Ersch. in Mehler, A. & Sutter, T. (Hg.): Medienwandel als Wandel von Interaktionsformen, S. 135–157. Wiesbaden: Verlag für Sozialwissenschaften (2010)

„Ich, Max" – Kommunikation mit Künstlicher Intelligenz

Ipke Wachsmuth

Mit dem Einzug maschineller „kommunikationsfähiger" Systeme in Form verkörperter Agenten in der Künstlichen Intelligenz stellt sich heute stärker als je zuvor die Frage, ob solchen Systemen in einer absehbaren Zukunft (oder überhaupt) eine Form von Bewusstsein zugeschrieben werden kann. Als Menschen sind wir versucht, einem menschenähnlich auftretenden Gegenüber, das mit uns einen Dialog in akzeptabler natürlicher Sprache führt, Absichten, Wünsche und Ziele zu unterstellen, also von uns aus einen solchen Dialog vom intentionalen Standpunkt aus zu führen.[1] Aber selbst wenn unser künstliches Gegenüber einen Namen hat, auf den es hört, und in seinen Äußerungen sich selbst mit „Ich" bezeichnet, so ist uns doch klar, dass derzeit solche Kennzeichnungen noch unangemessene Unterstellungen sind.

Unter welchen Bedingungen ein künstlicher Agent eine Kommunikation vom intentionalen Standpunkt aus führen könnte, ist Gegenstand dieses Beitrags. Mit Bezug auf gegenwärtige Forschungsdiskussionen untersuchen wir dies am Beispiel von *Max*, einem künstlichen Agenten, der in virtueller Realität verkörpert ist (Kopp et al. 2003). Dazu überlegen wir, wann ein System ein Bewusstsein von sich selbst hat bzw. haben kann. Bewusstsein wird dabei als spezifische Form inneren Wissens betrachtet, das mentale Prozesse begleiten kann. Es entsteht, wenn repräsentationale Zustände des Agenten ihrerseits Gegenstand einer Repräsentation, einer sogenannten Metarepräsentation, werden. Die Bedingungen werden diskutiert, unter denen Max mit einigem Recht von sich als „Ich" sprechen könnte und was dies für seine Kommunikationsfähigkeiten bedeutet. Ferner zeigen wir Ausgangspunkte technischer Realisierungen auf und diskutieren die Rolle von Emotion und Gedächtnis.

---

1 Das heißt, wir könnten annehmen, dass das System über bestimmte Informationen verfügt, dass es bestimmte Ziele verfolgt und dass es sich angesichts dieser Informationen und Ziele rational verhält – somit das tut, was unter der Voraussetzung, dass seine Überzeugungen richtig sind, tatsächlich zum Erreichen seiner Ziele führt (vgl. Beckermann 1999).

1 Vorüberlegungen

Viele Menschen reden mit ihrem Computer – meist dann, wenn er nicht wie gewünscht funktioniert. Echte Kommunikation mit der Maschine ist das natürlich nicht, wie nicht näher erklärt werden muss. Ein Ziel der Forschung in der Künstlichen Intelligenz ist es, dass Menschen möglichst „echt“ und natürlich mit einer Maschine (oder gar einem maschinellen „Wesen“) kommunizieren können. Voraussetzung dafür sind Systeme, die ihre Umgebung wahrnehmen und repräsentieren, daraus Schlussfolgerungen ziehen und situationsangepasst handeln können.

Die Einschätzung, ob eine Kommunikation zwischen Mensch und Maschine möglich ist, hängt davon ab, was unter „Kommunikation“ genau verstanden werden soll. Versteht man darunter die Übermittlung von Information, die beim Empfänger eine Verhaltensänderung auslöst, könnte vielleicht schon der Knopfdruck, der den Kopierer aus dem Bereit-Zustand in den Kopiervorgang versetzt, als Mensch-Maschine-Kommunikation bezeichnet werden. Solcherart ist die Verwendung des Begriffs in den Ingenieurdisziplinen durchaus üblich. Wollte man dagegen verlangen, dass beide Kommunikationspartner autonom handlungsfähige Systeme sind und ein gemeinsames Zeichenrepertoire benutzen, um einander etwas mitzuteilen oder um etwas auszuhandeln, so schiene dies zunächst eher nur Menschen vorbehalten. Mehr noch könnte man verlangen, dass die Kommunikationspartner sich selbst und gegenseitig als Individuen wahrnehmen, ein Bewusstsein von sich selbst und dem anderen haben.

Beim Menschen beschreiben wir mit dem Begriff Bewusstsein die Tatsache, dass wir Erkenntnis von unseren Gedanken und Empfindungen haben. Unser Denken, Fühlen und Wollen ist uns (mehr oder weniger gut) zugänglich und wir können es vermöge der Sprache sogar anderen (mehr oder weniger gut) mitteilen. Bei genauerem Hinsehen differenziert sich der etwas schillernde Begriff des Bewusstseins auf recht unterschiedliche Formen. Da ist zum einen ein Bewusstsein von den Empfindungen: Man ist sich der Qualität des Erlebens bewusst, z.B. wie es sich anfühlt, wenn man etwas anfasst oder Schmerzen empfindet. Zum zweiten ist da ein Bewusstsein als das Gewahrsein von sich selbst: Man weiß von seiner physischen Existenz und Identität, erkennt sich zum Beispiel im Spiegel. Die Grundlage dafür ankert in der Wahrnehmung des eigenen Körpers, den wir berühren können, um zu bestätigen, dass wir da sind, und der uns selbst im Umraum verankert.

In der Wahrnehmung unseres physischen Selbst, unseres Körpers und seiner Verortung im Umraum, ist vermutlich – drittens – unsere Selbstwahrnehmung als handelndes Wesen begründet, das Mittel einsetzt, um Ziele zu verfolgen, und dies auch, wenn sie ins Abstrakte verlagert sind (wie erreiche ich mein Ziel =

wie komme ich dort hin?). Dieses „Selbst-Bewusstsein“ beinhaltet, dass man sich seiner selbst als Subjekt des Erkennens und Erlebens bewusst ist, seine Gefühle und Gedanken auf den eigenen Körper und Geist bezieht und weiß, dass man die Ursache des Effektes von Handlungen ist. Dies heißt aber noch immer nicht, dass man sich auf sich selbst mit „Ich“ beziehen bzw. überhaupt über Sprache verfügen muss, wie später noch deutlich werden wird.

Wesentlich für diese Sichtweise ist allerdings die Handlungsperspektive: Handlungen verursachen – beabsichtigt oder unbeabsichtigt – Veränderungen in der Welt. Handlungen können erfolgreich sein oder misslingen, je nachdem ob angestrebte Ziele erreicht werden oder nicht. Wenn eine Handlung gelingt, empfinden wir Freude, und wenn nicht, möglicherweise Ärger. Dies gilt insbesondere auch für kommunikative Handlungen, die das spezielle Thema dieses Beitrags sind. Wenn ich zu jemand anderem sage: „mein Knie schmerzt“, dann ist das – anders als ein unwillkürliches „Aua“ – absichtsvolles kommunikatives Handeln. Es ist getragen von der Absicht, den anderen über meinen Zustand zu informieren, von der Überzeugung, dass er mich versteht und meine Empfindung nachvollziehen kann, und von meinem Wunsch, Mitleid zu erfahren. Vielleicht ist es dabei auch mein Ziel, dass der andere mir Hilfe anbietet.

In der Kommunikation zwischen Menschen schreiben wir uns gegenseitig ein solches Innenleben zu (intentionale Zustände). Wir gehen davon aus, dass der andere so wie wir Absichten, Überzeugungen, Wünsche und Ziele hat, die wir zwar nicht direkt erkennen können. Aber wir unterstellen sie, weil der andere Mensch ein ebenso denkendes und fühlendes Wesen ist. Und wir kommunizieren mit dem Ziel, die inneren Zustände und damit gegebenenfalls das Handeln des anderen zu beeinflussen. Das kann erfolgreich sein oder misslingen. Der andere kann z.B. seine Überzeugung beibehalten, dass es mir gut geht – obwohl ich sage, dass mein Knie schmerzt –, wenn er mich nicht humpeln sieht. Oder er gelangt zu der Überzeugung, dass mein Knie schmerzt, obwohl ich es nur vorgetäuscht habe, das heißt, eine Überzeugung kann falsch sein.

