Zum Selbstverständnis der Wissenschaft der Künstlichen Intelligenz

Künstliche Intelligenz ist eine Wissenschaft, die die Synthese und Analyse intelligenter Teilsysteme von Anwendungen untersucht (Teilsysteme werden in der Künstlichen Intelligenz Agenten genannt, siehe auch Russell und Norvig sowie Poole und Macworth). Agenten können bloße Softwaresysteme oder auch z.B. humanoide Roboter mit einer entsprechenden Hardware sein. Auch das Verhalten eines Menschen in einem Gesamtsystem kann aus technischer Sicht über einen Agenten abstrahiert werden. Das **wesentliche Element der Künstlichen Intelligenz** ist die Annahme einer **Situiertheit von Agenten in einer Umgebung**, so dass durch Wahrnehmung der Umgebung Agenten ein Modell über ihre Umgebung (und ggf. über andere Agenten in der Umgebung) aufbauen bzw. lernen können. **Agenten handeln rational**, indem sie aufgrund ihres jeweiligen (ggf. anzupassenden) Ziels aus ihrer lokalen Sicht heraus optimale Handlungen berechnen, die bei Ausführung einen Einfluss auf die Umgebung haben, deren Zustand dann wieder vom Agenten erfasst werden muss, so dass er weiterhin optimal handeln kann (Interaktion mit der Umgebung). Ziele können hierarchisch strukturiert und auch priorisiert werden.

Die Erfassung der Umgebung durch einen Agenten kann auch eine Rückkopplung bzgl. der Wahl von früheren Handlungen des Agenten umfassen und den Agenten zur Neufassung seines internen Zustandes und seiner zukünftigen Handlungsplanung veranlassen (Reinforcement). Die sequentielle Entscheidungstheorie löst das Problem der Bestimmung der besten Aktion, geht aber davon aus, dass sowohl die Verteilung von Ereignissen in der Umgebung bekannt ist als auch die möglichen Aktionen a priori bekannt sind, was in der Regel in Realweltsystemen eine nicht geeignete Annahme ist. Ein Agent muss die Verteilung(en) zur Beschreibung der Umgebung erst im Laufe der Zeit schätzen, und er muss auch mögliche neue Handlungen (Aktionen) entdecken. Bei der Handlungsplanung ist **jeder Agent in seiner Rationalität den Gesetzen der Komplexitätstheorie unterworfen**, d.h. die Rationalität von Agenten ist nicht nur unter starken Zeitbeschränkungen recht stark begrenzt. Optimale Handlungen können nicht unbedingt mit den zur Verfügung stehenden Berechnungsressourcen (Zeit und Speicher) bestimmt werden. Da optimale Handlungen nur approximiert berechnet werden können (bounded optimality), werden die theoretisch erreichbaren Ziele eines Agenten praktisch auch nicht immer oder nur verspätet erreicht.

Die beschriebene Form der Konstruktion von Systemen mit situierten Agenten, die in einer Umgebung aus ihrer lokalen Sicht optimal agieren, also rational handeln, bildet die Basis für einen **künstlichen, technischen Intelligenzbegriff**. Ein Agent ist intelligent, wenn er unter Berücksichtigung der begrenzten Rationalität durch situierte Interaktion mit seiner Umgebung seine Ziele in unerwartet kurzer Zeit erreicht. Manche fordern sogar, dass Ziele in wechselnden Kontexten (oder Umgebungen) erreicht werden, und zwar auf der Basis von abstrakten Beschreibungen der "Regeln" (z.B. Regeln eines Spiels, wie z.B. Schach), die eine Umgebung einem Agenten bereitstellt (Artificial General Intelligence). Für Spiele wie Go, Schach oder Shigo wurde gezeigt, dass durch Reinforcement-Lernen ein System aus einer Beschreibung der Spielregeln generiert werden kann, das menschliche Spieler ohne Probleme schlagen kann. Es ist hervorzuheben, dass es sich bei den genannten Spielen um Spezialprobleme handelt. Das generierte System wird gegen Menschen zwar fast immer gewinnen, kann aber nicht einem Anfänger erklären, wie man ein guter Spieler wird. Sehr generell ist diese Art der Intelligenz nicht. Aus den speziellen gelösten Aufgaben lassen sich aber durchaus sehr generelle Techniken für die Analyse und Synthese von intelligenten Systemen herleiten.

