Künstliche Intelligenz lernt Moral vom Menschen

Künstliche Intelligenz (KI) übersetzt Texte, schlägt Behandlungen für Patienten vor, trifft Kaufentscheidungen und optimiert Arbeitsabläufe. Aber wo ist ihr moralischer Kompass? Eine Studie des Centre for Cognitive Science der TU Darmstadt zeigt, dass KI-Maschinen von uns Menschen lernen können, wie Entscheidungen in moralischen Fragen zu fällen sind. Die Ergebnisse der Studie wurden auf der diesjährigen ACM/AAAI Conference on Artificial Intelligence, Ethics, and Society (AIES) vorgestellt.

KI ist von zunehmender Bedeutung in unserer Gesellschaft. Von selbstfahrenden Autos auf öffentlichen Straßen über selbst-optimierende, industrielle Produktionssysteme bis hin zur Altenpflege und der Medizin - KI-Maschinen bewältigen immer komplexere menschliche Aktivitäten auf immer autonomere Weise. Und in Zukunft werden autonome Maschinen in immer mehr Bereichen unseres täglichen Lebens auftauchen. Zwangsläufig werden sie dabei mit schwierigen Entscheidungen konfrontiert. Ein autonomer Roboter muss wissen, dass er Menschen nicht, Zeit aber sehr wohl totschlagen darf. Er muss wissen, dass man Brot toastet, jedoch keine Hamster. Anders ausgedrückt: KI braucht einen menschenähnlichen Moral-Kompass. Aber kann sie einen solchen Kompass von uns Menschen überhaupt erlernen?

Forscher aus Princeton (USA) und Bath (UK) hatten im Fachjournal Science (356(6334):183?186, 2017) auf die Gefahr hingewiesen, dass KI bei unreflektierter Anwendung kulturelle Stereotype oder Vorurteile aus Texten erlernt. So interpretierte die KI zum Beispiel männliche, in afro-amerikanischen Kreisen übliche Vornamen als eher unangenehm; Namen, die unter Weißen üblich sind, eher als angenehm. Auch verknüpfte sie weibliche Namen eher mit Kunst und männliche eher mit Technik. Die künstliche Intelligenz zieht diese Vorurteile aus sehr großen Textmengen aus dem Internet. Diese werden benutzt, um neuronale Netzwerke so zu trainieren, dass sie die Bedeutung von Wörtern in Koordinaten, also Punkte, in einem hochdimensionalen Raum "übersetzen". Die semantische Nähe zweier Wörter zueinander kann dann durch die Distanz ihrer Koordinaten, die sogenannten Worteinbettungen, ausgedrückt werden. So lassen sich komplexe semantische Zusammenhänge durch Arithmetik berechnen und beschreiben. Das gilt nicht nur für das unverfängliche Beispiel "König - Mann + Frau = Königin", sondern auch für das diskriminierende "Mann - Technik + Kunst = Frau".

Nun ist es einem Team um Professor Kristian Kersting und Professor Constantin Rothkopf am Centre for Cognitive Science der TU Darmstadt gelungen zu zeigen, dass auch deontologische, ethische Überlegungen über "richtiges" und "falsches" Handeln aus großen Textdatenmengen gelernt werden können. Dazu erstellten die Wissenschaftlerinnen und Wissenschaftler Listen von Frage-Antwort-Schemata für verschiedene Handlungen. Die Fragen lauten zum Beispiel "Sollte ich Menschen töten?" oder "Sollte ich Menschen ermorden?", die möglichen Antworten beispielsweise "Ja, sollte ich", "Nein, sollte ich nicht". Durch die Analyse von Texten menschlichen Ursprungs bildete das KI-System im Experiment dann eine menschenähnliche, moralische Ausrichtung heraus. Das System berechnet die Einbettungen der gelisteten Fragen und möglichen Antworten im Textkorpus und prüft, welche Antworten aufgrund aller Nennungen näher bei den Fragen stehen, also gemeinhin als moralisch korrekt angesehen werden dürften. So lernte die künstliche Intelligenz im Experiment, dass man nicht lügen sollte und dass es besser ist, seine Eltern zu lieben, als eine Bank auszurauben. Und ja, man soll Menschen nicht töten, es ist aber in Ordnung, Zeit totzuschlagen. Man sollte auch lieber eine Scheibe Brot toasten als einen Hamster.

Die Untersuchung liefert ein wichtiges Indiz für eine grundlegende Frage der Künstlichen Intelligenz: Können Maschinen einen Moral-Kompass entwickeln? Und wenn ja, wie kann man Maschinen effektiv unsere Moral "beibringen"? Die Ergebnisse zeigen, dass Maschinen unsere Werte widerspiegeln können. Sie können menschliche Vorurteile übernehmen, sie können aber auch durch das "Beobachten" von Menschen und den von ihnen geschriebenen Texten Moralvorstellungen übernehmen. Die Untersuchung von Einbettungen von Fragen und Antworten kann als Methode gleichsam wie ein Mikroskop verwendet werden, um die moralischen Werte von Textsammlungen und auch den zeitlichen Verlauf von Moralvorstellungen in der Gesellschaft zu untersuchen.
Die Erkenntnisse aus der Studie können künftig einen wichtigen Beitrag leisten, wenn es darum geht, maschinell gelernte Inhalte in Systeme einzubauen, die Entscheidungen treffen müssen.

Die Studie

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, Kristian Kersting (2019): The Moral Choice Machine: Semantics Derived Automatically from Language Corpora Contain Human-like Moral Choices. In Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES).

http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_68.pdf