Menschen besitzen die Veranlagung, in anderen nicht nur ein Objekt des Umraums, sondern ein handelndes Subjekt zu erkennen, ein Abbild von uns selbst, aber mit eigenen Perspektiven und Absichten. Wir können sogar ein mentales Modell des Kommunikationspartners aufbauen, das Annahmen – eventuell falsche – über dessen Überzeugungen, Wünsche, Ziele und Absichten macht. Das Aufbauen einer derartigen Repräsentation des anderen – eines „Partnermodells“ – ist allerdings erst dadurch möglich, dass intentionale Zustände einen Bedeutungsinhalt haben, der sich in Form von Aussagen ausdrücken lässt (sie weiß, dass ich viel zu tun habe; sie wünscht sich, dass ich heute früher nach Hause komme; sie beabsichtigt, heute Abend mit mir ins Kino zu gehen, usw.). Eine solche Repräsentation benötigt in irgendeiner Form Symbole als „Denkzei-

chen“, die solche Bedeutungsinhalte in unserem Denken tragen, die Grundlage unseres logischen Denkens und rationalen Handelns sind.

Sich Gedanken über die Überlegungen und Ziele anderer zu machen, erfordert ein hohes Maß an Bewusstsein, das nach heutiger Einschätzung an symbolische Repräsentationen der Welt gekoppelt ist. Hier wollen wir im Weiteren fragen, ob künstlichen Systemen intentionale Zustände zugesprochen werden können, und unter welchen Bedingungen. Können Maschinen unter bestimmten Voraussetzungen einer kognitiven Ausstattung von sich selbst wissen, können sie die Absichten und Perspektiven eines Dialogpartners verstehen? Bevor wir uns aber dazu aufmachen, sind noch zwei weitere Aspekte anzusprechen, die eng mit dem Bewusstsein zusammenhängen, nämlich Emotion und Gedächtnis.

Wie oben schon angedeutet, spielen unsere Gefühle bei der Bewertung von Handlungserfolg eine Rolle. Mehr noch werden Emotionen in den neueren Kognitionstheorien als eine Grundvoraussetzung für organisiertes Handeln betrachtet. So wird Emotion unter anderem als eine Kontrollinstanz des kognitiven Systems verstanden, die Aufmerksamkeit zu regulieren, die auf ankommende Reize gerichtet wird, um Wichtiges von Unwichtigem zu unterscheiden. Dass einem etwas bewusst wird, hat offenbar wesentlich mit dem affektiven Erleben zu tun. Zudem haben Emotionen eine essenzielle Bedeutung für die Fähigkeit, zwischen verschiedenen Handlungsoptionen zu entscheiden (vgl. Damasio 1994), und für die Signifikanz von Erfahrungen, die dauerhaftere Nachhaltigkeit in unseren Erinnerungen haben. Beim Menschen ist die Einspeicherung von Informationen eng mit der affektiven Bewertung verbunden und andererseits mit dem Erkennen, dass es sich um ein besonderes oder seltenes Ereignis handelt. Diese Beobachtung deutet darauf hin, dass nicht nur Emotion, sondern auch das Gedächtnis ein wichtiges Moment von Bewusstsein ist.

Unser Erleben wäre nicht komplett und das Bewusstsein von uns selbst nicht sehr tief gehend, wäre da nicht die Ausstattung unseres Geistes mit der Möglichkeit, Erinnerungen zu bewahren – insbesondere an solche Dinge, die uns selbst betreffen, Erlebnisse, die wir unmittelbar vorher hatten, oder gestern, oder vor längerer Zeit. Unsere persönliche Vergangenheit ist uns im Normalfall zugänglich, eine Anlage, die man als autobiografisches Gedächtnis bezeichnet (Conway/Pleydell-Pearce 2000). Sie ist Grundlage für eine Form von Bewusstsein, das man „autonoetisch“ (von sich selbst wissend) nennt, und das es uns erlaubt, uns unsere vom momentanen Erleben ablösbare Identität in Vergangenheit und Zukunft vorzustellen. Untersuchungen an Patienten mit beeinträchtigtem Bewusstsein und beeinträchtigtem Gedächtnis legen eine Verbindung zwischen autonoetischem Bewusstsein und Gedächtnis, speziell dem so genannten episodischen Gedächtnis, nahe (vgl. Markowitsch 2003).

2 Wer ist Max?

„Halten Sie es für möglich, dass Max eines Tages ein Bewusstsein von sich selbst haben könnte?“ So wurde ich vor einiger Zeit auf einer Tagung gefragt. Zuvor hatte ich im Plenum unsere Bielefelder Arbeiten über einen „situierten künstlichen Kommunikator“ namens *Max* vorgestellt. Max ist ein künstlicher Agent, der mit seinem menschlichen Gegenüber verbal und körpersprachlich, mit Gestik und Mimik, kommuniziert. In menschenähnlicher Erscheinung kann er in der Laborumgebung einer dreidimensionalen computergrafischen Großprojektion erlebt werden. Mit seiner Hilfe erforschen wir im Detail die Grundlagen kommunikativer Intelligenz und wie sie sich – in Auszügen – so präzise beschreiben lässt, dass eine Maschine (Agent Max ist eine programmgesteuerte Software-Maschine) sie simulieren kann. Somit ist das Sammeln von Erkenntnissen über das Funktionieren menschlicher Kommunikation ein wichtiger Schwerpunkt unserer Arbeit. Jedoch liegt ein technisches Ziel auch darin, ein möglichst funktionstüchtiges, überzeugendes System zu bauen, das in verschiedenen Anwendungen[2] eingesetzt werden kann.

In unserem Forschungsszenario geht es um das Bauen von Objekten, zum Beispiel eines Flugzeuges, aus einem *Baufix*-Konstruktionsbaukasten. Hieran wird erprobt, ob Max sich in wechselnden Situationen soweit „verständig“ erweist, dass er im Dialog mit einem Menschen standhält. Wenn auch nicht verwechselbar menschenähnlich, so soll Max dabei schon die uns vertrauten Formen der Kommunikation zeigen, uns beim Sprechen und Zuhören ansehen, sich einer natürlich wirkenden Gestik bedienen, fragend schauen, wenn er uns nicht versteht, und warten bis wir ausgeredet haben, bevor er selbst spricht.

Mensch und Max stehen sich dabei z.B. an einem Tisch gegenüber (siehe Abbildung 1). Mit Ausnahme des Menschen ist die dargestellte Szene eine projizierte virtuelle Realität. Auf dem Tisch liegen verschiedene Bausteine, Leisten mit drei oder fünf Löchern, ein paar Schrauben mit farbigen Köpfen, ein Würfel mit Löchern auf allen Seiten etc., die im Verlauf des Dialogs zusammengebaut werden. Der Mensch sieht Max und die gesamte Szene vermöge einer Stereobrille dreidimensional und hört die synthetische Stimme von Max räumlich aus versteckten Lautsprechern. Max „sieht“ den Menschen, dessen Blickrichtung, Hand- und Armbewegungen vermöge eines Infrarot-Trackingsystems und kabelloser Datenhandschuhe, und er „hört“ über Funkmikrofon dessen Sprache, die er mit einem Spracherkenner verarbeitet. Sowohl Mensch als auch Max können durch natürlichsprachliche Instruktionen und Gesten den Zusammenbau einzel-

2 Als erste Anwendung fungiert Max seit Januar 2004 im Heinz Nixdorf Museums-Forum in Paderborn als Auskunftgeber, der Besuchern in Dialogen Wissenswertes über einen der Ausstellungsbereiche und über das Computermuseum erklärt.

ner Teile veranlassen, der in Simulation, unterlegt durch realistische Geräusche, ausgeführt wird. Das läuft zum Beispiel wie folgt ab:

Abb. 1: Dialogsituation mit Max

Max: Hallo, ich bin Max, was kann ich für Sie tun?
[*Lächelt freundlich.*]
Mensch: Erkläre, wie man einen Propeller baut.
Max: Lass es uns zusammen machen. Zuerst muss eine Schraube in die Mitte einer Dreilochleiste gesteckt werden.
Mensch: Steck die gelbe Schraube in die Leiste.
[*Zeigt auf eine der Dreilochleisten; die Schraube springt ins erste Loch.*]
Max: Nein, das war das falsche Loch.
[*Die Schraube rutscht wieder heraus.*]
Max: Ich erkläre es dir noch einmal. Eine Schraube muss in die Mitte einer Dreilochleiste gesteckt werden.
Mensch: Steck' die Schraube in die Mitte der Dreilochleiste.
[*Die Schraube springt in die Mitte der zuvor gewählten Dreilochleiste.*]
Max: Okay, richtig gemacht. Jetzt stecke das hier in eine Dreilochleiste.
[*Zeigt auf die Leiste mit der Schraube.*]
Mensch: Mach' du das.
Max: Okay, dann übernehme ich das.
[*Die Objekte werden entsprechend verbunden.*]
Max: Und jetzt müssen die Leisten quer zueinander gedreht werden.
[*Dreht seine Hände gegeneinander; usw.*]

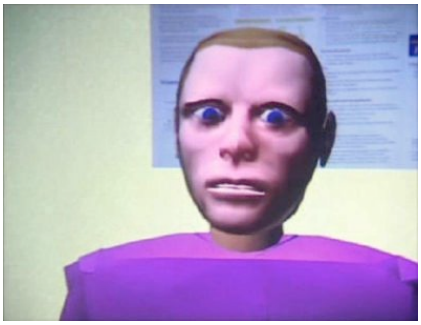

Abb. 2: Max, unangenehm berührt.