Wir betonen, dass die Zubilligung der Intelligenz von der Erwartungshaltung des Menschen abhängt, der das Verhalten maschineller Agenten interpretiert. Wenn z.B. ein Schachagent (oder Schachprogramm) das Ziel hat, 1000 Spiele zu gewinnen, und man nicht erwartet, dass er dieses Ziel schon nach 1000 Partien erreicht, mag man ein Schachprogramm als intelligent bezeichnen. Bei einem Taschenrechner würde man erwarten, dass er 1000 von 1000 Additionen richtig berechnet, und daher wird das Taschenrechnerprogramm auch kaum als intelligent bezeichnet. Wenn man erwartet, dass ein Schachprogramm immer gewinnt, wird die Tendenz, es als intelligent zu bezeichnen, stark

abnehmen. **Die Erwartungshaltung kann sich durchaus auch mit der Zeit ändern und vormals als intelligent bezeichnete Systeme erscheinen den Nutzern mit der Zeit profan**. Wenn man den Intelligenzbegriff in diesem Rahmen ernst nimmt, kann man auch verschiedene Formen der Intelligenz unterscheiden und stößt dabei zu einem weiteren Kernthema der Künstlichen Intelligenz vor, das ohne den Agenten-Begriff verborgen geblieben wäre.

Bei großen Sprachmodellen mit Diaglogkomponente (wie z.B. GPT-4) erwarten die meisten Menschen die gebotenen Leistungen bezüglich relativ einfacher Eingabeaufforderungen (Prompts) derzeit nicht. **GPT-ähnliche Dialogsysteme werden als intelligent betrachtet**. Interessanterweise sind die Nutzer sogar bereit, mit Mühen Aufforderungen so zu erzeugen, dass eine gewünschte Funktionalität erreicht wird (Prompt Engineering), und sie sind immer noch beeindruckt. Für manche ist es eine Art Sport, blockierte Funktionalitäten durch geschickt gewählte Aufforderungen (und gegebenenfalls Beispiele) den Systemen dennoch zu entlocken. AgentGPT werden auch Agenten für jeder mann angeboten, wobei Agenten automatisch komplexe Aufgabenbeschreibungen zerlegen, so dass Unterprompts automatisch generiert werden. Die Nutzer können die den Agenten gegebenen Aufgabenbeschreibungen interaktiv verfeinern. Bestimmte Gruppen werden aber nicht müde, über Nachteile bzw. Halluzinationen oder gar Fehler zu berichten. **Doch, findet in der Forschung eine breite systematische Diskussion statt über das, was GPT eigentlich weiß, statt?** Leider nicht. Eine sehr löbliche Ausnahme ist das Essay meines Kollegen PD Dr. Özgür Özcep über wirklich **fundamentale philosophische Aspekte der KI** auf das ich gerne **hiermit** verweise. Nur, wenn wir besser verstehen, was hinter den angesprochenen Systeme, wie z.B. GPT-4, steht, wird der durch die Systeme erzielte bedeutsame Nutzen für Gesellschaft, Wirtschaft und Wissenschaft erst deutlich, und erst dann werden wir die Systeme verantwortungsvoll einsetzen können.

