Wie eingebautes Misstrauen KI-Systeme sicherer machen kann

Originalartikel

Backup

<html> <header class=„article-header“><h1 class=„articleheading“>Wie eingebautes Misstrauen KI-Systeme sicherer machen kann</h1> Karen Hao </header><figure class=„aufmacherbild“><figcaption class=„akwa-caption“>(Bild: ACM)</figcaption></figure> Ayanna Howard, Robotikerin und Pädagogin, hält es für gefährlich, automatisierten Systemen zu vertrauen. Im TR-Interview erläutert sie Lösungsmöglichkeiten. Ayanna Howard sieht den Nutzen von Robotern und KI vor allem darin, Menschen zu helfen. In ihrer fast 30-jährigen Karriere hat sie unzählige Roboter gebaut: zur Erforschung des Mars, zur Reinigung von Sondermüll und zur Unterstützung von Kindern, die bestimmte Hilfen brauchen. Dabei hat sie eine beindruckende Palette an Techniken entwickelt zur Robotermanipulation, autonomen Navigation und Computer Vision. Und sie ist führend bei der Erforschung eines typisch menschlichen Verhaltens: Wir vertrauen automatisierten Systemen zu sehr. Im März trat sie nach 16 Jahren als Professorin am Georgia Institute of Technology als erste Frau ihre Stelle als Dekanin des College of Engineering an der Ohio State University an. Am Tag der ACM-Preisverleihung sprach TR mit ihr über ihre Karriere und ihre neueste Forschung. <p class=„frage rteabs–frage“>TR: Sie verwenden für Ihre Forschung den Begriff „humanisierte Intelligenz“ anstatt „künstliche Intelligenz“. Warum? <p class=„antwort rteabs–antwort“>Ayanna Howard: Ja, 2004 habe ich den Begriff in einer Veröffentlichung verwendet. Denn wir wollen Robotik und KI-Systeme ja nicht losgelöst von Menschen erschaffen. Wir bauen dabei auf die Erfahrung, die Daten und die Eingaben von Menschen auf. „Künstliche Intelligenz“ impliziert, dass es sich um eine andere Art von Intelligenz handelt. „Humanisierte Intelligenz“ besagt hingegen, dass diese Intelligenzen von Menschen geschaffene Konstrukte sind. <p class=„frage rteabs–frage“>Wie sind Sie zu dieser Arbeit gekommen? <p class=„antwort rteabs–antwort“>In erster Linie durch meine Doktorarbeit. Damals habe ich einen Roboter dazu trainiert, Gefahren in einem Krankenhaus zu verringern. Zu der Zeit wurden etwa noch Nadeln in denselben Müll geworfen wie alles andere, und es kam vor, dass dadurch Krankenhausmitarbeiter krank wurden. Also habe ich überlegt: Wie könnten Roboter da helfen? <figure class=„branding“><img alt=„Mehr von MIT Technology Review“ height=„693“ src=„https://static.wallabag.it/7862d1b7aff4c3b00f37212fefade4e0e2c4cf00/64656e6965643a646174613a696d6167652f7376672b786d6c2c25334373766725323078 6d6c6e733d27687474703a2f2f777

7772e77332e6f72672f323030302f737667272532307769647468 3d27363936707827253230686569676874 3d27333931707827253230766965774 26f783d27302532303025323036393625323033393127253345253343726563742532307 83d273027253230793d27302725323077696474683d27363936272532306865696768743d27333931272532306669 6c6c3d272532336632663266322725334525334326f726563742533452533432f737667253345/“ class=„c1“ width=„1200“ referrerpolicy=„no-referrer“ /> [1]</figure> Es ging mir damals schon um Roboter, die für Menschen nützlich sind. Wir wussten noch nicht, wie man Roboter baut, die solche Aufgaben erledigen könnten. Aber Menschen verrichten die Tätigkeiten ständig, also sollten wir sie nachahmen. So fing es an. Mit der NASA habe ich dann an der Navigation von Mars-Rovern gearbeitet. Da war es das Gleiche: Wissenschaftler können das sehr, sehr gut. Also ließ ich Wissenschaftler die Rover steuern und schaute mir an, was sie auf den Kameras der Rover sahen und wie sie darauf reagierten. Das war immer mein Thema: Warum gehe ich nicht einfach zu menschlichen Experten, kodiere das, was sie tun, in einen Algorithmus und bringe dann den Roboter dazu, das zu verstehen und umzusetzen? <p class=„frage rteabs–frage“>Haben andere Leute damals auch so gedacht oder waren Sie eine Außenseiterin? <p class=„antwort rteabs–antwort“>Eine total verrückte Außenseiterin. Ich habe die Dinge anders betrachtet als alle anderen. Und damals gab es noch keine Anleitung, wie man so eine Forschung betreibt. Wenn ich jetzt zurückblicke, würde ich es nach all den Erfahrungen in der Praxis ganz anders machen. <p class=„frage rteabs–frage“>Seit wann haben Sie, statt weiter Roboter für Menschen zu konstruieren, die Beziehung zwischen Robotern und Menschen in den Mittelpunkt Ihrer Arbeit gestellt? <p class=„antwort rteabs–antwort“>Angestoßen wurde das durch eine Studie, in der wir herausfinden wollten, ob sich Menschen in riskanten Situationen der Führung eines Roboters anvertrauen würden. Also brachten wir die Teilnehmer in ein verlassenes Bürogebäude, in das sie ein Roboter hineinließ. Wir haben ihnen erzählt, es würde sich um eine Umfrage handeln. Während sie dort drinnen waren, haben wir in dem Gebäude Rauch entfacht und den Feueralarm ausgelöst. Wir wollten sehen, ob die Menschen beim Verlassen des Gebäudes zur Eingangstür gehen oder dem Hinweisschild zum Notausgang folgen, oder ob sie sich dem Roboter anschließen würden, der sich in eine andere Richtung bewegt hat. Wir dachten, die Leute würden zur Eingangstür gehen, weil sie dort hereingekommen waren. Frühere Forschungen haben gezeigt, dass Menschen in einer Notsituation meist vertraute Wege gehen. Oder dass sie den Schildern folgen würden, weil das ein einstudiertes Verhalten ist. Aber die

