Max, unser Agent in der virtuellen Welt

Eine Maschine, die mit dem Menschen kommuniziert

Ipke Wachsmuth

Technische Fakultät
Arbeitsgruppe Wissensbasierte Systeme
(Künstliche Intelligenz)

Mit Data in Gene Roddenberrys „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, 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?

Agent Data aus Gene Roddenberrys „Star Trek Next Generation“

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 komplexen Systemen ein anthropomorpher Ansprechpartner zur Verfügung stünde, dessen Umgangsformen denen des Menschen gleichen.

„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 „Agent“ des Systems Leistungen zu vermitteln und uns dabei zu begleiten und zu assistieren?

Im Sonderforschungsbereich 360 „Situierte Künstliche Kommunikatoren“ entwickeln wir zu Forschungszwecken einen solchen Agenten. In unserem Labor – in der virtuellen Realität einer computergrafischen Großprojektion – ist er in menschenähnlicher Gestalt verkörpert. In seiner virtuellen Welt kann er bestimmte Aktionen ausführen und darüber einen Dialog mit einem menschlichen Benutzer führen. In unserem ersten Anwendungsbeispiel hilft der Agent 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. Der Agent sagt zum Beispiel: „Jetzt nimm diese Schraube und steck sie in diese Leiste“, und zeigt dabei auf die entsprechenden Bauteile, das heißt, er kann sich mit Sprache und Gestik – multimodal – äußern (Bild 2). 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.

Bild 2: Mit freundlichem Lächeln und einer Zweifingerzeigegeste kann Max zum Beispiel die Positionierung eines Bauteils andeuten und mit seiner synthetischen Stimme etwas dazu sagen.

An dieser Forschungsarbeit sind viele Mitarbeiter, Studentinnen und Studenten beteiligt. Weil unser Agent sich einerseits multimodal (mit Sprache, Gestik und auch Gesichtsmimik) äußern kann und er sich andererseits mit der Assemblierung, das heißt dem Zusammenbau virtueller Objekte auskennt, wurde er auf MAX (für „Multimodaler Assemblierungs-eXperte“) getauft.

In unserer Forschung geht es somit um Agenten mit kommunikativen Fähigkeiten. Und wir fragen uns damit im Detail, was eigentlich kommunikative Intelligenz genauer ist, ja wie sie sich – in Auszügen – so präzise beschreiben lässt, dass eine Maschine (auch unser Agent Max ist eine programmgesteuerte Software-Maschine) sie simulieren kann. Es ist dabei nicht unser Anliegen, Max verwechselbar menschenähnlich zu gestalten. Aber er soll die dem Menschen vertrauten Formen der Kommunikation an den Tag legen, uns beim Sprechen und Zuhören ansehen, sich einer natürlich wirkenden Gestik bedienen, verständnislos schauen, wenn er uns nicht versteht, warten, bis wir ausgeredet haben, bevor er selbst spricht, und so weiter.

■ 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 so genannten 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 Lexi-

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 3: 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.

kon 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 3). 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 (Bild 3) 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 – zwei Wimpernschläge lang!

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 so genannte *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. Mit der Hochgeschwindigkeit moderner Rechner lassen sich heute synthetische Stimmen durch Software, also Computerprogramme, in Echtzeit erzeugen. Grundlage ist die Erkenntnis, dass der Sprechschallstrom in Komponenten zerlegt werden kann: in die Grundfrequenz, die die Sprechmelodie 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 so genannte Diphone (Verbindungen aufeinander folgender Lautkomplexe) 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 so genannte *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.

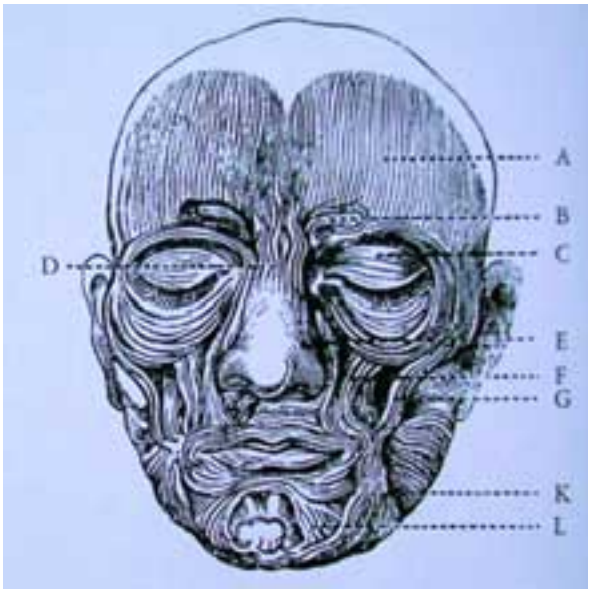

Bild 4: Im linken Bild (angefertigt von Sir Ch. Bell und entnommen Darwins Buch: Ausdruck der Gemütsbewegungen bei dem Menschen und den Tieren) sind die wichtigsten der über 40 Muskeln dargestellt, mit denen wir unserem Gesicht Ausdruck verleihen.