Agenten in einem sozialen Wirkkontext

Der langfristige Forschungsgegenstand der KI besteht also darin, die Entwicklung flexibler Systeme (Agenten) zu ermöglichen, die aufgrund einer vorgegebenen Aufgabenbeschreibung und des repräsentierten Weltwissens in der Lage sind, dynamisch gestellte Aufgabenbeschreibungen richtig zu interpretieren sowie Handlungen selbsttätig zu bestimmen und in der realen Welt auszuführen, um die gestellten Aufgaben bestmöglich zu lösen. In der Natur der Sache liegt, dass Aufgabenbeschreibungen in einem sozialen Wirkungskontext von den Agenten selbst durch Rückkopplungen bei der Ausführung von Handlungen zum Wohle aller Beteiligten im Kontext sinnvoll weiterentwickelt werden müssen. Die Forschung in der Wissenschaft der Künstlichen Intelligenz schließt die Analyse der sozialen Wirkungskontexte mit u.U. sehr vielen Agenten und deren Wechselwirkungen untereinander und den Nutzen für mit Menschen in Hinblick auf durch die Gesellschaft vorgegebene Ziele, Einschränkungen und Regularien ein. Aufgrund von generellen Komplexitätsüberlegungen, die unabhängig von ontologischen und epistemologischen Annahmen greifen, kann ein Agent bekanntermaßen nur in begrenzter Weise optimal handeln (bounded optimality). In den naturgegebenen Grenzen so gut wie möglich zu handeln, definiert den in der KI zu verwendenden Intelligenzbegriff, der aber ohne Bezug auf einen Agenten sinnfrei ist. Der verwendete Intelligenzbegriff „Optimales Handeln unter (starken) Ressourcenbeschränkungen unter ständiger Anpassung der Aufgabenbeschreibungen, die das Handeln steuern, an die Präferenzen aller Menschen in einem sozialen Mechanismus" definiert die Forschungsziele der KI unter Berücksichtigung vielfältiger ontologischer und epistemologischer Annahmen und die Betrachtung des sozialen Wechselwirkungskontextes, in dem Menschen mit intelligenten, nicht autonomen Agenten interagieren.

Missverständnisse

Ohne Berücksichtigung der Situiertheit, der begrenzten Rationalität und des sozialen Wirkungskontexts ist der Begriff der Intelligenz in jedem Fall schlecht definiert. Es ist wichtig zu verstehen, dass sich die **Intelligenz von Agenten nur "von außen" zeigt**. Soll heißen: Die Intelligenz des Agenten, die wir Menschen in das Verhalten von Agenten als Software- oder Hardwaresysteme hineininterpretieren, wird durchaus mit klassischen Techniken der Informatik erbracht (auch bei GPT-4). Eine Repräsentations- oder Berechnungsmethode als KI-Methode zu klassifizieren, führt nicht zu tiefen Einsichten. Von "intelligenten Algorithmen" oder "intelligenten Daten" oder gar der "Integration einer KI in ein System" bzw. "Erzeugung einer KI" kann keine Rede sein. Vermutlich wäre der von Bibel

vorgeschlagene Begriff Intellektik passender für das Wissenschaftsgebiet der Künstlichen Intelligenz gewesen (Bibel 2017, Bibel und Furbach 1992). Auf Basis der Tatsache, dass viele ein intelligentes System oder auch nur ein System, dessen Entwicklung entscheidend auf Lernen basiert, leider mittlerweile als "Intelligenz" oder "Künstliche Intelligenz" bezeichnen, ist es meines Erachtens nunmehr dringend an der Zeit, für das Fachgebiet der Künstlichen Intelligenz einen angemessenen Namen zu verwenden: Intellektik (engl. Intellectics).

