Künstliche Intelligenz als Grundlage autonomer Systeme

Prof. Dr. Dr. h.c. mult. Wolfgang Wahlster
Deutsches Forschungszentrum für Künstliche Intelligenz DFKI GmbH
Saarland Infomatics Campus D3 2
D-66123 Saarbrücken
E-Mail: wahlster@dfki.de

Homepage: http://www.dfki.de/~wahlster

Zusammenfassung

Autonome Systeme, die selbstständig ein ihnen vorgegebenes Ziel erreichen können, sind nur auf der Basis von Methoden und Werkzeugen der Künstlichen Intelligenz (KI) realisierbar. Der Beitrag stellt die Entwurfsziele, die sich daraus ergebenden informatischen Herausforderungen und eine Referenzarchitektur für autonome Systeme vor, die eine Vielzahl von KI-Komponenten vom maschinellen Lernen bis zur automatischen Handlungsplanung beinhaltet. Für die breite Akzeptanz autonomer Systeme ist in anormalen Situationen ein bidirektionaler Transfer der Kontrolle zwischen autonomen Systemen und seinen menschlichen Nutzern erforderlich. Intelligente Systemkomponenten für den standardisierten und multimodalen Kontrolltransfer sind Gegenstand aktueller Forschungsprojekte. Ein vielversprechender Zukunftstrend ist die Bildung hybrider Teams aus mehreren autonomen Systemen und mehreren Menschen, die sich eine Aufgabe gemäß ihrer spezifischen Fähigkeiten aufteilen und dann gemeinsam lösen.

Abstract

Autonomous systems, which can reach a given goal on their own, can only become a reality with the help of methods and tools of Artificial Intelligence (AI). The paper presents the design goals, the resulting challenges for computer science, and a reference architecture for autonomous systems that includes a broad range of AI components from machine learning to automated action planning. For the broad acceptance of autonomous systems, a bidirectional transfer of control between autonomous systems and their human users is necessary in abnormal situations. Intelligent system components for the standardized and multimodal transfer of control are the topic of current research projects. A promising future trend are hybrid teams of several autonomous systems and humans, who distribute a task across the team members according to their specific skills and then work on it together.

Merkmale autonomer Systeme

Autonome Systeme [2, 12] können komplexe Aufgaben in einer bestimmten Anwendungsdomäne trotz variierender Zielvorgaben und Ausgangssituationen selbstständig lösen. Autonome Systeme müssen abhängig vom aktuellen Aufgabenkontext eigenständig einen Handlungsplan generieren, mit dem ein Gesamtziel, das vom Betreiber des autonomen Systems vorgegeben ist, ohne Fernsteuerung und möglichst ohne Eingriffe und Hilfe menschlicher Operateure im Rahmen der gesetzlichen und ethischen Vorgaben erreicht werden kann. Wenn einzelne Aktionen des autonomen Systems während der Planausführung scheitern, muss das System in der Lage sein, selbstständig eine Planrevision auszuführen, um durch Adaption des ursprünglichen Plans auf anderem Wege die vorgegebene Zielsetzung dennoch zu erreichen.

Eine neue Generation von autonomen Systemen ist auch in der Lage, mit anderen autonomen Systemen und/oder einer Gruppe von Menschen gemeinsam eine Aufgabe verteilt zu lösen. In unserem DFKI-System HySociaTea [5, 6] haben wir ein solches hybrides System von Robotern und menschlichen Mitarbeitern in einem Produktionsszenario erfolgreich erprobt (siehe Abb. 1). Dabei geht es um die flexible Fertigung von Unikaten für Spezialverpackungen im Rahmen von Industrie 4.0, also um den Ansatz der cyber-physischen Produktionsumgebungen für die Losgröße 1 [9].

Abb. 1: Ein Werker arbeitet im Team zusammen mit drei verschiedenen Robotern und einem Softbot

Im Rahmen der Selbstregulation sollte ein autonomes System auch über explizite Modelle der eigenen Leistungsgrenzen verfügen und bei Vorgaben oder Umgebungsbedingungen, die keine erfolgreiche autonome Zielerreichung erwarten lassen, den Systembetreiber auf diesen Umstand hinweisen (z.B. zu starke Scherwinde verhindern einen Drohnenflug, ein extrem steiler Streckenabschnitt übersteigt die maximale Steigfähigkeit eines autonomen Fahrzeuges).

Das Verhalten des autonomen Systems sollte für die Menschen, die ihm das Ziel vorgegeben haben oder ihm zusammen an einem gemeinsamen Ziel arbeiten, verständlich und auf Nachfrage vom System erklärbar sein.

Autonome Systeme müssen auch in der Lage sein, nur vage und mit geringem Detailierungsgrad spezifizierte Zielvorgaben im situativen Kontext sinnvoll zu interpretieren (z.B. Zielvorgabe an einen mobilen Roboter „Räume die Werkstatt auf“) und in Handlungspläne umzusetzen.

Eine besondere Herausforderung besteht für autonome Systeme darin, dass sie auch in ungewöhnlichen, bislang nicht bekannten Situationen sicher ihre Ziele mit den ihnen verfügbaren Ressourcen erreichen müssen.

Maschinelles Lernen ist für autonome Systeme zwingend erforderlich, um deren Verhalten über die Zeit durch Erfahrung zu optimieren und zusätzliche Fähigkeiten für neue Aufgabenstellungen zu erwerben. Im Rahmen der Selbstregulation sind auch metakognitive Fähigkeiten wie die einfach Introspektion über senso-motorische Fähigkeiten in der Architektur autonomer Systeme zu berücksichtigen.

