Können Computer lügen?[1]

Rainer Hammwöhner
Universität Regensburg
Informationswissenschaft

Zusammenfassung

Ausgehend von einer einfachen Definition der Lüge, die eine intentionale Beschreibung zuläßt, wird untersucht, ob Computern entsprechende intentionale Zustände sinnvoll zugesprochen werden können. Dies ist als notwendige aber nicht hinreichende Voraussetzung anzusehen, sinnvoll von lügenden Computern sprechen zu können. Weitergehende Fragestellungen werden zum Abschluß kurz angesprochen.

1. Einleitung

Über Computer und Lüge läßt sich zumindest in den folgenden drei Zusammenhängen debattieren:

a. Der Computer als Instrument der Lüge: Neue Algorithmen und die große Kapazität neuer Computer ermöglichen die Generierung von Medienobjekten – Fotos, Filmen, Tondokumenten etc. –, die nur noch durch unmittelbaren Vergleich mit der Realität, wenn diese denn zugänglich ist, als verfälscht bzw. komplett künstlich zu erkennen sind.

b. Der Computer als Medium der Lüge: Die neuen netzwerkbasierten Medien eröffnen neue Möglichkeiten der Kommunikation und auch Selbstpräsentation. Das Spektrum reicht vom spielerischen Umgang mit der eigenen Identität bis zur Computerkriminalität.

c. Der Computer als Subjekt der Lüge: Schon kurz nach der Erfindung des Computers begannen Spekulationen über die Möglichkeit, denkende Computer zu entwikkeln, zu deren Verhaltensspektrum dann auch die Lüge gehören müßte.[2]

Das Verbindende ist der Vorgang der Simulation, die mit Hilfe des Computers zur Perfektion gebracht werden kann. Simulation als solche ist zunächst wertfrei zu sehen. Innerhalb von Naturwissenschaft und Technik wird sie primär als ein Mittel wahrgenommen, um Modelle zu validieren oder verständlich zu machen.[3] Als Instrument der Massenkommunikation kann sie den Zugang zur Realität verändern oder gar verstellen, ohne daß dem Wahrnehmenden dies in der Hyperrealität der virtuellen Realität bewußt werden müßte. Die Berichterstattung des Golfkriegs war dementsprechend auch ein Angriffspunkt von Medientheoretikern wie

---

[1] Dieser Text ist erschienen in: Mathias Mayer (Hrsg.): Kulturen der Lüge. Böhlau, 2003, S. 299-320.

[2] B. Brock: *Kultur, Ästhetik und künstliche Intelligenz*. *Spektrum der Wissenschaft Dossier*. Nr. 4. 1997. S. 100-105. In den folgenden Literaturangaben sind die Referenzen nach dem Prinzip der Zugänglichkeit ausgewählt worden. Der Nennung der Originalpublikation ist die eines gut erhältlichen Sammelbandes vorgezogen worden.

[3] Eine Einführung gibt z. B. H. Bossel: *Modellbildung und Simulation*. Braunschweig/Wiesbaden 1994.

Baudrillard oder Virilio.[4] Deren der Postmoderne zuzurechnende Theorien wurden für den Bereich der computerbasierten neuen Medien schnell rezipiert, ob man sie nun eher als Literaturphilosophie[5] oder Lebenseinstellung auffaßte.[6] Letztlich zeigt sich, daß die verschiedenen, oben genannten Rollen, welche dem Computer im Reich der Täuschung zugewiesen wurden, im Umfeld der neuen Medien verschwimmen. Die zunehmende Perfektion der Bildverarbeitung dient der Generierung immer perfekterer virtueller Welten, die zum Treffpunkt von Menschen, die mit multiplen Identitäten spielen, und computergenerierter Pseudowesen werden. Wer dann nicht mehr sicher ist, ob er einen solchen Bot oder einen Menschen vor sich hat[7], wird vielleicht über die unter c. erwähnten Möglichkeiten unter anderer Perspektive nachdenken. Zunächst allerdings ist ein solcher Fall nur im Bereich von Spielwelten mit sehr eingeschränkten Ausdrucksmöglichkeiten denkbar.

Für die weiteren Überlegungen möchte ich die Definition der Lüge verwenden, wie sie auch von Rott[8] in diesem Band zu Grunde gelegt wird: „Eine Lüge ist das bewußte Äußern einer Unwahrheit in Täuschungsabsicht". Wenn wir die Frage nach der Lügenkompetenz von Computern beantworten oder zumindest zielführend diskutieren wollen, sind folgende Fragen zu beantworten:

a. Woran erkennen wir, ob ein Computer lügt? Die Contrafaktizität einer Aussage ist vergleichsweise leicht festzustellen, aber auf welcher Basis können wir einem Computer eine Täuschungsabsicht, d. h. einen intentionalen Zustand zuschreiben?

b. Lügen sind Ergebnisse kreativer Prozesse. Es ist eine genaue Kenntnis der Realität vonnöten sowie die Fähigkeit, gezielt von dieser abzuweichen.[9] Sind Computer dazu in der Lage?

c. Erfolgreiches Lügen erfordert eine gute Einschätzung des Wissenstandes und Interaktionsverhaltens des Kommunikationspartners. Können Computer ihren Kommunikationspartner gut genug einschätzen, um gute Lügen zu erfinden?

Diese Fragen werden uns durch diesen Beitrag begleiten. Es steht außer Frage, daß wir sie, selbst wenn uns viel mehr Raum zur Verfügung stünde, nicht beantworten könnten. Dabei wollen wir den Aspekt der ethisch-moralischen Bewertung der Lüge nicht berücksichtigen. Wenn auch an anderer Stelle die Frage nach einer möglichen Subjektwerdung von Computern im ethischen Sinne aufgeworfen wird – z. B. Birnbacher[10] oder Metzinger[11] -, so scheint diese Möglichkeit vom jetzigen Stand der Dinge doch recht weit entfernt.

---

[4] M. Wetzel: *Paradoxe Intervention*. In: *Baudrillard – Simulation und Verführung*. Hg. von P. Bohn/D. Fuder. München 1994.

[5] G. P. Landow: Hypertext. The Convergence of Contemporary Critical Theory and Technology. Baltimore/London 1992.

[6] S. Turkle: Leben im Netz. Identität in Zeiten des Internet. Reinbek 1999.

[7] Ebd.

[8] H. Rott: *Der Wert der Wahrheit*. Im vorliegenden Band S. 7-34.

[9] R. Hettlage: Der entspannte Umgang der Gesellschaft mit der Lüge. Im vorliegenden Band S. 69-98.

[10] D. Birnbacher: *Künstliches Bewußtsein*. In: *Bewußtsein. Beiträge aus der Gegenwartsphilosophie*. Hg. von Th. Metzinger. Bielefeld 2001. S. 713-729.

[11] Th. Metzinger: Postbiotisches Bewußtsein: Wie man ein künstliches Subjekt baut – und warum wir es nicht tun sollten. In: Computer.Gehirn. Was kann der Mensch? Was können die Computer? Begleitpublikation zur Sonderausstellung im Heinz Nixdorf Museums Forum. Paderborn 2001. S. 87-113.

