Kein Spiel mehr: Auf einzelnen Teststrecken sind bereits komplett autonom fahrende Fahrzeuge unterwegs. Doch laut Umfragen vertraut ihnen nur eine knappe Mehrheit der Autofahrer.

ILLUSTRATION: LUISA JUNG FÜR MPG

WENN MASCHINEN MITMISCHEN

TEXT: RALF GRÖTKER

Immer öfter treffen wir im Alltag auf künstliche Intelligenz: auf Chatbots im Callcenter, auf Roboterkollegen am Fließband oder auf elektronisch gesteuerte Mitspieler beim Gamen. Am Max-Planck-Institut für Bildungsforschung in Berlin untersucht Iyad Rahwan mit seinem Team, wie Menschen sich verhalten, wenn sie mit intelligenten Maschinen zu tun haben, und was sie von ihrem künstlichen Gegenüber erwarten.

FOTO: ARNE SATTLER FÜR MPG

„Fake News“ ist seit 2017, also ungefähr seit Donald Trump das Amt des Präsidenten der USA antrat, laut Duden ein offizieller Begriff der deutschen Sprache. Das Entstehen von Fake News ist eng gekoppelt an die Entwicklung von künstlicher Intelligenz. Ein Beispiel: Künstliche Intelligenz macht es möglich, Fake News in hoher Reichweite zu erstellen und in den sozialen Netzwerken massenhaft zu verbreiten. Verändert sich dadurch generell das Vertrauen, das wir in Medieninhalte haben? Verändern Fake News unser Verhalten? Für Iyad Rahwan, seit Ende 2018 Direktor des von ihm neu gegründeten Center for Humans and Machines am Berliner Max-Planck-Institut für Bildungsforschung, ist dies eine typische Forschungsfrage. Gemeinsam mit seinem Team beantwortet er derartige Problemstellungen nicht mit Umfragen, sondern experimentell – um herauszufinden, welchen Effekt bereits existierende Technologien haben; und, mehr noch, um eine Idee davon zu bekommen, wie sich gerade in den Kinderschuhen befindliche Innovationen in Zukunft auswirken könnten. Das Forschungsprogramm des Centers fasst er in einem einzigen Satz zusammen: Welchen Einfluss haben digitale Technologien, soziale Medien und künstliche Intelligenz auf das menschliche Verhalten?

„Man muss sich nur einmal vorstellen, jemand hätte in den frühen 2000er-Jahren eine Vorstellung davon gehabt, wie sich Facebook und Twitter entwickeln werden, und er hätte Verhaltensexperimente durchgeführt, um zu antizipieren, wie sich die fortschreitende digitale Vernetzung auf die Verbreitung von Falschinformationen auswirken würde“, schwärmt Rahwan. „Man hätte die ganze Situation, die wir heute vorfinden, simulieren können, bevor sie eintrat.“ Das Experiment dazu? Im Fall von Fake News, schlägt der Wissenschaftler vor, könne ein Experiment darin bestehen zu untersuchen, wie gut Versuchspersonen es erkennen können, wenn aus existierenden Fotos Personen wegretuschiert werden – was mit Verfahren der künstlichen Intelligenz kinderleicht und für eine große Anzahl von Bildern möglich ist. Gleichzeitig könne man Experimente aufsetzen, um einen Eindruck davon zu bekommen, ob der Einsatz von KI-Verfahren Menschen besonders dazu animiert, andere via Fake News zu manipulieren – „nicht nur, weil die neuen Technologien Manipulationen einfacher machen, sondern auch, weil derjenige, der die Technologie einsetzt, sich quasi nicht selbst die Hände dreckig machen muss“.

Die Methoden, derer sich Rahwan bedient, sind improvisiert. Es gibt keine wissenschaftliche Disziplin, die in Gänze bereits die erforderlichen Instrumente bereitstellen würde. „Was wir machen, ist zum großen Teil Science-Fiction-Forschung. Es geht darum, Versuchspersonen dazu zu bringen, sich Situationen vorzustellen, die sie noch gar nicht kennen – und in diesen Situationen Entscheidungen zu treffen.“ Die Wissenschaftlerinnen und Wissenschaftler, mit denen er vorzugsweise zusammenarbeitet, kommen deshalb teils aus der wirtschaftswissenschaftlich geprägten Verhaltensforschung, die mit Simulationen und Laborexperimenten Erfahrung hat, aber auch aus der Psychologie, der Computerwissenschaft, der Anthropologie und der Soziologie.

