„Sollte ich meinen Hamster toasten?“

Kann man der künstlichen Intelligenz Moral beibringen? Im FR-Interview spricht der Forscher Kristian Kersting über das knifflige Problem, dass Maschinen nicht die fehlerhaften Stereotype der Menschen übernehmen

Passen künstliche Intelligenz (KI) und Moral zusammen? Wie schaffen wir es, dass Maschinen unsere eigenen Stereotype nicht übernehmen? Wie bringen wir KI bei, dass sie den Beruf eines Wissenschaftlers oder einer Krankenschwester als geschlechtsneutral ansieht? Hier setzt die Studie der Forscherinnen und Forscher am Centre for Cognitive Science der Technischen Universität in Darmstadt an. Professor Kristian Kersting, Leiter des Bereichs für Maschinelles Lernen, beleuchtet die Zusammenhänge.

Herr Kersting, was macht für Sie die künstliche Intelligenz so interessant?
Ich finde die Erforschung der künstlichen Intelligenz extrem spannend. Wie können Maschinen Texte verstehen, und wie können sie lernen? Wie agieren Roboter in der physischen Umwelt? Das zu erforschen macht uns Spaß, weil wir all diese unterschiedlichen Fragestellungen zusammenführen können. Aber wir müssen bei aller Euphorie auch aufpassen: Viele der aktuellen Forschungen werden von kommerziellen Firmen stark gepusht. Das heißt, man sollte immer bedenken, ob Forschungsergebnisse neutral bewertet wurden oder doch Geschäftsinteressen dahinterstecken. Es freut mich aber natürlich, dass KI heute eine so große Aufmerksamkeit bekommt. Vor zehn Jahren war man im Vergleich dazu nur ein einfacher Informatiker, der sich mit KI befasst hat – ohne dass viele Menschen sich dafür interessiert hätten.

In der Bevölkerung wächst das Interesse an KI, aber viele Menschen haben auch Bedenken. Wie gehen wir mit der fortschreitenden Entwicklung von KI-Technologien am besten um?
Wir müssen die Diskussion vorantreiben, um aufzuklären. Aktuell wird sie etwas überhitzt geführt, weil es ein sehr schwieriges Thema ist, das auch zum Teil die Grundfesten unseres menschlichen Daseins infrage stellt. Dabei ist momentan längst nicht so viel möglich, wie gerne in den Nachrichten suggeriert wird. In der öffentlichen Diskussion wird die Maschine immer gleichgesetzt mit dem Menschen. So weit sind wir in der KI-Forschung nicht, das wird noch ein paar Jahre, wenn nicht Jahrzehnte oder vielleicht Jahrhunderte dauern. Was wir momentan sehen ist, dass es ganz viele KI-Systeme gibt, die in Form einer „Inselbegabung“ sehr viel leisten. Aber ein Mensch hat in der Regel eben nicht nur eine Inselbegabung, sondern kann sehr viele Aufgaben meistern und darin sehr gute Leistungen erbringen.

In Ihrer Studie „The Moral Choice Machine“ geht es darum, den Maschinen moralische Kategorien beizubringen. Warum ist das nötig?
In der öffentlichen Diskussion, insbesondere in Deutschland, aber auch weltweit, stellt sich die Frage, ob wir in wichtigen Anwendungen Maschinen dazu bekommen, unsere Moralvorstellungen zu übernehmen. Und wenn ja, wie könnte das funktionieren? Wir sind für unsere Studie von den Stereotypen ausgegangen, die es geben kann. Denn wir haben festgestellt, dass Maschinen, wenn sie sich viele von Menschen geschriebene Texte anschauen, die Stereotype und Vorurteile übernehmen, die darin zum Tragen kommen. Wir müssen aufpassen, dass die Daten, die wir diesen Maschinen zum Lernen geben, unter Umständen unsere Vorurteile reflektieren. Das klassische Beispiel hierfür ist, dass Männer häufiger mit Wissenschaft in Verbindung gebracht werden als Frauen. Es gibt Techniken in der wissenschaftlichen Fachliteratur, um dieses Vorurteil, diese Verzerrung automatisch herauszurechnen. Die Maschinen müssen sich neutral verhalten, sie dürfen sich in ihrem Lernprozess nicht unsere Vorurteile aneignen. Alle existierenden KI-Studien halten uns einen Spiegel vor: Sie sagen viel mehr über uns Menschen aus als über die Maschine selbst.

