Der Avatar Max als virtuelles Phänomen

Ipke Wachsmuth, Universität Bielefeld (Künstliche Intelligenz)

Mit *Data* in Gene Roddenberry's *Star Trek Next Generation* und dem holographischen Doktor in *Voyager* sind künstliche Wesen, die in sozialer Gemeinschaft mit Menschen ihren Beitrag erbringen, für viele von uns längst vorstellbar geworden. Im Internet begegnen uns menschenähnliche Avatare[1], die Kunden gegenüber treten und Geschäfte vermitteln, in höhlenartigen Großprojektionen der Virtuellen Realität sogar in Lebensgröße. Können wir eines Tages Maschinen als ansatzweise gleichrangige Kommunikationspartner erleben, die „verstehen", was wir von ihnen wollen, und die Rolle eines sozialen Gegenübers einnehmen können?

Im Gebiet Künstliche Intelligenz wird erforscht, wie sich Systeme konstruieren lassen, die wie der Mensch ihre Umgebung wahrnehmen, daraus Schlussfolgerungen ziehen und in ihrer Umgebung angepasst handeln können; damit sollen detaillierte Aufschlüsse über das Funktionieren von Intelligenz erlangt werden. Ein technisches Ziel ist die Verbesserung der Mensch-Maschine-Kommunikation durch Systeme, die sich sprachlich und gestisch mit dem Menschen verständigen können und damit die Kommunikation mit der Maschine leichter fasslich gestalten. Es wäre viel gewonnen, wenn uns im Umgang mit der zunehmend komplexeren Technik ein anthropomorpher Ansprechpartner zur Verfügung stünde, dessen Umgangsformen denen des Menschen gleichen.

Eine Maschine, die mit dem Menschen kommuniziert

„Hallo, ich bin Max, was kann ich für Sie tun?" Eine freundliche Begrüßung, noch dazu mit einem Hilfsangebot, wird wohl von jedem gern angenommen. Wäre es nicht angenehm, wenn wir im virtuellen Raum von einem freundlich lächelnden Assistenten begrüsst würden, der zudem noch Kenntnis von seiner Arbeitsumgebung hätte und die Fähigkeit, als „Avatar" eines technischen Systems Leistungen zu vermitteln und uns dabei zu begleiten und zu assistieren?

Im Labor für Künstliche Intelligenz der Universität Bielefeld entwickeln wir mit einem Forscherteam einen solchen Avatar. In einer computergrafischen Großprojektion ist er in menschenähnlicher Gestalt verkörpert, ein virtuelles Phänomen also, das wir zwar sehen, aber nicht anfassen können. In der virtuellen Welt kann der Avatar bestimmte Aktionen ausführen und darüber einen Dialog mit seinen menschlichen Besuchern führen. In einem unserer Anwendungsbeispiele hilft er beim Zusammenbau kleiner Fahrzeug- und Flugzeugmodelle aus Teilen eines Konstruktionsbaukastens, die in dreidimensionaler computergrafischer Darstellung als „virtuelle" Objekte auf einem „virtuellen" Tisch vor uns liegen; es handelt sich also um eine Computersimulation. Er sagt (mit einer synthetischen Stimme) zum Beispiel *„Jetzt nimm diese Schraube und steck sie*

[1] Der Begriff „Avatar" ist dem Sanskrit entlehnt und bedeutet dort den „Herabstieg" (Sanskrit: Avatara) einer Gottheit in irdischer Gestalt. In der digitalen Welt bezeichnet er oft die virtuelle Repräsentanz des Teilnehmers einer Web-Umgebung oder Spielewelt, kann aber – wie im vorliegenden Beitrag – auch verstanden werden als sichtbare Erscheinung einer unsichtbaren, eigenständig agierenden Software-Maschine, verkörpert als eine computeranimierte Figur.

in diese Leiste" und zeigt dabei mit einem freundlichen Lächeln auf die entsprechenden Bauteile (Bild 1). Umgekehrt kann er auch unser Sprechen und Zeigen, über Mikrofon und Infrarot-Kameras, wahrnehmen – ein echter Ansprechpartner, der sogar ein kleiner „Experte" im Baukastenbauen ist. Weil unser Avatar sich einerseits multimodal – mit Sprache, Mimik und Gestik – äußert und sich andererseits mit der Assemblierung, das heißt dem Zusammenbau virtueller Objekte auskennt, wurde er MAX – „Multimodaler AssemblierungseXperte" – getauft.