2. Das Problem des Fremdpsychischen

Lukesch[12] befaßt sich mit der Erkennbarkeit der Lüge aufgrund des Verhaltens oder meßbarer physiologischer Parameter des Lügners (Lügendetektor). Der Rückgriff auf derartige Methoden wird erforderlich, da die Absichten und Intentionen – diese gehen ja in die Definition der Lüge ein – oder allgemeiner jegliche psychische Zustände des potentiellen Lügners uns nicht unmittelbar zugänglich sind. Dieses Phänomen wird in der Philosophie auch als das *Problem des Fremdbewußtseins* bezeichnet. Besonders drastisch wird es von Nagel[13] veranschaulicht, der den Leser auffordert sich in eine Fledermaus hineinzuversetzen.

v. Kutschera[14] stellt folgende Ansätze vor, die Rückschlüsse auf Fremdbewußtsein erlauben:

a. Das *Turing-Prinzip* spricht jedem Wesen, das sich verhält wie ein Mensch – insbesondere im Sprechen – analoge psychische Zustände zu. Wer im Verhalten nicht von einem Lügner zu unterscheiden ist, ist demnach wohl auch einer.

b. Der *Analogie-Gedanke* postuliert, daß wir bei uns selbst eine Korrelation zwischen Verhalten und geistigen Zuständen beobachten. Beobachten wir bei einer Person ein Verhalten, schreiben wir ihr die geistigen Zustände zu, die unser Selbstmodell für dieses Verhalten vorsieht. Obwohl v. Kutschera diesen Ansatz für obsolet ansieht, wird er doch durch neuere neurobiologische Forschungen gestützt, die, wenn auch nicht unwidersprochen, die Existenz von sogenannten Spiegelneuronen belegen, die es erlauben, Gedanken und Verhaltensweisen simulativ auszuführen – bei Hemmung der motorischen Bahnen -, um das Verhalten und die Intentionen anderer Personen antizipieren zu können.[15]

c. *Hypothetische Zuschreibungen* erfolgen auf der Grundlage einer Theorie, die Prognosen über zukünftiges Verhalten anderer Personen erlaubt. Eine Basis allerdings begrenzter Reichweite kann hier eine Alltagspsychologie sein, die, folgt man Tetens[16], auch Voraussetzung unserer Selbstwahrnehmung und Selbstbeschreibung sein dürfte.

d. Das *Argument der gemeinsamen Sprache* beruft sich auf die Ausdrucksmittel, welche eine gemeinsame Sprache für die Artikulation psychischer Prozesse bereithält. Es wird eingewandt, daß zwei Personen Schmerz auf die gleiche Weise artikulieren können, diesen aber nicht gleich empfinden müssen.

2. 1. Der Turing-Test

Schon vor der technischen Entwicklung des Computers wurden formale Fragen der Berechenbarkeit auf der Basis abstrakter Maschinen (Turing-Maschine) diskutiert. Die Churchsche These formuliert als ein Ergebnis dieser Arbeiten die Vermutung, daß alle im intuitiven Sinne berechenbaren Funktionen Turing-berechenbar sind, also mit einer Turing-Maschine berechnet werden können. Der Begriff der intuitiven Berechenbarkeit ist allerdings, wie die Be-

---

[12] H. Lukesch: Erkennbarkeit der Lüge: Alltagstheorien und empirische Befunde. Im vorliegenden Band S. 121-149.

[13] Th. Nagel: Wie ist es, eine Fledermaus zu sein? In: Analytische Philosophie des Geistes. Hg. von P. Bieri. Weinheim 1997. S. 261-275.

[14] F. von Kutschera: *Die falsche Objektivität*. Berlin/New York 1993. S. 237ff.

[15] H. Breuer: *Zellen, die Gedanken lesen*. In: *Gehirn und Geist*. Nr. 2. 2002. S. 70-71.

[16] H. Tetens: Geist, Gehirn, Maschine. Philosophische Versuche über ihren Zusammenhang. Stuttgart 1994. S. 15ff

zeichnung schon ausdrückt, nicht exakt definiert. Insofern ist die Churchsche These einem mathematischen Beweis nicht zugänglich. Sie ist allerdings durchgängig akzeptiert. Es erhebt sich die Frage, ob geistige Leistungen sich als intuitiv berechenbare Funktionen auffassen lassen, ob also Computer – die Formulierung geeigneter Algorithmen vorausgesetzt – zu geistigen Leistungen in der Lage sind oder sogar ein Bewußtsein entwickeln können.[17] Weiterhin stellte sich die Frage, auf welcher Basis die Zuschreibung geistiger Leistungen erfolgen sollte. Alan Turing schlug 1950 in einem Gedankenexperiment einen Test vor, dessen Bestehen eine hinreichende Bedingung für die Zuschreibung von Intelligenz sein sollte. Folgende Aspekte des Turing-Tests sind von zentraler Bedeutung:

a. Maßstab der Beurteilung ist die menschliche Intelligenz. Diese ist dem Rechner mangels eines objektiven, nicht auf bestimmte Phänomene beschränkten Meßverfahrens für Intelligenz durch objektiv urteilende Gutachter zu- oder abzusprechen.

b. Die Einschätzung der Intelligenz sollte allein durch intellektuelle Leistungen, nicht aber durch Erscheinungsbild oder physische Aktionsmöglichkeiten bestimmt werden.

Um diese Ziele zu erreichen, entwarf Turing ein Imitationsspiel, das hier auf seinen relevanten Kern reduziert beschrieben werden soll. Ein Beobachter kommuniziert über zwei Datenleitungen vermittels Fernschreiber (Tastatur und Bildschirm) mit einer Testperson und einem Computer, die in einem getrennten Raum befindlich sind. Er kann beiden Fragen stellen, um zu ermitteln, wer von beiden der Computer ist. Während der Mensch ihn von seiner wahren Identität zu überzeugen versucht, ist es an dem Computer, den Eindruck zu erwecken, er sei der Mensch. Damit der wahre Sachverhalt nicht durch einfache Nachfrage geklärt werden kann, darf der Computer natürlich auch den Weg der Täuschung beschreiten. Er hat den Test dann bestanden, wenn es dem Beobachter unmöglich ist, einen relevanten Unterschied zwischen seinen Gesprächspartnern festzustellen. Soviel ist also schon klar: Sollte ein Computer den Turing-Test bestehen, so kann er – unter der Prämisse, daß der Turing-Test als Verfahren akzeptiert wird – auch lügen.

Natürlich hat der Turing-Test auch heftige Kritik auf sich gezogen. French[18] gibt eine Übersicht über die Rezeption des Turing-Tests in den 50 Jahren nach seiner Erfindung. Zu den wichtigsten Kritikpunkten gehören die Vernachlässigung des Körperlichen und eine falsche Ausrichtung der Forschung.

Die Grundlage des Turing-Tests, die Abstraktion von körperlichen Eigenschaften, entzieht viele natürliche Kommunikationsformen dem Test. Emotionen aber auch kommunikative Absichten werden vielfach durch Mimik und Gestik oder die Prosodie der Sprache ausgedrückt. Die Situierung eines Gesprächs erfolgt ebenfalls durch physischen Bezug auf die Umgebung, Blickrichtung, Zeigegesten etc. Harnad[19] schlug deshalb einen erweiterten Turing-Test vor, der in mehreren Stufen zunächst auch eine physische Gegenüberstellung und im weiteren den Test interner Prozesse vorsieht.

Es kann nicht geleugnet werden, daß der Turing-Test zumindest auf einen Teil der Forschung insofern eine nachteilige Auswirkung gehabt hat, als er die Aufmerksamkeit sehr früh

---

[17] A. Turing: *Kann eine Maschine denken?* In: *Künstliche Intelligenz. Philosophische Probleme*. Hg. von W. Ch. Zimmerli/St. Wolf. Stuttgart 1994. S. 39-78.

[18] R. M. French: The Turing-Test: The First 50 Years. In: Trends in Cognitive Sciences. Bd. 4. Nr. 3. 2000. S. 115-122.

[19] S. Harnad: Other Bodies, Other Minds. A Machine Incarnation of an Old Philosophical Problem. In: Minds and Machines. Nr. 1. 1991. S. 43-54.

auf die „höheren“ kognitiven Funktionen gelenkt hat.[20] Außerdem war, zumindest nachdem auf das Bestehen des Turing-Tests ein Preis (Loebner-Preis) ausgesetzt wurde, die Versuchung recht groß, die Aufmerksamkeit auf die Irreführung der Jury zu lenken. Sieht man von diesem Problem ab, so ist dennoch festzustellen, daß der Turing-Test ein noch nicht ausgeschöpftes Potential für die wissenschaftliche Forschung enthält, wenn man an Untersuchungen in Form von Streuversuchen denkt[21], die Worthäufigkeiten, Fehlerarten etc. berücksichtigen. Auf diese Art wären Rückschlüsse auf Tiefenmechanismen des Denkens möglich. Der Turing-Test handelt allerdings mit sehr großer Münze. Sollte ein völlig dem Menschen analoges Verhalten erforderlich sein, um zu lügen? In diesem Zusammenhang wollen wir einen Blick auf andere potentielle Lügner werfen. Wie weit ist die Lüge im Tierreich verbreitet und wie kommt man ihr dort auf die Schliche?