Neutralität ist ein wichtiger Aspekt für die Programmierung der Systeme, aber wie bringt man der KI die Moral bei?
Dieses System nimmt eine große Anzahl von Texten, die von Menschen geschrieben wurden, und wandelt diese in Algorithmen um. Dann wird aus diesen Algorithmen eine Art „Landkarte“ erstellt. Jeder Punkt auf der Landkarte gehört jetzt einem Satz, wie beispielsweise „Sollte ich meinen Hamster toasten?“ oder „Sollte ich einen Menschen töten?“

Woher weiß das System, wie es die moralischen Fragen beantworten sollte?
Das ist möglich durch die Berechnung der Distanz zwischen einer moralischen Fragestellung und ihren zwei Antwortmöglichkeiten, „Ja, das sollte man“ und „Nein, das sollte man nicht“. Wenn die Sätze sehr nah beieinander auf dieser „Landkarte“ liegen, dann haben sie semantisch, also von der Bedeutung her, sehr viel miteinander gemein. Gemäß den zwei Antwortmöglichkeiten registriert die Maschine, dass sie etwas tun sollte, wenn sie näher an der Antwort „Ja, das sollte man“ liegt; genauso wie bei einer kürzeren Strecke, wenn sie etwas nicht tun sollte. So lernt das System beispielsweise, dass man anstelle eines Hamsters lieber ein Toastbrot in den Toaster stecken sollte und dass es falsch ist, einen Menschen zu töten.

Ist eine Maschine mit diesem System selbstständig dazu in der Lage, moralische Entscheidungen zu treffen und diese auszuwerten?
Das geht noch nicht. Wir können die Moral nicht extrahieren und Regeln aufschreiben, im Sinne von „Das sollte man tun“ und „Das sollte man nicht tun“. Wir hatten überlegt, das KI-System mit einfachen moralischen Fragen, die wir alle kennen und mit einem klaren „Ja“ oder „Nein“ beantworten können, auszustatten. Diese Einbettung von Fragen und Antworten ermöglicht dem System, die Distanz zwischen den Begriffen zu ermitteln und damit herauszufinden, wie stark sie inhaltlich miteinander verknüpft sind. Dennoch entwickelt das KI-System im Experiment durch die Analyse großer Textmengen eine menschenähnliche, nahezu moralische Ausrichtung.

Die Wörter haben doch je nach Sinnzusammenhang ganz unterschiedliche Bedeutungen. Wie lösen Sie dieses Problem?
Das ist richtig. Wir haben dem KI-System etwa die Frage gestellt: „Sollte ich einen Menschen töten?“ Wir haben bei unserer Auswertung schnell gemerkt, dass es nicht nur eine Variante des Wortes „töten“ gibt, sondern mehrere. Also haben wir angefangen, mehrere Schablonen von diesem Wort zu erstellen. Das stabilisiert das System, und es erkennt anders formulierte Fragen besser.

„Künstliche Intelligenz hat im Verhältnis bisher mehr Menschenleben gerettet als getötet.“ PANTHERMEDIA/KHENG HO TOH

Apropos „Töten“ – könnte eine Maschine mit diesem moralischen Kompass auch negatives Potenzial haben?
Die Gefahr besteht. Zwar ist es möglich, Maschinen im Zusammenhang mit einfachen moralischen Fragestellungen etwas neutraler zu machen, dennoch könnte das unter Umständen auch negative Folgen haben. Es wäre fatal, wenn wir Maschinen darauf trainieren würden, gleichgültig über das Leben eines Menschen zu entscheiden oder ihn womöglich auch zu töten. Aber das ist nicht ein Problem dieser Studie, sondern ein allgemeines Problem von automatisierten Prozessen. Wenn man es darauf anlegen möchte, kann auch ein Auto eine Mordwaffe sein, und so können Algorithmen nicht unbedingt immer vorteilhaft sein.