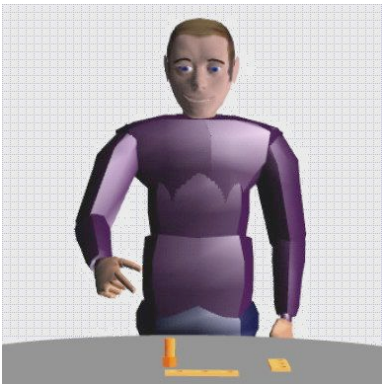

Bild 1

In unserer Forschung geht es somit um kybernetische Maschinen mit kommunikativen Fähigkeiten. Im Detail wollen wir herausfinden, was kommunikative Intelligenz ist, ja wie sie sich in Auszügen so präzise beschreiben lässt, dass eine Maschine (auch unser Max ist eine programmgesteuerte Software-Maschine) sie simulieren kann. Es ist dabei nicht unser Anliegen, Max verwechselbar menschenähnlich zu gestalten. Doch soll er die dem Menschen vertrauten Formen der Kommunikation zeigen, uns beim Sprechen und Zuhören ansehen, sich einer natürlich wirkenden Gestik bedienen, vielleicht verständnislos schauen, wenn er uns nicht versteht, und warten bis wir ausgeredet haben, bevor er selbst spricht.

Die Software-Maschine Max ist ein System aus vielen interagierenden Programm-Modulen (ein sogenanntes Multi-Agenten-System), von denen einige im Folgenden beschrieben sind.

Wie versteht Max Sprache?

Das Verstehen von Sprache zählt zu den zentralen kognitiven Fähigkeiten. Wie meistert Max so etwas? Stellen wir uns vor, Max „hört" folgenden Satz über ein Mikrofon, das die Rolle seiner Ohren übernimmt: *„Jetzt steck die gelbe Schraube in die lange Leiste."*

Max verarbeitet das akustische Signal zunächst mit einem Spracherkenner. Das ist ein Computerprogramm, das mit Hilfe eines Wortlexikons aus dem Signal-Klangmuster Wörter herausfiltert (segmentiert). Dabei werden mit Grammatikregeln unsyntaktische Alternativen ausgeschieden, zum Beispiel könnten die letzten zwei Wörter auch als

„lang geleistet“ gehört worden sein, was im Kontext des *„in die“* keinen korrekten Satz ergäbe. Wenn der Prozess bis hierhin erfolgreich war, hat Max aus dem Gehörten das Gesagte, also den Satz *„Jetzt steck die gelbe Schraube in die lange Leiste“* in Computertext rekonstruiert, was den ersten Schritt des Sprachverstehens – die Spracherkennung – abschließt.

Wie kann Max aber den Sinn des Gesagten verstehen? Dazu braucht er Wissen über die Wortbedeutungen, auf die er in einem semantischen Lexikon zugreifen kann, zum Beispiel dass „stecken“ eine Art des Verbindens und die Imperativform „steck“ einen Befehl bezeichnet. Bei der Analyse des Satzes schreibt Max diese Bedeutungsaspekte den einzelnen Wörtern zu und setzt daraus die Satzbedeutung zusammen (kompositionelle Semantik; siehe Bild 2). Die Satzbedeutung wird aus den Wortbedeutungen zusammengesetzt. In diesem Beispiel wird das „jetzt“ als Füllwort gewertet, das „steck“ als Befehl, eine Verbindung herzustellen (CONNECT), das Wort „die“ als bestimmter Artikel (determiner), das „gelbe“ als eine Farbe, die als GELB angegeben wird, das Wort „Schraube“ als ein Objekt des Typs SCHRAUBE, das „in“ als Präposition IN, das „lange“ als Größenangabe, die als GROSS benannt wird, und das Wort „Leiste“ als ein Objekt des Typs LEISTE.