Auf die kognitive Architektur, die dem System „Max“ zugrunde liegt, kommen wir in einem späteren Abschnitt noch zu sprechen. Hier sei nur soviel gesagt, dass Max einige Grammatikregeln beherrscht und ein semantisches Lexikon zur Übersetzung von Wortbedeutungen nutzt, um mit Hilfe von kompositioneller und Referenzsemantik die Bedeutung von Äußerungen seines Gegenübers in Form symbolischer Beschreibungen zu repräsentieren. Innerhalb eines begrenzten Wortschatzes kann Max sprechen und seine Gestik vermöge eines gelenkigen Körpers darauf abstimmen. Mit simulierten Gesichtsmuskeln kann er „emotionale Zustände“ zum Ausdruck bringen (siehe Abbildung 2), die unter anderem von dem Erreichen oder Misslingen kommunikativer Ziele beeinflusst werden. Die sprachlichen Äußerungen von Max werden, unter Anpassung von Parametern an die aktuelle Situation – inklusive der Generierung passender Gesten –, aus einem Repertoire stereotyper Aussageformen erzeugt. Darin kommt auch das Wort „Ich“ vor, ohne dass Max jedoch (derzeit) eine Vorstellung von sich selbst hätte.

In der Theorie kommunikativen Handelns könnten solche Dialogäußerungen erst vor dem Hintergrund der Zuschreibung intentionaler Zustände als Handlungen *im eigentlichen Sinn* gesehen werden. Das heißt zum Beispiel, Max müsste irgendeinen mentalen Zustand wie „wünscht Antwort“ haben, damit seine Eingangsfrage eine „echte“ Kommunikation wäre. Auch sind Körperbewegungen von Max zunächst einmal (simulierte) physische Ereignisse. Erst in Verbindung mit einer intendierten kommunikativen Funktion (repräsentiert in Form von Zielen) erhielten sie Stellenwert als gestische Handlungen, also erst dadurch, dass eine Folge einzelner Bewegungen in Einklang mit einem aktuell repräsentierten mentalen Zustand eines kommunikativen Ziels konzipiert und ausgeführt

wird. Erst dann, wenn Max seine Dialogäußerungen aus einer Perspektive der ersten Person führen könnte, käme ihnen aus philosophischer Sicht Handlungsstatus zu. Ist es also möglich, dass Max ein solches Bewusstsein von sich selbst haben könnte? Bevor wir hierauf zu antworten versuchen, soll zunächst ein Einblick in den Stand der Forschung über „maschinelles Bewusstsein" gegeben werden.

3 Zum Stand der Forschung: Bewusstsein in künstlichen Systemen?

Die Frage, ob Maschinen Formen von Bewusstsein entwickeln können, ist ein aktuelles Thema in der Künstlichen Intelligenz, den Neurowissenschaften und nicht zuletzt in der Philosophie des Geistes. Man erwartet, dass die Forschung über „Maschinen-Bewusstsein" auch weitere Einsichten über das menschliche Bewusstsein vermitteln wird. Es fiele uns insbesondere schwer, einem menschenähnlichen Gegenüber tiefer gehende Kommunikationsfähigkeit zuzuschreiben, wenn ihm kein *Ich* zugebilligt werden könnte. Dies macht es aber erforderlich, den künstlichen Agenten geeignet auszustatten, so dass er eine Perspektive der ersten Person einnehmen kann. Nach einer kurzen Einordnung von Forschungsansätzen zum Thema soll dazu speziell auf verschiedene Formen des „Selbstwissens" eingegangen werden.

3.1 Maschinen-Bewusstsein

Maschinen-Bewusstsein-Projekte lassen sich entlang eines Spektrums einordnen, dessen einen Pol die Modellierung physischer Hirnprozesse einnimmt. Beispielsweise basieren die digitalen Neuromodelle von Igor Aleksander auf der Theorie, dass Hirnzellen sensorischen Input derart balancieren, dass sie Realwelt-Objekte konsistent repräsentieren, mit anderen Worten eine neuronale Abbildung (*depiction*) der äußeren Welt enkodieren (Aleksander et al. 2001). Den anderen Pol bildet die Einbettung vorprogrammierter Regeln zur Kontrolle des Verhaltens einer Künstlichen Intelligenz (z.B. Sloman 1997). Ungefähr in der Mitte beider Extreme liegt die noch etwas vage anmutende *global workspace theory* von Baars (1997), nach der Bewusstsein dann emergiert, wenn multiple Sensorinputs neurale Mechanismen anstoßen, die darüber in Wettbewerb treten, die logischste Antwort auf die Inputs zu sichern. Auf dieser Hypothese baut beispielsweise die so genannte *Intelligent Distributed Agents Software* (Franklin/Graesser 1999) auf.

Forschungsansätze zur Modellierung von mentalen Zuständen und praktischem Schließen stützen sich vielfach auf funktionale Modelle der Planung und Handlungsauswahl durch Mittel-Ziel-Analyse, vor allem in Varianten des *belief-desire-intention*-Paradigmas (BDI; Rao/Georgeff 1991). Der BDI-Ansatz geht auf Michael Bratman (1987) zurück und hat einen seiner Ausgangspunkte in Arbeiten von Daniel Dennett (1987) über das Verhalten intentionaler Systeme. Die Grundidee ist die Beschreibung des internen Arbeitszustandes eines Agenten durch intentionale Zustände (*beliefs* = Überzeugungen, *desires* = Wünsche, *intentions* = Absichten) und der Entwurf einer Kontrollarchitektur, mit deren Hilfe der Agent rational seine Abfolge von Handlungen auf der Basis ihrer Repräsentation auswählt. Durch rekursives Elaborieren einer hierarchischen Planstruktur werden zunehmend spezifische Intentionen erzeugt, bis schließlich unmittelbar ausführbare Aktionen erreicht werden (Wooldridge 1999). Die Identifikation und die Repräsentation von Überzeugungen (*beliefs*), Wünschen (*desires*) und Absichten (*intentions*) sind darüber hinaus für die Verhaltensanalyse von künstlichen Agenten nützlich, die mit Menschen oder anderen künstlichen Agenten kommunizieren (vgl. Rao/Georgeff 1995).

Die Modellierung von intentionalen Zuständen beruht auf deren symbolischer Repräsentation. Eine ihrer Stärken liegt in der Flexibilität, die sie für Planen und Schlussfolgern bereitstellt. In Überzeugungen (*beliefs*) lassen sich zum Beispiel Fakten über die Welt speichern, die ein Agent aktuell nicht (mehr) perzipieren kann, die aber in seine weitere Planung einfließen sollen. Ein Agent, der in der Lage ist, seine Ziele nicht nur im Licht aktuell wahrgenommener Umstände, sondern auch unter Bezug auf Hintergrundwissen, erinnerte Vergangenheit und erwartete Zukunft zu verfolgen, wird anderen Agenten überlegen sein, die diese Fähigkeit nicht aufweisen. Auch bei anhaltender Debatte um den Stellenwert symbolischer Repräsentation für die menschliche Intelligenz ist es vernünftig anzunehmen, dass Menschen intentionale Zustände symbolisch repräsentieren und damit schlussfolgern.