Teilnehmer taten das nicht. Sie folgten tatsächlich dem Roboter. Dann haben wir einige Fehler eingebaut. Wir ließen den Roboter ausfallen, wir ließen ihn im Kreis fahren oder einen Weg fahren, bei dem man erst Möbel beiseiteschieben musste. Wir dachten, irgendwann würde jemand sagen: „Lass uns zur Eingangstür oder zum Ausgangsschild dort gehen.“ Aber sie folgten ihm buchstäblich bis zum Ende des Experiments. Damit lagen wir zum ersten Mal mit unseren Hypothesen völlig falsch. Wir konnten kaum glauben, dass die Leute einem System so vertrauen. Das ist interessant und faszinierend – aber es ist ein Problem. <p class=„frage rteabs–frage“>Haben Sie dieses Phänomen seitdem in der realen Welt auch beobachten können? <p class=„antwort rteabs–antwort“>Jedes Mal bei einem Tesla-Unfall. Besonders bei den ersten Autopilot-Modellen. Die Leute vertrauen diesen Systemen zu sehr. So nach dem Motto: „Jetzt musst du das Lenkrad noch mindestens fünf Sekunden lang halten. Wenn du die Hand nicht am Lenkrad hast, schaltet sich das System ab.“ Aber sie haben nie mit mir oder meinem Team darüber gesprochen, weil das gar nicht funktioniert. Und das ist so, weil sich das System so leicht manipulieren lässt. Wenn Sie auf Ihr Handy schauen und dann den Piepston des Tesla hören, nehmen Sie einfach die Hand hoch, richtig? Das ist unterbewusst so. Sie sind aber immer noch unaufmerksam. Und zwar deshalb, weil Sie denken, dass das System einfach weiter läuft und dass Sie eigentlich immer noch das tun können, was Sie gerade getan haben – ein Buch lesen, fernsehen oder eben auf Ihr Handy schauen. Es funktioniert also nicht, weil sich für Sie das Risiko, die Unsicherheit oder das Misstrauen gegenüber dem automatischen System nicht erhöht haben. Es reicht nicht, dass Sie sich wieder mit der Situation beschäftigen. <p class=„frage rteabs–frage“>Sie meinen, dass man aktiv Misstrauen in das System einbauen müsste, um es sicherer zu machen? <p class=„antwort rteabs–antwort“>Genau das. Wir machen gerade ein Experiment zum Thema Denial of Service. Wir ringen allerdings noch mit ethischen Fragen. Denn sobald wir darüber sprechen und die Ergebnisse veröffentlichen, werden wir erklären müssen, warum man manchmal der KI die Möglichkeit geben will, ihren Dienst auch zu verweigern. Was etwa, wenn ein Dienst verweigert wird, wenn jemand ihn wirklich braucht? Aber nehmen wir mal den Autopilot des Tesla als Beispiel. Denial of Service wäre: Ich erstelle ein Profil Ihres Vertrauens – darauf basierend, wie oft Sie das Lenkrad deaktivieren oder das Steuer loslassen. Anhand dieser Profile kann ich dann modellieren, ab wann jemand dem System völlig vertraut. Wir haben das nicht mit Tesla-Daten, sondern mit unseren eigenen gemacht. Wenn Sie dann das nächste Mal das Autofahren, würden sich in bestimmten Situationen ein Denial-of-Service einschalten. Das heißt, Sie haben in einer Zeitspanne von X keinen Zugriff auf das System. So, als würden Sie einen Teenager durch das Wegnehmen seines Smartphones bestrafen, wenn Sie es anders nicht erreichen, dass er das macht, was Sie wollen. <p class=„frage rteabs–frage“>Welche anderen Mechanismen haben Sie erforscht, um das Misstrauen in Systeme zu erhöhen? <p class=„antwort rteabs–antwort“>Zum Beispiel durch eine Art „erklärende KI“. Das System erklärt selbst seine Risiken oder Unsicherheiten. Denn alle diese Systeme bergen Unsicherheiten - keines ist 100%ig sicher. Und ein System weiß, wann es unsicher ist. Es könnte also diese Informationen in einer verständlichen Form bereitstellen, so dass die Nutzer ihr Verhalten ändern. Ein Beispiel: Ich bin ein selbstfahrendes Auto und habe alle Karteninformationen. Ich weiß, dass bestimmte Kreuzungen unfallträchtiger sind als andere. Wenn wir uns einer von ihnen nähern, würde ich sagen: „Wir nähern uns einer Kreuzung, an der letztes Jahr 10 Menschen gestorben sind. Dann kann der Fahrer sich sagen: „Oh, vielleicht sollte ich aufmerksamer sein.“ <p class=„frage rteabs–frage“>So viel zu Ihren Bedenken bezüglich unserer Tendenz, diesen Systemen zu sehr zu vertrauen. Kann das denn nicht auch Vorteile haben? <p class=„antwort rteabs–antwort“>Positiv ist, dass automatisierte Systeme im Allgemeinen besser sind als Menschen. Ich selbst würde in manchen Situationen lieber mit einem KI-System interagieren als mit bestimmten Menschen. Die Systeme haben mehr Daten; sie sind genauer. Vor allem, wenn