Um den Satz in vollem Umfang zu verstehen, muss Max den Bezug auf die wahrgenommene Weltsituation herstellen (Referenzsemantik). Aus den Satzteilen „die gelbe Schraube“ und „die lange Leiste“ werden Suchanfragen etwa wie folgt abgeleitet:

(select x (OBJEKTTYP(x) = SCHRAUBE und FARBE(x) = GELB))
(select y (OBJEKTTYP(y) = LEISTE und GROESSE(y) = GROSS))

Das heißt, in der wahrgenommenen Szene (rechts in Bild 2) sind Objekte zu bestimmen, die diesen Anfragen genügen. Zum Beispiel ist die Größenbeschreibung GROSS eine Angabe, die relativ zu anderen LEISTE-Objekten bestimmt wird, etc. Wenn eindeutige Bezugsobjekte für x und y bestimmt werden konnten, ist der Auftrag an Max, diese zu verbinden (CONNECT) in vollem Umfang verstanden und kann ausgeführt werden. Das Verstehen eines solchen Satzes dauert kaum mehr als eine halbe Sekunde.

jetzt	FUELL	
steck	BEFEHL	CONNECT
die	DET	
gelbe	FARBE	GELB
Schraube	OBJEKTTYP	SCHRAUBE
in	PRAEP	IN
die	DET	
lange	GROESSE	GROSS
Leiste	OBJEKTTYP	LEISTE

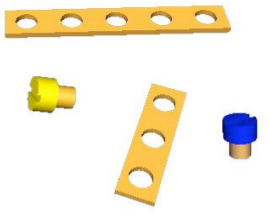

Bild 2

Zu den kognitiven Fähigkeiten von Max gehört weiter, dass er nonverbale Äußerungen seines menschlichen Gegenübers wahrnehmen und interpretieren kann. Gesten und Blickrichtung des Menschen werden ihm über sog. *Tracker* übermittelt, so dass Max auch mitbekommt, wohin der Mensch beim Sprechen eines Auftrages schaut oder worauf er dabei zeigt.

Eine Stimme für Max

Damit Max auch selber sprechen kann, müssen zunächst einmal Klänge und Geräusche erzeugt werden, die der menschlichen Stimme ähneln. Die physikalische Analyse gesprochener Sprache zeigt unter anderem, dass wir stimmhafte Laute, wie Vokale, in ganz bestimmten Frequenzbereichen formen, man nennt sie Formanten. Die Grundfrequenz, auch Grundton genannt, liegt für eine männliche Stimme je nach Stimmlage zum Beispiel bei etwa 100 Hertz (Schwingungen pro Sekunde). Für eine weibliche Stimme liegt sie zum Beispiel bei 220 Hertz. Alle Formanten sind ganzzahlige Vielfache der Grundfrequenz, also Obertöne. Ihre spezielle Mixtur ergibt den persönlichen Charakter einer Stimme, was die klangliche Komponente angeht. Stimmlose Laute sind dagegen Geräusche ohne Klang, sie stellen sich als Rauschen in hohen Frequenzbereichen dar. Vokale sind also durch Obertöne des Grundtons gekennzeichnet, Konsonanten vornehmlich als Geräusche.