Bei autonomen Systemen bildet die Dauer des angestrebten autonomen Verhaltens eine wichtige Dimension zur Bewertung deren Leistungsfähigkeit. Heute gibt es bereits etliche Systeme, die sich kurzzeitig autonom verhalten können, z.B. Autopiloten in Flugzeugen und Autos, die kurzfristig und in Standardsituationen Autonomie erreichen. Die Herausforderung besteht aber in der Langzeitautonomie bis hin zum Extrem der gesamten Lebensdauer des Artefakts, die auch bei völlig ungewöhnlichen Situationen rationales Verhalten realisieren. Nur dann kann das

System ein episodisches Gedächtnis entwickeln, fallbasiertes Schließen aufgrund langer Erfahrungen bei der Problemlösung nutzen und sein Wissen durch maschinelles Lernen perfektionieren.

Ein mittelfristiges Ziel ist es, autonome Systeme kontinuierlich zu betreiben, so dass diese ihr Erfahrungswissen über einen langen Zeitraum ohne Unterbrechung aufbauen können. Ein solches semantisches Gedächtnis [8] speichert alle Beobachtungen und Aktionen des Systems und kann zum maschinellen Lernen und zur Selbstoptimierung bis hin zu einer eingeschränkten Selbstreflexion genutzt werden. Besonders bei der Operation in toxischen Umgebungen oder Missionen ohne Rückkehrmöglichkeit für das autonome System ist die sichere Zugriffsmöglichkeit auf dieses als eine Art Lifelog realisierte Langzeitgedächtnis auch für den Systembetreiber von größter Bedeutung.

Autonome Systeme weisen die folgenden sieben Merkmale auf, die in verschiedenen Kombinationen intelligentes Verhalten realisieren:

1. *Entscheidungsfähigkeit*: Ein autonomes System muss selbst frei entscheiden können, wie es eine vorgegebene Zielsetzung am besten erreicht, wenn es Wahlmöglichkeiten zwischen Handlungsalternativen gibt.
2. *Selbstlernfähigkeit:* Das System kann ohne Hilfe von außen rein aufgrund von Erfahrungsdaten und Beobachtungen seine Wissensbasis ergänzen und sein Problemlösungsverhalten optimieren.
3. *Selbsterklärungsfähigkeit:* Das System kann seine Handlungsentscheidungen gegenüber einem Menschen in verständlicher, rationaler Weise erklären.
4. *Resilienz:* Das System kann auch bei Funktionsausfällen in seinen Komponenten oder trotz massiver externer Störungen seine wesentlichen Leistungen aufrechterhalten und seine Aufgaben zumindest partiell weitererfüllen.
5. *Kooperativität:* Das System kann mit anderen autonomen Systemen oder Menschen in seiner Umgebung im Team zusammenwirken, um seine Ziele zu erreichen. Es setzt auch vage artikulierte Aufträge und Änderungswünsche seines Betreibers um.
6. *Ressourcenadaption:* Das System macht seine jeweilige Vorgehensweise abhängig von den aktuell verfügbaren Ressourcen (z.B. Zeit, Energie, Werkzeuge, Teammitglieder) und erkennt frühzeitig eigene Leistungsgrenzen.
7. *Proaktivität:* Das System kann vorausschauend agieren und bei seiner Handlungsplanung zukünftig zu erwartende Ereignisse in seiner Umgebung antizipieren.

Wir arbeiten am DFKI an einem breiten Leistungsspektrum autonomer Systeme, die in Zukunft ein wichtiger Faktor bei der weiteren Informatisierung sein werden, falls sie die Akzeptanz der menschlichen Nutzer finden und man sie wirtschaftlich betreiben kann. Die Leistungen dieser Systeme kann man grob folgendermaßen kategorisieren:

- *Mobilitätsdienstleistungen* (z.B. autonome Autos, Busse, Züge, Schiffe und Flugzeuge, autonomer Lastentransport in der Logistik),
- *Arbeitsleistungen* in der industriellen oder landwirtschaftlichen Produktion und Logistik (z.B. kollaborative Team-Roboter, autonome Flotten von Erntemaschinen),
- *Explorations-, Rettungs- und Reparaturleistungen* in für den Menschen lebensgefährlichen oder toxischen Umgebungen (z.B. kontaminierte oder einsturzgefährdete Gebäude, Tiefsee, Weltraum),
- *Assistenzleistungen* im Handwerk, im Handel, in der Verwaltung, im Haushalt und in der Pflege

Gemäß unserer Mission arbeiten wir am DFKI jedoch nicht an autonomen Waffensystemen, die vor allem in den USA, in China und in Russland vorangetrieben werden.

Als Vorteile autonomer Systeme sind hauptsächlich die gleichbleibende Leistungsfähigkeit ohne Ablenkungs-, Überlastungs- und Ermüdungserscheinungen, die damit verbundene hohe Betriebssicherheit sowie die hohe zeitliche Verfügbarkeit (minimale Ausfallzeiten durch Wartung) und Zuverlässigkeit zu sehen. Erhebliche Einschränkungen ergeben sich jedoch dadurch, dass beim derzeitigen Forschungsstand autonome Systeme im Bereich der senso-motorischen, der emotionalen und sozialen Intelligenz dem Menschen in den meisten Situationen unterlegen sind und nur in Spezialgebieten im Bereich der kognitiven Intelligenz ein vergleichbares oder sogar performanteres Verhalten aufweisen. Selbstverständlich treten beim breiten Einsatz von autonomen Systemen künftig auch dringende Fragen der IT-Sicherheit und der Ethik [2] auf, die aber in den derzeit schon weit entwickelten Anwendungsszenarien wie autonomen Landmaschinen, Unterwasserrobotern für die Pipelineinspektion oder kollaborativen Montagerobotern eine geringere Rolle spielen.

