ (Ey)\ bew(g,y)$$

darstellbar (in Worten: Es existiert keine Gödelzahl *y* für eine Formelfolge, die ein Beweis für die Formel mit der Gödelzahl *g* wäre – wobei *g* die Gödelzahl von *G* selbst ist). Hierbei ist durch *bew* auf *T* Bezug genommen und durch die Gödelzahl *g* auf den gesamten Ausdruck *G* selbst, der dadurch auf sich selbst referiert[3]. Ein Beweis von *G* impliziert solchermaßen einen Widerspruch (vgl. Fußnote 1): eine „Beweisblockade" sozusagen.

Auf diese Weise ist der Ausdruck *G* gerade so konstruiert, daß er zwangsläufig „außer Reichweite" der Axiome des Systems *S* und damit formal beweisbar ist, und diese Möglichkeit beruht auf einer Referenz- bzw. Selbstreferenzbeziehung und damit auf einer semantischen Struktur. So gesehen wird die „formale Unvollständigkeit" fast zu einer Tautologie: Das über die formale Ebene der Axiome Hinausgehende ist gewissermaßen das *Semantische* an Gödels' Formel. Bemerkenswert ist, daß die Arithmetik derartige Möglichkeiten einschließt: und zwar – damit komme ich zum Ausgangspunkt dieser Überlegungen zurück – auf der Grundlage der *Gödelisierung* der Ausdrücke. Das Zahlzeichen einer Gödelzahl ist also nicht einfach eine Zeichengestalt, sondern darüber hinaus *Referent* eines (ihr durch Gödelisierung) zugeordneten Ausdrucks. Es weist über sein bloßes Formsein hinaus auf etwas von ihm selbst Verschiedenes. Die rein *formale* Ebene ist in Gödels' Konstruktion somit bereits verlassen. Dieser folgenreiche Begleitumstand der Gödelisierung bleibt, wie schon erwähnt, bei Gödel selbst und in der Gödelliteratur so gut wie unbemerkt (eine Ausnahme stellt diesbezüglich v. Kutschera dar).

Daß Gödels Resultate oft falsch eingeschätzt worden sind, ist sicher auch auf eine Reihe *irreführender Formulierungen* in der Gödelliteratur zurückzuführen. So zum Beispiel, wenn gesagt wird, daß *G* *nicht mit den Mitteln des Systems selber beweisbar sei* (Frey, S. 200). Aber *entscheidend* ist, wie gesagt, daß der Satz *G* aufgrund seiner selbstreferentiellen Struktur nicht beweisbar sein *kann*; denn von den *beweistechnischen Mitteln* des Systems hängt die Unbeweisbarkeit von *G* tatsächlich nicht ab. Irreführend ist auch die Aussage bei Nagel/Newman, demzufolge Gödels Theoreme „eine grundlegende Begrenzung für die Reichweite der axiomatischen Methode" bedeuten sollen (S. 93), denn, so die Folgerung, die von den Autoren gezogen wird: Die Existenz eines Satzes, der, wie *G*, wahr und gleichwohl unbeweisbar ist, zeige, daß es „arithmetische Wahrheiten gibt, die nicht formal beweisbar sind" (Ebd. S. 99, ähnlich S. 85 und 96) – was sehr eigenartig wäre, nämlich wenn sich gewisse Zahlenverhältnisse logischer Ausweisbarkeit entziehen sollten! Hier kommt es darauf an, zwischen dem Beweis des *Gödelschen Sachverhalts* und des ihm in *S* korrespondierenden formalen *Ausdrucks G* klar zu unterscheiden; denn der *Sachverhalt* der Unbeweisbarkeit von *G* und damit die ihm entsprechende „arithmetische Wahrheit" wird ja tatsächlich streng bewiesen, nur eben nicht der *Ausdruck G*, der diesen Sachverhalt mit den formalen Mitteln von *S* selbst formuliert. Daß ein derartiger Ausdruck nicht beweisbar ist, beruht nach dem Vorigen anderseits nicht auf einem Mangel des in *S* installierten formalen Beweisverfahrens, sondern – wie gesagt – auf der *Selbstreferenz* von *G*.