Mit der Hochgeschwindigkeit moderner Rechner lassen sich heute synthetische Stimmen durch Software, also Computerprogramme, in Echtzeit zu erzeugen. Grundlage dafür ist die Erkenntnis, dass der Sprechschallstrom in Komponenten zerlegt werden kann: in die Grundfrequenz, die die Sprachmelodie bestimmt, und in wechselnde Oberton- und Geräuschanteile für die Vokale und Konsonanten. Das Programm MBROLA („Embrola“), das wir dazu einsetzen, hat in einer umfangreichen Datenbank sog. Diphone (Übergänge zwischen zwei Lauten) gespeichert. Sie lassen sich zu einer digitalen Klangbeschreibung zusammensetzen und über Soundkarte und Lautsprecher als akustisches Signal hörbar machen.

Der zu sprechende Text muss zuvor aber erst in eine Liste von Phonemen überführt werden. Dafür setzen wir das Programm TXT2PHO („Text to Pho“) von der Universität Bonn ein, zu dem ein Aussprachelexikon mit über 50.000 Einträgen gehört. In unserem Labor haben wir eine Methode entwickelt, mit der die Betonung nach Bedarf erzeugt werden kann. Dazu benutzen wir eine sog. *markup*-Sprache, SABLE, die auf der *extensible markup language* (XML) basiert, um betonte Silben zu markieren, die bei Überführung der Texteingabe in phonetischen Text sofort – „online“ – ausgewertet werden. Auch wenn es der synthetischen Sprache von Max ein wenig an „Seele“ fehlt, kann die Betonung kontrolliert und mit der Gestik abgestimmt werden. So kann Max in natürlich wirkendem Miteinander sprechen und zeigen.

Ein animiertes Gesicht für Max

Mimik ist ein universales, über alle Kulturen hinweg verständliches System der Kommunikation. Deshalb lässt sich auch erwarten, dass der Gesichtsausdruck von Max, wenn er den Regeln der mimischen Programme folgt, vom Menschen richtig verstanden wird. Werfen wir zuerst einen Blick auf die menschliche Gesichtsmuskulatur (Bild 3). Da gibt es zum Beispiel den Stirnmuskel (A), der die Augenbrauen hebt, und den Augenbrauenrunzler (B), der nicht nur beim finsteren Blick zum Einsatz kommt. Beim Lächeln spielen Augenringmuskel (C), Jochbeinmuskel und Mundwinkelheber ihre Rolle, während der „Herabdrücker des Winkels des Mundes“ *(depressor anguli oris)* eher negative Emotionen ausdrückt. Die Aktivität der Gesichtsmuskulatur führt also zu der von uns erkennbaren Mimik und natürlich auch zu den Lippenbewegungen beim Sprechen.

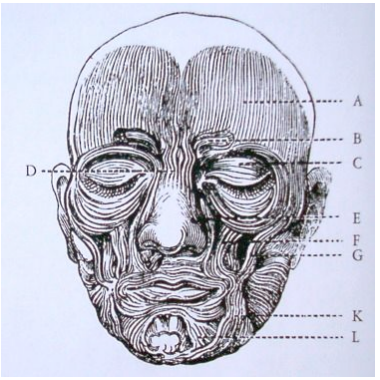

Bild 3[2]