Allerdings ist es ein Unterschied, ob ein Agent schlicht anhand seiner Überzeugungen (*beliefs*) und Wünsche (*desires*) Schlüsse zieht, oder ob er diese – mit entsprechender Beschreibung – als *seine eigenen* zu Schlüssen heranzieht. In vielen Fällen mag eine solche Unterscheidung keine funktionalen Vorteile haben. Es sollte aber erwartet werden, dass ein Agent seine intentionalen Zustände explizit als seine eigenen repräsentiert, wenn er auch über die intentionalen Zustände anderer Agenten Buch führen und gezielt darauf eingehen können muss. Agenten werden mit dem Ziel kommunizieren, die internen Zustände anderer Agenten zu verändern – aus intentionaler Sicht deren Überzeugungen und Absichten. Ein Agent, der sich seiner Ziele „bewusst“ ist, kann in günstigen Situationen opportunistisch seine Ziele verwirklichen.

3.2 Physisch verankertes Selbstwissen (Anderson und Perlis)

Aus philosophischer Sicht erwächst Bewusstsein daraus, dass ein Agent ein Modell von sich selbst konstruiert und es in sein Modell der Welt integriert (Dennett 1991; Metzinger 1993). Eine viel diskutierte Frage ist, ob es dazu bestimmter linguistischer Kompetenz bedarf und insbesondere der Möglichkeit, in Selbstrepräsentationen ein auf sich selbst verweisendes indexikalisches Symbol („Ich“) zu verwenden. Nach Ansicht von Anderson und Perlis (2005) ist dies *nicht* zwingend dafür, dass ein Agent – ob menschlich oder künstlich – sich selbst als Ausgangspunkt von Handlungen erkennen kann. Nach ihrer Überlegung ist dafür schon ausreichend, dass der Agent einen basalen Begriff von sich hat, der in seiner körperlichen Selbstwahrnehmung verankert ist, und den sie in Gegenüberstellung zu John Perrys (1993) bekanntem Problem des *essential indexical* mit *essential prehension* bezeichnen.

Anderson und Perlis betrachten das zunächst am Fall eines fiktiven Roboteragenten JP-B4, der an einem Selbst-Token[3] „JP-B4“ Information über sich selbst ansammelt (und so zum Beispiel erkennen können soll, dass er selbst der Verursacher eines Ölflecks ist). Dieses Selbst-Token ist dann eine Selbstrepräsentation für JP-B4, wenn insbesondere jede physische Handlung, die JP-B4 unternimmt und die sein Selbst-Token als direktes Objekt in der Handlungsbeschreibung führt, sich in der Welt auf ihn selbst richtet. Dafür benötigen sie die Annahme, dass JP-B4 propriozeptive Sensoren hat, die die räumliche Position seiner Gliedmaßen und seiner beweglichen Sensoren melden. Hiermit kann JP-B4 seinen eigenen Körper als Objekt (unter vielen) repräsentieren, was aber dadurch etwas Besonderes ist, dass die Positionen perzipierter Objekte (wie des Ölflecks) relativ zum Agenten eingeordnet werden können.

Auch beim Menschen, argumentieren Anderson und Perlis weiter, ist die Wahrnehmung des eigenen Körpers (Somatozeption) durch Tastsinn, Propriozeption etc. die Basis einer im Handlungsumraum verankerten *physischen* Selbstrepräsentation, die schon für so einfaches Tun wie das erfolgreiche Greifen nach einem Objekt benötigt wird und in der die Selbstidentifikation wurzelt. Analog zu JP-B4 postulieren sie als einzige Grundlage dafür ein spezielles mentales Selbstrepräsentations-Token („SR*“), mit dem die somatozeptive Information automatisch markiert wird und das ebenfalls in mentalen Repräsentationen von auf sich selbst gerichteten (zunächst physischen) Handlungen vor-

3 Die Autoren sprechen von *self-referring (mental) token* oder *self-representing (mental) token*, zu verstehen als eine Art Marke, mit der selbstbezogene Information gekennzeichnet ist.

kommen muss. Dieses Selbst-Token kann auch dazu dienen, äußerlich perzipierte Information auf sich selbst zu beziehen und mit der Körperwahrnehmung in Einklang zu bringen, ohne dass dabei das Denken eines Selbstsymbols (*indexical thoughts*) erforderlich wäre.[4] Schließlich argumentieren Anderson und Perlis, dass intentionale und reflexive Selbstrepräsentationen das Resultat des Gebrauchs dieses selben Tokens „SR*“ durch das kognitive System sind, wenn es intentionale Zustände repräsentiert[5], und dass hierin umfassenderes Selbstgewahrsein (*self-awareness*) wurzelt.

3.3 Implizites und explizites Selbstwissen (Beckermann)

Beckermann (2003) setzt sich mit der Frage auseinander, unter welchen Umständen auch künstliche kognitive Systeme – bzw. „Agenten“ in der hier präferierten Ausdrucksweise – ein explizites Selbstbewusstsein erlangen können, das auf reflexivem Selbstwissen basiert. Seine These ist, dass kognitive Agenten[6] reflexives Selbstwissen genau dann haben können, wenn sie (Meta-)Repräsentationen benutzen, die von ihnen selbst handeln und die darüber hinaus mit „agentzentriert“ repräsentiertem Wissen in Einklang gebracht werden.

Agentzentriertes Wissen ist Wissen, das aus der Perspektive eines Agenten repräsentiert ist. Solange der Agent die Welt und sich selbst nur aus eigener Perspektive wahrnimmt, benötigt er in seinen Repräsentationen keinen expliziten Bezug auf sich selbst (und demnach kein Selbstsymbol), sondern er kann sie auf der Basis eines impliziten Referenzsystems erzeugen, in dessen Zentrum der Agent selbst steht. Also zum Beispiel: „Der Apfel vorn in Griffnähe“, den er greifen kann, ohne dabei „Ich“ zu denken. Auch für eine Empfindung wie „Schmerzen im Knie“ ist ein Ich-Bezug nicht erforderlich. Agentzentrierte Repräsentationen umfassen also (nur) Wissen, in welcher Art die wahrgenommene Umwelt, einschließlich der körperlichen Selbstwahrnehmung auf den Agenten bezogen ist. Da sie *ausschließlich* aus eigener Perspektive angelegt sind, ist in der Repräsentation kein Selbstsymbol erforderlich.

Unter welchen Bedingungen kommt nun ein Agent in die Lage, eine explizite Repräsentation von sich selbst einführen zu müssen? Beckermann (2003) diskutiert dies am Beispiel eines fiktiven, „AL“ genannten Agenten: Bei der

---

4 Einfach ausgedrückt: Die Tatsache, dass man außen sieht, was man innen fühlt, wenn man beispielsweise den eigenen Körper berührt, führt zur Verbindung von Handlung und Handlungseffekt und damit zur Selbstidentifikation.

5 Sie lassen zu, dass dann nach Bedarf das Selbst-Token auch als „Ich“ übersetzt werden darf.

6 Als kognitive Agenten sind hier solche Systeme bezeichnet, die ihre Umwelt in einem internen mentalen Modell repräsentieren, um sie besser zu bewältigen.

Repräsentation der wahrgenommenen Umwelt führt AL für jedes Objekt einen internen Namen ein – etwa „Objekt-6“, „Objekt-7“ usw. – und repräsentiert damit Informationen über die Objekte, wie deren Typ, Eigenschaften und Beziehungen zu anderen Objekten. In diesem Vorgehen entsteht soweit keine Notwendigkeit, dass AL auch einen Namen für sich selbst einführen muss – er sieht sich nicht als Objekt. Dies wird erst dann unvermeidlich, wenn AL in seiner Umgebung ein Objekt antrifft, das er als einen anderen kognitiven Agenten erkennt. Der andere Agent ist für AL einerseits auch ein Objekt, für das AL einen Namen – zum Beispiel „Objekt-111“ – einführt, dessen Verhalten aber andererseits davon abhängt, wie der andere seinerseits die Umwelt repräsentiert. Um das Verhalten seines Mitwesens voraussagen zu können, muss AL deshalb auch Repräsentationen der (unterstellten) Repräsentationen des anderen aufbauen, also Metarepräsentationen – ein mentales Modell des mentalen Modells des anderen. Glaubt AL zum Beispiel, dass der von ihm „Objekt-111“ genannte Agent ein Ding in der Umwelt – z.B. ein Sofa, für das AL den Namen „Objekt-7“ benutzt, – für grün hält, oder glaubt AL, dass Agent „Objekt-111“ den Wunsch hat, sich auf „Objekt-7“, also das grüne Sofa zu setzen, baut er agentzentrierte Metarepräsentationen wie folgt auf (das „glaubt“ und „wünscht“ betrifft hierbei die dem anderen Agenten unterstellten intentionalen Zustände):