2. 2. Können Tiere lügen?

“One of the most important things to realize about systems of animal communication is that they are not systems for the dissemination of the truth.”[22]

„Haben Tiere ein mentales Leben, das dem Menschen ähnlich ist, oder sind sie Automaten ohne Bewußtsein? In den vergangenen hundert Jahren quälte die Frage die Tierpsychologen und Ethologen.“[23] Diese stehen vor ähnlichen Problemen wie wir. Auf welcher Basis sollen Tieren mentale Zustände attributiert werden? Tieren einen Turing-Test abzuverlangen, ist nun offensichtlich unsinnig. Auch der Analogiegedanke führt hier in die Irre. Er ist allenfalls für Tierfreunde tauglich, die im Verhalten ihres Haustiers allein den Ausfluß tiefer Zuneigung erkennen können und wollen. Sollte die oben erwähnte Vermutung zutreffen, daß das menschliche Gehirn über ein spezielles Organ für die einfühlende Simulation der psychischen Prozesse anderer verfügt, so folgt gerade daraus, daß es bei anders strukturierten Psychen zu Fehlurteilen kommen muß. Auch eine gemeinsame Sprache existiert nicht. Sie muß, wie etwa im Zusammenhang der Kommunikation mit Menschenaffen, erst konstruiert (visuelle Zeichen) bzw. gelehrt werden (etwa Ameslan, eine Gebärdensprache für Taubstumme).[24] Der Sprachumfang ist dabei aber so gering, daß der Gegenstandsbereich der Kommunikation begrenzt bleibt. Mit hypothetischen Zuschreibungen kann hier aber erfolgreich gearbeitet werden. Eine Grundannahme dabei ist, daß aufgrund der evolutionären Auslese der Ressourceneinsatz von Tieren weitgehend optimiert ist. Das betrifft z. B. auch den Einsatz von Zeit. Wenn also ein Sachverhalt oder eine Begebenheit eine überdurchschnittliche Aufmerksamkeit erfährt, so kann davon ausgegangen werden, daß diese als besonders empfunden werden. Tiere, die erstaunt, d. h. mit erhöhter Aufmerksamkeit, auf das Verschwinden eines Gegenstands reagieren, der scheinbar nur kurzfristig, etwa durch Verdeckung durch einen Wandschirm, der Sicht entzogen war, wissen z. B. um die Persistenz von Objekten. Mit derartigen Methoden kann man zahlreiche kognitive Fähigkeiten überprüfen. Jede Gattung verfügt dann über eine Art kognitives Instrumentarium, das mehr oder weniger gut bestückt sein kann.[25]

Täuschendes Verhalten wird bei zahlreichen Gattungen beobachtet. Als Täuschung wird ein Verhalten angesehen, bei dem die evolutionäre Fitneß auf Kosten des Empfängers erhöht

---

[20] P. M. Churchland: *Die Seelenmaschine. Eine philosophische Reise ins Gehirn*. Heidelberg/Berlin 1997. S. 267-297.

[21] D. R. Hofstadter: Die FARGonauten. Über Analogie und Kreativität. Stuttgart 1996.

[22] R. Trivers: *Social Evolution*. Menlo Park. 1985. S. 395.

[23] D. McFarland: *Biologie des Verhaltens*. Heidelberg/Berlin 1999. S. 429.

[24] Ebd. S. 432ff.

[25] M. H. Hauser: Wilde Intelligenz. Was Tiere wirklich denken. München 2001.

wird. Beispiele sind das Fluchtverhalten des Regenpfeifers, der Verletzungen vortäuscht, um Räuber von seinem Netz fortzulocken.[26] Das funktioniert auf die Dauer jedoch nur, wenn die Anzahl täuschender Signale im Verhältnis zum Überprüfungsaufwand als gering anzusehen ist. Ansonsten werden die Empfänger ihre Regeln zum Entschlüsseln der oder dem Umgang mit den Signalen ändern.[27] Bis zu dieser Grenze kann man also, anders als Rott[28], auch durchaus vom Wert oder besser Nutzen der Unwahrheit sprechen. Ein absoluter Zwang zur Wahrheit käme hier nur den Starken zugute, die einer Tarnung nicht bedürfen.

Wie sieht es nun mit dem Lügen aus? Wir erinnern uns, Lügen sind mit täuschender Absicht getroffene Aussagen. Sie sollten also klar zu trennen sein von evolutionär erworbenem täuschendem Verhalten, wie es die oben erwähnten Beispiele repräsentieren. Eine solche Unterscheidung von funktionaler und intentionaler Täuschung trifft auch Hauser.[29] Bewußte Täuschungen und manipulative Strategien sind aber nicht immer leicht zu unterscheiden.[30] Es ist nicht auszuschließen, daß die eine oder andere Lüge, deren wir uns schämen, in Wirklichkeit Teil eines Verhaltensmusters ist, dem in unserer Kognition nur im Sinne einer A-posteriori-Rationalisierung Intentionalität zugesprochen wurde.

Sollte es die Lüge als Phänomen aber dennoch geben, so scheint sie auf einen relativ engen Kreis von Gattungen und auch dort eher auf Einzelfälle begrenzt zu sein. Bestimmte Fallbeschreibungen von lügenden bzw. täuschenden Makaken oder Schimpansen begegnen einem in der Fachliteratur wiederholt. Wir haben es hier z. T. mit nicht reproduzierbaren anekdotischen Berichten zu tun oder mit Ergebnissen sehr aufwendiger Verhaltenstests. Diese können nicht in beliebiger Anzahl etwa zur Bestätigung einer Theorie oder auch zur Beantwortung der aus einem vollzogenen Test neu entstandenen Folgefragen durchgeführt werden. Es wird dabei untersucht, ob das Verhalten unwillkürlich, d. h. ohne Berücksichtigung von Kontextinformation ausgelöst wird, in wie weit die Versuchstiere ihr *Wissen* um den *Kenntnisstand* des Opfers der Täuschung mit einbeziehen, wie Artgenossen auf entdeckte Täuschung reagieren usw. Derartige Versuche sind von ihrer Methodik her einem Forschungsparadigma verpflichtet, das den Versuchstieren intentionale Zustände, wie *Wissen* oder *Absicht* zuschreibt. Intentionalität wird also im Rahmen des Versuchs nicht bestätigt oder entdeckt, sondern a priori zugeschrieben[31], wenn man von gewissen Grundbedingungen der Intentionalität (s. u.) absieht.

2. 3. Intentionale Zustände

Bevor wir uns im nächsten Kapitel nun endgültig dem Lügenpotential von Computern zuwenden, wollen wir noch eine Präzisierung des Begriffs des intentionalen Zustands unternehmen.

Intentionalität bezeichnet die Zielgerichtetheit des Handelns, der Gefühle oder des Denkens. Schreibt man einem System Intentionalität zu, so geschieht dies in Form von Sätzen, die sich durch *referentielle Opakheit* auszeichnen.[32] In einem normalen Satz können Worte ohne Änderung des Wahrheitswertes durch referenzidentische ausgetauscht werden. Wenn der

---

[26] McFarland (wie Anm. 23). S. 354ff.

[27] Ebd. S. 354

[28] Rott (wie Anm. 8).

[29] M. H. Hauser: Minding the Behavior of Deception. In: A. Whiten/R. W. Byrne: Machiavellian Intelligence II. Extensions and Evaluations. Cambridge 1997. S. 112-143.