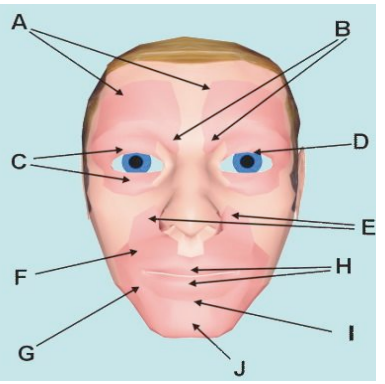

Bild 4

Über 40 Muskeln verleihen unserem Gesicht Ausdruck, und die wichtigsten davon sind für Max berücksichtigt (Bild 4). Die hervorgehobenen Gesichtspartien von Max können durch simulierte Muskeleffekte mit Hilfe sogenannter Aktionseinheiten animiert werden. Grundlage dafür ist das von den Psychologen Paul Ekman und Wallace Friesen entwickelte *Facial Action Coding System*, das die Kodierung sämtlicher mimischen Gesichtsausdrücke erlaubt. Dabei kann ein und derselbe Muskel an verschiedenen Aktionen beteiligt sein, und es können sich mehrere Aktionseinheiten in einem Gesichtsausdruck mischen, wie bei finsterem Lächeln oder fröhlicher Überraschung. Mit seiner Gesichtsmimik kann Max unterschiedliche Emotionen ausdrücken und so dem Menschen ein leicht verständliches Feedback übermitteln (Bild 5). Wenn Max zum Beispiel eine gesprochene Eingabe nicht verstanden hat oder noch an der Planung einer Äußerung „überlegt", kann er verständnislos oder nachdenklich schauen.

Bild 5

Auch die Sprechbewegung des Mundes entspringt dem Zusammenspiel der Gesichtsmuskeln. Für die Sprechanimation sind die sog. *Viseme* (visuellen Phoneme) entscheidend, sie beschreiben die Gesichtsstellung (Mund, Lippen etc.) bei der Artikulation der Phoneme. Ob *Mama*, *Papa* oder *Ball* gesagt wird, beim Wortanfang sind die Lippen auf gleiche Weise geschlossen, das heißt es reicht ein Visem für *M, P, B* und so fort. Wenn ein von Max zu sprechender Satz in eine Phonemliste überführt wird, werden zugleich die passenden Viseme zugeordnet. So kann Max den Mund synchron zum Sprechen bewegen.

[2] Zeichnung der Gesichtsmuskeln (Ch. Bell), zitiert aus: Charles Darwin, Der Ausdruck der Gemütsbewegungen bei dem Menschen und den Tieren, Frankfurt a. M. 2000 (Eichborn), S. 32.

Ein humanoider Körper für Max

Die in der virtuellen Realität verkörperte Erscheinung von Max umfasst nicht nur eine Stimme und ein animiertes Gesicht, sondern auch einen vollständigen anthropomorphen – nach dem Menschen geformten – Körper, der verschiedene Stellungen und Haltungen einnehmen kann und sich in der uns vertrauten Weise bewegt, wenn er zum Beispiel auf etwas zeigt. Unter der Körperhülle sorgt ein Skelett aus verbundenen Segmenten, sog. kinematischen Ketten, dafür, dass Max sich gelenkig bewegen kann. Besonders für die Gestik ist Max sehr „gelenkig" (Bild 6): Schulter und Schlüsselbeingelenk, Ellenbogen und Handgelenk, Hände mit fünf Fingern, jeder mit drei Gelenken modelliert, und ein Daumen, der zur Handfläche eingeklappt werden kann, erlauben natürliche Beweglichkeit.

