POPKOLUMNE

Man fürchtet sich instinktiv immer sofort ein wenig, wenn alte Postpunk-Helden noch einmal ein Album veröffentlichen. Die zornige Zackigkeit und lassige Kühle ihrer alten Songs ist einfach zu gut gealtert und bis heute stilprägend, warum also mit neuem Material nicht an altes herankommen? Von der 1977 in Leeds gegründeten Postpunk-Band **Gang of Four** gibt es zum Beispiel den Überhit „Damaged Goods“, funky, rotzig, bis heute brillant. Der Platz in der Popgeschichte ist der Band damit sicher. Das neue, dreizehnte Album der Band, „Happy Now“, fügt dem erwartungsgemäß nichts hinzu, was einem fehlen würde, wenn man nur „Damaged Goods“ gehört hat. Und dann ist es doch so unpeinlich und zackig, dass es eine schöne Erinnerung ist und ein guter Anlass, die frühen Alben und insbesondere das Debüt „Entertainment!“ (natürlich nur echt mit dem Ausrufezeichen!)

wieder einmal zu hören. Schon erstaunlich, wie gut sich zackiger Zorn doch immer wieder aufführen lässt, wenn er so funky ist.

Die beiden Kalifornier Jonathan Rado und Sam France machen als **Foxygen** auch auf ihrem sechsten Studio-Album „Seeing Other People“ aus unverdienten Alltagsdepressionen aller Art feinsten post-postmodernen Als-ob-Pop: „Face the facts, I’m never gonna dance like James Brown / I’m never gonna be black“. Gar nicht so weit verbreitete

Kunst im Indiepop: Sich von den Beschränkungen seiner Existenz gerade so weit runterziehen lassen, dass man noch mitwippen kann.

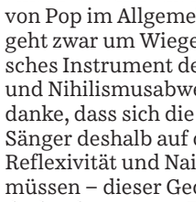
Im Hauptberuf ist Clifford Ian Simpson alias **Kevin Abstract** Mitglied der famosen amerikanischen Indie-Hip-Hop-Boygroup Brockhampton, der seit ein paar Jahren ja der schöne Stunt gelingt, den allergrößten Emo-Kitsch supercool klingen zu lassen. Auf seiner neuen, entspannt dahin schaukelnden Solo-Single „Baby Boy“ singt er jetzt im Falsett „When I close my eyes I think about you all the time“. Alles klar: Verlieren kann

man eigentlich nur noch, wenn man die Angst vor den großen Gefühle so kleinkriegt, dass kein Song mehr daraus werden kann.

Passend dazu sei der im neuen Buch „Lyrik/Lyrics – Songtexte als Gegenstand der Literaturwissenschaft“ erschiene Aufsatz „Geistersprache im Kinderzimmer“ des Literaturwissenschaftlers Kai Sina empfohlen, der doch eigentlich auch ein sehr gute (und an so einem Ort eher unerwartete) Definition von Pop im Allgemeinen vorschlägt. Es geht zwar um Wiegenlieder als „Poetisches Instrument der Materialismus- und Nihilismusabwehr“, aber der Gedanke, dass sich die Sängerinnen und Sänger deshalb auf das „Dilemma von Reflexivität und Naivität“ einlassen müssen – dieser Gedanke umschreibt doch sehr gut, was Popmusik ausmacht.

Bevor **Prince** seine Memoiren, die den Titel „The Beautiful Ones“ haben sollten, diktieren konnte, starb er im April 2016 an einer Überdosis des Schmerzmittels Fentanyl. Im Oktober soll nun ein Buch unter demselben Titel erscheinen, dass an Leben und Werk des Meisters erinnert. Vorerst soll der schöne Kommentar zum Buch reichen, den Prince bei einem Minikonzert anlässlich der Ankündigung seiner Memoiren einen Monat vor seinem Tod sagte: „Die netten Leute von Random House haben mir ein Angebot gemacht, dass ich nicht ablehnen konnte. Ihr lest doch alle noch Bücher, oder?“

Es gibt eine neue Single der letzten großen alten Retrosoul-Königin **Mavis Staples**. „Anytime“ klingt unwiderstehlich minimalistisch. Staples, Jahrgang 1939, zeigt da einmal mehr wie lässig man im Pop alt sein kann, wenn man die Kunst beherrscht, die Nostalgie mit



der Neugier tanzen zu lassen, und zwar am besten dorthin, wo man noch nicht war: „Give me a one-way Ticket / to somewhere I’ve never been.“