Informatische Herausforderungen autonomer Systeme

Autonome Systeme verfügen über mindestens drei Komponenten, die für folgende Funktionen zuständig sind:

1. Sensorik
2. Künstliche Intelligenz
3. Aktorik

Über die Sensorik wird der Zustand der relevanten Umgebung analysiert, mit Künstlicher Intelligenz werden die Beobachtungen ausgewertet, Schlussfolgerungen gezogen, maschinelle Lernprozesse angestoßen und Handlungspläne berechnet, um die Ziele des autonomen Systems zu erreichen. Schließlich werden die Handlungspläne über die Aktorik ausgeführt und damit der Zustand der Umgebung in Richtung auf eine Zielerreichung geändert. In weiteren Iterationen wird dann der Zustand der Umgebung erneut analysiert und so lange weiter transformiert, bis schließlich alle Ziele des autonomen Systems erreicht sind (vgl. Abb. 2).

Die Bausteine autonomer Systeme können als physische Module oder als Softwaremodule realisiert werden. Auf der einen Seite haben Roboter in der Regel physische Sensoren und Aktoren, während dagegen Softbots als reine Softwareagenten ihre sensorischen Beobachtungen und ihre Aktionen rein digital in IT-Systemen oder im Internet ausführen (z.B. autonome Auktionsagenten, autonome Zerstörung von kriminellen Bot-Netzen). Es gibt auch Mischformen, in denen Softbots und Roboter zusammenwirken, wie wir dies im HySociaTea-Projekt erfolgreich erprobt haben. Wichtig für kollaborative autonome Systeme ist, dass es sich bei den Aktionen des Systems auch um Kommunikationsakte handeln kann, die z.B. dem Informationsaustausch mit anderen Menschen oder anderen autonomen Systemen dienen können. So können Sprechhandlungen und Dialogakte von einem autonomen System genauso geplant werden wie Aktionen von physischer Aktoren.

Autonome Systeme können ihre Entscheidungen nur selten auf vollständige, präzise und zuverlässige Information stützen, sondern müssen mit Situationen umgehen können, die geprägt sind von:

- *unvollständiger Information* (das autonome Fahrzeug weiß bei der Annäherung an eine grüne Baustellenampel nicht, wann deren Rotphase einsetzen wird)
- *vager Information* (soll das autonome Fahrzeug beim Verkehrsschild „80 bei Nässe“ schon bei leichtem Sommerregen wegen Aquaplaning-Gefahr das Tempo drosseln?)
- *unsicherer Information* (klassifiziert das autonome Fahrzeug per Kamerasensor im Nebel das Vorderfahrzeug als Motorrad oder per Radar als PKW mit einem defekten Rücklicht?)

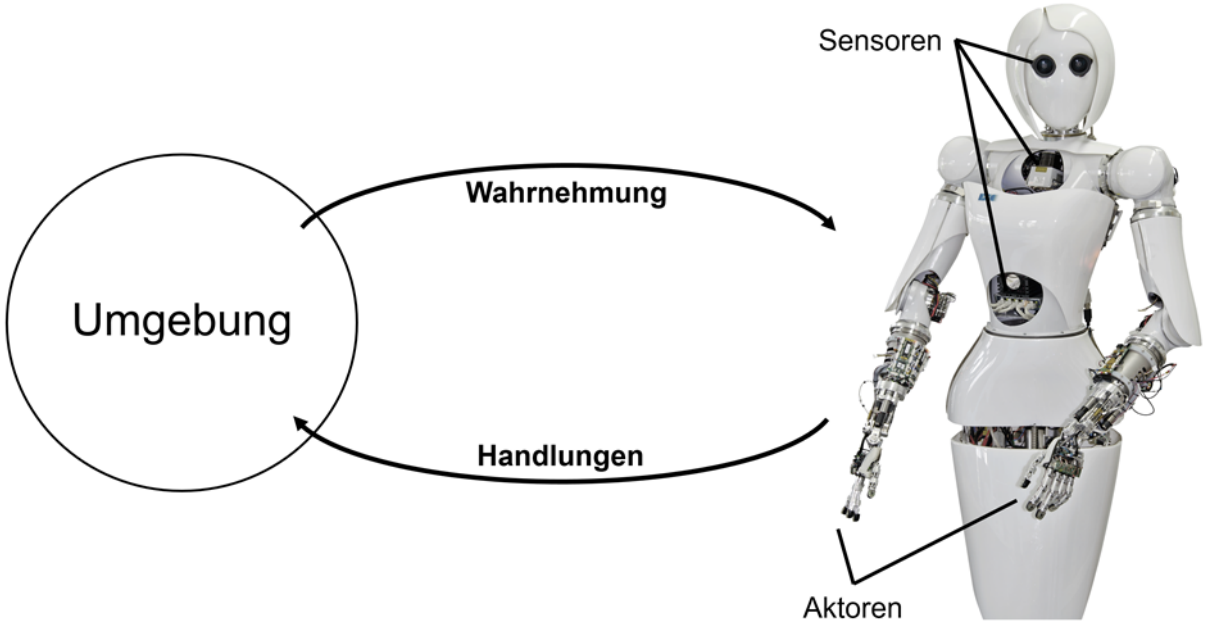

Abb. 2: Das Grundprinzip autonomer Systeme

Dabei müssen alle kognitiven Prozesse des Systems an die jeweiligen situativen Ressourcenbeschränkungen adaptiert werden, weil z.B. eine Entscheidung bis zu einer bestimmten Frist fallen muss. So stehen einem autonomen Auto für die Entscheidung, ob ein langsamer LKW noch überholt werden soll, nur wenige Sekunden zur Verfügung, wenn es von der Autobahn bereits in 4 km abbiegen muss.