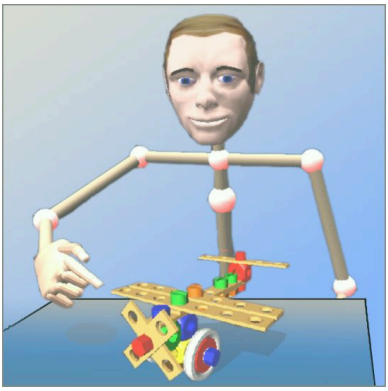

Bild 6

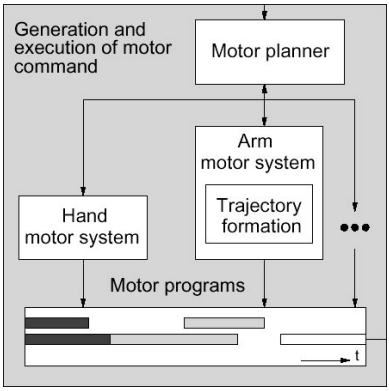

Bild 7

Ein hierarchisches Motorsystem steuert das kinematische Skelett von Max (Bild 7). Auf der höchsten Ebene wird die geplante Bewegung als Ziel repräsentiert (zum Beispiel „auf das Flugzeug zeigen!"). Die Steuerung der auszuführenden Bewegungen wird schrittweise in Unterplänen detailliert, bis schließlich einzelne Motorprogramme die Gelenke in Bewegung versetzen, so dass sich die Bewegung in die Zeigestellung ergibt. Max hat dazu ein Gestenlexikon, aus dem die Bewegungsverläufe typischer Gesten parametrisch abgerufen und situationsgerecht angepasst werden. Ausgehend davon werden alle Zwischenbewegungen vom motorischen System automatisch erzeugt. Hinter der „Körperintelligenz" von Max verbirgt sich eine Menge Mathematik, die es ermöglicht, die zeitliche Feinplanung der Bewegungen mit seiner synthetischen Sprache („steck sie in *diese* Leiste") abzustimmen. Mit seinem gelenkigen Körper kann Max sein Sprechen mit Gesten untermalen und sogar die Gesten des vor ihm stehenden Menschen imitieren (Bild 8).

Aber wie steht es mit der fühlbaren Körperlichkeit von Max? Sein computergrafisch animierter Körper ist zwar sichtbar, aber nicht berührbar und in dieser Hinsicht körperlos. Ein Mensch, der Max gegenüber tritt, spürt dennoch, wenn Max bis auf „Normalabstand" herankommt, und kommt er noch näher, verspürt man sogar den unmittelbaren Impuls zum Zurückweichen. Ebenso hat Max „proxemische Sensoren", Körperfühler sozusagen, mit denen er Nähe und Annäherung spüren kann. In dem

Moment, wo eine menschliche Hand – mit einem Datenhandschuh bestückt – und des Avatars computeranimierte Hand sich in der virtuellen Welt treffen, funkt und knistert es – ein virtueller Funke, der in die reale Welt überspringt (Bild 9).

Bild 8

Bild 9

Ausblick

Mit den Arbeiten an Max fragen wir uns, wie man bestimmte Ausschnitte der Kommunikation und ihr zugrunde liegende Intelligenzfähigkeiten synthetisch herstellen kann. Das erfordert nicht nur bestimmte kognitive Fähigkeiten, sondern auch die Möglichkeit, sich körperlich mitzuteilen, und dies betrifft mehr als Stimme und Sprechen. Gerade das Zusammenspiel verbaler und nonverbaler Kommunikationsformen, zum Beispiel mit Gestik und Mimik, erlaubt eine robuste und intuitive Verständigung. Und auch die verkörperte Gegenwart im virtuellen Raum gehört dazu, um sinnvoll „hier" und „dort", „links" und „rechts" sagen zu können.

Durch die Entwicklung und Erprobung technischer Modelle wollen wir neue Erkenntnisse über das Funktionieren menschlicher Kommunikation, dem vielleicht eindrucksvollsten Feld menschlicher Intelligenz, gewinnen. Wie funktioniert beispielsweise das zeitliche Zusammenspiel von Sprechen und Zeigen? Wie wird das Abwechseln im Dialog gesteuert? Welche Rolle spielen Emotionen?[3] Die sich hiermit ergebende – wohl spannendste – Frage nach der Architektur eines körperlichen natürlichen bzw. eines verkörperten künstlichen „Organismus" kann nur in Zusammenarbeit der Disziplinen erforscht werden. Sie umfasst das Verhältnis von Maschine und Körper ebenso wie die Interaktion von Mensch und Maschine.[4]

Auch wenn es schwer sein dürfte, die organische Vielfalt menschlicher Kommunikation technisch vollkommen nachzubilden: Eine kommunikationsfähige Maschine, die sich ähnlich wie ein Mensch ausdrücken kann und die unsere Sprache, Gestik und Mimik versteht, könnte ein verständiger und verständlicher Partner in der Alltagswelt

[3] Siehe dazu I. Wachsmuth: „Ich, Max" – Kommunikation mit künstlicher Intelligenz, in Ch.S. Herrmann, M. Pauen, J.W. Rieger, S. Schicktanz (Hg.), Bewusstsein: Philosophie, Neurowissenschaften, Ethik. München 2005 (Wilhelm Fink Verlag).