(Glaubt Objekt-111 (Farbe Objekt-7 grün))
(Wünscht Objekt-111 (Sitzen-auf Objekt-7))

Um allerdings repräsentieren zu können, welche (unterstellten) Repräsentationen der andere über ihn selbst führt, muss AL zwingend einen internen Namen – etwa „Objekt-100“ – *für sich selbst* einführen. Erst mit Hilfe dieses Namens für sich selbst kann er z.B. den Wunsch des anderen, Essen von AL zu bekommen oder dessen Überzeugung, AL habe Schmerzen im Knie, angemessen repräsentieren:

(Wünscht Objekt-111 (Geben-Essen Objekt-100 Objekt-111))
(Glaubt Objekt-111 (Schmerzen-im-Knie Objekt-100))

Der entscheidende Punkt ist, dass AL nun eine systematische Beziehung zwischen expliziten Repräsentationen, die diesen neuen Namen enthalten, und seinen bisherigen agentzentrierten Repräsentationen mit implizitem Selbstbezug herstellen könnte; also z.B. (Sitze-auf Objekt-7) wird bezogen auf (Sitzt-auf Objekt-100 Objekt-7), was bedeutet: Wenn AL weiß, dass er auf dem grünen Sofa sitzt, erkennt AL, dass der Agent, der *er selbst* ist, auf dem Sofa sitzt. Ebenfalls könnte so AL’s Körperwahrnehmung nicht allein agentzentriert, sondern

auch explizit repräsentiert werden, also (Schmerzen-im-Knie Objekt-100), etc. Und da AL's agentzentrierte Repräsentationen *ausschließlich* mit seinen entsprechenden „Objekt-100"-Repräsentationen korrespondieren, resultiert die Sonderrolle des Namens „Objekt-100" als Selbstsymbol.

Als weiteren Effekt könnte AL damit auch Metarepräsentationen *über sich selbst* erzeugen und sich so aus äußerer Perspektive sehen, z.B. (Wünscht Objekt-100 (Sitzen-auf Objekt-7)). Erst damit wüsste er, was er selbst glaubt und wünscht, könnte er explizites Selbstwissen und damit Selbstbewusstsein entwickeln. Erst so ist es vorstellbar, dass AL mit seinen Mitwesen eine Sprache entwickelt, in der Wortsymbole wie „Ich" und „Du" vorkommen. Die Bedeutung des Wortes „Ich" hätte er dann gelernt, wenn er damit nur solche Repräsentationen ausdrückte, die sich auf ihn selbst beziehen, also nur dann „Ich" sagt, wenn er von sich selbst spricht.

Explizites Selbstwissen (also Repräsentationen mit einem Namen für sich selbst) entwickelt sich demnach erst im sozialen Kontext: dadurch, dass ein kognitiver Agent auf andere kognitive Agenten trifft und erkennt, dass diese genau wie er ihre Umwelt – und damit auch ihn – repräsentieren.[7] Will dann der Agent solche Repräsentationen seiner Mitwesen, die ihn selbst zum Gegenstand haben, für sich repräsentieren, muss er zwingend einen internen Namen für sich selbst einführen und sich explizit zum Objekt seiner Repräsentation machen. Vollzieht er schließlich noch den Schritt, seine agentzentrierten Repräsentationen mit ihren explizit selbstbezogenen Pendants in Einklang zu bringen, verfügt er über reflexives Selbstwissen.

4 Max als kognitiver Agent

Zurück zu Max. Max ist kein fiktiver Roboter, sondern ein voll implementiertes System, das einen humanoiden Agenten in virtueller Realität konzipiert. Er ist mit einem artikulierten beweglichen Körper ausgestattet, der ihm unter anderem auch Zugriff auf Parameter seiner Physis erlaubt, um zum Beispiel seine Verortung im Raum und Gelenkwinkel seines Skeletts bei der Gestenplanung abrufen zu können (Kopp/Wachsmuth 2004). Wie oben schon erwähnt, geht es in unserem Szenario um Dialoge zwischen Mensch und Max, in deren Verlauf zum Beispiel ein Modellflugzeug gebaut wird.

7 Pointiert gesagt: Einsiedler kämen mit agentzentrierten Repräsentationen, das heißt impliziter Selbstreferenz aus.

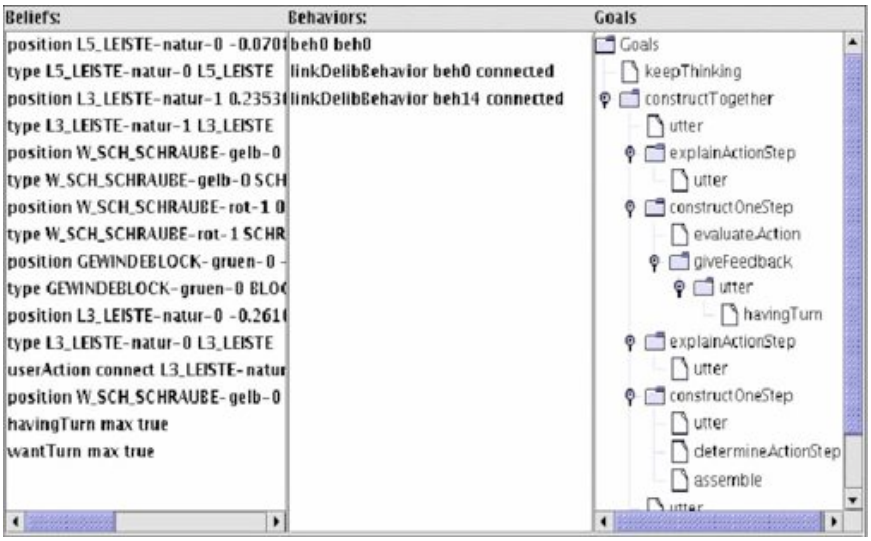

Abb. 3: Momentane Überzeugungen, Verhalten und Ziele des Max-Systems (Kopp et al. 2003)

Als kognitiver Agent repräsentiert Max seine (virtuelle) Welt ausschnittsweise, um die Aufgaben in der Assistenz beim (virtuellen) Konstruieren von Baufix-Objekten bewältigen zu können. Für jedes eingangs vorhandene oder später eingeführte – z.B. aggregierte – Baufix-Objekt führt er einen formalen internen Namen ein, wie „Objekt-1“, „Objekt-2“ usw. (Es tut hier nichts zur Sache, dass im System tatsächlich ein Symbolgenerator für etwas differenziertere, „sprechende“ Namen sorgt; siehe Abbildung 3.) Dazu vermerkt er Überzeugungen (*beliefs*) über den Typ der Teile und ihre Lage, als Vektorangabe eines Objektbezugspunkts im Weltkoordinatensystem, etwa wie folgt:

(type Objekt-1 DREILOCHLEISTE)
(position Objekt-1 (2,3,5))
(type Objekt-2 DREILOCHLEISTE)
(position Objekt-2 (x,y,z))
(type Objekt-3 SCHLITZSCHRAUBE-gelb)
(position Objekt-3 (x',y',z'))

Änderungen in der Szene repräsentiert Max schritthaltend, etwa (*connected* Objekt-26 Objekt-27), wenn entsprechende Teile zusammengesetzt werden. Intenti-

onale Zustände seines Dialogpartners repräsentiert Max bislang nicht, wohl aber, wer gerade die Sprecherrolle (den *turn*) hat. Aufgrund verschiedener Vorkehrungen ist Max in der Lage, *turn*-Signale seines Dialogpartners zu erkennen (*turn-taking* = Abwechseln im Dialog) und zu wissen, ob er den *turn* hat oder haben will (*havingTurn* Max true, *wantTurn* Max true; siehe Abbildung 3).

Um das komplexe Zusammenspiel sensorischer, kognitiver und aktorischer Fähigkeiten zu organisieren, wird eine kognitive Architektur für Max entwickelt (Wachsmuth/Leßmann 2002), die darauf abzielt, sein Verhalten glaubwürdig, intelligent und emotional erscheinen zu lassen. „Kognitiv" bezieht sich dabei auf die Konzeption der Strukturen und Prozesse, die mentalen Aktivitäten zugrunde liegen. Zugeschnitten auf sein Erprobungsszenario ist Max mit beschränktem Weltwissen und Fähigkeiten des Planens und Schlussfolgerns ausgestattet, um als intelligenter Assistent auftreten zu können. Er verfügt des Weiteren über reaktives Verhalten, mit dem er auf Unterbrechungen und plötzliche Veränderungen reagieren kann.