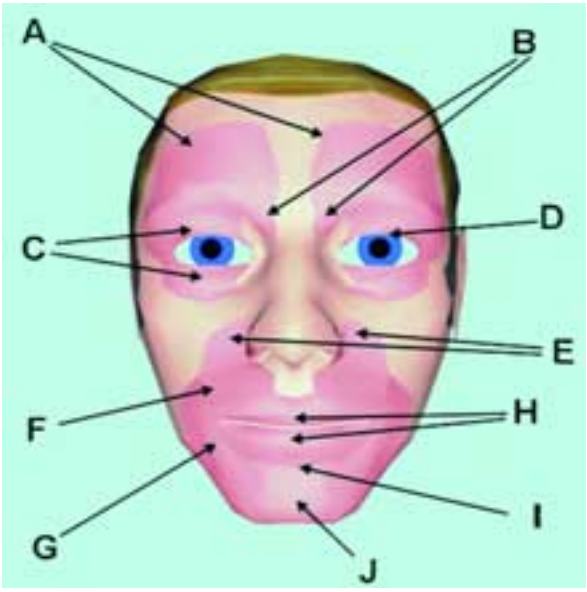

Bild 5: Das rechte Bild zeigt Gesichtspartien von Max, die mit „virtuellen Muskeln" animiert werden können.

■ 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 4). 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 der Lippenbewegung beim Sprechen.

Über 40 Muskeln verleihen unserem Gesicht Ausdruck, und die wichtigsten davon sind für Max berücksichtigt (Bild 5). Die Gesichtsoberfläche von Max kann durch simulierte Muskeleffekte mit Hilfe so genannter 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 Mimik kann Max Emotionen ausdrücken und so dem Menschen ein leicht verständliches Feedback übermitteln (Bild 6). Wenn Max zum Beispiel eine gesprochene Eingabe nicht verstanden hat oder noch an der Planung einer Äußerung „überlegt", kann er verständnislos oder nachdenklich schauen.

Auch die Sprechbewegung des Mundes entspringt dem Zusammenspiel der Gesichtsmuskeln. Für die Sprechanimation sind die so genannten *Viseme* (visuellen Phoneme) entscheidend. Sie beschreiben die Gesichtsstellung (Mund, Lippen etc.) bei der Artikulation der Phoneme. Ob *Mama, Papa* oder

Bild 6: Mit seiner Gesichtsmimik kann Max unterschiedliche Emotionen ausdrücken.

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.

■ 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. Besonders für die Gestik ist Max sehr „gelenkig“ (Bild 7): Schulter und Schlüsselbeingelenk, Ellenbogen und Handgelenk, Hände mit fünf Fingern, jeder mit drei Gelenken modelliert, ein Daumen, der zur Handfläche eingeklappt werden kann, erlauben natürliche Beweglichkeit.

Ein hierarchisches Steuerungssystem kontrolliert das kinematische Skelett von Max (Bild 8). Auf der höchsten Ebene wird die geplante Bewegung als Ziel repräsentiert (z.B. „auf das Flugzeug zeigen!“). Die Kontrolle 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 selbst ermöglicht, die zeitliche Feinplanung der Bewegungen mit seiner synthetischen Sprache („steck sie in

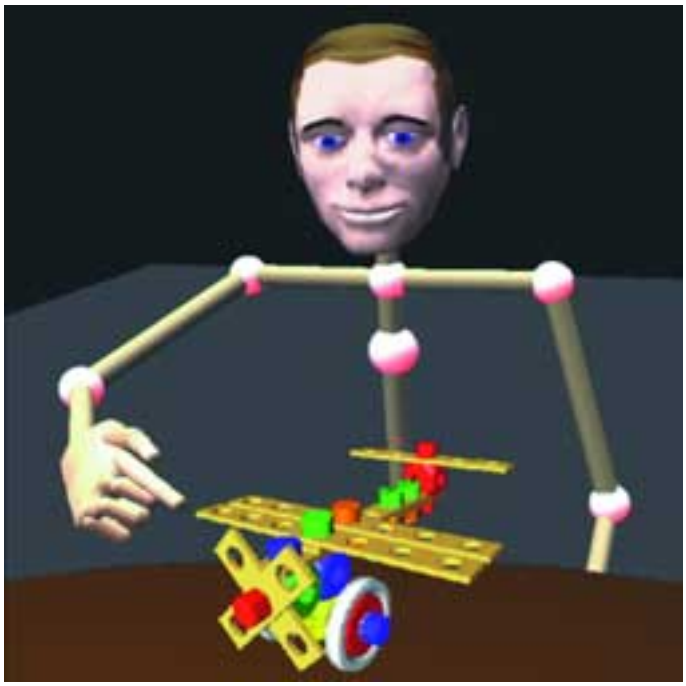

Bild 7: Das „Darunter“ von Max: Unter der Körperhülle sorgt ein Skelett aus verbundenen Segmenten, so genannten kinematischen Ketten, dafür, dass Max sich gelenkig bewegen kann.