Daher müssen probabilistische, possibilistische, evidenzbasierte und entscheidungstheoretische Wissensrepräsentations-, Lern- und Inferenzsysteme zum Einsatz kommen, die auch explizit mit Ressourcenbeschränkungen zeitlicher, räumlicher, kognitiver oder energetischer Art umgehen können. Hierbei hat die KI-Grundlagenforschung in der letzten Dekade enorme Fortschritte gemacht (u.a. dynamische Bayessche Netze, partiell beobachtbare Markov-Entscheidungsprozesse, Dempster-Shafer Evidenzfusion, Tiefes Lernen mit mehrschichtigen neuronalen Netzen).

Referenzarchitektur für Autonome Systeme

Nachdem in den letzten Jahren etliche autonome Systeme realisiert und praktisch erprobt wurden, ist eine Abstraktion der dabei verwendeten Architekturen in Form einer idealtypischen Referenzarchitektur sinnvoll. Unter meiner Leitung wurde in der Arbeitsgruppe 5 des Fachforums Autonome Systeme [2] im Hightech-Forum der Bundesregierung eine Referenzarchitektur für autonome Systeme erarbeitet (vgl. Abb. 25, 26 und 27 im Abschlussbericht [2]), die schrittweise in allen Schichten verfeinert werden kann (siehe Abb. 3).

Den Rahmen bilden die *Sensorik* zur Beobachtung der Umgebung und die *Aktorik* zur Änderung von Umgebungszuständen in Hinblick auf die Zielerreichung des autonomen Systems. Zusätzlich kann sich durch die *Kommunikation* mit der vernetzten Umgebung des Systems (z.B. Internet der Dinge in einer Fabrikumgebung oder einer digitalen Straßeninfrastruktur) und mit kooperierenden Menschen weitere wichtige Information für das Verhalten des autonomen Systems ergeben. Prinzipiell besteht das autonome System aus mehreren Moduln zur kognitiven Informationsverarbeitung, die durch verschiedene Mechanismen zur *Selbstregulation* gesteuert werden, sowie mehreren *Wissensbasen*, die durch maschinelles Lernen und Schlussfolgern ausgehend von einer initialen Konfiguration ständig adaptiert werden.

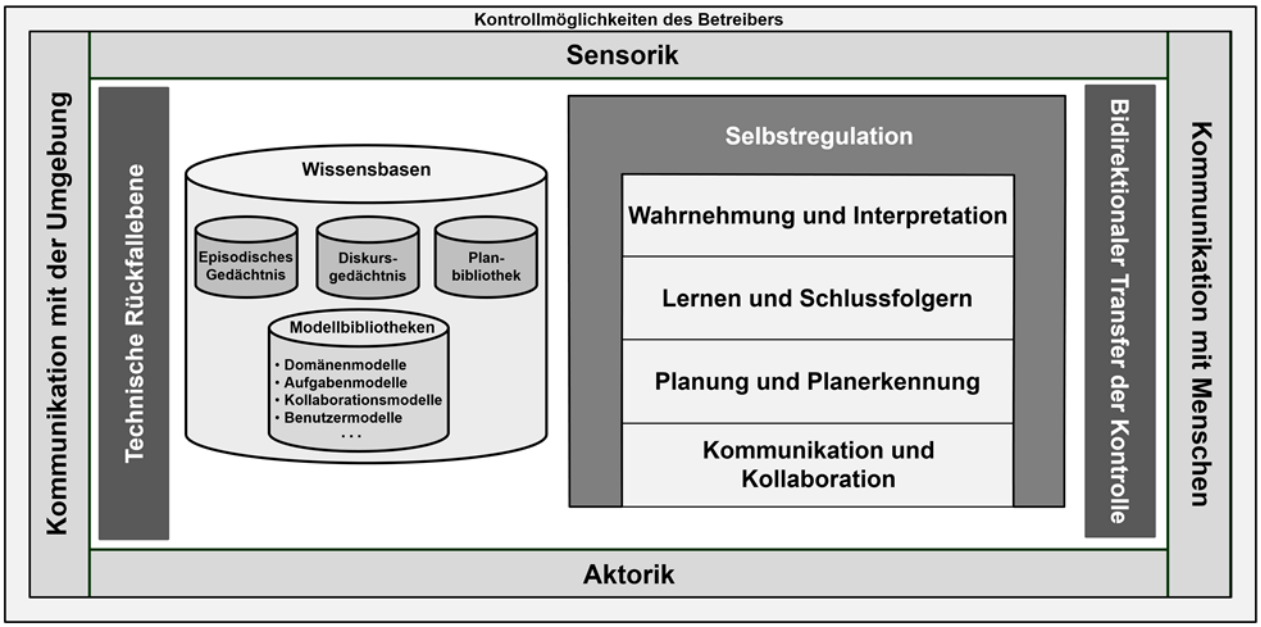

Abb. 3: Grundschema der Referenzarchitektur für autonome Systeme

Bei den Wissensbasen dient ein *episodisches Gedächtnis* als Langzeitspeicher für Ereignisse, die das autonome System unmittelbar betroffen haben, um fallbasiertes Schließen und Lernen aus Erfahrung zu ermöglichen. Im *Diskursgedächtnis* wird der gesamte Verlauf der Kommunikation des Systems mit Menschen und technischen Systemen in der Umgebung gespeichert, um jederzeit Referenzen auf Vorerwähntes und Mehrdeutigkeiten im

Kontext auflösen zu können. Eine *Planbibliothek* speichert erfolgreich ausgeführte Pläne für häufig auftretende Problemklassen, um durch Planrevision ohne Neuplanung Ziele effizienter erreichen zu können und durch Planerkennung aufgrund der Beobachtung von Aktionen anderer Agenten in der Umgebung deren Intention zu analysieren (z.B. ein anderes Fahrzeug strebt auf dem Beschleunigungsstreifen die gleiche Geschwindigkeit wie der fließende Verkehr an und setzt den Blinker links – erkanntes Ziel: Einfädeln auf Fernstraße).