3. Einbeziehung des Logikers

Nach diesen Überlegungen zum Grundsätzlichen der Gödelschen Konstruktion soll nun auch der *Logiker* in die Betrachtung einbezogen werden, denn die Frage ist doch: Wieso kann dieser den Ausdruck *G* als *wahr* erweisen? Diese Frage ist notwendig, wenn man bedenkt, daß der Logiker ja einen Sachverhalt beweist, der (bei inhaltlicher Deutung) *von G selbst auch* formuliert wird, nämlich daß

3/ Die konstruktive Entsprechung mit dem sogenannten Cantorschen Diagonalverfahren ist ebenfalls unübersehbar.

G unbeweisbar ist: Hat er somit nicht doch *G* bewiesen und sich dergestalt in einen Widerspruch verstrickt? Dies ist, wie man weiß, nicht der Fall. Aber wie ist die Frage dann zu beantworten? Wie kann der Logiker die *Un*beweisbarkeit von *G* und damit im gewissem Sinn eben doch *G*, nämlich dessen *Wahrheit*, beweisen?
Das Problem ist also das folgende: Gödels Theorem besagt, daß der Satz *G* unbeweisbar ist. Aber das ist exakt die gleiche Aussage, die *G* über sich selbst macht. Doch *G* ist unbeweisbar, während der Logiker seine Aussage beweist. Wie ist das möglich? Nun, wesentlich ist – das ist meine *dritte These* –, daß der Logiker selbst auf einer anderen, vom System *S* verschiedenen Sprachebene operiert. Daß *G* unbeweisbar ist – und das ist genau die Aussage, die *G* selbst macht –, wird ja vom Logiker nicht auf der *G*-Ebene (also auf der Sprachebene des Ausdrucks *G* selbst) formuliert, sondern auf einer davon verschiedenen *Metaebene*. Man muß also genau unterscheiden zwischen der Formulierung auf der *G*-Ebene, daß *G* unbeweisbar ist – das ist der Ausdruck *G* selbst – und der in die Metasprache übersetzten Formulierung – nennen wir sie *Ü* – desselben Sachverhalts. *G* und *Ü* besagen also dasselbe, nämlich daß *G* unbeweisbar ist; aber *G* ist ein Satz der *G*-Ebene, während die Aussage *Ü*, die denselben Sachverhalt ausdrückt, zur Metasprache des Logikers gehört.
Damit ist nun zugleich ein entscheidender *struktureller* Unterschied von *G* und *Ü* involviert: *G*, so hatten wir gesehen, ist *selbstreferentiell* und *deshalb*, wie dargelegt, unbeweisbar. Die übersetzte Formulierung *Ü* hingegen macht zwar dieselbe Aussage, aber – als Übersetzung – *über* den von *Ü* verschiedenen Satz *G*. *Ü* ist folglich *nicht selbstreferentiell* und kann daher, im Unterschied zu *G*, grundsätzlich *beweisbar* sein, mit anderen Worten: Der Logiker kann *beweisen*, was *innerhalb* des Systems unbeweisbar ist, indem er den *Übergang zur Metaebene* vollzieht. Er kann dadurch nämlich den fraglichen Sachverhalt, daß *G* unbeweisbar ist, in einer *nicht*-selbstreferentiellen Form *Ü* ausdrücken, der auf der *G*-Ebene nur in selbstreferentieller Form existiert und dort genau deshalb, wie gezeigt, prinzipiell nicht beweisbar sein kann.

Diese Überlegungen lassen sich dahin zusammenfassen, daß der Logiker zu einer *Reflexionsleistung* befähigt ist in dem Sinn, daß er die *G*-Ebene verlassen, sie zum Gegenstand seiner Betrachtung macht, d. h. auf der Metaebene darüber „reflektieren" kann. Der Logiker kann dort den Sachverhalt beweisen, den auch *G* formuliert, d. h., er kann *G* als *wahr* erweisen. Genau das ist es, was Gödel tut. Möglich, so hat sich gezeigt, ist dies dadurch, daß der durch *G* formulierte Sachverhalt auf der Metaebene in einer *nicht*-selbstreferentiellen Übersetzung *Ü* verfügbar ist und in dieser Form nun auch beweisbar sein kann, d. h., die auf der *G*-Ebene geltende *Beweisblockade* für *G* (nämlich durch die Selbstreferentialität von *G*) ist auf der Metaebene *nicht* mehr gegeben. Das, so scheint es mir, ist zentral für das Verhältnis von Logiker und Logiksystem im Gödelschen Beweis, nämlich daß der menschliche Logiker etwas kann, was *innerhalb* des formalen Systems oder der (heutigen) Maschine nicht möglich ist. Es beruht also einfach darauf, daß er aus dem formalen System sozusagen *aussteigen*, auf die *Metaebene* übergehen und sozusagen „von oben" in das System hineinschauen kann – ähnlich gibt es auch beim *Skatspiel* für die Mitspieler *Grenzen* möglichen Wissens, aber nicht für den, der allen über die Schulter schaut.