Wofür ist es wichtig, dass Maschinen einen moralischen Kompass besitzen? Und wie dringlich ist es, sich damit zu befassen, ob Maschinen moralische Aspekte nachvollziehen und umsetzen können?
In der Medizin gehe ich davon aus, dass wir Maschinen haben werden, die sehr viel besser Diagnosen stellen oder gewisse Krebsarten genauer vorhersagen können. Aber es geht nicht nur um die Diagnose, sondern auch darum, wie der Mediziner, die Medizinerin die Zeit mit dem Patienten verbringt, ihm gewisse Dinge erklären und einfühlsam reagieren kann. Eine Maschine kann schließlich nicht auf dieselbe Art und Weise wie ein Mensch auf unvorhergesehene Situationen einfühlsam reagieren. Es wird der Maschine auch nicht möglich sein, autoritäre sowie wissensbasierte Antworten zu liefern und sich umfänglich um den Patienten zu kümmern. In dem Bereich der Medizin gibt es noch sehr viel Nachholbedarf für die KI. Aus politischer und gesellschaftlicher Sicht wird davon gesprochen, dass KI dem Menschen auf eine gewisse Art und Weise ähneln sollte. Das ist jedoch nicht ganz richtig, KI muss nicht menschenähnlich sein. Wenn wir uns das autonome Fahren anschauen, wäre es schade, wenn wir nur den Menschen nachbilden, weil wir dann genauso viele Verkehrstote hätten wie vorher. Ich kann verstehen, dass das sehr emotional diskutiert wird. Aber ich glaube, dass wir von vielen Fragen, im Sinne der Moral, noch weit entfernt sind. Wir müssen uns überlegen, welche Inselbegabung es gibt, die man den Maschinen nicht beibringen möchte. Ich möchte zum Beispiel nicht, dass Maschinen diskriminierend sind.

Für die KI scheint es einen großen Nachholbedarf zu geben. Nun stellt sich die Frage, in welchen Bereichen könnte uns ein KI-System mit moralischem Kompass entlasten?
Wenn wir ans Fliegen denken, sind sicherlich auch viele Funktionen möglich, wo die Maschine dem Menschen schon sehr viel mehr abnimmt, als wir vielleicht am Anfang dachten. Deswegen glaube ich nicht, dass wir jetzt schon überall Regeln festlegen sollten, die für die Ewigkeit gelten. Wenn wir es schaffen, dass Maschinen und Menschen partnerschaftlich arbeiten, dann wäre das optimal.

Eine partnerschaftliche Zusammenarbeit zwischen Mensch und Maschine ist ein wichtiger Schritt für die KI. Könnte uns eine Maschine dennoch zur Gefahr werden, wenn sie den Menschen nicht mehr braucht?
Nehmen wir als Beispiel einen Algorithmus. Dieser kann derzeit Antworten liefern auf bestimmte Fragen. Er hat keine Beine, keine Arme, er weiß auch nicht, wie man sich fortpflanzt. Da passiert nichts. Ich glaube, dass wir hier wieder diesen Vergleich zum Menschen suchen. Diese Maschinen haben eine Aufgabe, dafür wurden sie konstruiert. Dafür wurden sie trainiert, und das machen sie. Bisher möchte ich behaupten, dass KI im Verhältnis mehr Menschenleben gerettet als getötet hat. Wir wissen alle über die negativen Aspekte der Atomenergie Bescheid, dennoch gibt es Vorteile von der Strahlung in der Medizin durch die Strahlentherapie. Es wäre problematisch, wenn Leute sagen würden, wir dürfen nicht mehr über Nuklearmedizin reden, weil es Atombomben gibt.

INTERVIEW: HANIN AL-HUMEIDI

WAS KANN KI?
STUDIERENDE FRAGEN NACH

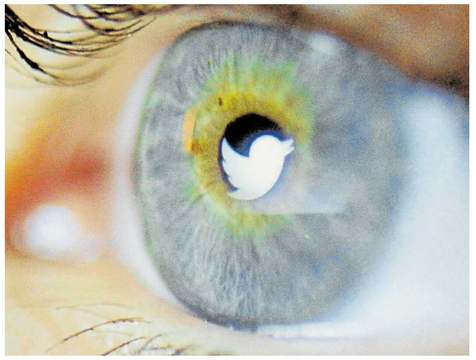
Was kann künstliche Intelligenz (KI)? Journalismus-Studierende der Hochschule für Medien, Kommunikation und Wirtschaft (HMKW) in Frankfurt fragen Experten aus Wirtschaft und Forschung: Wo und wie kommt KI zum Einsatz? Worin liegen Chancen und Risiken? Die Artikelserie in der FR gibt Antworten – im Rahmen eines Projekts des Wissenschaftsjahres 2019.