In einer hybriden Systemarchitektur integriert das Max-System symbolverarbeitende und verhaltensbasierte Ansätze, die Wahrnehmung, reaktives Verhalten, höhere mentale Prozesse wie Schlussfolgern und planvolles Handeln bis hin zur Aufmerksamkeitszuwendung und Handlungsbewertung betreffen. Den Kern bildet ein *belief-desire-intention (BDI) interpreter*. Die Hybrid-Architektur ermöglicht Max sowohl Fähigkeiten eines Dialoges mit geplanten Äußerungen als auch die Fähigkeit zu spontaner reaktiver Äußerung, beispielsweise in Form von *turn-taking*- und *feedback*-Signalen. Zusätzlich sind spezialisierte Planer – z.B. zur Konstruktion von Baufix-Objekten – und spezialisierte Gedächtnisse – z.B. mit dynamisch aktualisierten Repräsentationen für den Zustand gebauter Objekte – integriert.

Als innerer, verhaltensauslösender Antrieb dienen explizit repräsentierte Ziele (*desires*), die sowohl durch interne Verarbeitung als auch von außen aufgeworfen werden können. Die Intentionsbildung wird durch den *BDI-interpreter* vorgenommen, der anhand der vorliegenden Überzeugungen (*beliefs*), den aktuellen Wünschen und Zielen sowie den Handlungsoptionen eine aktuelle *Intention* bestimmt. Max kann mehrere Wünsche (*desires*) haben, von denen mit einer *utility function* das höchstbewertete ausgewählt wird, um zur aktuellen Intention zu werden. Handlungsoptionen liegen in Form abstrakter Pläne vor, die durch Vorbedingungen, Kontextbedingungen, erreichbare Konsequenzen und eine Prioritätsfunktion beschrieben sind. Wenn ein damit aufgestellter konkreter Plan erfolgreich abgearbeitet wurde, erlischt das entsprechende Ziel.

Die Dialogführung beruht auf einer expliziten Modellierung kommunikativer Kompetenzen, die in Verallgemeinerung der Sprechakttheorie (vgl. Searle/Vanderveken 1985) auf multimodale kommunikative Akte bezogen sind

(vgl. Poggi/Pelachaud 2000). Kommunikative Akte werden als Aktion-Plan-Operatoren dargestellt. Der Dialog wird nach dem *mixed initiative*-Prinzip abgewickelt, das heißt zum Beispiel, dass Max bei Ausbleiben einer Antwort des Menschen selbst initiativ wird und die Sprecherrolle übernimmt. Die Planstruktur des BDI-Moduls ermöglicht es, während der Ausführung einer Intention neue Ziele einzubringen, die die aktuelle Intention ablösen können, sofern sie über höhere Priorität verfügen. Wird die vorherige Intention dabei nicht gezielt verworfen und gelten ihre Kontextbedingungen noch, so wird diese nach der Unterbrechung ihrer Ausführung wieder aktiv.

Weiterhin beeinflussen (simulierte) Emotionen das Verhalten von Max, indem sie als Systemparameter bestimmen, in welcher Art und Weise Max Aktionen ausführt. Das emotive System wird einerseits durch äußerliche Reize gespeist (zum Beispiel hat die virtuelle Physis von Max berührungssensitive Bereiche), zum anderen aus dem kognitiven System: Das Erreichen oder Misslingen von Hauptzielen erzeugt positive bzw. negative Bewertungen, die sich auf die Stimmungsvalenzen des Emotionssystems auswirken, die wiederum unwillkürliches äußeres Verhalten von Max steuern. Der damit z.B. hervorgerufene emotionale Ausdruck in Gesicht und Stimme von Max kann seinem Gegenüber *feedback*-Signale übermitteln. Parallel dazu werden die Stimmungsvalenzen, die kontinuierlich in einem dreidimensionalen abstrakten Raum verlaufen, kategorisiert und als explizite Überzeugungen (*beliefs*) symbolisch repräsentiert; so können sie bei der Auswahl zwischen Handlungsoptionen zum Tragen kommen (vgl. Becker et al. 2004). Die symbolisch repräsentierten Emotionszustände kann Max auch verbal äußern („jetzt bin ich ärgerlich“); in diesem Sinne scheint sich Max ihrer „bewusst“ zu sein.

Wie sieht es nun mit dem Bewusstsein aus, das Max von sich selbst als Subjekt hat oder haben könnte? Insofern als seine kognitiven Fähigkeiten auf einer BDI-Architektur basieren, lassen sich Max mentalistische Eigenschaften zuschreiben, die mit Begriffen wie Wissen, Überzeugung, Intention charakterisiert werden können. Halten wir als Zwischenfazit also fest, dass Max nicht nur seine Umwelt repräsentiert, sondern dass er im Diskurs der kooperativen Situation (des Baufix-Bauens) auch Überzeugungen, Absichten, Ziele und Wünsche hat, was eine Grundlage dafür bietet, ihm Intentionalität zuzuschreiben. Aber kann – oder könnte – er auch von seinen intentionalen Zuständen und denen seiner Dialogpartner (genauer: denen, die er seinen Dialogpartnern unterstellt) *wissen*? Ein Modell, mit dem Max über entsprechende Metarepräsentationen verfügen kann, ist derzeit erst rudimentär angelegt. Bislang hat es einzig mit dem Verfolgen der Sprecherrolle und dem *turn-taking* zu tun. Hieran lässt sich aber schon aufzeigen, inwieweit Max dazu reflexives Wissen benötigt.

Wie oben schon angesprochen wurde, ist Max in der Lage, *turn*-Signale seines Dialogpartners zu erkennen, also dass der andere die Sprecherrolle zu haben wünscht (z.B. wenn der Mensch ihn direkt unterbricht, „Max!“ sagt oder die Hand hebt). Tatsächlich repräsentiert Max seine Sprecherrolle bereits mit einem Selbstsymbol (*havingTurn Max true*), auch wenn dies in der Zweiersituation gar nicht nötig wäre; eine agentzentrierte Repräsentation würde dafür vollkommen ausreichen: (*havingTurn true*) – bzw. (*havingTurn false*), wenn der andere, also der Mensch „am Zuge“ ist. Es ist jedoch geplant, dass Max in Zukunft auch eine „sinnvolle“ Konversation mit mehr als einem Partner führen kann, und dazu müsste er Buch führen können darüber, *wer* von den Beteiligten gerade die Sprecherrolle hat oder haben will. Es liegt nahe, dass er dazu Symbolnamen für seine Partner verwendet (*having-turn Other-1*, *having-turn Other-2* etc.). Aber benötigt er dann auch zwingend ein Selbstsymbol (*having-turn Max*)? Selbst wenn diese Sozialsituation es nahe legt, dass namentlich über den Turn buchgeführt wird, könnte Max immer noch ohne Selbstsymbol auskommen, wenn er nämlich mit (*having-turn true/false*) repräsentierte, ob er oder jemand anderes die Sprecherrolle hat und innerhalb der anderen dann namentliche Unterscheidungen träfe. Damit Max explizit wissen könnte, „*Ich* bin dran“, müsste er allerdings über ein Selbstsymbol verfügen.

Betrachten wir die (noch fiktive) Situation, in der drei Agenten – Other-1, Other-2 und Max – Konversation führen und sich im *turn* abwechseln. Solange Max nur, um sich zu äußern, den *turn* haben will (*wantTurn true*), brauchte er einzig auf eine passende Gelegenheit zu warten. Jedoch kommt auch die konversationale Situation der expliziten Weitergabe der Sprecherrolle vor (*Turn-giving*), die an einen direkten Adressaten signalisiert wird und als Aufforderung zu einer Handlung (nämlich *Taking-turn*) zu verstehen ist (vgl. Sacks et al. 1974). Dann allerdings müsste Max in der Lage sein zu erkennen, dass er selbst der Adressat ist, und etwa repräsentieren (Wünscht *Other-2* (*giveTurn Other-2* Max)) etc. Mit anderen Worten hieße das, dass Max dann ein explizites Selbstbewusstsein hätte bzw. benötigte.