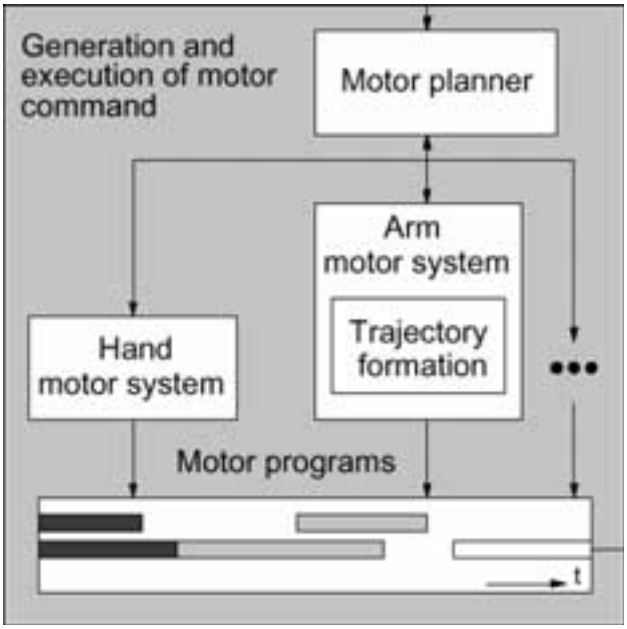

Bild 8: Ein strukturiertes Motorsystem macht Max beweglich.

diese Leiste“) abzustimmen. Mit seinem gelenkigen Körper kann er sein Sprechen mit Gesten untermalen und sogar die Gesten des vor ihm stehenden Menschen imitieren (Bild 9, nächste Seite).

Aber wie steht es mit der fühlbaren Körperlichkeit von Max? Sein computergrafisch animierter Körper ist nicht berührbar und in dieser Hinsicht körperlos. Dem Menschen, der Max gegenüber tritt, ist dennoch direkt spürbar, wenn Max bis auf „Normalabstand“ herankommt, und kommt er noch näher, verspürt man selbst den unmittelbaren Impuls zum Zurückweichen. Und genauso hat Max proxemische Sensoren, Körperfühler sozusagen, mit der er Nähe und Annäherung spüren kann. In dem Moment, wo eine menschliche Hand – mit einem Datenhandschuh

Bild 9: Max kann die Gesten seines menschlichen Gegenübers imitieren.

Bild 10: Max hat proxemische Sensoren („Körperfühler"), mit denen er Nähe und Annäherung spüren kann. Wenn eine menschliche Hand – mit einem Datenhandschuh bestückt – und Max' computeranimierte Hand sich in der virtuellen Welt treffen, knistert und funkt es.

bestückt – und Max' computeranimierte Hand sich in der virtuellen Welt treffen, funkt und knistert es (Bild 10).

■ Ausblick

Mit den Arbeiten an Max fragen wir uns, wie man bestimmte Aspekte der Kommunikation und ihr zugrunde liegende Intelligenzfähigkeiten synthetisch herstellen kann. Das erfordert nicht nur bestimmte „geistige" (kognitive) Fähigkeiten, sondern auch die Möglichkeit, sich körperlich mitzuteilen, und dies betrifft nicht nur 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 physische Gegenwart am räumlichen Ort gehört dazu, um sinnvoll „hier" und „dort", „links" und „rechts" sagen zu können.

Neben dem Aspekt der technischen Machbarkeit sind unsere Forschungsarbeiten auch mit der Erwartung verbunden, durch die Entwicklung und den Test konkreter Modelle neue Erkenntnisse über das Funktionieren menschlicher Kommunikation, dem vielleicht eindrucksvollsten Feld menschlicher Intelligenz, zu gewinnen. Wie funktioniert beispielsweise das zeitliche Zusammenspiel von Sprechen und Zeigen? Wie wird das Abwechseln im Dialog gesteuert? Die sich hiermit ergebende – wohl spannendste – Frage nach der Architektur eines körperlichen natürlichen bzw. verkörperten künstlichen „Organismus" kann nur in interdisziplinärer Zusammenarbeit erforscht werden. Mit der starken Verzahnung linguistischer, psycholinguistischer und informatischer Forschungsmethoden, die die „situierte Kommunikation" empirisch und technisch untersuchen, bietet der DFG-Sonderforschungsbereich 360 hierfür ein hervorragendes Umfeld.

Prof. Dr. Ipke Wachsmuth, geboren 1950, studierte Mathematik und Informatik in Hannover. Nach Lehr- und Forschungstätigkeiten an der Universität Osnabrück, der Northern Illinois University und bei IBM Deutschland habilitierte er sich 1989 an der Universität Osnabrück; im selben Jahr wurde er auf die Professur für Wissensbasierte Systeme (Künstliche Intelligenz) an die Universität Bielefeld berufen. Er ist Gründungsmitglied der Technischen Fakultät, Mitinitiator und stellvertretender Sprecher des Sonderforschungsbereichs 360 „Situierte Künstliche Kommunikatoren" und derzeitiger Vorsitzender der Gesellschaft für Kognitionswissenschaft. Seit Oktober ist er der neue geschäftsführende Direktor des Zentrums für interdisziplinäre Forschung der Universität Bielefeld.