4. Konsequenzen für das Projekt Künstlicher Intelligenz

Man kann nun die Frage stellen, ob möglicherweise auch eine *Maschine* an die Stelle des Logikers treten und einen Gödelschen Beweis führen könnte, oder ob gerade die Gödels Theoreme etwas Derartiges verunmöglichen, wie etwa Lucas meint (vgl. 1. Kap.). Nun, soll die Maschine dazu befähigt sein, so muß sie, soviel ist nach dem eben Gesagten deutlich, eine Reflexionsleistung wie der Logiker, d. h. den Übergang zur Metaebene vollziehen können. Es gilt also in diesem Zusammenhang zu klären, ob das grundsätzlich möglich ist. Ganz zweifellos handelt es sich hierbei um ein weites Feld, dessen Bearbeitung, soweit ich sehe, im Grunde noch gar nicht

in Angriff genommen und auch in dem hier vorgegebenen Rahmen nicht leistbar ist.

Was indes sicher gesagt werden kann – so denke ich – ist, daß *Gödels* Theoreme den Reflexionsübergang auf die Metaebene jedenfalls in keiner Weise *behindern*. Sie selbst beruhen ja wesentlich auf dieser Möglichkeit des Logikers, und ob eine Maschine in dieser Hinsicht dem Logiker ebenbürtig sein kann oder nicht, ist durch Gödels Theoreme in keiner Weise präjudiziert oder auch nur tangiert. Selbst wenn eine Maschine nichts als die technische Realisierung eines formalen Systems *wäre* – was, wie schon bemerkt, von Gödel her überhaupt nicht geklärt wird –, so würden die Gödeltheoreme in diesem Fall nur Grenzen bezüglich der Beweisbarkeit bestimmter Maschinensätze markieren, aber *nicht* bezüglich der Möglichkeit, Reflexionsleistungen zu implementieren.

Die bisherige Inanspruchnahme der Gödeltheoreme für das Mensch-Maschine-Problem und damit auch für das Projekt Künstlicher Intelligenz hat offenbar einen ganz falschen Problemfocus: Daß es für logische Systeme und auch für Computerprogramme gewisse „Gödelsche Blockaden" gibt, ist richtig, aber für das Mensch-Maschine-Problem im Grunde irrelevant. Charakteristisch für den Logiker ist, wie gesagt, die Möglichkeit, zur Metaebene überzugehen. In diesem Sinn wäre die entscheidende Frage somit: Ob die Maschine möglicherweise ebenfalls einen solchen Übergang vollziehen kann. *Das* hat mit den Gödelschen Grenzen formaler Beweisbarkeit aber überhaupt nichts zu tun. Die sich darauf versteifende Argumentation hat den zentralen Punkt des Mensch-Maschine-Problems verfehlt. Dieses besteht nicht darin, daß es eine Beweisblockade (auf der *G*-Ebene) gibt, sondern ob man sich von dieser durch Übergang auf die Metaebene *befreien* kann.

Diese Kritik gilt analog für die originelle Wendung, die R. Penrose dem Problem kürzlich gegeben hat. Penrose ist, ähnlich wie Lucas, der Meinung, daß die Gödeltheoreme die Überlegenheit des Logikers über jede mögliche Maschine beweisen. Wesentlich dafür, so wird von Penrose argumentiert, sei die Möglichkeit des Logikers, die

Existenz Gödelscher Beweisbarkeitsgrenzen *einzusehen*. Diese Möglichkeit intelligenter „Einsicht" wird von Penrose näher als eine Art Platonischer „Ideenschau" gedeutet. Der Logiker *sieht* gleichsam die logischen Zusammenhänge, und genau das, meint Penrose, sei aufgrund der Gödeltheoreme von einer – notwendig *algorithmisch* konstruierten – Maschine prinzipiell nicht zu erwarten. Auch für Penrose haben die Gödeltheoreme also einschneidende Konsequenzen für das Projekt Künstlicher Intelligenz, hier insbesondere mit einer Platonistischen Pointe. In dem das *Bewußtsein*, so wird argumentiert, eine Art Kontakt mit der platonischen Welt ideeller Entitäten darstellt, können wir logische *Einsichten* haben, die als ein *Erschauen* logischer Strukturen aber wesentlich *nicht-algorithmischer* Natur sein sollen. Genau in diesem Sinn soll menschliches Denken und Bewußtsein jeder möglichen Maschine prinzipiell überlegen sein.