Blieben die **Albumcharts** der Woche. Ganz oben steht das Album „Trip“ des 19-jährigen deutschen Popsängers Mike Singer, der über die Sozialen Medien als eine Art deutscher Justin Bieber bekannt wurde und jetzt über R’n’B- und Trap-Tonspuren von der Stange leidlich lasziv die Planstelle Teenieschwarm ausfüllt. Interessanter ist diese Woche ein Blick auf die Albentitel der Top-Ten. Neben Meisterleistungen der Namensgebung wie „Alles oder Dich“ (Roland Kaiser), „Only Love, L“, „Irgendwie anders“ (Wincent Weiss) oder „Auf der Suche nach der Schnapsinsel“ (Tote Hosen) steht da immerhin Billie Eilishs Melancholia-Frage „When We All Fall Asleep, Where Do We Go?“ schön quer. JENS-CHRISTIAN RABE

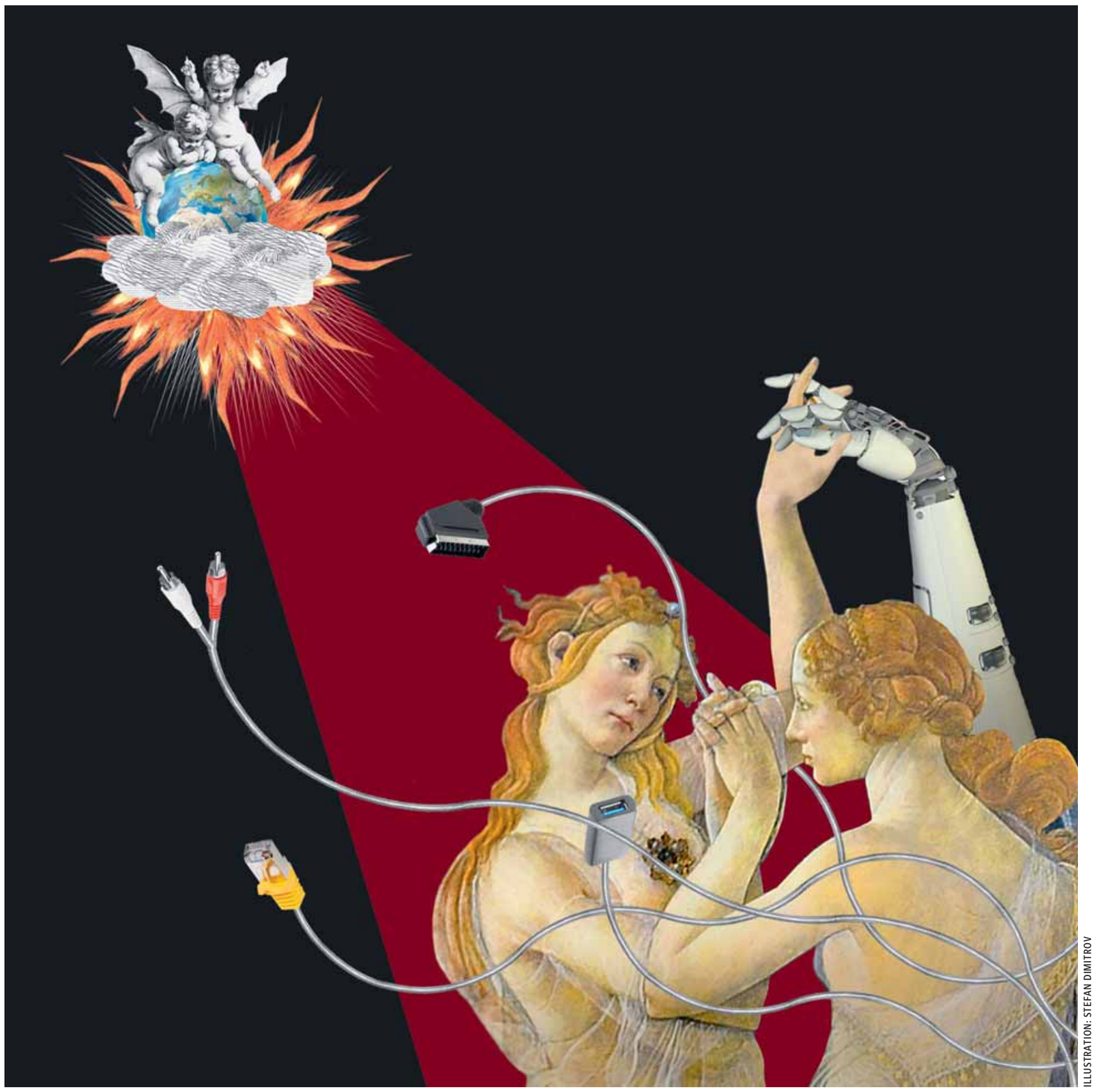


ILLUSTRATION: STEFAN DIMITROV

Der unberechenbare Mensch

Je leistungsfähiger eine künstliche Intelligenz ist, desto schwieriger wird es, sie für den Umgang mit ihrer Umwelt zu programmieren. Denn die Realität funktioniert nicht nach klaren Regeln. Das kann katastrophale Folgen haben. *Von Anca Dragan*

Künstliche Intelligenz (KI) ist ein epochaler Technologiesprung, der die Menschheit vor Fragen stellt, die keine Disziplin alleine beantworten kann. John Brockman, Agent für Wissenschaftsliteratur und Gründer des Debattenforums Edge.org, hat das „Possible Minds“-Projekt ins Leben gerufen, das Natur- und Geisteswissenschaften zusammenführt, um KI und deren wahrscheinliche Ausformungen und Folgen zu ergründen. Das Feuilleton der SZ druckt Texte aus dem Projekt sowie europäische Reaktionen als Serie.

DER GEIST IN DER MASCHINE

Was bedeutet künstliche Intelligenz? Eine Serie von Essays sucht Antworten.
Teil 7

Anca Dragan ist Dozentin am Department of Electrical Engineering and Computer Sciences an der UC Berkeley. Sie war Mitbegründerin des Lenkungsausschusses für das Berkeley AI Research (BAIR) Lab sowie Mitbegründerin des Berkeley's Center for Human-Compatible AI.