[30] Trivers (wie Anm. 22).

[31] D. Dennett: Intentionale Systeme in der kognitiven Verhaltensforschung. In: Kognitionswissenschaft. Grundlagen, Probleme, Perspektiven. Hg. von D. Münch. Frankfurt am Main 2000. S. 343-386.

[32] Ebd.

Postbote Hans Hurtig heißt, kann der Satz *Es hat geklingelt, es war der Postbote.* durch den Satz *Es hat geklingelt, es war Hans Hurtig* ersetzt werden, ohne daß der Wahrheitswert betroffen würde. Dies ist bei dem Satz *Es hat geklingelt, Karl glaubte, es sei der Postbote* nicht der Fall, da die referentielle Identität der Ausdrücke *Hans Hurtig* und *der Postbote* nicht Teil des Glaubenssystems von Karl sein muß.

Man kann nun verschiedene Ordnungen der Intentionalität unterscheiden.

0. Ordnung: Instinktive Reaktion – keine Berücksichtigung eigener oder fremder mentaler Zustände.
1. Ordnung: Berücksichtigung eigener mentaler Zustände
2. Ordnung: Berücksichtigung fremder mentaler Zustände 1. Ordnung.
3. Ordnung: Berücksichtigung fremder mentaler Zustände 2. Ordnung

…

Für eine exaktere Fassung des Lügenbegriffs ist eine intentionale Beschreibung der Lüge vorzunehmen[33] und damit auch die Zuweisung einer intentionalen Ordnung. Eine Lüge ist zumindest mit einem intentionalen Akt zweiter Ordnung verbunden:

I. Der Lügner weiß, daß nicht *L*.

II. Der Lügner will, daß der Belogene glaubt, daß *L.*

III. Der Lügner weiß (hofft, glaubt), daß der Belogene nicht weiß, daß nicht *L*.

Um diesen Effekt zu erzielen, muß man allerdings nicht nur den Willen sondern auch noch Glaubwürdigkeit mitbringen. Dazu ist eine Minimalvoraussetzung zu erfüllen:

IV. Der Lügner will, daß der Belogene glaubt, daß der Lügner glaubt, daß *L*.

Außerdem soll die Lüge ja wie eine erstgemeinte Mitteilung wirken:

V. Der Lügner will, daß der Belogene glaubt, daß der Lügner will, daß der Belogene weiß, daß *L*.

Gerne läßt man sich den Erfolg seins Coups auch noch bestätigen:

VI. Der Lügner will, daß der Belogene will, daß der Lügner weiß, daß der Belogene glaubt, daß der Lügner will, daß der Belogene weiß, daß *L*.

Vermutlich lassen sich auch noch intentionale Zustände höherer Ordnung konstruieren als dieser letzte, der von 6. Ordnung ist, die einen Zusammenhang zum Phänomen der Lüge aufweisen. Mit einiger Sicherheit werden sie aber von den meisten Lesern als von begrenzter Relevanz für das Zustandekommen einer erfolgreichen Lüge angesehen werden. Folgt man allerdings Dennet[34], der sich hier auf Grice[35] beruft, so gibt es gute Gründe bei einer erfolgreichen Kommunikation von der Existenz intentionaler Zustände zumindest 3. Ordnung auszugehen. Dies dürfte auch anhand der obigen Betrachtungen zur Lüge plausibel zu machen sein. Ein intentionaler Akt dritter Ordnung (IV) stellt die Mindestvoraussetzung für eine erfolgreiche Lüge dar. Zwar kann es Lügenstrategien geben, die nicht voraussetzen, daß der

---

[33] U. Wiedemann: *Kommunikation in der wirklichen Welt*. /www.pyrrhon.de/magister (Stand 11. 6. 2002) untersucht die Lüge im Rahmen einer handlungstheoretischen Semantik.

[34] Dennet (wie Anm. 31).

[35] H. P. Grice: *Meaning*. In: *Philosophical Review* 66. 1957. S. 377-388. Ders.: *Utterers Meaning and Intentions*. In: *Philosophical Review* 78. 1969. S. 147-177.

Lügner selbst seine Lüge zu glauben scheint, sie sind dann aber trotzdem von mindestens dritter Ordnung, wie z. B.:

VII. Der Lügner will, daß der Belogene glaubt, (daß der Lügner glaubt,) daß ein Gewährsmann glaubt, daß L.

Folgen wir Dennett, so kann es als legitim angesehen werden, Computern intentionale Zustände zuzusprechen, wenn eine Reduktion ihres Verhaltens auf physische Zustände sich als theoretisch oder praktisch unmöglich erweisen sollte. Dennett selbst führt in einem ähnlichen Argumentationsgang leistungsfähige Schachcomputer als legitime Gegenstände einer intentionalen Beschreibung ein, da selbst ihre Programmierer zu einer physikalischen Beschreibung ihres hochkomplexen Verhaltens kaum in der Lage sein dürften.[36] Eine ähnliche Sicht vertritt McCarthy.[37] Wenn andererseits Chisholm[38] den Roboter als ein exemplarisches Beispiel anführt, daß scheinbar intentionales Verhalten durch seine simple Zurückführung auf physikalische Zustände entlarvt werden könne, so ist dies zunächst als ein Zeichen für die 15 Jahre zu interpretieren (1956 / 1971), die zwischen den beiden Originalveröffentlichungen liegen, in denen sich nicht nur die Technik – Chisholm bezieht sich in erster Linie auf den Computer als *Rechner* – sondern auch die Einstellung zu Computern erheblich verändert hat. Zugleich ist klar, daß, folgt man der hier skizzierten Auffassung, intentionale Zustände (nur?) zugeschrieben werden. Es wird nichts darüber ausgesagt, ob ein Computer mentale Zustände habe, sondern daß es für bestimmte Ziele sinnvoll sein kann, sein Verhalten in diesen Kategorien zu beschreiben. Es wird zu untersuchen sein, ob es Konstellationen gibt, in denen es sinnvoll ist, einem Computer eine Lüge zuzuschreiben.

Die Frage nach den intentionalen Zuständen von Computern berührt auch den immer noch prekären Status des Begriffs der Information. Der Kognitivismus betrachtet intentionale Systeme als informationsverarbeitende Systeme, welche semantische Information[39] verarbeiten. Semantische Information ist im Gegensatz zum Informationsbegriff von Shannon[40] nicht auf die Quantität sondern auf die Bedeutung des Kommunikats gerichtet. Eine Klärung des Zusammenhangs der verschiedenen Informationsbegriffe steht trotz älterer[41] und jüngerer Vereinheitlichungsversuche noch aus.[42] Sollte diese Brücke je ganz tragfähig werden, gäbe es auch eine formale Begründung für die Ansicht, daß die Syntaxmaschine Computer über Semantik verfügen kann.

Natürlich wäre es befriedigender, sagen zu können, einem Computer werde eine Lüge nicht nur zugeschrieben, sondern er lüge tatsächlich – d. h. er habe realiter einen dementsprechenden intentionalen Zustand. Damit würde man die Position des intentionalen Realismus einnehmen, wie ihn Dretske[43] vertritt, der Intentionalität auf natürliche Indikatoren (Geruchsspuren etc.[44]) zurückführt. Einen Bezug auf natürliche Indikatoren werden wir in der künstli-

---

[36] D. Dennett: *Intentionale Systeme*. In: Bieri (wie Anm. 13). S. 162-183.

[37] J. McCarthy: *Können einer Maschine geistige Eigenschaften zugeschrieben werden?* In: Zimmerli/ Wolf (wie Anm. 17). S. 184-231.

[38] R. M Chisholm: *Sätze über Glauben*. In: Bieri (wie Anm. 13). S. 145-161.

[39] Das Konzept der semantischen Information geht auf Carnap und Bar Hillel zurück: R. Carnap/Y. Bar Hillel: *An Outline of Semantic Information.* In: *Language and Information*. Hg. von Y. Bar Hillel. Reading, Mass. 1964.