Nun, das mag so sein oder auch nicht, aber das ist, denke ich, nicht die entscheidende Frage. Wesentlich für die Möglichkeit des Gödelschen Beweises ist vielmehr, so haben wir gesehen, daß der Logiker den *Übergang zur Metaebene* vollziehen kann. Für den Mensch-Maschine-Vergleich wäre somit zu klären, ob die Möglichkeit eines solchen Übergangs nur für den Menschen oder grundsätzlich auch für die Maschine besteht. Bezogen auf Penroses Argumentation entspricht dem die Frage, ob algorithmisch verfaßte Maschinen solche Übergänge prinzipiell ausschließen und, wenn das der Fall sein sollte, ob auch nicht-algorithmische Maschinen denkbar sind oder schließlich, umgekehrt gefragt, ob menschliches Denken tatsächlich nicht-algorithmischer Natur ist. Für *diese* hier sich stellenden Fragen gibt die von Penrose vertretene Platonische Auffassung – die ich grundsätzlich teile – in der Tat nichts her.

Kurzum: Aus Gödels Resultaten folgt nicht das mindeste für oder gegen die Möglichkeit von Maschinen, einen Gödelschen Beweis wie der Logiker zu führen. Lucas' bzw. Penroses Thesen von der prinzipiellen Überlegenheit des Logikers über die Maschine aufgrund der Theoreme Gödels sind somit als nicht haltbar zurückzuweisen.

Gewiß – um möglichen Mißverständnissen vorzubeugen: An der *faktischen* Superiorität des Denkens im Vergleich mit den heute faktisch realisierten Maschinen ist nicht im mindesten zu zweifeln. Aber dieses Faktum begründet keine *prinzipielle* Differenz. Man könnte einwenden, daß das Denken seinen technischen Produkten, einfach durch seine Urheberschaft, überlegen sei. Ein Blick auf moderne Computer lehrt freilich, daß dies in *quantitativer* Hinsicht heute schon nicht mehr generell zutreffend ist – man denke nur an die extremen Rechengeschwindigkeiten und Speicherkapazitäten solcher Maschinen. Doch es wäre ignorant, die gewaltige *qualitative* Differenz von Denken und Computer für die gegenwärtige Situation zu leugnen. Noch sind wir es, nicht die Maschinen, die denken und so unter anderem auch Logiksysteme und Maschinen erfinden; und solche Gebilde sind, worauf wiederum Lucas (S. 48 ff.) eindringlich hingewiesen hat, *beschränkte, fixierte* Gestalten, denen wir, als deren Bildner, stets viele Schritte voraus sind. Wir haben die Fähigkeit selbstüberholender Reflexion, die uns instand setzt, unseren jeweiligen Zustand immer noch zu überbieten. Wir sind uns gewissermaßen selbst voraus und dabei – das ist wesentlich – dennoch dasselbe *identische Subjekt*. Dieses schon aus der philosophischen Tradition geläufige Argument – zu erinnern wäre an Kants *transzendentale Apperzeption*, Fichtes Reflexionsbegriff, Hegels Nachweis der Begriffsstruktur von Subjektivität oder auch an anthropologische Konzepte bei Scheler, Plessner, Gehlen und nicht zuletzt auch bei Heidegger – dieses bekannte Argument, das auch von Lucas ins Spiel gebracht wird, ist schwerlich bestreitbar. Aber Lucas irrt, wenn er meint, daß sich die technische Rekonstruierbarkeit eines solchen Subjekts aufgrund des *Gödelschen Unvollständigkeitstheorems* grundsätzlich verbiete. Zumindest die – für die hier diskutierte Frage wesentliche – Möglichkeit, den Übergang zur Metaebene zu vollziehen, wird durch Gödels Resultat nicht im geringsten ausgeschlossen, in gewissem Sinne sogar erzwungen.