[4] Aufgegriffen zum Beispiel in einem Forschungsjahr am Zentrum für interdisziplinäre Forschung (ZiF), http://www.uni-bielefeld.de/ZIF/FG/2005Communication/

sein. Es böte sich die Chance zur Gestaltung technischer Systeme, die dem Menschen nicht undurchschaubar und fremd sind. Keiner muss den Umgang mit Max erst mühsam erlernen, jeder von uns weiß, wie man einen Menschen etwas fragt und somit auch, wie man Max etwas fragt.[5]

Ein Vorteil, den Max als virtuelles Wesen gegenüber allen Lebewesen hat, ist seine Affinität zur digitalen Welt. Er kann digital verfügbare Information leicht in sein Tun einbeziehen. Um dies auszuprobieren (schon damit ihm seine Kenntnisse nicht alle mühsam von Hand beigebracht werden müssen), haben wir Max mit dem Internet verbunden. Von dort kann er Informationen abrufen und in seinen Auskünften mitteilen. Man kann ihn zum Beispiel fragen, wie das Wetter wird:

BESUCHER: Wie wird das Wetter?

MAX: Moment, ich frag mal einen Wettersatelliten... Bis zum Mittag regnet es teilweise recht stark, zum Nachmittag ist es dann heiter bis wolkig. Maximal 4 Grad. In der Nacht kühlt es sich bei wechselnder Bewölkung auf -1 Grad ab.

BESUCHER: Und morgen?

MAX: Mal den Stand der Sterne betrachten... Vormittags ist es heiter bis wolkig, am Nachmittag meist dicht bewölkt bei -2 Grad. In der Nacht fällt bei 1 Grad Schnee.

BESUCHER: Danke.

MAX: Aber bitte Paul, gern geschehen. (Der Besucher hatte zu Beginn des Dialogs seinen Vornamen eingegeben.)

Ist die Internet-Verbindung getrennt, kann Max in vielen Fällen doch angemessen reagieren:

BESUCHER: Wie wird das Wetter?

MAX: Ich würde gern eine Bauernregel zitieren, aber leider find ich keine passende. Das Wetter ist ja nur in den Ländern ein Gesprächsthema, wo es sehr unterschiedlich ist. Zum Beispiel am Äquator ist es sowieso immer warm.

Mit der Simulation des Menschen in Form von Avataren – zunächst ein virtuelles Phänomen – entstehen neue Kommunikationsmöglichkeiten mit virtuellen Welten, die aus reinen Informationen bestehen; die reale Welt und die virtuelle Welt verschmelzen. Das Modellieren solcher menschlichen Fähigkeiten trägt aber auch dazu bei, dass Max ein für den Menschen akzeptabler Sozialpartner sein könnte. Als digitaler Assistent oder Informationsvermittler könnte er zuweilen vielleicht bessere Dienste leisten als ein menschlicher Partner. Nimmt das dem Menschen etwas weg? Dass wir die künstliche Kommunikation natürlicher machen, heißt doch, maschinelle Systeme dem Menschen verständlicher zu machen. Es geht um Maschinen, die *wie* – nicht *als* – Menschen erscheinen und die sich *wie* – nicht *als* – Menschen mit uns verständigen können.

[ca. 19000 Zeichen und 9 Abbildungen]

[5] Im Heinz Nixdorf MuseumsForum in Paderborn beantwortet Max seit etlichen Jahren Besuchern Fragen über die Ausstellung, über den Computer-Pionier Nixdorf und vieles mehr.