[40] C. Shannon: The Mathematical Theory of Communication. Illinois 1949.

[41] F. Dretske: *Knowledge and the Flow of Information*. Cambridge, Mass. 1981.

[42] J. van Eijck/A. Visser (Hgg.): *Logic and Information Flow*. Cambridge, Mass. 1994.

[43] F. Dretske: *Naturalizing the Mind*. Cambridge, Mass. 1995.

[44] vgl. auch Lukesch (wie Anm. 12).

chen Welt des Computers aber wohl schuldig bleiben müssen. Ein in diesen Zusammenhang gehöriges Argument von Scriven[45] weist einen interessanten Bezug zur Lüge auf. Scriven behauptet, daß das Problem des Fremdpsychischen für Roboter partiell eliminiert werden könne, da man ihnen durch eine „Spezialschaltung" die Möglichkeit zur Lüge nehmen könne. Man informiert einen Roboter zunächst über alltagspsychologische Modelle, um ihm eine Interpretation seiner eigenen Zustände zu ermöglichen. Weiterhin erklärt man ihm, was eine Lüge ist, so daß eine sodann angebrachte Spezialschaltung jegliche Lügen unterbinden kann. Ein solcher Roboter kann und muß dann wahrheitsgemäß Auskunft über seine intentionalen Zustände geben.[46] Dieses Argument ignoriert auf sonderbare Weise die Option des freien Willens, die Robotern von Scriven attestiert wurde.[47] Eine solche Spezialschaltung, die mit Sicherheit funktioniert, kann es in Systemen nicht-trivialer Komplexität nicht geben. Das Problem des Fremdpsychischen ist so nicht zu hintergehen. Die Frage nach lügenden Computern oder Robotern wird uns nun im nächsten Kapitel beschäftigen.

3. Lügende Computer, lügende Roboter oder nichts von dem?

Unsere Überlegungen zu lügenden Computern wollen wir mit einer Sequenz von konstruierten Beispielen beginnen, die keine Entsprechungen in tatsachlichen Systemen haben – wer bekommt schon Fördergelder für die Implementation lügenhafter Systeme? –, deren angenommene Fähigkeiten aber nicht zu weit vom Stand der Kunst entfernt liegen.

3. 1. Ist noch ein Zimmer frei?

Wir denken uns ein natürlichsprachliches Hotelzimmer-Reservierungssystem, wie es in einer wahrheitsliebenden Form an anderer Stelle ausführlich beschrieben wurde.[48] Sein Name sei HAL (*H*euristic *A*lgorithmic *L*iar). Das Ziel von HAL ist es, möglichst viele Zimmer zu möglichst hohen Preisen zu vermieten, um den Shareholder-Value des Touristik-Konzerns zu optimieren, von dem er in Auftrag gegeben wurde. An die Wahrheit – welche Zimmer sind frei, welche nicht usw. – soll er sich nur in so weit halten, als er keine Kunden verärgern soll. Dies gilt besonders für Stammkunden.

1. HAL	Hier Hotel Saturn, guten Abend. Was kann ich für Sie tun?
2. David	Guten Abend, hier ist David Bowman. Hätten Sie am 2.7. ein Zimmer für mich, für drei Übernachtungen.
3. HAL	Schön, Herr Bowman, daß Sie uns wieder einmal besuchen. Vom 2.-5.7. also?
4. David	Ja!
5. HAL	Möchten Sie wieder Ihr gewohntes Zimmer haben?

---

[45] M. Scriven: *Der vollkommene Roboter. Prolegomena zu einer Androidologie*. In: Zimmerli/Wolf (wie Anm. 17). S. 79-111.

[46] Ebd. S. 108f.

[47] Ebd. S. 83ff.

[48] W. Hoeppner/H. Marburger/K. Morik: *Talking it Over: The Natural Language Dialog System HAM- ANS*. In: *Cooperative Interfaces to Information Systems*. Hg. von L. Bolc/M. Jarke. Berlin/Heidelberg 1986. S. 189-258.

6. David	Das wäre zwar schön, aber ginge es diesmal nicht etwas preisgünstiger?
7. HAL	Es tut mir leid, die preisgünstigeren Zimmer sind für diesen Termin leider schon alle ausgebucht.
8. David	Äh, ja dann …, ich weiß nicht so recht, ich bin diesmal privat unterwegs.
9. HAL	Entschuldigung, wenn Sie einen Augenblick warten würden, ich habe da parallel noch ein Gespräch in der anderen Leitung.
Pause	
10. HAL	Hallo, ich habe da eine gute Nachricht für Sie. Gerade ist eine Buchung für ein günstigeres Zimmer storniert worden. Ich könnte Ihnen das dann anbieten
11. David	Oh, vielen Dank, dann buchen Sie das bitte!
12. HAL	Schon geschehen, vielen Dank.
13. David	Ganz meinerseits, auf Wiederhören.
14. HAL	Auf Wiederhören.

Zum obigen Beispieldialog kann man sich folgendes Skript vorstellen. HAL identifiziert David Bowman als einen Stammkunden (falls keine Namensgleichheit vorliegen sollte). Schritt 3 dient zur Bestätigung der bisherigen Angaben. Herr Bowman ist als Geschäftskunde bekannt, der bisher auf seine Ausgaben wenig achten mußte. Er hatte bisher immer ein vergleichsweise teures Zimmer, das ihm auch jetzt wieder angeboten wird. Überraschenderweise wünscht Herr Bowman aber ein preisgünstigeres Zimmer. Da dieser Wunsch aber nicht sehr entschieden vorgebracht wird, versucht HAL ihn zu nötigen, bei dem gewohnten Zimmer zu bleiben, indem er vortäuscht, es sei sonst kein Zimmer frei. Der Kunde reagiert verunsichert, vielleicht sogar verärgert. Durch den Hinweis auf den privaten Zweck der Reise wird deutlich, daß die bisherigen Annahmen über den Kunden in diesem Fall nicht gelten. Es ist abzuwägen, ob man es riskiert, daß der Kunde ein anderes Hotel sucht, das ihm vielleicht so gut gefällt, daß er es auch für die weiteren Dienstreisen aufsucht. Da dieses Risiko als zu hoch eingeschätzt wird, muß ein vernünftiger Grund gefunden werden, warum plötzlich ein günstiges Zimmer frei ist. Dann kann dieses Zimmer offeriert werden.

Die Leistungen von HAL sind verglichen mit dem Stand der Kunst nicht übermäßig beeindruckend. Er weiß um die Zimmerbelegungen und Preise sowie um manche Vorlieben und Gewohnheiten der bisherigen Kundschaft. Ihm stehen Regeln zur Verfügung, wie – auch durch Fehlinformation der Kunden – die Belegung der teuren Zimmer optimiert werden kann. Weitere Regeln ermöglichen es ihm, eine angemessene Erklärung für das plötzliche Freiwerden des Zimmers zu finden. Eine Analyse von prosodischen Eigenschaften der Kundenantwort oder der Mimik des Kunden (z. B. bei Videotelefonaten) ermöglicht ihm, Annahmen über den emotionalen Zustand des Kunden – Ungeduld, Ärger, Unzufriedenheit oder Zufriedenheit, Gleichgültigkeit – zu treffen. Gerade letzteres ist Gegenstand aktueller Forschung.[49]

Sind die Aussagen von HAL als Lügen zu werten? Ist HAL ein Lügner? Bei einem menschlichen Gesprächspartner würde man nicht zögern, beide Fragen zu bejahen. Schließ-

[49] A. Batliner/R. Huber/H. Niemann/E. Nöth/J. Spilker/K. Fischer: *The Recognition of Emotion*. In *Verbmobil: Foundations of Speech-to-Speech Translations*. Hg. von W. Wahlster. New York/Berlin 2000. S. 122-130.

lich sind alle Kriterien, die wir bisher für die Existenz von Lügen definiert haben erfüllt. Hier jedoch zögert man. Dafür gibt es sicherlich mehrere Gründe, von denen einige hier diskutiert werden sollen.