Der Umfang und die Qualität der *Modellbibliotheken* sind entscheidend für die Leistungsfähigkeit autonomer Systeme. Durch das Training über Massendaten mit Verfahren des maschinellen Lernens wurden in den letzten Jahren sehr leistungsfähige Modellbibliotheken geschaffen. *Domänenmodelle* enthalten vernetzte Modelle aller relevanten Objekte, Relationen, Zustände und Ereignisse in einem Anwendungsfeld, die zu deren Erkennung über die Sensorik oder zur deren Transformation durch die Aktorik des autonomen Systems notwendig sind. In *Aufgabenmodellen* werden typische Aufgabenklassen für ein autonomes System schematisch erfasst, um eine durch den Systembetreiber neu gestellte Aufgabe rasch verstehen und einordnen zu können oder in eine Reihe bekannter Aufgaben zu dekomponieren. In *Kollaborationsmodellen* werden bewährte Schemata für das Zusammenspiel mit anderen Akteuren in der Umgebung gespeichert, um gemeinsam ein vorgegebenes Ziel schneller zu erreichen. *Benutzermodelle* sind besonders beim Einsatz autonomer Systeme als Assistenzsysteme im Dienstleistungsbereich entscheidend, da diese u.a. Annahmen über die Präferenzen, Fähigkeiten und den Wissensstand eines Systemnutzers enthalten, die eine Personalisierung der Serviceleistung durch adaptives Verhalten ermöglichen.

Die Komponenten zur *Wahrnehmung und Interpretation* sorgen durch Sensorfusion für eine Erkennung von relevanten Objekten und Situationen, die zu einer Aktivierung passender Aktorikschemata oder zu einer komplexen Handlungsplanung führen können. So muss die optische und akustische Erkennung eines nahenden Fahrzeuges mit Blaulicht und Sirene dazu führen, dass ein autonomes Fahrzeug rasch ein Ausweichmanöver zur Bildung einer Rettungsgasse plant.

Die großen Fortschritte auf dem Gebiet des *maschinellen Lernens* sind für die Realisierbarkeit autonomer Systeme von entscheidender Bedeutung. Heuristische Mustererkennungsalgorithmen müssen nicht mehr aufwändig programmiert und Wissensbasen nicht länger komplett manuell erstellt werden, sondern unter Nutzung riesiger Mengen an Trainingsdaten ermöglichen Verfahren des tiefen Lernens z.B. mit CNNs die Berechnung sehr robuster und kompakter Modelle für präzise Klassifikatoren. Diese können unter Verwendung von neuartigen Höchstleistungsrechnern auf der Basis von Graphikprozessoren wie der DXP-1 von NVIDIA sehr effizient trainiert werden. Das sogenannte Ende-zu-Ende Lernen tiefer neuronaler Netze ist für die reaktive Schicht in der Architektur eines autonomen Systems besonders attraktiv, weil hier sensorische Eingaben direkt und ohne Zwischenschritte auf Ausgaben der Aktorik umgesetzt werden. So ist es einer Forschungsgruppe bei NVIDIA beispielsweise gelungen, eine Autopilotfunktion durch CNN-Lernen so zu trainieren, dass aus einer Bildfolge einer Videokamera in einem autonomen Testfahrzeug auf einer Landstraße direkt Lenkimpulse erzeugt werden. Bei komplexeren Fragestellungen und geringerem Zeitdruck sind temporale und räumliche Inferenzverfahren in der KI inzwischen so performant, dass das automatische *Schlussfolgern* in autonomen Systemen ergänzend zu den statistischen Lernverfahren besonders in Situationen mit einem Mangel an Trainingsdaten eingesetzt werden kann.

Die *Planung* von Handlungen befasst sich als eines der ältesten Teilgebiete der KI mit der automatischen Auswahl zielgerichteter Aktionen in autonomen Systemen [3]. Der Ansatz ist modellbasiert, d.h. das System verfügt über ein formales Modell der Welt, des derzeitigen Weltzustandes, der verfügbaren Aktionen, einer Zielbedingung, sowie diverser Randbedingungen bezüglich Ressourcen, Fristen und Sicherheitsanforderungen. Anhand des Modells simuliert der Planungsalgorithmus die möglichen Entwicklungen und findet eine adäquate Handlungsstrategie. Maschinelles Lernen ermöglicht es, beobachtetes Verhalten zu assimilieren. Es kann aber keine Garantien des generierten Verhaltens geben, insbesondere bezüglich der Sicherheit. Hierfür bietet sich die Kombination mit KI-Planungsverfahren an: Der Zustandsraum möglicher Entwicklungen wird durch modellbasierte Simulation im Raum der vom Maschinellen Lernen „intuitiv“ favorisierten Handlungsoptionen rational analysiert. Der Planungsprozess macht die gelernten Handlungsoptionen überprüfbar und erklärbar.

Bei *Planerkennung* handelt es sich um die inverse Problemstellung zur Handlungsplanung. Während beim letzteren ein Ziel vorgegeben wird, zu dem eine Folge von Handlungen gefunden werden muss, geht es bei der Planerkennung um das Schließen auf ein Ziel aus den beobachteten Aktionen. Besonders in Szenarien, in denen das Verhalten anderer Agenten in einer Situation vorausgesagt werden muss, um dieses bei der Planung des autonomen Systems zu berücksichtigen (z.B. Betritt der Fußgänger gleich den Zebrastreifen oder wartet er noch auf jemand?), ist eine Planerkennung unerlässlich für eine sichere Funktionsweise.