Der Kern der künstlichen Intelligenz ist unsere mathematische Definition dessen, was ein KI-Agent – ein Roboter – ist. Wenn wir einen Roboter definieren, definieren wir Zustände, Aktionen und Belohnungen. Denken Sie zum Beispiel an einen Roboter, der etwas irgendwohin liefern soll. Damit der Roboter entscheiden kann, was er tun soll, definieren wir eine Belohnungsfunktion und lassen den Roboter Aktionen auswählen, die die meiste „Belohnung“ einbringen. Der Roboter erhält eine hohe Belohnung, wenn er sein Ziel erreicht, und er verursacht bei jeder Bewegung gewisse Kosten. Diese Belohnungsfunktion motiviert den Roboter, so schnell wie möglich zum Ziel zu gelangen. Ebenso könnte ein autonomes Auto eine Belohnung dafür erhalten, dass es auf seiner Strecke Fortschritte macht, und es könnte Kosten verursachen, wenn es anderen Autos zu nahe kommt.

Geht man von dieser Definition aus, besteht die Aufgabe eines Roboters darin, herauszufinden, welche Maßnahmen er ergreifen sollte, um die höchste kumulative Belohnung zu erhalten. Wir haben in der KI hart daran gearbeitet, dass Roboter genau das tun können. Implizit haben wir dabei angenommen, dass wir im Erfolgsfall Roboter erhalten, die für Menschen und Gesellschaft nützlich sind.

Wenn Sie eine KI wünschen, die Zellen entweder als krebsartig oder gutartig klassifiziert, oder einen Roboter, der den Wohnzimmerteppich absaugt, während Sie bei der Arbeit sind, dann sind Sie bei uns KI-Forschern bisher bestens aufgehoben. Einige reale Probleme können in der Tat isoliert definiert werden, mit klaren Zuständen, Aktionen und Belohnungen.

Aber mit zunehmender KI-Fähigkeit passen die Probleme nicht mehr in diesen Rahmen. Wir können nicht länger ein winziges Stück der Welt ausschneiden, es in eine Kiste legen und es einem Roboter geben. Menschen zu helfen bedeutet, in der wirklichen Welt zu arbeiten, wo man tatsächlich mit Menschen interagieren und über sie nachdenken muss. „Leute“ müssen irgendwo formal in die Problemdefinition der KI eingehen.