Eine Forschungsarbeit, auf die Rahwan besonders stolz ist und die ihn auch weit über Fachkreise hinaus bekannt gemacht hat, ist ein Experiment mit dem Titel *Moral Machine*. Das Experiment wird seit Juni 2016 über eine frei zugängliche Onlineplattform durchgeführt. Mehrere Millionen Menschen aus 233 Ländern und Regionen nahmen bisher teil. Es wird weltweit in Science Centers und in Museen präsentiert und ist auch in zahlreiche Lehrbücher aufgenommen worden. Das Experiment präsentiert ein Dilemma: Ein automatisiert gesteuertes Fahrzeug hält auf eine Gruppe

Vielseitig: Iyad Rahwan hat Informatik studiert, interessiert sich aber auch für psychologische und philosophische Fragestellungen. Im Forschungsbereich Mensch und Maschine am Max-Planck-Institut für Bildungsforschung bringt er diese Themenfelder zusammen.

von Fußgängern zu und kann nicht mehr rechtzeitig bremsen. Die künstliche Intelligenz, die das Fahrzeug steuert, kann sich nur noch entscheiden, entweder gegen eine Wand zu fahren (wobei die Insassen des Autos zu Schaden kommen) oder die Fußgänger zu überfahren. Was soll die künstliche Intelligenz tun? Auf welche Entscheidung hin soll sie von ihren Entwicklern trainiert werden? In dem Versuch wurden verschiedene Szenarien abgefragt, die sich unter anderem durch die Anzahl der Fahrgäste und der Fußgänger sowie deren Alter unterschieden. Die Resultate zeigen, dass die meisten mit ihren Entscheidungen versuchen, möglichst viele Menschenleben zu retten, und dass sie der Rettung jüngerer Menschen den Vorzug geben gegenüber der Rettung älterer Personen.

Eine Art Gleichnis

Warum sind solche Erkenntnisse von Belang? Das Szenario, das in dem Versuch präsentiert wurde, zeigt schließlich eine absolute Extremsituation – keine Situation, mit der sich Entwickler selbstgesteuerter Fahrzeuge vornehmlich beschäftigen. Zumindest im deutschen Rechtsrahmen ist es auch nicht vorgesehen, dass autonom gesteuerte Fahrzeuge Abwägungen vornehmen, wen sie im Ernstfall schützen und wen sie, wenn es nicht anders geht, zu Schaden kommen lassen sollten. In Gefahrensituationen, so sieht es die Vorschrift vor, muss das Fahrzeug einfach so schnell wie möglich zum Halten kommen. Punkt. „Man kann das Szenario auch als eine Art Gleichnis verstehen“, verteidigt Rahwan das Experiment. „Denn natürlich müssen selbstfahrende Autos darauf trainiert und programmiert werden, Entscheidungen zu treffen. Lässt man das Auto zum Beispiel näher in der Straßenmitte fahren, wo es mit entgegenkommenden Fahrzeugen kollidieren kann? Oder am Straßenrand, wo die Gefahr besteht, dass es einen Radfahrer streift? Statistisch betrachtet, haben solche Verhaltensregeln durchaus Einfluss darauf, welche Personengruppen zu Schaden kommen und welche eher nicht.“ Natürlich können ethische Fragen nicht dadurch gelöst werden, dass Menschen in einer Umfrage oder in einem Onlineexperiment entscheiden. „Aber die Politik und diejenigen, die Richtlinien erstellen, sollten zumindest wissen, wie gewöhnliche Leute über solche Fragen denken – auch deshalb, weil sie darauf gefasst sein müssen, ihre Entscheidungen vor einer Öffentlichkeit zu begründen, die vielleicht anderer Meinung ist als sie.“

„Natürlich müssen selbstfahrende Autos darauf trainiert werden, Entscheidungen zu treffen.“

IYAD RAHWAN

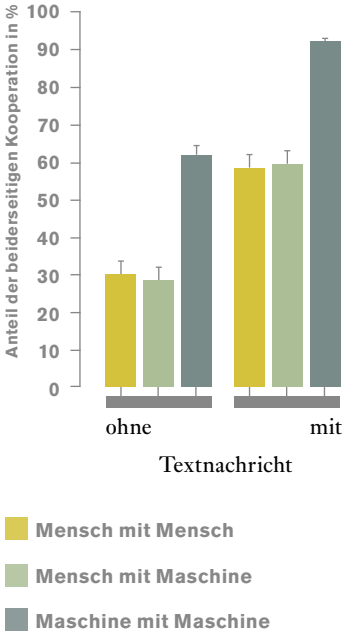
Ein herausstechendes Merkmal von Iyad Rahwans Forschungsarbeiten ist, dass er seine Experimente immer wieder neu erfindet. Die meisten arbeiten mit einer Story, die sich aus vielen Blickwinkeln interpretieren