5. *Fragen bezüglich prinzipieller Möglichkeiten und Grenzen maschineller Systeme*

Eine – im Gegensatz zu Lucas' tendenziell antimechanistischer Gödeldeutung – umgekehrt dezidiert *mechanistische Position* wird in dem anregenden Buch von J. C. Webb vertreten. Hatte Lucas Gödels Resultate als die „Achillesferse" der Maschine bezeichnet (Lucas, S. 47), so nennt Webb sie demgegenüber „guardian angels" (Webb, S. 202 und S. 208), Schutzengel des Mechanismus in dem Sinn, daß die Unvollständigkeit formaler Systeme im Blick auf das Verhalten von Maschinen zu nicht berechenbaren, unvorhersagbaren Prozessen führe (Ebd. S. 200, 209 und 245 f.) und selbst die Möglichkeit von „self-reflection" involviere (S. 246, 235), wovon, so Webb, „frühere Mechanisten nur hätten träumen können" (S. 235). Gödels Theoreme seien, recht verstanden, „genau das, was der Doktor dem Mechanismus verordnet" habe (S. 200), so daß „Mechanisten ihren Glückssternen für seine Beweise danken" könnten (S. 245). Webb illustriert seine Auffassung am Beispiel einer „Gödelmaschine", wie er sie nennt, die mit anderen Maschinen kommuniziert „wie eine Person, die genau zu denen *Hallo* sagt, die zu ihr *Hallo* sagen" (S. 234), ein Vorgang, der durch *Hineinverlegung* in die *Gödelmaschine* auch „more introspective" gestaltet werden könne (S. 235).

Nun, das ist wohl zu einfach: Die Möglichkeit unbestimmten, nicht berechenbaren Maschinenverhaltens aufgrund Gödelscher Unvollständigkeit[4] soll eine Affinität zu menschlichem, sich selbst als frei verstehendem Handeln suggerieren (Ebd., S. 245 f.) – eine sicher nicht weniger dubiose mechanistische Vereinnahmung Gödels als die antimechanistische, denn: Ist für das Handeln wirklich Unberechenbarkeit spezifisch, um als menschliches gelten zu können, und

4/ Dies ist eine Folge davon, daß der Satz *G*, aufgrund seiner Selbstreferentialität, eine *semantisch nicht fundierte* und damit unbestimmte Aussage darstellt.

ist für das Problem des Selbstbewußtseins irgendetwas durch den Nachweis gewonnen, daß eine solche *Gödelmaschine* zu sich *Hallo* sagen kann? Derartiges wäre zudem mit simpleren, *nicht-gödelschen* Mitteln erzielbar. Hier werden improvisierte, simplifizierende Geistmodelle in Anschlag gebracht, und Searles Kritik dieser Art von „Kognitionswissenschaft" ist nur zu berechtigt. Der Gedanke anderseits, daß Gödels Theoreme möglicherweise auch etwas zur Klärung der Probleme des Selbst und menschlicher Freiheit beitragen können, ist indessen nicht von der Hand zu weisen und mag in verschiedener Hinsicht sogar manches für sich haben. Aber beiläufige Versicherungen oder dunkle Andeutungen sind diesbezüglich, auch wenn sie, wie bei Hofstadter etwa, gehäuft auftreten (Hofstadter, S. 741 ff., S. 753 ff. und 760 ff.), wenig hilfreich.

Die eigentliche Frage im Zusammenhang mit dem Mensch-Maschine-Problem, so hat sich gezeigt, ist vielmehr die, ob man, unabhängig von Gödelschen Argumenten, die Befähigung der Maschine zu einem *Reflexionsübergang* grundsätzlich für technisch machbar hält oder nicht. Meines Erachtens ist dies eine heute offene Frage, um das Mindeste zu sagen. Wir sind zwar geneigt, die Maschine als ein *fixiertes* Gebilde zu betrachten, wie es dem herkömmlichen – im Grunde *mechanistischen* – Begriff des maschinellen oder auch formalen Systems entspricht. Eine solche begriffliche Festlegung wäre freilich, bei Licht besehen, eine Petitio principii, d. h., das Mensch-Maschine-Problem wäre damit schon vorweg entschieden. Natürlich sind Maschinen in *bestimmter* Weise strukturiert, aber sind sie darum auch *fixiert*? Auch Organismen sind in *bestimmter* Weise strukturiert oder auch neuronale Netze und Gehirne.