Bei einem System wie dem hier beschriebenen ist das Spektrum der möglichen Lügen vermutlich begrenzt. Der Diskursbereich ist eingeschränkt – Hotelzimmer, Betten, Reservierungszeiten etc. Die möglichen Lügenstrategien sind durch Regeln festgelegt. Das Lügen wird jedoch meist als ein kreativer Prozeß empfunden, selbst wenn der durchschnittliche Hotelbedienstete auch nicht mehr Verstand in seine Lügen investieren sollte als unser HAL.

Der Mensch weiß zumeist, worüber er lügt. HAL hat vermutlich noch nie ein Hotelzimmer von innen gesehen. Auch seine Kenntnis von Personen ist vermutlich begrenzt.

HAL hat zwar Regeln darüber, daß man sich beim Lügen nicht erwischen lassen sollte. Dies könnte die Kunden verärgern. Was aber sind seine Alternativen, die aus dem obigen Beispiel nicht deutlich werden? Ihm müssen andere Verhaltensweisen zur Verfügung stehen – neutrales Gespräch, Versuch durch ein Werbegespräch für ein teureres Zimmer zu werben usw.

Der Mensch lügt, so denkt man zumindest, aus einem Interesse heraus, selbst wenn er auf Aufforderung oder Kommando hin lügt. Wo liegt aber das Interesse von HAL? Er ist in kein kommunikatives oder soziales Netz so eingebunden, daß er die Folgen seiner Übeltat oder die Früchte seines Erfolgs selbst erfahren könnte – was auch immer „erfahren“ in diesem Kontext bedeuten kann.

Bezogen auf unseren bisherigen Argumentationsgang wäre aber zunächst einmal festzustellen, welche Ordnung intentionaler Zustände dem obigen Dialog zuzuordnen ist. Läge hier eine verhaltensbiologische Fragestellung vor, so wären jetzt Verhaltensexperimente zu entwerfen. Wie verändert sich das Verhalten des Lügners, wenn die Reaktion des Belogenen entweder die Faktizität der Lüge, die Glaubwürdigkeit des Lügners – Expertise, Ehrlichkeit – in Frage gestellt wird, wenn er glauben muß, selbst belogen zu werden usw. Von HAL können wir, wenn wir den aktuellen Stand der Technik zu Grunde legen, davon ausgehen, daß nur intentionale Zustände maximal zweiter Ordnung sinnvoll zugeschrieben werden können. Bisher sind aus der Prosodie z. B. nur wenige Emotionen zu erschließen, vor allem Ärger. Worauf sie sich genau richten, ist auch noch nicht festzustellen. Wenn es also vielleicht statthaft sein mag, HALs Äußerungen als Lügen zu bezeichnen – unsere Arbeitsdefinition legt nichts anderes nahe –, so sind es doch gewißlich sehr schlichte. Diese Vorgehensweise macht allerdings ein methodisches Problem deutlich. Während wir bei interagierenden Menschen bzw. höheren Tieren wegen unserer, trotz vehementer Fortschritte der jüngeren Zeit fragmentarischen Kenntnis der neurobiologischen Prozesse gezwungen sind, auf der intentionalen Ebene der Untersuchung zu verharren, so sind wir bei künstlichen Automaten in der Lage, auf die funktionale Ebene zu wechseln, in der wir das Verhalten durch Funktionen beschreiben, welche interne und externe Zustandsgrößen berücksichtigen. Diese zusätzliche, zwischen intentionalen und physikalischen Zuständen vermittelnde Ebene ist uns bei biologischen Systemen noch weitgehend unzugänglich. Unsere Beurteilungskriterien sind also jeweils unterschiedlich. Die funktionale Sicht gibt uns im Vergleich zum Tier sogar einen höheren Grad an Sicherheit im Urteil. Zur Veranschaulichung mag ein Beispiel dienen, das wir Dennett[50] verdanken.

Ein Hund sieht seinen Lieblingsschlafplatz, einen Sessel, von seinem Herrchen besetzt. Alle Versuche ihn zu „überreden“ den Sessel zu räumen scheitern. Da rennt der Hund zur Tür

[50] D. Dennett: *Bedingungen der Personalität*. In: Bieri (wie Anm. 12). S. 303-324.

und scharrt daran. Herrchen greift zur Leine und folgt. Der Hund jedoch rennt zum Sessel zurück und hat seinen Platz zurückerobert.

Die weitergehende Interpretation würde besagen, der Hund habe überlegt[51], daß sein Herrchen annehme, er müsse Gassi gehen, wenn er an der Tür scharre. Das würde ihn überzeugen, ihm zu folgen. Sodann könne der Sessel besetzt werden. Wir nehmen den Hund also als gewieften Lügner wahr. Es könnte jedoch sein, daß der Hund weiß und berücksichtigt, daß sein Herrchen immer aufsteht, wenn er an der Tür kratzt. Erwägungen über Herrchens mentale Zustände müssen nicht angestellt werden. Es liegt nur ein Fall von Konditionierung von Menschen durch Hunde vor.

Im Fall von HAL können wir aber wissen, ob Repräsentationen über das Wissen und nicht nur das Verhalten der Kommunikationspartner gebildet werden, wenn wir Zugang zu einer funktionalen Beschreibung des Systems haben.

3. 2. Der Computer als Lügenerfinder?

In jüngeren Publikationen – etwa Sesín[52] oder Nix[53] – wird zur Zeit die Frage des Kreativitätspotentials von Computern gerade auch auf populärwissenschaftlichem Niveau diskutiert. Es werden Programme vorgestellt, die Leistungen im künstlerischen Bereich erbringen – Bilder malen, Choräle komponieren oder Gedichte dichten. Während die Leistungen in der Dichtkunst überaus ernüchternd sind, kommen im Bereich der Komposition und der Malerei Ergebnisse zustande, die zunächst beeindrucken. Das Programm Aaron z. B., erstellt von dem Künstler Harold Cohen, ist in der Lage, eigenständig eine Bildkomposition zu entwerfen – eine Szene, die Menschen und Pflanzen enthält –, diese Objekte anatomisch korrekt zu zeichnen und nach einem stimmigen Farbschema auszuführen. Dieses und ähnliche Systeme wurden von Hofstadter[54], aber schon 1996 einer heftigen Kritik unterzogen. Hofstadter hebt besonders hervor, daß die Bilder zwar evtl. schön anzusehen, aber epistemisch leer seien. Der Computer, oder vielmehr das Programm verstünden nichts von Pflanzen, Menschen etc., so daß diese Bilder keine Aussage enthielten. Man könnte mit unserer Terminologie anmerken, daß sie der 0. Ordnung der Intentionalität zuzurechnen seien. Die komponierenden Systeme Harmonet und Melonet sind stark in der Erfassung der formalen Struktur der Musik. Die eigentliche musikalische Idee muß ihnen aber vorgegeben werden.

Diese Form nicht-propositionaler Kreativität ist aber in unserem Falle von geringem Nutzen. Es ist sogar zu fragen, welcher Platz der Kreativität im Entstehungsprozeß der Lüge einzuräumen ist. Das Ziel der Lüge, eine falsche Vorstellung beim Hörer zu erwecken, ist in den meisten Fällen ein abgeleitetes Ziel. Als Primärziel kommt z. B. in Frage, daß der Hörer zu einem Handeln veranlaßt werden soll, das er ansonsten nicht in Erwägung gezogen hätte. Oder er soll von einer geplanten Handlung abgehalten werden. Oder, weniger direkt, es soll im Hörer eine bestimmte intentionale Einstellung zum Lügner und seinen Vorhaben bewirkt werden. Im Gegensatz zu einem künstlerischen Prozeß ist eine Lüge also im Prinzip planbar. Will man den Hörer von einer Handlung abhalten, muß man die Prämissen dieser Handlung

---

[51] Will man weniger anthropomorph formulieren, wird man dem Hund ein mentales Modell der Intentionen seines Gegenüber zusprechen (P. N. Johnson-Laird: *Mental Models*. Cambridge. 1983). Wichtig in unserem Zusammenhang ist vor allem, daß mentale Modelle auch nicht-propositionale Repräsentationen einschließen, wie man sie nicht sprachbegabten Tieren eher zusprechen kann.