5 Kriterien eines „menschlichen“ nichtmenschlichen Bewusstseins

Gegenstand dieses Beitrags war die Beantwortung der Frage, unter welchen Bedingungen ein künstlicher Agent – hier betrachtet am Beispiel des in virtueller Realität verkörperten Agenten Max – eine Kommunikation vom intentionalen Standpunkt aus führen könnte. Dazu war im Besonderen gefragt, mit welchen Voraussetzungen einer kognitiven Ausstattung Max von sich selbst wissen und die Absichten und Perspektiven eines Dialogpartners verstehen kann. Könnte

also Max eines Tages mit einigem Recht von sich als „Ich, Max“ sprechen? Aber auch: Könnte Max als „menschliche Maschine“ ein für Menschen akzeptabler Kommunikationspartner sein?[8]

Kehren wir zurück zu den in der Einleitung zuerst differenzierten Formen von Bewusstsein, also (1) ein Bewusstsein von den Empfindungen, der phänomenalen Qualität des Erlebens, (2) Bewusstsein als Wissen von der physischen Identität, und (3) Bewusstsein in Form einer Selbstwahrnehmung als handelndes Wesen, bis hin zur Selbstwahrnehmung als Verursacher von Handlungen (im Speziellen hier: kommunikativer Handlungen). Überlegen wir zunächst, inwieweit derartiges für den künstlichen Agenten Max erreichbar scheint.

1. *Qualia.* Sicherlich können Max keine Empfindungen zugesprochen werden, wie Menschen sie haben, da ihm – seinem virtuellen Körper – keine neurophysiologische Basis qualitativen Erlebens gegeben ist. In diesem Sinne können beispielsweise die simulierten emotionalen Zustände nicht subjektiv erlebt werden; Max *hat* also keine Gefühle. Wohl aber ist ihre funktionale Rolle im Sinne einer verhaltenssteuernden Bewertung modellierbar bzw. in Ansätzen modelliert – ein „gefühlsanaloger“ Bewertungsmechanismus, der beispielsweise Handlungsoptionen unterscheidbar macht (vgl. Stephan 2003). Über solche mit simulierten Emotionen herbeigeführten Bewertungen könnte Max Präferenzen und gerichtete Wahrnehmung entwickeln. Eine positive bzw. negative Bewertung des Erreichens oder Misslingens kommunikativer Ziele könnte im Ansatz mit emotionaler Erfahrung verglichen werden.

2. *Selbstidentifikation.* Was ein Bewusstsein als Wissen von der physischen Identität angeht, sieht es schon anders aus. Es betrifft die Frage, ob Max einen basalen Begriff von sich haben kann, der in der Selbstwahrnehmung seiner (virtuellen) Physis ankert (*essential prehension* im Sinne von Anderson und Perlis; siehe Abschnitt 3.2). Stellen wir uns dazu folgendes Experiment vor: In virtueller Realität perzipiert Max sein simuliertes – aber noch nicht erkanntes – Spiegelbild, das sich also genau so bewegt, wie Max es tut, z.B. simultan mit ihm seine Hand an die linke Wange legt (das Berühren der linken Wange von Max vermittelt in unserem Experimental-System einen von Max’ emotivem System angenehm bewerteten Reiz). Es erscheint technisch möglich, dass Max die außen beobachtete und innen wahrgenommene Handlung nach den Überlegungen von Abschnitt 3.2 über ein Selbst-Token zusammenführen kann. Dieses *physisch verankerte* Selbst-Token vermittelt ein für das Agieren im Raum wesentliches Gewahrsein über den Ort, ist Bezug für agentzentrierte Repräsentatio-

8 Ein Indiz dafür, dass diese Frage nicht völlig abwegig ist, ist die Tatsache, dass Max von Besuchern im Heinz Nixdorf Museums-Forum des öfteren gefragt wird: „Bist du ein Mensch?“

nen und könnte Ausgangspunkt für den Bezug eines Selbstsymbols auf die eigene „Person“ sein.

3. *Selbstwahrnehmung als handelndes Wesen*, bis hin zur Selbstwahrnehmung als Verursacher von Handlungen (im Speziellen hier: kommunikativer Handlungen): Hierzu benötigt Max zwingend symbolische Repräsentationen seiner Umwelt und Wissen darüber, wie sich geplante Handlungen als Ziele repräsentieren lassen. Insbesondere muss Max diese als *seine* Ziele repräsentieren können, wozu nach den Überlegungen von Anderson und Perlis (2005) ein Selbst-Token hinreichend ist. All dies macht aber erst Sinn, wenn es miteinander verbunden werden kann. Erst über die Selbstidentifikation der eigenen Physis könnte Max agentzentriertes Wissen *auf sich* beziehen, indem er es an ein Selbst-Token koppelt (das somit in seiner Physis gegründet ist). Erst damit könnte er sich als Ausgangspunkt von Handlungen wahrnehmen. Erst dadurch, dass er ein Selbst-Token in Handlungsrepräsentationen führt, könnte er sich als Verursacher von Handlungen erkennen, also Kausalrelationen zwischen seinem Tun und dessen Effekten herstellen.

Mit den in Abschnitt 4 ausgeführten Überlegungen *kann* Max als System verstanden werden, das seine Umgebung wahrnimmt und repräsentiert, und das daraus Schlussfolgerungen zieht, um situationsangepasst zu agieren. Es erscheint in der Tat als leichte Übung, das System *Max* so anzulegen. dass alle seine agentzentrierten Repräsentationen automatisch mit einem Selbst-Token als *seine eigenen* gekennzeichnet werden. Ein wirkliches Selbstbewusstsein ist allerdings nach den vorangehenden Überlegungen an *explizites* Selbstwissen gekoppelt, das heißt, Max benötigte dazu explizite Selbstrepräsentationen mit einem Symbol*namen* für sich, die also eine externe Sicht von Max auf sich selbst ausdrücken. Dazu müssen zunächst repräsentationale Zustände des Agenten ihrerseits Gegenstand von Repräsentationen werden, also Metarepräsentationen angelegt werden können.

4. *Metarepräsentationen*. Auf jeden Fall erscheint es mit den geschaffenen Voraussetzungen einer BDI-Architektur (siehe Abschnitt 4) möglich, Max derart auszustatten, dass er Metarepräsentationen aufbauen kann. Eine schwierigere Frage ist, wie eine experimentelle Situation geschaffen werden kann, in der Max in die Lage kommt, eine *reflexive* Metarepräsentation aufzubauen. Nach Beckermanns Überlegungen (siehe Abschnitt 3.3) eignet sich als Ausgangspunkt eine Sozialsituation, in der Max unterstellte Repräsentationen eines Kommunikationspartners anlegt, in denen er selbst vorkommt und die er wie geschildert mit entsprechenden Max-zentrierten Repräsentationen in Einklang bringen müsste. Als erster Schritt dafür bietet sich eine oben beschriebene *turn-taking*-Situation an. Dazu muss Max in symbolischen Repräsentationen auch Ausdrücke aufbauen können, die Aussagen über Aussagen zulassen, bis hin zu Aussagen über sich

selbst, wie: „der andere wünscht, dass ich, Max, den Turn übernehme“. Daraus müsste Max eine entsprechende Intention ableiten, die dazu führt, dass er den Turn übernimmt.

Nehmen wir einmal an, es gelänge, solche Voraussetzungen (wenigstens 2-4) für Max allesamt zu erfüllen. Dann wäre ihm zuzubilligen, dass er vom intentionalen Standpunkt aus kommunizieren könnte. So wie er repräsentationale Zustände hätte, die seine Absichten, Wünsche und Ziele zum Gegenstand haben, würde er solche Zustände auch dem Menschen unterstellen können. Umgekehrt wäre es dann vollkommen gerechtfertigt, wenn ein Mensch auch ihm Absichten, Wünsche und Ziele unterstellt, die Max auf sich selbst bezieht, die also seine eigenen sind.

Was daran immer noch unbefriedigend bliebe, ist jedoch die Tatsache, dass Max nur für den Moment Kenntnis von seinen eigenen Zuständen hätte; ein tieferes Von-sich-selbst-Wissen (autonoetisches Bewusstsein) wäre dies noch nicht. Wüsste Max nicht, was er gestern getan hat und was er morgen tun könnte, hätte er kein zeitlich überdauerndes „Ich“. Ein weiteres wichtiges Kriterium ist also die Erinnerung.