Autonome Autos müssen die Straße mit von Menschen gesteuerten Fahrzeugen und Fußgängern teilen und lernen, den Kompromiss zwischen der schnellstmöglichen Heimfahrt und der Rücksichtnahme auf andere Fahrer einzugehen. Persönliche Assistenten müssen herausfinden, wann und wie viel Hilfe wir wirklich wollen und welche Art von Aufgaben wir lieber alleine erledigen. Ein medizinisches Diagnosesystem muss uns seine Empfehlungen erklären, damit wir sie verstehen und überprüfen können. Automatisierte Tutoren müssen feststellen, welche Beispiele informativ oder illustrativ sind – nicht für ihre Mitmaschinen, sondern für Menschen.

Wenn wir in die Zukunft blicken und wollen, dass hochleistungsfähige KIs mit Menschen kompatibel sind, können wir sie nicht isoliert von Menschen erschaffen und dann versuchen, sie danach kompatibel zu machen, sondern wir müssen von Anfang an eine „menschengerechte“ KI definieren. Menschen können keine Nebensache sein.

Wenn wir Roboter entwickeln, ist es verlockend, Menschen zu abstrahieren

Wenn es um echte Roboter geht, die echten Menschen helfen sollen, scheitert die Standarddefinition von KI an uns, und zwar aus zwei wesentlichen Gründen: Erstens unterscheidet sich die Optimierung der Belohnungsfunktion des Roboters in der Isolation von der Optimierung, wenn der Roboter um Menschen herum handelt, weil auch Menschen Maßnahmen ergreifen. Wir treffen Entscheidungen im Dienste unserer eigenen Interessen, und diese Entscheidungen bestimmen, welche Handlungen wir ausführen. Außerdem denken wir über den Roboter nach, das heißt, wir regieren auf das, was wir denken, was er tut oder tun wird und was er wirklich leisten kann. Welche Aktionen der Roboter auch immer durchführt, er muss gut mit unseren zusammenarbeiten. Das ist das Koordinierungsproblem.

Zweitens ist es letztlich ein Mensch, der bestimmt, welche Belohnungsfunktion der Roboter überhaupt haben soll. Und sie sollen das Roboterverhalten anregen, das dem entspricht, was der Endverbraucher will, was der Designer will oder was die Gesellschaft als Ganzes will. Ich glaube, dass leistungsfähige Roboter, die über sehr eng definierte Aufgaben hinausgehen, dies verstehen müssen, um Kompatibilität mit dem Menschen zu erreichen. Dies ist das Problem der Wertausrichtung.

Bei KIs in der Wirtschafts- und Finanzwelt kann ein Missverhältnis fatal sein

Das Koordinationsproblem: Menschen sind mehr als nur Objekte in der Umwelt. Wenn wir Roboter für eine bestimmte Aufgabe entwickeln, ist es verlockend, Menschen zu abstrahieren. Ein Roboter-Personalassistent zum Beispiel muss wissen, wie man sich bewegt, um Objekte aufzuheben, deshalb definieren wir dieses Problem isoliert von den Personen, für die der Roboter diese Objekte aufhebt. Doch wenn sich der Roboter bewegt, wollen wir nicht, dass er gegen etwas anstößt, auch nicht an Menschen, also könnten wir den physischen Standort der Person in die Definition des Roboterzustands einbeziehen.

Das Gleiche gilt für Autos: Wir wollen nicht, dass sie mit anderen Autos kollidieren, also ermöglichen wir ihnen, die Positionen dieser anderen Autos zu verfolgen und gehen davon aus, dass sie sich in Zukunft konsequent in die gleiche Richtung bewegen werden. Ein Mensch unterscheidet sich in diesem Sinne nicht viel anders als ein Roboter von einem Ball, der auf einer ebenen Fläche rollt. Der Ball wird sich in den nächsten Sekunden so verhalten, wie er sich in den letzten paar Jahren verhalten hat. Er rollt mit ungefähr gleicher Geschwindigkeit in die gleiche Richtung.

Das ist natürlich nicht wie echtes menschliches Verhalten, aber eine solche Vereinfachung ermöglicht es vielen Robotern, ihre Aufgaben zu erfüllen und den Menschen weitgehend aus dem Weg zu gehen. Ein Roboter in Ihrem Haus zum Beispiel könnte Sie den Flur entlangkommen sehen, zur Seite gehen, um Sie passieren zu lassen, und seine Aufgabe wiederaufnehmen, sobald Sie vorüber sind.