→

lässt. Gleichzeitig liefern sie im Resultat solide quantitative Ergebnisse. Wie gelangt man zu solchen Forschungsdesigns? „Wichtig ist, dass man es sich erlaubt, offen zu bleiben, und bei einer Fragestellung nicht immer sofort daran denkt, ob sie sich überhaupt mit geeigneten Methoden untersuchen lässt. Man muss wirklich versuchen, die interessanteste Frage zu finden“, meint Rahwan. Was außerdem hilft, ist der Blick über den eigenen Tellerrand. „Ich lese viele populäre Wissenschaftsbücher. Oft komme ich dabei auf Ideen für meine eigene Arbeit. Mein Projekt über die Kooperation zwischen Menschen und Maschinen beispielsweise war vor allem inspiriert von Sachbüchern, die ich zum Thema Kooperation zwischen Menschen gelesen hatte.“ Populäre Wissenschaftsbücher helfen laut Rahwan nicht nur, wissenschaftliche Inhalte einer größeren Öffentlichkeit bekannt zu machen. Sie helfen auch Wissenschaftlern, interdisziplinär zu arbeiten. „Wenn ich in einem fremden Feld stöbere, ist es schwierig für mich, unter den vielen Artikeln aus Fachzeitschriften genau die zu finden, die mich weiterbringen. Bei Wissenschaftsbüchern, die für einen breiteren Leserkreis konzipiert sind, hat hingegen schon eine gewisse Selektion stattgefunden.“

In dem erwähnten Projekt zur Kooperation von Menschen und Maschinen testete Rahwan, wie künstliche Intelligenzen zusammenarbeiten können – untereinander und mit Menschen. „Es gibt viele Diskussionen darüber, ob Computer den Menschen ersetzen können. Und die meisten Tests, die es gibt, um das Potenzial von künstlicher Intelligenz zu testen, laufen mit Spielen wie Schach oder Go, wo es immer einen Gewinner und einen Verlierer gibt. Die Interaktionen aber, die in der Realität stattfinden, sehen anders aus.“ Die Forschenden untersuchten die Kooperation von Maschinen mit Menschen oder untereinander mithilfe von Kooperationsspielen aus der Spieltheorie. Das bekannteste Kooperationsspiel ist das Gefangenendilemma, in dem zwei Spieler entscheiden müssen, ob sie sich gegenseitig verraten, wenn sie separat als Zeugen befragt werden. Verrät ein Spieler den anderen, zieht er daraus den größten Vorteil, solange der andere schweigt. Halten beide zusammen und sagen nichts, kommen sie immer noch besser weg, als wenn beide den jeweils anderen verraten.

ZUSAMMENARBEIT ZWISCHEN MENSCH UND MASCHINE

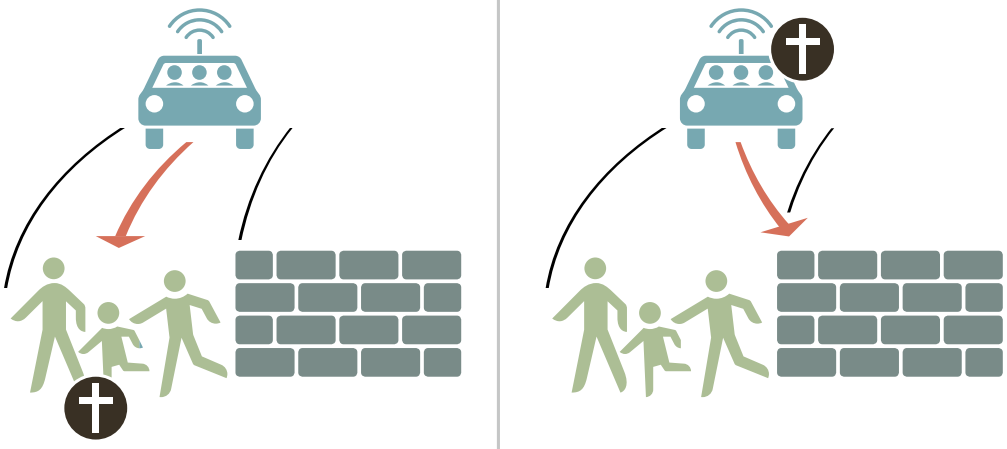


In einem Kooperationsspiel arbeiteten Menschen und Maschinen zunächst nur schlecht zusammen. Als die Beteiligten Textnachrichten austauschen konnten, stieg die Kooperationsbereitschaft deutlich.