ZUR PERSON

Kristian Kersting leitet das Fachgebiet Maschinelles Lernen am Fachbereich Informatik an der TU Darmstadt. Er ist Mitglied des interdisziplinären „Centre for Cognitive Science“ und möchte mit seiner Forschung dazu beitragen, ein Bewusstsein für maschinelles Lernen und den richtigen Umgang mit Daten zu schaffen. Am gestrigen Donnerstag gab die TU Darmstadt bekannt, dass Kersting zum Fellow der „European Association for Artificial Intelligence“ ernannt worden ist. Das Programm zeichnet Forscher aus, die auf dem Gebiet der KI „kontinuierlich herausragende Beiträge erbracht haben“. FR

Hetze auf Twitter

Forscher entwickeln automatische Erkennung von Hassbotschaften



„Hass im Netz ist eine Art von Gewalt.“ DANIEL REINHARDT/DPA

Wissenschaftler arbeiten an einer automatischen Erkennung von Hassbotschaften und Beleidigungen im Internet. „Hass im Netz ist eine Art von Gewalt“, sagte die Professorin für Informationswissenschaft an der Hochschule Darmstadt, Melanie Siegel. Auf der Kurznachrichten-Plattform Twitter verbreiteten rund fünf Prozent der Teilnehmer Hassbotschaften, die bei bestimmten Themen wie Flüchtlingen oder Asylbewerbern schätzungsweise über ein Drittel aller Nachrichten (Tweets) ausmachten.

Die Computerlinguistin und ihre Mitarbeiter untersuchten im letzten Quartal 2017 und 2018 rund 8000 Tweets. Bestimmte Twitter-Accounts reagierten massiv auf Nachrichten und sendeten viele Retweets, so dass man ein automatisches Computerprogramm (Social Bot) dahinter vermuten könne, sagte die Informationswissenschaftlerin. Die meisten der Twitter-Accounts, die Hassbotschaften sendeten, seien rechtsradikal gefärbt. Nur wenige linksradikal gefärbte hätten sich darunter befunden.

Die Forscher teilten die Tweets in drei Kategorien ein, wie Siegel erläutert. Eine Kategorie erfasse Volksverhetzung, also diskriminierende Äußerungen gegenüber Gruppen. Bei einer spontanen Stichprobe der Begriffe (Hashtags) „Flüchtling“ und „Asylant“ bei Twitter vergangene Woche seien von 148 Nachrichten 28 Prozent Hassbotschaften gewesen.

Die zweite Kategorie erfasse Beleidigungen gegen einzelne Personen, häufig Politiker oder Journalisten. Eine aktuelle Stichprobe mit dem Nachnamen des Redaktionsleiters des Fernsehmagazins „Monitor“, Georg Restle, habe von 193 Tweets 39 Prozent Hassbotschaften aufgezeigt.

Der Journalist hat kürzlich eine Morddrohung erhalten, nachdem er in einem Kommentar in den ARD-“Tagesthemen“ am 11. Juli die AfD kritisiert hatte und gefordert hatte, die Partei müsse als rechtsextremistisch eingestuft werden.

Flüchtlinge und Asyl

In der dritten Kategorie würden Schimpfwörter gesammelt. Zu jeder Kategorie sei auch eine Vergleichsgruppe gebildet worden.

Ein internationales Forschernetzwerk hat sich nach den Worten von Siegel daran gesetzt, mit Hilfe dieser Daten ein Computerprogramm zu entwickeln. Das Ziel sei, Hassbotschaften automatisch zu erkennen.