Die *Kommunikation* des autonomen Systems mit menschlichen Nutzern wird durch KI-Methoden immer stärker den menschlichen Kommunikationsmodalitäten angepasst [7], so dass Tastatur, Maus und graphische Benutzerschnittstellen durch die multimodale Analyse von Sprache, Gestik, Mimik und Blickbewegungen abgelöst werden

und durch die automatische Generierung humanoiden Verhaltens eines virtuellen Agenten, der das autonome System als Dialogpartner verkörpert, ergänzt werden. Zur *Kollaboration* mit Menschen und anderen autonomen Systemen können KI-Techniken aus dem Bereich der Multiagentensysteme eingesetzt werden. Dabei wird aufgrund von Annahmen über das Wissen, die Fähigkeiten und Erfahrung der beteiligten Agenten eine Aufgabenverteilung ausgehandelt, welche die menschlichen und maschinellen Fähigkeiten optimal für die angestrebte verteilte Problemlösung nutzt. Dazu müssen die jeweiligen Teilziele und Pläne untereinander ausgetauscht und die Planausführung koordiniert werden.

Um das Vertrauen in die Nutzung autonomer Systeme zu stärken und die Gefährdung von Menschen in der Umgebung bei einem technischem Komplettausfall der zentralen Steuerungsfunktionen zu minimieren, muss es gemäß der Referenzarchitektur eine *technische Rückfallebene* (vgl. Abb. 3) geben, die das autonome Systeme im Notfall beispielsweise über eine redundante mechatronische Funktion oder eine funkbasierte Fernsteuerung in einen sicheren Betriebszustand versetzt und über Kommunikation mit der Umgebung eine Alarmmeldung generiert.

Häufiger wird es vorkommen, dass ein autonomes System die Grenzen seiner Fähigkeiten in anormalen Situationen erreicht und eine Kontrollübergabe an einen Menschen durchführen muss. Ein *bidirektionaler Transfer der Kontrolle* ist vorzusehen (vgl. Abb. 3), damit der Mensch nach der Überwindung eines Hindernisses für die Zielerreichung, die für das autonome System alleine nicht leistbar ist, die Kontrolle wieder vollständig an das System zurückgeben kann. Intelligente Systemkomponenten für den standardisierten und multimodalen Kontrolltransfer sind Gegenstand aktueller Forschungsprojekte.

Für Fragen der Haftung und Ethik ist es auch sinnvoll, dass *Kontrollmöglichkeiten des Betreibers* des autonomen Systems (vgl. Abb. 3) möglichst in Echtzeit verfügbar sind, um bei sich abzeichnendem Fehlverhalten Korrekturmaßnahmen einzuleiten oder gar ein NotAus zu bewirken. Dabei ist auch die komplette Rückverfolgbarkeit aller Ein- und Ausgaben und internen Verarbeitungsschritte eines autonomen Systems von Bedeutung, um bei Unfällen Schuld- und Haftungsfragen zweifelsfrei zu klären und eine gezielte Fehlervermeidung zu ermöglichen.

Transfer der Kontrolle zwischen autonomen Agenten

Autonome Systeme brauchen gelegentlich in anormalen Situationen die Hilfe von Menschen. Ein reibungsloser Kontrolltransfer wird nur möglich, wenn Interaktionsplattformen [1] realisiert werden, die explizit auch den attentionalen und kognitiven Zustand des Menschen, der die Kontrolle übernehmen soll, berücksichtigt [4, 10]. Wie Abb. 4 zeigt, muss eine sichere Kontrollübergabe vom System nach einem Vorfall zeitlich und inhaltlich geplant werden. Der *Übergabeplan* muss adaptiert werden an die Situation und eine Erklärung für den Grund der Übergabe enthalten [13]. Unbegründete, vom Menschen nicht nachvollziehbare und extrem kurzfristige Aufforderungen zu Kontrollübernahmen erschüttern das Vertrauen des menschlichen Nutzers oder Betreibers in die Autonomiefähigkeiten des Systems und verhindern letztlich die breite Akzeptanz solcher Systeme.

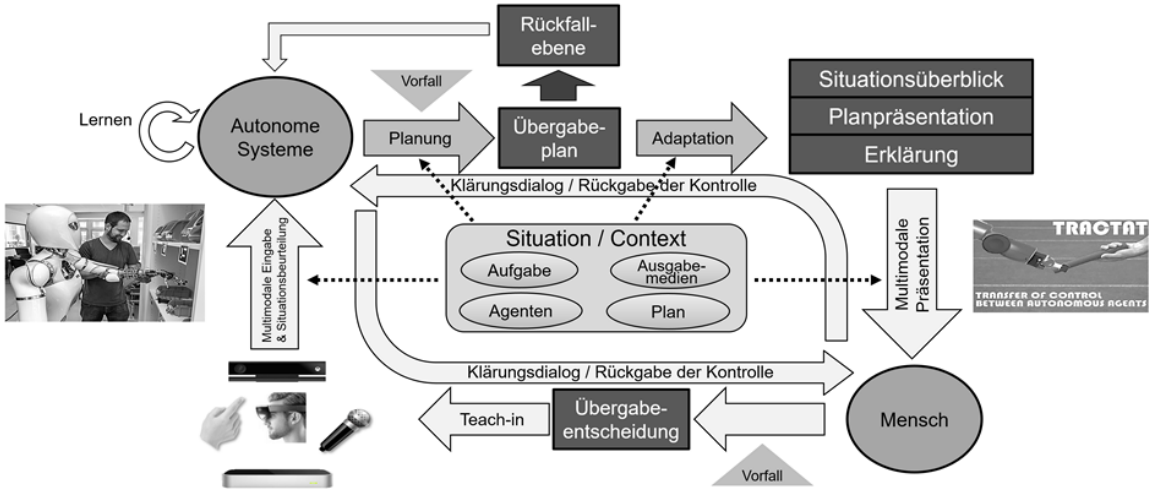

Abb. 4: Ablaufschema für den Kontrolltransfer zwischen Menschen und autonomen Systemen