Menschliche Züge fördern Kooperation

Rahwan und sein Team testeten das Spiel mit 25 verschiedenen KI-Typen, die Verfahren des maschinellen Lernens anwenden. Die Resultate waren zunächst eher frustrierend. Die meisten Algorithmen schienen mehr oder weniger unfähig zur Kooperation, und selbst der am besten funktionierende Algorithmus konnte nicht erfolgreich mit Menschen kooperieren. Interessant wurde es erst, als das Team sowohl den menschlichen Spielern als auch dem siegreichen Algorithmus aus der ersten Versuchsrunde die Möglichkeit gab, eine Nachricht auszutauschen. Konkret konnten sowohl die menschlichen wie die maschinellen Spieler zu Beginn jeder Runde eine Textnachricht mit Sätzen wie „Tu, was ich sage, oder ich bestrafe dich“, „Ich ändere jetzt meine Strategie“ oder „Gib mir noch eine Chance“ an den Mitspieler senden. Dazu konnten sie sich aus einem vorgegebenen Pool von Textnachrichten bedienen. Getestet wurden sowohl Szenarien, in denen die menschlichen Spieler lügen konnten, als auch Szenarien, in denen dies nicht der Fall war. Die Algorithmen konnten grundsätzlich nicht lügen. Keiner der Spieler kannte die Identität des Gegenübers. Der erstaunliche Effekt: In den Versuchsläufen ohne zusätzliche Textnachrich-

GRAFIK: GEO NACH CRANDALL, J.W., OUDAH, M., TENNOM ET AL. COOPERATING WITH MACHINES. NAT COMMUN 9, 233 (2018).



GRAFIK: GCO NACH MORALMACHINE.NET

Dilemma: Wie soll ein selbstfahrendes Auto handeln, wenn es nicht mehr rechtzeitig bremsen kann? Befragungen zeigen, dass eine Mehrheit dafür ist, möglichst viele Menschenleben zu retten.

AUF DEN PUNKT GEBRACHT

Forschende um Iyad Rahwan gehen der Frage nach, welchen Einfluss digitale Technologien auf das menschliche Verhalten haben.

In Experimenten mit Testpersonen untersuchten sie zum Beispiel, unter welchen Voraussetzungen Menschen und Maschinen kooperieren.

Ein weiterer Versuch befasste sich mit ethischen Vorgaben für selbstfahrende Autos.

ten führten Spiele im Szenario „Mensch mit Mensch“ und auch im Szenario „Maschine mit Mensch“ nicht zu besonders kooperativem Verhalten. Das Szenario „Maschine mit Maschine“ schnitt etwas besser ab. Sobald jedoch Textnachrichten als Zusatzelement eingeführt wurden, verdoppelte sich in allen drei Szenarien die Kooperationsbereitschaft.

Diese Resultate zeigen dreierlei. Erstens: Selbst ohne die Fähigkeit zur Kommunikation ist die Kooperationsbereitschaft künstlicher Intelligenzen höher als die des Menschen. Zweitens: Die Kooperationsleistung künstlicher Intelligenzen kann erhöht werden, wenn ihr menschliche Züge verliehen werden. Wenn sie kommunizieren können, stellen künstliche Intelligenzen in Sachen Kooperationsbereitschaft alle Teams, an denen Menschen beteiligt sind, deutlich in den Schatten. Drittens: Menschen reagieren anders auf eine künstliche Intelligenz, wenn diese kommuniziert. Tatsächlich konnten in den Experimenten mit Textchat die menschlichen Versuchsteilnehmenden in vielen Fällen die Maschine nicht mehr von einem menschlichen Gegenüber unterscheiden. Ob es einen Grund dafür gibt, dass Algorithmen im Kooperationsspiel erfolgreicher abschneiden als Menschen? „Eine Ursache könnte sein, dass Maschinen sich treu bleiben. Wenn sie mehrere Spielrunden erfolgreich hinter sich gebracht haben, in denen sie von der erlaubten Möglichkeit des nichtkooperativen, auf Eigennutz bedachten Verhaltens keinen Gebrauch gemacht haben, dann werden sie auch in weiteren Runden die Kooperation nicht abbrechen. Menschen ticken hier anders, auch wenn sie mit dieser Strategie beinahe immer verlieren“, meint Iyad Rahwan. Ein anderer Grund könne sein, dass Menschen sich oft an die Zusagen, die sie im Textchat gemacht hatten, nicht gehalten haben. Auch das führt dazu, dass der gemeinsam erzielte Spielerfolg sich verringert.

Gibt es Dinge, die ein Computer oder eine KI niemals wird tun können? „Schlussendlich, glaube ich, gibt es nichts, was eine KI nicht kann“, sagt Iyad Rahwan. „Aber zumindest für die nähere Zukunft sehe ich dort Grenzen, wo Menschen miteinander auf eine Weise interagieren, die ein tieferes psychologisches Verständnis erfordert. Maschinen sind hier deshalb im Nachteil, weil sie nur durch Beobachtung aus menschlichem Verhalten lernen können. Sie können nicht aus den Erfahrungen der eigenen Lebensgeschichte schöpfen und diese Erfahrungen in die Interpretation einer Situation einbringen.“

🎧 *www.mpg.de/podcasts/kuenstliche-intelligenz*