man sie mit einer unerfahrenen Person vergleicht <p class=„frage rteabs–frage“>Sie haben bereits vor 20 Jahren ein Mentorenprogramm für Mädchen an der High School ins Leben gerufen, lange bevor sich andere Gedanken dazu gemacht haben. Warum ist das für Sie wichtig, und warum für die Branche? <p class=„antwort rteabs–antwort“>Es gab in meinem Leben immer Personen, die mir den Zugang zu Technik und Informatik ermöglicht haben. Deshalb hatte ich auch später nie ein Problem damit, in diesem Berufsfeld zu arbeiten. Ich wollte gern für andere das Gleiche tun. Später fielen mir die vielen Menschen auf, die nicht so aussahen wie ich. Da wurde mir klar: Moment, hier gibt es definitiv ein Problem, denn die Leute haben einfach keine Vorbilder, und sie haben keinen Zugang zu bestimmten Kreisen. Dabei ist es in der Praxis so wichtig, dass jeder eine andere Erfahrung mitbringt. So wie ich bereits über Mensch-Roboter-Interaktion nachgedacht habe, bevor das überhaupt ein Thema war. Nicht, weil ich brillant war. Sondern einfach, weil ich das Problem auf eine andere Art und Weise betrachtet habe. Spreche ich mit jemand, der eine andere Sicht auf die Dinge hat, können wir uns zusammentun und sagen: „Lass uns versuchen, das Beste aus beiden Welten zu kombinieren.“ Zum Beispiel töten Airbags häufiger Frauen und Kinder als Männer. Warum? Nun, wahrscheinlich gab es damals niemanden, der sagte: „Hey, warum testen wir das nicht auch an Frauen auf dem Vordersitz?“ Es gibt viele Probleme, durch die bestimmte Gruppen von Menschen in Gefahr gebracht oder sogar getötet wurden. Ich glaube, dass einfach nicht genug Leute gesagt haben: „Hey, habt ihr auch darüber nachgedacht?“ – einfach aufgrund ihrer speziellen Erfahrungen und ihres anderen Umfelds. <p class=„frage rteabs–frage“>Wie hoffen Sie, dass sich die KI- und Robotikforschung im Laufe der Zeit weiterentwickeln wird? Was ist Ihre Vision? <p class=„antwort rteabs–antwort“>Programmieren kann so ziemlich jeder. Es gibt jetzt so viele Organisationen wie Code.org. Die Ressourcen und Werkzeuge sind da. Ich würde gerne eines Tages einen Studenten fragen: „Kennst du dich mit KI und maschinellem Lernen aus?“ und er sagt: „Dr. H, das mache ich schon seit der dritten Klasse!“ Das wäre wunderbar. Natürlich müsste mir dann wohl einen neuen Job suchen, aber das ist eine ganz andere Geschichte. Wenn man die richtigen Werkzeuge und das Wissen hat, mit KI und maschinellem Lernen umzugehen, kann man sich seine eigenen Jobs, seine eigene Zukunft, seine eigene Lösung selbst erschaffen. Das wäre mein Traum. () <hr /> URL dieses Artikels:
<small><code>https://www.heise.de/-6049154</code></small> Links in diesem Artikel:
<small><code>[1] https://www.heise.de/tr/</code></small>
<small><code>[2] mailto:bsc@heise.de</code></small>
 Copyright © 2021 Heise Medien </html>