Wichtig ist es auch zu verstehen, dass in der Künstlichen Intelligenz der Begriff Lernen sich darauf bezieht, dass ein Agent zur Laufzeit ein Modell seiner Umgebung aufbaut, also Lernalgorithmen auf Daten von seinen "Sensoren" anwendet, um seine Ziele besser zu erreichen. Obwohl schon lange zentraler Gedanke der KI, wird in jüngster Zeit der Begriff "Adaptive AI" zur Betonung dieses wichtigen Prinzips verwendet. Der Einsatz von Lerntechniken zur Entwicklungszeit von Agenten zur Generierung eines initialen Modells bzw. Verhaltens ist möglich, trifft aber nicht den Kern der Künstlichen Intelligenz, sondern kann als Data Science zur Systementwicklung verstanden werden. Lernen unter Verwendung von Vorwissen ist wesentlich zur Reduktion der notwendigen Trainingsdaten, denn für die allermeisten Lernprobleme können insbesondere zur Laufzeit kaum ausreichend Trainingsdaten bereitgestellt werden - eine Tatsache, die heute kaum beachtet wird. Mit Vorwissen wird es möglich, den Hypothesenraum für mögliche Lernergebnisse automatisch zu strukturieren, was eine notwendige Voraussetzung für effektives Handeln eines Agenten darstellt.

Ein weitverbreiteter Irrtum ist es anzunehmen, dass in der KI das Lernen (zur Laufzeit) nur in Bezug auf die verwendeten Modelle relevant ist (z.B. für Funktionalitäten wie die Klassifikation von Bildern, Erkennung von Objekten, usw.). Genauso wichtig ist es, dass Agenten ihre Eingabedaten (Perzepte) analysieren, um bestimmte Strategien zu entwickeln, mit den sehr begrenzten Zeit- und Platzressourcen für die Bestimmung der jeweils besten nächsten Aktion effizient umzugehen. Nehmen wir an, ein Agent löst hierzu verschiedene Probleme, deren Klasse sich über die Historie der Eingaben definiert. Mit Techniken wie AutoMLoder Automatic Algorithm Configuration lassen sich (z.B. für SAT-Probleme) nach kurzer Trainingsphase sehr effiziente Strategien zur Bearbeitung von Eingaben automatisch generieren bzw. lernen. Es gibt Einsatzszenarien, in den Agenten zunächst einfache Instanzen einer Problemklasse bearbeiten, um geeignete Strategien für die Lösung von Problemen aus der jeweiligen Problemklasse zu identifizieren. Erst danach können harte Instanzen einer Problemklasse angegangen werden. Ohne "Anlernen" durch einfache Probleminstanzen könnten harte Instanzen vielleicht bei gegebenen Zeitressourcen gar nicht bearbeitet werden, ohne auf ein Timeout zu laufen. Strategielernen kann wesentlich bedeutsamer in praktischen Anwendungen sein als Funktionalitätslernen.

Anwendungen

Nicht in jeder Anwendung sind Agenten in der Software explizit modelliert. In einigen Anwendungen sind auch nur Teile des oben dargestellten Zyklus "Perzipieren, Schlussfolgern, Handeln" erkennbar. Es hat sich herausgestellt, dass die Perspektive der KI, die sich in der Agenten-Metapher manifestiert, verallgemeinerbare Prinzipien zur System-Entwicklung hervorgebracht hat, die sich für die Standard-Software-Entwicklung als nützlich herausgestellt haben. Insbesondere die entwickelten Techniken zur Programmierung von Systemen durch Lerntechniken erfreuen sich großer Beliebtheit. In aktuellen Anwendungen spiegelt sich meist nur ein Teil des Grundgedankens der Künstlichen Intelligenz explizit wider. So sind z.B. Techniken der signalnahen Sprachverarbeitung (Perzipieren) nützlich, um Chatbots zu entwickeln, die vielleicht kaum Anwendungswissen in expliziter Form vorliegen haben (um zielfokussiert zu schlussfolgern) und kaum Handlungen optimal bestimmen, aber dennoch in Web-Anwendungen nützliche Dienste erbringen. Audiologische Signalverarbeitung kann in Robotern zum Einsatz kommen, aber auch in technischen Lösungen für neuartige Hörhilfen. Techniken der Auswertung von radiologischen Bilddaten (Perzipieren) zusammen mit implizit in die Software integriertem Anwendungswissen ermöglichen beispielsweise das Hervorheben kritischer Bereiche im Bild, so dass eine neue Anwendung zur optimierten Unterstützung bei der Erstellung von Diagnosen aufgebaut werden kann. Schlussfolgern und Handeln im klassischen Sinne werden die genannten Systeme (außer den Robotern) nicht. In der Software bzw. auch Hardware sind nur Teile des Zyklus manifestiert. Man kann aber dennoch von einer KI-Anwendung sprechen, also einer Anwendung von in der Künstlichen Intelligenz entwickelten Datenverarbeitunstechniken.