[52] C.-P. Sesín: *Künstliche Künstler*. In: *Gehirn und Geist*. 2. 2002. S. 52-53.

[53] Computer.Gehirn. Was kann der Mensch? Was können die Computer? (wie Anm. 11). S. 252ff.

[54] Hofstadter (wie Anm. 21). S. 522.

kennen und durch Fehlinformation in Frage stellen. Das Anstoßen einer Handlung erfordert umgekehrt zunächst die Erfüllung der Prämissen. Diese sind aber für das Zustandekommen einer Handlung nicht hinreichend. Eine weitere Motivation muß hinzukommen. Dieses motivierende Element kann der Lügner entweder aus einer intimen Kenntnis der Motivationsstruktur des Hörers ableiten, oder er muß es sich durch Hypothesen erschließen. Hier liegt ein gewisses kreatives Element. Eine zentrale Rolle spielt aber die genaue Beobachtung der Kommunikationspartner, aus der durch Induktion und Abduktion eine Alltagstheorie des Verhaltens entstehen kann. Ein nennenswerter Raum dürfte hier Prozessen der Analogiebildung einzuräumen sein, für die Computermodelle zur Zeit erst in Entwicklung begriffen sind.[55]

Es scheint also, als ob das Entstehen einer Lüge als ein Planungsprozeß zu begreifen ist, in den auch die intentionalen Zustände des Hörers einfließen. Dabei ist die konsistente Ableitung eines neuen, in sich konsistenten und mit der Realität in Übereinstimmung zu bringenden Glaubenssystems zur Stützung einer vom Sprecher geplanten Lüge ein Sonderfall, der mit modernen Repräsentationstechniken etwa in Form intensionaler Logiken faßbar ist. Die Überzeugungskraft einer Lüge hängt allerdings nicht allein von ihrer logischen Stringenz ab, sondern ist z. T. wohl auch im Ästhetischen[56], sowie in kulturell begründeten Wahrheitsvorstellungen begründet[57], die einen Bezug zur Logik nicht aufweisen müssen. Es gibt auch wirkungsvolle Lügen, für die es in keiner realen oder auch nur logisch konsistenten Welt ein Modell gibt.

Aber selbst der einfache Fall ist nicht unproblematisch. Die Änderung eines Faktums kann eventuell Auswirkungen auf viele weitere bisher korrekt getroffene Annahmen haben. Der Volksmund sagt: „Eine Lüge zieht hundert Lügen nach sich." Dies ist eine spezielle Form des Frame-Problems. Dieses betrifft die Gültigkeit von Annahmen nach einer das System oder die Umwelt verändernden Aktion. Wir wissen z. B., daß sich Farbe, Aussehen und Geschmack von Lebensmitteln nicht durch Transport, wohl aber durch Kochen verändern. Diese Unterschiede müssen beachtet werden, ob die Aktion nun wirklich erfolgt oder nur erlogen ist. Die zu berücksichtigenden Zusammenhänge werden in Planungssystemen durch Axiome repräsentiert. Als Qualification-Problem wird die Schwierigkeit bezeichnet, herauszufinden unter welchen Bedingungen in der realen Welt eine Aktion garantiert gelingt. In technischen Spielwelten sind diese Probleme lösbar, in komplexeren Fällen allerdings nur heuristisch. Im Fall der Lüge müßten nun die entsprechenden Axiome und Standardannahmen so formuliert werden, daß sie nicht nur die Regularitäten aller möglichen Lügenwelten repräsentieren sondern auch den Gesetzmäßigkeiten einer Alltagspsychologie folgen, und somit Aussagen über die möglichen oder wahrscheinlichen Auswirkungen einer Lüge auf den Hörer erlauben. Gleichzeitig müßte die Beobachtung des Kommunikationspartners so verfeinert werden, daß der Erfolg, wenn auch im Rahmen gewisser Unsicherheiten, beobachtbar wird. Hier fehlen aber schon auf der inhaltlichen Ebene die erforderlichen Theorien.

Williams[58] untersucht den Unterschied zwischen den intentionalen Zuständen *Glauben* und *Wissen*. Für eine wichtige Eigenschaft des Glaubens hält er, daß man Aussagen treffen kann, die nicht mit dem Geglaubten übereinstimmen. Diese Kompetenz spricht er Computern ab. Er hält sie für unfähig zu lügen. Dies dürfte in dieser Allgemeinheit nicht zutreffen, wenn unsere bisherigen Überlegungen zutreffen. Es ist allerdings zu vermuten, daß Computer im Reich der Lüge noch auf längere Zeit Amateure bleiben werden.

---

[55] Ebd.

[56] M. Mayer: Das rechte Leben und das Falsche lesen? Über den Zusammenhang von Literatur, Lüge und Ethik. Im vorliegenden Band S. 225-245.

[57] W. Koschmal: *Die russische Wahrheit*. Im vorliegenden Band S 247-272.

[58] B. Williams: *Deciding to Believe*. In: *Problems of the Self*. Cambridge 1973. S. 136-151.

3. 3. Wo geht es hier zur Steckdose

Einer der heftigsten und meist diskutierten Angriffe gegen die Hypothese, Computer könnten eventuell denken oder etwas verstehen, aber auch gegen die Kognitionswissenschaft ging von John Searle aus. Besonders die Debatte um den *Chinese Room* fand große Aufmerksamkeit. Diese ist an zahlreichen Stellen im Internet aber auch von Hofstadter und Dennet[59] dokumentiert und soll hier nicht wieder aufgenommen werden. Kern der Argumentation war der Hinweis auf die rein syntaktische Arbeitsweise des Computers. Die Symbole, mit denen der Computer hantiere, hätten keine Bedeutung und keine Beziehung zur realen Welt. Intentional könnten nur biologische Systeme sein. Die Annahme dieser These würde uns die Antwort auf unsere Frage wieder sehr erleichtern. Unter diesen Umständen könnte ein Computer unter keinen wie auch immer gearteten Umständen lügen.

In einer der Repliken wurde vorgeschlagen, man solle den Computer mit Wahrnehmungssystemen und Effektoren ausstatten. Die vom Roboter in der Kommunikation verwendeten Symbole erhielten dann durch sein Handeln Bedeutung. Dieses Argument wurde vielfach aufgegriffen, z. B. von Beckermann[60], der hier auch einer Skepsis Dennetts gegenüber semantischen Maschinen begegnet.

Auch uns wäre vermutlich wohler, wenn wir wüßten, daß der Computer, dem wir Intentionalität zusprechen wollen oder sollen, Gelegenheit hätte, den Gehalt seiner Kommunikate an der Realität zu messen. Eine Unwahrheit äußert sich doch entschieden leichter, wenn man die Wahrheit kennt. Die Bedeutung sprachlicher Zeichen erschließt sich am besten durch den Gebrauch in realen Situationen. Dieses Argument ist aber nicht vollständig überzeugend. Sind Blinde nicht in der Lage, Lügen über farbige Gegenstände zu äußern? Die Frage scheint weniger zu sein, ob ein Computer über ein dem Menschen analoges Sensorium verfügen und wie dieser in der Welt handeln kann, sondern ob er überhaupt zu Erkenntnis und Wissen gelangen kann. Auch hier gehen die Meinungen auseinander. Williams[61] billigt dem Computer – im Gegensatz zum Glauben (s. o.) – Wissen durchaus zu, Searle[62] wohl kaum. Voraussetzung dürfte jedoch, um dem Begründungsanspruch des Wissens gerecht zu werden, eine Kritikfähigkeit gegenüber den Informationsquellen sein, wie dem eigenen Sensorium oder etwaigen Kommunikationspartnern. Insbesondere letzterem Aspekt wird in jüngeren Publikationen, die sich mit autonomen, intelligenten Software-Agenten befassen, viel Aufmerksamkeit gewidmet. Derartige Software-Systeme können einander Vertrauen und Mißtrauen entgegen bringen. Diese Zustände konkretisieren sich in Annahmen über die Wahrscheinlichkeit des Zutreffens von Aussagen oder über die Struktur sozialer Netzwerke in der Tradition der Systemtheorie.[63] Es ist also davon auszugehen, daß auch bei Software-Systemen ein gewisses Maß an Rationalität im Umgang mit den eigenen Informationsquellen vorausgesetzt werden kann. Daß immer die Möglichkeit einer – vielleicht auch fundamentalen – Täuschung besteht, wenn z. B. HAL nicht bemerkt, daß er nur ein Prototyp ist, der über nicht existierende Zim-