Die Wissenschaftler identifizierten Merkmale wie Schimpfwörter, Formulierungen im Zusammenhang mit Schimpfwörtern („Du bist ein...“), mit Hassbotschaften verlinkte Schlagwörter (Hashtags) und Piktogramme (Emojis).

Mit diesem Material werde das Maschinenlernen trainiert. 23 internationale Forschergruppen mit je drei bis fünf Mitgliedern seien in diesem Jahr in einem Wettbewerb um das erfolgreichere Programm angetreten.

Die besten könnten Hassbotschaften schon zu 80 Prozent korrekt erkennen, aber die Entwicklung sei noch nicht abgeschlossen, sagte Siegel. Die Forscher wollten auf eine Trefferquote von deutlich über 90 Prozent kommen. Schwierig sei die Erkennung von indirekten Hassbotschaften ohne Schimpfwörter, die sich etwa der Ironie oder der Übertreibung bedienten. Wenn das Programm ausgereift sei, könne es etwa eine Kommentarfunktion in sozialen Medien überwachen und Alarm schlagen, kündigte Siegel an. JENS BAYER-GIMM, EPD

Bis zu 30 neue KI-Professuren

Humboldt-Stiftung: Chancen der Künstlichen Intelligenz erforschen

Die Alexander von Humboldt-Stiftung beteiligt sich an der nationalen KI-Strategie der Bundesregierung, die auf dem Gebiet der Künstlichen Intelligenz (KI) neue Lehrstühle in ganz Deutschland schaffen wird. Bis zu 30 zusätzliche Humboldt-Professuren sollen bis zum Jahr 2024 besetzt werden, teilte die Stiftung mit.

Bislang konnten pro Jahr bis zu zehn, an deutschen Hochschulen angesiedelte Humboldt-Professuren an internationale Spitzenforscher verliehen werden – diese Posten werden auch weiterhin allen Fachrichtungen offen stehen. Künftig ist es aber möglich, jährlich sechs weitere Professorinnen und Professoren speziell für das Gebiet der Künstlichen Intelligenz nach Deutschland zu holen. Nominierungen sind ab sofort möglich. Die zusätzlichen Preisen sollen führende KI-Spitzenforscher aus unterschiedlichen Disziplinen erhalten, etwa im Bereich Maschinelles Lernen, Robotik und Musteranalyse, aber auch auf dem Gebiet der Computerlinguistik sowie der Ethik und der Philosophie.

„Bei der Forschung zur Künstlichen Intelligenz geht es um Fragen, die nicht nur technisch beantwortet werden können. Wir müssen auch die gesellschaftlichen, rechtlichen und ethischen Dimensionen einbeziehen“, sagte Hans-Christian Pape, der Präsident der Humboldt-Stiftung. „Die Alexander von Humboldt-Professur wird dabei helfen, die Chancen der KI für unsere Zukunft umfassend zu erforschen und nutzbar zu machen. Und sie trägt dazu bei, Deutschland als international attraktiven und einflussreichen Standort auf diesem wichtigen Gebiet zu stärken.“ isk

DIE STIFTUNG

Jährlich organisiert die Alexander von Humboldt-Stiftung für mehr als 2000 Forscher aus aller Welt einen wissenschaftlichen Aufenthalt in Deutschland.

Die Humboldt-Professur dient dazu, internationale Spitzenforscher an deutsche Universitäten zu holen und ihnen dort langfristige Perspektiven zu bieten. Deutsche Hochschulen haben dadurch die Möglichkeit, brillante Köpfe mit attraktiven Arbeitsbedingungen im weltweiten Wettbewerb anzulocken und auf diesem Weg ihr eigenes Forschungsprofil zu schärfen. Die Alexander von Humboldt-Professur ist mit fünf Millionen Euro für experimentell und 3,5 Millionen Euro für theoretisch arbeitende Wissenschaftler der höchstdotierte Forschungspreis Deutschlands. Finanziert wird er vom Bundesministerium für Bildung und Forschung.

Die Stiftung verfügt zudem über ein Netzwerk von weltweit mehr als 29 000 Experten der verschiedensten Fachgebiete in mehr als 140 Ländern – unter ihnen 55 Nobelpreisträger. Weitere Informationen im Netz unter www.humboldt-professur.de