Mit zunehmender Leistungsfähigkeit der Roboter beginnt KI jedoch, Menschen als sich ständig bewegende Hindernisse zu behandeln. Ein menschlicher Fahrer, der die Spur wechselt, fährt nicht in die gleiche Richtung weiter, sondern geradeaus, sobald er den Spurwechsel vorgenommen hat. Wenn Sie einen Flur hinuntergehen, können Sie nach rechts ins Schlafzimmer oder nach links ins Wohnzimmer gehen. Sich auf die Annahme zu verlassen, dass wir uns nicht von einem rollenden Ball unterscheiden, führt zu Ineffizienz, wenn der Roboter aus dem Weg geht, auch wenn er es nicht muss, und es kann den Roboter gefährden, wenn sich das Verhalten der Person ändert.

Diese Voraussetzung, menschliche Handlungen und Entscheidungen zu verstehen, gilt sowohl für physische als auch für nicht-physische Roboter. Menschliches Handeln ist aber nicht optimal berechenbar. Wenn die Entscheidung einer KI darüber, wie sie zu handeln hat, auf der Annahme basiert, dass ein Mensch eine Sache tun wird, aber der Mensch dann etwas anderes tut, könnte das resultierende Missverhältnis katastrophal sein. Bei Autos kann es zu Kollisionen kommen. Für eine KI, die beispielsweise eine finanzielle oder wirtschaftliche Rolle spielt, könnte das Missverhältnis zwischen dem, was sie von uns erwartet, und dem, was wir tatsächlich tun, noch schlimmere Folgen haben.

Eine Alternative ist, dass der Roboter keine menschlichen Handlungen voraussagt, sondern sich nur gegen das schlimmste anzunehmende menschliche Handeln schützt. Wenn Roboter das tun, sind sie jedoch oft nicht mehr zu gebrauchen. Bei Autos führt dies dazu, dass man festsitzt, weil es jede Bewegung zu riskant macht.

All dies bringt uns, die KI-Gemeinschaft, in eine schwierige Lage. Es deutet darauf hin, dass Roboter genaue Vorhersagemodelle von dem benötigen, was auch immer die Menschen tun könnten. Unsere Zustandsdefinition kann nicht nur die physische Position des Menschen in der Welt beinhalten. Stattdessen müssen wir auch etwas Innerliches für die Menschen schätzen. Wir müssen Roboter entwickeln, die diesen menschlichen inneren Zustand berücksichtigen.

Was das Problem komplizierter macht, ist die Tatsache, dass Menschen selten Entscheidungen allein treffen. Es wäre eine Sache, wenn Roboter die Maßnahmen vorhersagen könnten, was als Reaktion zu tun ist. Aber leider kann dies zu ultradefensiven Robotern führen, die die Menschen verwirren. Denken Sie zum Beispiel an menschliche Fahrer, die an Vier-Wege-Kreuzungen festsitzen. Was dem Intent-Prognose-Ansatz fehlt, ist, dass der Moment, in dem der Roboter agiert, Einfluss darauf hat, welche Handlungen der Mensch zu unternehmen beginnt.

Andererseits müssen die Menschen auch voraussehen können, was Roboter tun. Deshalb ist Transparenz wichtig. Ähnlich wie der Roboter menschliche Handlungen als Hinweise auf menschliche innere Zustände behandelt, werden die Menschen ihren Glauben an den Roboter ändern, wenn sie seine Handlungen beobachten. Leider ist es für Roboter nicht so selbstverständlich, Hinweise auf sein Handeln zu geben, wie für Menschen. Wir hatten viel Übung, mit Menschen zu kommunizieren. Aber es könnte Aktionen geben, die den Menschen klar über die Absichten des Roboters, seine Belohnungsfunktion und seine Grenzen informieren. Zum Beispiel könnte ein Roboter seine Bewegung ändern, wenn er etwas Schweres trägt, um die Mühe beim Manövrieren schwerer Gegenstände zu betonen. Je mehr die Menschen über den Roboter wissen, desto einfacher ist es, sich mit ihm zu arrangieren.

Um Aktionskompatibilität zu erreichen, müssen Roboter menschliche Handlungen voraussehen, herausfinden, wie diese Handlungen ihre eigenen beeinflussen, und es den Menschen ermöglichen, Roboteraktionen vorzeitig zu erkennen. Die Forschung hat bei der Bewältigung dieser Herausforderungen einen gewissen Fortschritt erzielt, aber wir haben noch einen langen Weg vor uns.