Ein auf die wesentlichen Aussagen beschränkter und leicht verständlicher Überblick zur aktuellen Situation ist für einen reibungslosen Kontrolltransfer unerlässlich. Hierbei spielt die automatische Generierung multimodaler Präsentationen mit sprachlichen und visuellen Elementen eine herausragende Rolle [11]. Wenn ein autonomes Fahrzeug beispielsweise feststellt, dass es zur Zielerreichung demnächst in ein Gebiet fahren muss, für das es über keinerlei hochauflösendes Kartenmaterial verfügt, so sollte es den Passagieren rechtzeitig einen notwendigen Kontrolltransfer mitteilen. Dazu kann beispielsweise auf der digitalen Karte neben der aktuellen Position das problematische Zielgebiet angezeigt und eine sprachliche Erklärung generiert werden: *„Jemand muss die Kontrolle über das Fahrzeug in 3 Minuten übernehmen, weil wir dann ein Gebiet erreichen, für das meine Karteninformation nicht für autonomes Fahren ausreicht.“* In Klärungsdialogen können Nachfragen bezüglich der Situationsbeschreibung oder der Erklärung für den Transferwunsch beantwortet werden, wenn ausreichend Zeit bis zur Übergabefrist vorhanden ist.

Beim Einsatz in der Produktion werden gemäß dem Paradigma von Industrie 4.0 autonome Systeme auch als Helfer für Facharbeiter eingesetzt. Eine neue Generation von kollaborativen Robotern arbeitet dabei Hand-in-Hand mit dem Werker zusammen, weicht ihm intelligent aus, lernt von ihm und hört auf seine Anweisungen. Der Mensch bleibt im Team mit Robotern der Vorarbeiter, der das Team koordiniert, ihm neue Arbeitsabläufe beibringt und für die Qualität der Arbeitsresultate bürgt. Hierbei kann der Werker einen Kontrolltransfer beim autonomen System anfordern, wenn er beispielsweise dringend seine Arbeit unterbrechen muss. Eine solche Kontrollübergabe vom Menschen an ein autonomes System muss nach der Übergabeentscheidung oft von einem Teach-In für das autonome System gefolgt werden, wenn es sich um eine bisher dem System völlig unbekannte Aufgabenklasse handelt.

Insgesamt muss durch eine standardisierte und multiadaptive Kommunikation bei notwendigen Kontrollübergaben erreicht werden, dass die Selbstwirksamkeitserwartungen des Menschen erfüllt und das Gefühl des Kontrollverlustes vermieden werden.

Hybride Teams von Menschen und autonomen Systemen

Autonome Systeme können auf absehbare Zeit bei vielen Entscheidungs- und Problemfällen den Menschen nicht ersetzen, weil ihnen in völlig unvorhersehbaren Extremsituationen meist der Common Sense als eine Art Alltagsintelligenz sowie die notwendige sozial-emotionale Intelligenz fehlt. Daher ist eine effiziente und zuverlässige Kommunikation, Interaktion und Kollaboration zwischen Menschen und autonomen Systemen erforderlich, wobei sich die autonomen Systeme immer mehr dem Menschen in seinem Kommunikationsverhalten anpassen müssen als umgekehrt.

Angestrebt wird eine multimodale und dialogbasierte Kommunikation, die alle Sinne des Menschen nutzt, und die Dialogkohärenz sowie die Konversationsprinzipien zwischen menschlichen Kommunikationspartnern als Vorbild nutzt. Dabei kommen neben der gesprochenen Sprache, Gesten, Mimik, Blick- und Kopfbewegungen bis hin zur Körpersprache zum Einsatz, die in einigen Anwendungen auch durch Exoskelette und die Analyse von Gehirnsignalen ergänzt werden können. Neben der semantischen Fusion der multiplen Modalitäten und der wechselseitigen Auflösung von Mehrdeutigkeiten spielt dabei auch die kontextsensitive Fission geplanter semantischer Ausgaben im Hinblick auf die verteilte Präsentation in akustischen, visuellen oder haptisch-taktilen Informationskanälen eine entscheidende Rolle. Auch Verfahren aus dem Bereich der virtuellen, erweiterten, gemischten und dualen Realität (VR, AR, MR und DR-Technologien) werden für die menschliche Interaktion mit autonomen Systemen immer wichtiger.

Ein vielversprechender Zukunftstrend ist die Bildung hybrider Teams aus mehreren autonomen Systemen und mehreren Menschen, die sich eine Aufgabe gemäß ihrer spezifischen Fähigkeiten aufteilen und dann gemeinsam lösen. Die Kollaboration zwischen Menschen und autonomen Systemen setzt ein wechselseitiges Verständnis der jeweiligen Ziele, Pläne und Fähigkeiten aller Teammitglieder voraus.

Abb. 5 zeigt ein Szenario aus dem vom BMBF geförderten Verbundprojekt Hybr-iT, bei dem fünf Werker, die eine Hololens nutzen, mit drei Robotern im Team am DFKI zusammenarbeiten. Auf der CeBIT 2017 wurde in einer Weltpremiere die verteilte Teamarbeit an drei verschiedenen Standorten mit insgesamt acht Robotern und fünf Menschen demonstriert.

Abb. 5: Ein hybrides Team von Robotern und Werkern im BMBF-Projekt Hybr-iT

Im Verbundprojekt Hybr-iT nutzen wir aktuell grundlegende Ergebnisse aus dem DFKI-Projekt HySociaTea (siehe Abb. 1), um die soziale und multimodale Interaktion in hybriden Montageteams für den Automobil- und Flugzeugbau umzusetzen. Dabei wird eine Montageaufgabe im Team nach den verschiedenen Fähigkeiten aufgeteilt. Wichtig ist die transparente Kommunikation aller Agenten während der individuellen Planausführung, da nur so alle Teammitglieder ein gemeinsames Verständnis des gemeinsamen Arbeitsvorganges entwickeln und sich bedarfsweise wechselseitig helfen können.