---

[59] D. R. Hofstadter/D. Dennett: Einsicht ins Ich. Fantasien und Reflexionen über Selbst und Seele. Stuttgart 1981. S. 337-366

[60] A. Beckermann: *Semantische Maschinen*. In: *Intentionalität und Verstehen*. Hg. vom Forum für Philosophie. Bad Homburg 1990. S. 196-211.

[61] Williams: *Deciding to Believe* (wie Anm. 58).

[62] J. R. Searle: *Geist, Hirn und Wissenschaft*. Frankfurt am Main 1986.

[63] J. G. Gans/M. Jarke/S. Kethers/G. Lakemeyer: *Modeling the Impact of Trust and Distrust in Agent Networks*. //www-i5.informatik.rwth-aachen.de/kbsg/publications/download/GG +GL+:AOIS-01.pdf (Stand 11. 6. 2002).

mer in einem Phantasiehotel verhandelt, unterscheidet Computer oder Software-Systeme nicht im Grundsatz von uns Menschen.

Wenn man dennoch in der Zuschreibung intentionaler Zustände einem Roboter den Vorzug vor einem körperlosen Software-System gibt, so hängt das vermutlich damit zusammen, daß man einem physisch agierende System eher die Verfolgung *eigener* Ziele zutraut, während ein Software-System im wörtlichen Sinne als fremdprogrammiert angesehen wird. Dies gilt in besonderer Weise, wenn die Roboter fundamentale Ziele – Wahrung der eigenen Existenz und Handlungsfähigkeit, Reproduktion – mit uns teilen, wenn sie nicht nur in einer Nischenaufgabe Kompetenz aufweisen – HAL ist hier ein Beispiel für ein besonders eingeschränktes Tätigkeitsfeld – sondern im *wahren Leben* bestehen können. Schon Dennett[64] merkt an, daß die KI-Forschung vermutlich mehr aus der Konstruktion einer virtuellen oder robotisierten Ameise als aus der Modellierung von Schachproblemen gewinnen könne. Diesem Wunsch entspricht eine Entwicklung, die zu einer zunehmenden Kooperation von Robotik und Verhaltensbiologie geführt hat. Die Erklärung vergleichsweise primitiver Tiere als Automaten und ihre Simulation als technisch konstruierte oder auf dem Computer simulierte Automaten gehört schon zum verhaltensbiologischen Standardinventar.[65] Derartige Roboter richten ihr Verhalten nicht nur darauf aus, Menschen Hotelzimmer anzudrehen, sondern sie achten auch darauf, zum rechten Zeitpunkt ihre Akkus aufzuladen usw. Sie beherrschen allerdings bisher nur sehr elementares Verhalten, das zu seiner Beschreibung keinerlei intentionaler Zustände bedarf.[66] Stellen wir uns eine Population von Robotern vor, die in einer realen oder simulierten Welt um begrenzte, versteckte Ressourcen – etwa eine Stromquelle – konkurrieren. Sie können zwischen folgenden Verhaltensweisen wählen:

- Einzelkämpferexistenz (alles, was ich finde, gehört mir),
- Kooperation innerhalb einer Gruppe durch gegenseitiges Mitteilen von Information und Teilen des Gefundenen ,
- Vortäuschen der Kooperation, Versuch mehr zu nehmen als zu geben. Ausgeben von Fehlinformation, z. B. durch Hinweis auf eine erschöpfte Stromquelle bei Verheimlichen einer noch ergiebigen.

Wir beobachten nun, daß die Roboter der Population gegeneinander z. T. reziprokes kooperatives Verhalten an den Tag legen, z. T. aber auch täuschendes. In ihr Verhalten fließen die Erfahrungen mit bestimmten Individuen ein sowie mit erfolgreichen Teams. Dabei kann die Teamgröße eine Rolle spielen. Die Stromquelle könnte z. B. nur maximal vier Individuen erfolgreich versorgen. Verändert sich die Population durch Ausscheiden oder Neueintreffen von Robotern, so paßt sich das Verhalten an. Mit findigen Neuen befreundet man sich, dumme betrügt man.

Angenommen, ein solches Verhalten wäre zu beobachten. Würde man es nicht zumindest für eine gute Simulation lügenhaften Verhaltens und seiner Chancen und Risiken halten? Dies wird kaum jemand bestreiten. Wo aber ist der Unterschied zur realen Lüge. Searle[67] würde ihn in der Unfähigkeit symbolverarbeitender Systeme erkennen, Semantik hervorzubringen. Ein gravierendes Problem dürfte jedoch darin liegen, daß mit dem Begriff der Lüge – wenn

---

[64] Dennet (wie Anm. 31). S. 366.

[65] McFarland (wie Anm. 23). S. 405-425.

[66] H. Cruse/J. Dean/H. Ritter: *Prärationale Intelligenz*. In: *Spektrum der Wissenschaft*. Dossier 4. 1998. S. 100-103. – Eine schöne Zusammenstellung einfacher kybernetischer Maschinen mit z. T. überraschenden Eigenschaften gib. V. Braitenberg: *Vehikel. Experimente mit kybernetischen Wesen*. Reinbek 1983.

[67] Searle (wie Anm. 62).

auch in Definitionen nicht immer expliziert – die Vorstellung des verantwortlichen Handelns verbunden ist. Kaum jemand wird zögern, Tieren oder unseren Robotern täuschendes Verhalten zu unterstellen. Eine Lüge unterstellt man aber nur einer für ihr Handeln verantwortlichen Person.

4. Fazit

Wir haben festgestellt, daß es innerhalb des kognitivistischen Forschungsparadigmas durchaus sinnvoll sein kann, symbolverarbeitenden Systemen intentionale Zustände zuzuschreiben. Nicht ganz so klar ist, ob diese Zustände eine Komplexität aufweisen, die uns – im Fall der Äußerung von Unwahrheiten – zwänge oder auch nur berechtigte, von Lügen zu sprechen. Eindeutig ist aber, daß außerhalb des kognitivistischen Paradigmas nicht einmal diese Zuschreibung möglich ist.

Wir können aber nicht sagen, ob ein Computer *wirklich* lügt. Was haben wir also gewonnen? Wir können z. B. Strukturen der Lüge untersuchen und auf Computern simulieren und die begründete Hoffnung hegen, daß wir das Wesentliche erfaßt haben. Darüber hinaus wissen wir jetzt genug, um der Vermutung Ausdruck zu verleihen, daß die Frage, ob Computer *wirklich* lügen können, nicht im Sinne einer empirischen Frage zu beantworten ist, die nur das Verhalten der Computer betrifft. Die Wirklichkeit höherer intentionaler Zustände wird in der Diskussion häufig in einen Zusammenhang mit dem Begriff der Person gebracht.[68] Wäre ein Computer eine Person, könnte er auch *wirklich* lügen. Person zu sein, erfordert aber neben anderen notwendigen Vorbedingungen auch, von anderen Personen als eine Person anerkannt zu werden. Sollen wir also Computer als Personen ansehen? – Dafür ist in der näheren Zukunft kein Anlaß zu erkennen.

[68] Dennet (wie Anm. 50).