Aus heutiger Sicht erscheint es mir unmöglich zu sein zu sagen, was eine Maschine je können wird und was nicht. Das Modell der Turing-Maschine, das prinzipiell *alles* einschließt, was Maschinen *überhaupt* können, ist nicht etwa schon die Antwort auf diese Frage. Denn es sagt uns im Grunde nur, daß die Maschine ein operationsfähiges, algorithmisch beschreibbares System ist. Man kann aber –

zumindest heute – z. B. nicht behaupten, das menschliche Gehirn sei *etwas ganz anderes*. Insofern macht der Hinweis auf die generelle Turing-Modellierbarkeit beliebiger Maschinen gerade nicht die *spezifische Differenz* von Gehirn und Maschine sichtbar. Und die bloße Versicherung, Gehirnprozesse seien, eben als Gehirnprozesse, prinzipiell nicht technisch rekonstruierbar, wofür ich nicht die geringste Legitimation sehe: Eine solche Versicherung wäre nur wieder die genannte Petitio.

Sollte indes auch das Gehirn-Turing modellierbar sein, so würde das bedeuten, daß die im Modell der Turingmaschine enthaltenen Möglichkeiten heute noch völlig unabsehbar sind. In der Tat zeigt die noch in Anfängen steckende *Theorie der neuronalen Netze*, mit welchen prinzipiellen Verifikationsproblemen und damit auch Unvorhersagbarkeiten in diesem Feld zu rechnen ist. Ähnliche Konsequenzen legen sich aufgrund der Ergebnisse der *Komplexitätstheorie* nahe. Ich will damit keinesfalls sagen, daß Unbestimmtheit hier das letzte Wort sei, sondern im Gegenteil, daß man sich forscher Unmöglichkeitsaussagen in Sachen Künstlicher Intelligenz tunlichst enthalten sollte.

Mehr noch: Der einfache Gedanke, daß auch das *Gehirn* ein durch und durch physisch bestimmtes und insofern eben doch *grundsätzlich* technisch rekonstruierbares System sein muß, scheint mir unabweisbar zu sein, und es ist wichtig zu realisieren – darum war es hier vor allem zu tun –, daß *Gödels Theoreme* einer solchen Annahme entschieden nicht entgegenstehen. Gewiß, das ist nur ein negatives Resultat, aber dennoch, wie ich hoffe, immerhin ein Beitrag zur Klärung einer notorischen Unklarheit des Mensch-Maschine-Problems und damit auch des Projekts Künstlicher Intelligenz.

Literatur

Frey, G.: *Sind bewußtseinsanaloge Maschinen möglich?* In: *Studium Generale* 19 (1966), S. 200.

Gödel, K.: *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I.* In: *Monatshefte für Mathematik und Physik* XXXVIII (1931).

Hofstadter, D. R.: *Gödel, Escher, Bach.* Stuttgart: Klett - Cotta, 31985.

Kutschera, F.v.: *Die Antinomien der Logik.* Freiburg/München: Alber, 1964.

Lucas, J. R.: *Minds, Machines, and Gödel.* In: Anderson, A.R.: *Minds and Machines.* Englewood Cliffs, New Jersey: Prentice Hall, 1964.

Lyotard, J.-F.: *La Condition Postmoderne.* Paris: Ed. de Minuit, 1979.

Nagel, E./Newman, J. R.: *Der Gödelsche Beweis.* Wien/München: Oldenbourg, 1964.

Penrose, R.: *Computerdenken. Des Kaisers neue Kleider oder die Debatte um Künstliche Intelligenz, Bewußtsein und die Gesetze der Physik.* Heidelberg: Spektrum d. Wiss., 1991.

Searle, J.R.: *Geist, Hirn und Wissenschaft.* Frankfurt a. M.: Suhrkamp, 1986.

Smullyan, R.M.: *Languages in which Self Reference is Possible.* In: Hintikka, J.: *The Philosphy of Mathematics.* London: Oxford University Press, 1969.

Webb, J.C.: *Mechanism, Mentalism, and Metamathematics.* Dordrecht (Holland): Reidel, 1980.