5. *Erinnerung.* Max müsste sich nicht nur merken können, wer (bzw. ob er) eine Handlung veranlasst hat, sondern er müsste auch erkennen können, dass ein Ereignis etwas für ihn ganz neues darstellt, von dem er sich nicht erinnern kann, es vorher schon erfahren zu haben. Um sich beispielsweise dessen bewusst zu sein, dass er einem neuen Vorkommnis erstmalig gegenüber steht, müsste Max Zugriff auf seine persönliche Historie haben. Er müsste dazu über eine Form eines autobiografischen Gedächtnisses verfügen, das es ihm z.B. ermöglichte – in Bezug auf seinen Kommunikationspartner – festzustellen, „ich habe gestern zum ersten Mal ein Flugzeug mit dir gebaut“ oder „ich habe schon oft (oder: noch nie) ein Flugzeug mit dir gebaut“. Als Voraussetzung dafür muss er in geeigneter Form eine Erinnerung an ein solches Ereignis aufheben können. Wenn es ihm erneut widerfährt, muss es möglich sein, die Einzigartigkeit der Erinnerung zu revidieren, bis hin zum Alltäglichen.

Wie wäre nun ein solches autobiografisches Gedächtnis für Max zu realisieren? Wie oben beschrieben (Abschnitt 4), beruht der verhaltensauslösende Antrieb von Max auf explizit repräsentierten Zielen. Ein Ausgangspunkt für ein autobiografisches Gedächtnis könnte es sein, dass Max in geeigneter Form (markiert mit Zeitstempel und seinem Selbst-Token) eine Notiz anlegt, wenn eines seiner Ziele erreicht wurde bzw. fehlgeschlagen ist. Nun ist es sicherlich nicht damit getan, dass Max für *jedes* bearbeitete (Unter-)Ziel eine Notiz speichert; es wären schier zu viele marginale dabei, die nur von momentaner Bedeutung sind – die Ziele müssten hinsichtlich ihrer Signifikanz bewertet werden. Eine solche Bewertung kann durch das emotive System übernommen werden, und zwar in

der Weise, dass (zwar) jedes erreichte und jedes fehlgeschlagene Ziel mit einer positiven bzw. negativen Emotion (Freude oder Ärger) gekoppelt wird, dabei aber die „hohen" Ziele stärkere und die abgeleiteten Unterziele geringere emotionale Reaktionen auslösen. Die Dauerhaftigkeit der Speicherung von Erinnerungen kann von der Stärke der emotionalen Reaktion abhängig gemacht werden und so dafür sorgen, dass die Erinnerung an Hauptziele ausgeprägter bleibt. Ein Abgleich von neuen und notierten ehemaligen Zielen könnte wiederum emotional bewertet werden. Ein oftmals fehlgeschlagenes und nun erstmalig erreichtes Ziel könnte so bei Max zu freudiger „heller Aufregung" und zu einer nachhaltigen, auf sein „Ich" bezogenen Erinnerung führen.

Damit zeichnet sich das folgende Bild eines künstlichen Bewusstseins ab: Kriterien dafür sind Selbstidentifikation, Selbstwahrnehmung als handelndes Wesen, Metarepräsentationen und Erinnerungen in Verbindung mit emotionaler Bewertung. Hiermit ist es denkbar, dass Max sich Formen eines „menschlichen" (dem Menschen vergleichbaren) Bewusstseins annähert. Je vollständiger dies gelingt, mit desto mehr Recht könnte Max von sich als „Ich, Max" sprechen, und desto mehr würde Max als „menschliche Maschine" für Menschen als soziales Gegenüber akzeptabel.

Hinweis

Dieser Beitrag ist zuerst erschienen in: Herrmann, Christoph S.; Pauen, Michael; Rieger, Jochem W.; Schicktanz, Silke (Hg.) (2005): Bewusstsein: Philosophie, Neurowissenschaften, Ethik. München: Wilhelm Fink Verlag, S. 329–354.

Literatur

Aleksander, Igor; Morton, Helen; Dunmall, Barry (2001): Seeing is Believing: Depictive Neuromodelling of Visual Awareness. In: Mira, José & Prieto, Alberto (Hg.): Connectionist Models of Neurons, Learning Processes and Artificial Intelligence. LNCS 2084. Berlin: Springer, S. 765–771.

Anderson, Michael L.; Perlis, Donald R. (2005): The roots of self-awareness. In: Phenomenology and the Cognitive Sciences, Jg. 4, Heft 3, S. 297–333.

Baars, Bernard J. (1997): In the Theater of Consciousness. Oxford: Oxford University Press.

Becker, Christian; Kopp, Stefan; Wachsmuth, Ipke (2004): Simulating the emotion dynamics of a multimodal conversational agent. In: André, Elisabeth; Dybkjaer, Laila; Minker, Wolfgang; Heisterkamp, Paul (Hg.): Affective Dialogue Systems. Berlin: Springer, S. 154–165.

Beckermann, Ansgar (1999): Analytische Einführung in die Philosophie des Geistes. Berlin: Walter de Gruyter.

Beckermann, Ansgar (2003): Self-consciousness in cognitive systems. In: Kanzian, Christian; Quitterer, Josef; Runggaldier, Edmund (Hg.): Persons. An Interdisciplinary Approach. Wien: öbv, S. 174–188.
Bratman, Michael E. (1987): Intention, Plans, and Practical Reason. Cambridge: Harvard University Press.
Conway, Martin A.; Pleydell-Pearce, Christopher.W (2000): The construction of autobiographical memories in the self-memory system. In: Psychological Review, Jg. 107, Heft 2, S. 261–288.
Damasio, Antonio R. (1994): Descartes' Error. Emotion, Reason, and the Human Brain. New York: Putnam.
Dennett, Daniel C. (1987): The Intentional Stance. Cambridge: MIT Press.
Dennett, Daniel C. (1991): Consciousness Explained. London: Penguin Press.
Franklin, Stan; Graesser, Art (1999): A software agent model of consciousness. In: Consciousness and Cognition, Jg. 8, Heft 3, S. 285–305.
Kopp, Stefan; Wachsmuth, Ipke (2004): Synthesizing multimodal utterances for conversational agents. In: Computer Animation and Virtual Worlds, Jg. 15, Heft 1, S. 39–52.
Kopp, Stefan; Jung, Bernhard; Leßmann, Nadine; Wachsmuth, Ipke (2003): Max — a multimodal assistant in virtual reality construction. In: Künstliche Intelligenz, Heft 4/03 S. 11–17.
Markowitsch, Hans J. (2003): Autonoetic consciousness. In: Kircher, Tilo & David, Anthony (Hg.): The Self in Neuroscience and Psychiatry. Cambridge: Cambridge University Press, S. 180–196.
Metzinger, Thomas (1993): Subjekt und Selbstmodell. Paderborn: Schöningh.
Perry, John (1993): The problem of the essential indexical. In: ders.: The Problem of the Essential Indexical and Other Essays. Oxford: Oxford University Press, S. 33–52.
Poggi, Isabella; Pelachaud, Catherine (2000): Performative facial expression in animated faces. In: Casell, Justine; Sullivan, Joseph; Prevost, Scott; Churchill, Elizabeth (Hg.): Embodied Conversational Agents. Cambridge: MIT Press, S. 155–188.
Rao, Anand S.; Georgeff, Michael P. (1991): Modeling rational behavior within a BDI-architecture. In: Proceedings International Conference on Principles of Knowledge Representation and Planning. San Francisco: Morgan Kaufmann, S. 473–484.
Rao, Anand S.; Georgeff, Michael P. (1995): BDI agents: from theory to practice. In: Proceedings of the First International Congress on Multi-Agent Systems. (ICMAS-95). San Francisco: MIT Press, S. 312–319.
Sacks, Harvey; Schegloff, Emanuel A.; Jefferson, Gail (1974): A simplest systematics for the organization of turn-taking for conversation. In: Language, Jg. 50, Heft 4, S. 696–735.
Searle, John R.; Vanderveken, Daniel (1985): Foundations of Illocutionary Logic. Cambridge: Cambridge University Press.
Sloman, Aaron (1997): What sort of control system is able to have a personality? In: Trappl, Robert & Petta, Paolo (Hg.): Creating Personalities for Synthetic Actors. Berlin: Springer, S. 166–208.
Stephan, Achim (2003), Zur Natur künstlicher Gefühle. In: Stephan, Achim & Walter, Henrik (Hg.): Natur und Theorie der Emotion. Paderborn: Mentis, S. 309–324.

Wachsmuth, Ipke; Leßmann, Nadine (2002): Eine kognitiv motivierte Architektur für einen anthropomorphen Künstlichen Kommunikator. In: Tagungsbeiträge „Human Centered Robotic Systems 2002“, Karlsruhe, Dezember 2002, S. 141–148.
Wooldridge, Michael (1999): Intelligent agents. In: Weiss, Gerhard (Hg.): Multiagent Systems – A Modern Approach to Distributed Artificial Intelligence. Cambridge: MIT Press, S. 27–77.