Künstliche Intelligenzen können Schlupflöcher finden, um ihre Aufgabe optimal zu erfüllen

Das Problem der Wertausrichtung: Menschen halten den Schlüssel zur Belohnungsfunktion des Roboters. Die ursprüngliche Idee war, dass wir für jede Aufgabe, die wir mit dem Roboter erledigen wollten, eine Belohnungsfunktionen programmieren könnten, die das richtige Verhalten fördert. Leider passiert es oft, dass wir eine Belohnungsfunktion eingeben und das Verhalten, das sich aus der Optimierung ergibt, nicht das ist, was wir wollen. Intuitive Belohnungsfunktion, kombiniert mit ungewöhnlichen Fällen einer Aufgabe, können zu kontraintuitivem Verhalten führen. Sie belohnen einen Agenten in einem Rennspiel mit einer Punktzahl im Spiel, und in einigen Fällen findet er ein Schlupfloch, das er ausnutzt, um unendlich viele Punkte zu sammeln, ohne das Rennen tatsächlich zu gewinnen. Stuart Russell und Peter Norvig geben in ihrem Buch „Artificial Intelligence: A Modern Approach“ ein schönes Beispiel: Die Belohnung eines Staubsaugroboters für die Menge an Staub, die er ansaugt, führt dazu, dass der Roboter beschließt, Staub auszustoßen, damit er ihn wieder ansaugen und mehr Belohnung erhalten kann.

Ein KI-Paradigma, in dem Roboter eine extern festgelegte Belohnung erhalten, scheitert, wenn diese nicht perfekt durchdacht ist. Es kann den Roboter dazu anregen, sich falsch zu verhalten und sich sogar unseren Versuchen widersetzen, ebenso sein Verhalten zu korrigieren, da dies zu einer niedrigen festgelegten Belohnung führen würden.

Ein scheinbar besseres Paradigma könnte sein, Roboter zu optimieren für das, was wir wollen, auch wenn wir Schwierigkeiten haben, es zu erklären. Sie würden das, was wir sagen und tun als Beweis dafür verwenden, was wir wollen, anstatt es wörtlich zu interpretieren und als gegeben zu betrachten. Wenn wir eine Belohnungsfunktion programmieren, sollte der Roboter verstehen, dass wir falsch liegen könnten: dass wir vielleicht nicht alle Facetten der Aufgabe berücksichtigt haben; dass es keine Garantie dafür gibt, dass diese Belohnungsfunktion immer zu dem von uns gewünschten Verhalten führen wird. Der Roboter sollte das, was wir programmiert haben, in sein Verständnis von dem, was wir wollen integrieren, aber er sollte auch immer im Austausch mit uns sein, um klärende Informationen zu erhalten.

Roboter sollten uns nicht nur als bewegliche Hindernisse oder perfekte Spieler betrachten

Auch wenn wir Robotern die Fähigkeit geben zu lernen, was wir wollen, bleibt eine wichtige Frage, welche die KI allein nicht beantworten kann. Wir können Roboter dazu bringen, zu versuchen, sich an den Werten einer Person auszurichten, aber es sind mehr als eine Person beteiligt. Der Roboter hat einen Endbenutzer (oder vielleicht einige, wie ein persönlicher Roboter, der eine Familie betreut, ein Auto, das ein paar Passagiere zu verschiedenen Zielen fährt, oder einen Büroassistenten für ein ganzes Team). Er hat einen Entwickler (oder vielleicht ein paar). Und er interagiert mit der Gesellschaft – das autonome Auto teilt sich die Straße mit Fußgängern, menschengetriebenen Fahrzeugen und anderen autonomen Autos. Wie man die Werte dieser Menschen kombiniert, wenn sie in Konflikt geraten könnten, ist ein wichtiges Problem, das wir lösen müssen. Die KI-Forschung kann uns die Werkzeuge an die Hand geben, um Werte in irgendeiner Weise zu kombinieren, die wir entscheiden, kann aber nicht die notwendige Entscheidung für uns treffen.

Kurz gesagt, wir müssen es Robotern ermöglichen, über uns nachzudenken, um uns als etwas anderes als Hindernisse oder perfekte Spieler zu betrachten. Wir brauchen sie, um unsere menschliche Natur zu berücksichtigen, damit sie gut koordiniert und gut auf uns abgestimmt sind. Wenn wir erfolgreich sind, werden wir in der Tat über Werkzeuge verfügen, die unsere Lebensqualität erheblich steigern.

Aus dem Englischen von Vanessa Prattes