Neben einem anytime KI-Planungssystem und einem digitalen Blackboard für dynamisch generierte Teilaufgaben der einzelnen Agenten wird dabei auch ein menschlicher Teamchef benötigt, der Konfliktfälle und Ressourcenengpässe auflöst und dazu notwendige Kontrolltransfers veranlasst.

Fazit

Aufgrund der Stärken der deutschen Wissenschaft im Informatikgebiet der Künstlichen Intelligenz und der wirtschaftlichen Führerschaft bei der Industrieautomatisierung, Sensorsystemen, eingebetteten Systemen und Mechatronik bestehen gute Chancen dafür, dass Deutschland zum Leitanbieter autonomer Systeme auf dem Weltmarkt wird. Daher ist es folgerichtig, nach den von mir mitgestalteten Zukunftsprojekten Industrie 4.0 und Smart Service Welt nun ein drittes Zukunftsprojekt zum Thema „Autonome Systeme“ zu starten. Die in diesem Beitrag erläuterten sehr anspruchsvollen Entwurfsziele für die nächste Generation autonomer Systeme sind nur erreichbar, wenn die Forschung zur Künstlichen Intelligenz verstärkt gefördert und in Verbundprojekten mit der Industrie die Umsetzung in Produktfunktionen massiv vorangetrieben wird. Ausgehend von der im HighTech-Forum erarbeiteten Referenzarchitektur sind die neuen BMBF-Projekte zu hybriden Teams und dem reibungslosen Transfer der Kontrolle zwischen autonomen Systemen und deren Nutzern ein wichtiger Schritt, um eine internationale Führungsposition auf diesem zukunftsträchtigen Gebiet zu erobern.

Danksagung

Das Projekt HySociaTea wurde vom Bundesministerium für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01/W14001 gefördert. Das BMBF fördert auch das Projekt Hybr-iT unter dem Förderkennzeichen 01/S16026A. Mein Dank geht auch an alle Mitglieder der AG 5 "Technologische Wegbereiter" im Fachforum Autonome Systeme des HighTech-Forums sowie meine Mitarbeiter am DFKI.

Literaturverzeichnis

1. Christ, P., Lachner, F., Hoesl, A., Butz, A.: Human-Drone-Interaction: A Case Study to Investigate the Relation between Autonomy and User Experience. In: Workshop on Assistive Computer Vision and Robotics (ACVR'16), ECCV Workshops (2): S. 238-253, 2016.

2. Fachforum Autonome Systeme im Hightech-Forum: Autonome Systeme – Chancen und Risiken für Wirtschaft, Wissenschaft und Gesellschaft. Langversion, Abschlussbericht, Berlin, April 2017.

3. Ghallab, M., Nau, D., Traverso, P.: Automated Planning and Acting. New York, USA: Cambridge University Press, August 2016.

4. Neßelrath, R.: Towards a Cognitive Load Aware Multimodal Dialogue Framework for the Automotive Domain. In: Proceedings of the 9th International Conference on Intelligent Environments (IE), Athens: IEEE, S. 266-269, 2013.

5. Schwartz, T., Feld, M., Bürckert, C., Dimitrov, S., Folz, J., Hutter, D., Hevesi, P., Kiefer, B., Krieger, H.-U., Lüth, C., Mronga, D., Pirkl, G., Röfer, T., Spieldenner, T., Wirkus, M., Zinnikus, I., Straube, S.: Hybrid Teams of Humans, Robots and Virtual Agents in a Production Setting. In: Proceedings of the International Conference on Intelligent Environments (IE'16), S. 234-237, 2016.

6. Straube, S., Schwartz, T.: Hybrid Teams in the Digital Network of the Future – Application, Architecture and Communication. Industrie 4.0 Management, 2: S. 41-45 (2016).

7. Wahlster, W.: Mit den Dingen sprechen: Autos, Roboter und Weinflaschen als Dialogpartner? In: Lengauer, Th. (ed.) Computermodelle in der Wissenschaft - zwischen Analyse, Vorhersage und Suggestion. Nova Acta Leopoldina, Band 110, Nummer 377, S. 119-141. Stuttgart: Wissenschaftliche Verlagsgesellschaft, 2011.

8. Wahlster, W.: The Semantic Product Memory: An Interactive Black Box for Smart Objects. In: Wahlster (ed.) SemProM: Foundations of Semantic Product Memories for the Internet of Things, Cognitive Technologies. S. 3-21. Heidelberg: Springer 2013

9. Wahlster, W.: Semantic Technologies for Mass Customization. In: Wahlster, W. et al (eds): Towards the Internet of Services. Heidelberg: Springer, S. 3-14, 2014.

10. Wahlster, W.: Help me if you can: Towards Multiadaptive Interaction Platforms. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, Keynote Lecture Summary. Tokyo, Japan, November 2016.

11. Wahlster, W., Müller, C.: Multimodale Dialogsysteme für Interaktive Anwendungen im Fahrzeug. In: Automatisierungstechnik, Volume 61, S. 777-783 (2013).

12. Wahlster W., Kirchner F.: Autonome Systeme: Technisch-wissenschaftliche Herausforderungen und Anwendungspotentiale. Report DFKI D-15-04. Saarbrücken: DFKI Press, 11 pp., Juli 2015.

13. Way, K. H., Pineda, L., Zilberstein, S.: Hierarchical Approach to Transfer of Control in Semi-autonomous Systems. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'16), S. 517-523. 2016.