Es haben neue Entwicklungen auch nur für Teile des Agenten-Zyklus "Perzipieren, Schlussfolgern, Handeln" die Entwicklung "klassischer" Systeme deutlich vorangebracht, nicht zuletzt auch durch enorme Verbesserungen in der Hardware-Entwicklung. Nichtsdestotrotz ist es sehr vorteilhaft, die Agenten-Metapher zu beleuchten, um zu verstehen, was die Wissenschaft der Künstlichen Intelligenz insgesamt bewegt.

Ängste vor der Wissenschaft der Künstlichen Intelligenz?

Man beachte, dass technische Agenten, die durch Interaktion in einer speziellen Umgebung eine große Performanz erzielt haben, kopierbar sind (im Gegensatz zu Menschen) und in einer anderen, ähnlichen Umgebung weiter lernen können und noch besser werden. Die **Kompetenzerhöhung von Agenten mittels Lernen durch Interaktion mit einer Umgebung** ist durchaus ein sehr interessantes **Prinzip der Systemerstellung**. Ob ein durch Lerntechniken erstelltes System in einer bestimmten Umgebung ohne Weiterlernen betrieben werden kann, lässt sich nicht allgemein beantworten. Die Realität entwickelt sich allerdings weiter und ohne laufende Anpassung ist die Verwendbarkeit von durch Lerntechniken gewonnenen Systemen sicherlich begrenzt.

Ängste in der Bevölkerung vor den Ergebnissen der Wissenschaft der Künstlichen Intelligenz könnten sich aus einer Überschätzung der Leistungsfähigkeit der oben genannten Form des Kettenbrief-artigen Aufschaukelns der Intelligenz von Agenten durchaus ergeben, insbesondere wenn man die immanente Begrenztheit der Rationalität von Agenten in Betracht zieht. Der Punkt, an dem Systeme entstehen, die intelligenter sind als Menschen (wie auch immer das gemessen werden soll), wird manchmal als Singularität bezeichnet. Ein Aufschaukeln von Intelligenz, insbesondere auch durch automatische Kombination von selbstlernenden Teilsystemen, ist heute keine Realität, gerade auch weil die Rationalität von Agenten aus prinzipiellen Überlegungen heraus begrenzt sein muss. Das spontane Verwenden eines neu eingetragenen Kontaktes in einem Sprachassistenten, also eine Funktionalität, die wir heute schon vorfinden, ist aus dieser Perspektive wenig beeindruckend. Die korrekte Aussprache von neuen Namen ist aufgrund von großen Mengen an Sprachdaten heutzutage auch kein Hexenwerk.

Es ist allerdings deutlich zu machen, dass massenhafte Einflussnahme auf Menschen durch intelligente Agenten heute schon Realität ist und ohne Gegenmaßnahmen immer subtiler wird und auch noch deutlich zunehmen wird, so dass Demokratien in Gefahr geraten können. Man darf auch **Intelligenz nicht mit Menschen-kompatiblem Verhalten verwechseln**, denn das (unerwartet) schnelle Erreichen von Zielen heißt nicht, dass für Menschen sinnvolle Ziele auch in aus Menschensicht "vernünftiger" Weise durch einen Agenten eigenständig aufgesetzt werden. Sollte die menschliche Kontrolle über das Aufsetzen von Zielen in automatischen Systemen verloren gehen, ist wenig Gutes zu befürchten. Die Wissenschaft der Künstlichen Intelligenz muss ihre Forschung darauf konzentrieren, nicht nur die Erstellung von in einem gewissen Sinne korrekten aber unkontrollierbaren Systemen zu ermöglichen, sondern beweisbar Menschen-kompatible Systeme zu erzeugen (siehe Russell 2019). Es wird berechtigt argumentiert, dass Agenten die Ziele bzw. Präferenzen von Menschen bzw. der Menschheit berücksichtigen müssen (und nicht ihre eigenen), damit für beteiligte Menschen beweisbar ein positiver Nutzen erzielt werden kann. Agenten sind zunächst unsicher über die Präferenzen von Menschen und reduzieren diese Unsicherheit durch situierte Interaktion. Eine permanente Unsicherheit über die Präferenzen von Menschen sollte ein Agent beibehalten, um frühere Fehler zu antizipieren und Anpassungen an wechselnde menschliche Präferenzen zu ermöglichen (Human-Compatible AI).

Ethik der KI-Wissenschaft

Geisteswissenschaftliche Forschung und insbesondere Ethik spielt eine bedeutende Rolle für die Künstliche Intelligenz. Statt z.B. Ethik nur so zu begreifen, dass man Konstrukteuren von Systemen natürlichsprachlich formulierte moralische Leitsätze vorgibt, bietet die KI die Möglichkeit, mit formaler Ethik intelligente Agenten in die Lage zu versetzen, über Modelle des moralischen Handelns die Präferenzen von Menschen mit weniger Aufwand zu schätzen und gegeneinander abzuwägen, was einen enormen Gewinn bezüglich zu kurz greifender ethischer Ansätze nur für die Konstruktion von Systemen darstellt und wichtige Perspektiven eröffnet (vgl. Begriffe wie z.B. Value Alignment, siehe Russell 2019). Nicht nur die Konstrukteure von KI-Systemen müssen nach ethischen Leitlinien handeln, sondern die konstruierten Systeme (Agenten) selbst. Letzteres wird nur gelingen, wenn bei der

Entwicklung des Mechanismus eines Systems (Umgebung und Zusammenspiel der Agenten) berücksichtigt wird, dass das Handeln nach ethischen Leitlinien eine gewinnbringende Strategie für jeden einzelnen Agenten darstellt (Trustworthy AI)! Einen solchen Mechanismus zu gestalten, ist nicht einfach. Die Grundlagen hierfür untersucht das Forschungsgebiet Algorithmic Mechanism Design. Der Ansatz, bei der Konstruktion von Systemen Ethikleitlinen zu befolgen, ist nicht falsch, wird aber den Notwendigkeiten, die sich aus dem Potential der KI ergeben, nicht vollumfänglich gerecht. Es muss erreicht werden, dass rationales Handeln von Agenten bestimmten Ethikleitlinien genügt. Andere Aspekte von Trustworthy AI betreffen die Privatheit: Z.B. sollte vermieden werden, dass ermittelt werden kann, ob ein bestimmtes Datum (oder die Daten einer bestimmten Person) Bestandteil der Trainingsdaten oder zumindest Basis für die Beantwortung einer Anfrage war oder nicht. Techniken wie z.B. Differential Privacy sind für Agentensysteme hochrelevant und sind mit dem Ethikkonzept technisch zu kombinieren.

Die angesprochenen Ziele eines Agenten können unsicher sein, spiegeln aber die den Agenten übertragenen Aufgaben wider. Die Analyse eines Systems von Agenten (also eines sog. Mechanismus) und die Prüfung, ob eine gewünschte Funktionalität durch ein solches System von Agenten auch wirklich realisiert wird, stellt eine zentrale Fragestellung in der Wissenschaft der Künstlichen Intelligenz dar. Ohne eine Lösung dieser Frage der formalen Korrektheit eines Mechanismus im konkreten Fall sind intelligente Systeme in bestimmen Anwendungsbereichen schlichtweg nicht einsetzbar.