Rote Linien

Unternehmen setzen ethische Grenzen für maschinelles Lernen

So segensreich künstliche Intelligenz auch sein mag: Sie ist fehleranfällig, lässt sich in bisher nicht dagewesener Weise zur Manipulation einsetzen und schwer kontrollieren. Erste Firmen haben daher ethische Leitlinien für den Einsatz der Technik formuliert.

Von Jo Bager und Martin Fischer

Neue Machine-Learning-Techniken machen es immer einfacher, täuschend echt wirkende Video- und Audio-Inhalte zu generieren. Google hatte zum Beispiel bei seiner Entwicklerkonferenz I/O eine neue Funktion seines Assistenten vorgestellt, bei der dieser telefonisch Termine aushandelt. Seinem menschlichen Gesprächspartner machte der Assistent bei der Demonstration nicht klar, dass er mit einer KI spricht (c't 12/2018, S. 18).

In den letzten Monaten haben KI-generierte Inhalte vor allem in Form von Deep Fakes Furore gemacht, bei denen Politikern falsche Äußerungen in den Mund gelegt oder die Gesichter von Prominenten in Pornovideos montiert wurden (c't 8/2018, S. 100). Das US-Verteidigungsministerium finanziert ein Projekt, innerhalb dessen erforscht werden soll, ob sich derartige Deep Fakes eines Tages selbst von dafür spezialisierten KI-Systemen nicht mehr aufdecken lassen. Dafür sollen Video-, Foto- und Ton-Inhalte möglichst realistisch gefälscht und zugleich Werkzeuge entwickelt werden, mit denen sich diese Fälschungen automatisch identifizieren lassen.

Wolf oder Husky?

So gut KI in vielen Fällen bereits funktionieren mag (siehe S. 52) – sie macht aber auch immer noch Fehler: Christian Bauckhage vom Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) zeigte bei einem Vortrag auf der Cebit an mehreren Beispielen das Versagen von KI in bestimmten Situationen auf. So sei etwa ein Husky von einer Bilderkennung als Wolf identifiziert worden. Der Grund: Alle Wolfsbilder, mit denen jenes neuronale Netzwerk trainiert wurde, enthielten Schnee – wie auch das Husky-Bild.

Der Professor führte auch ein Bilderkennungsproblem von Google Photos aus dem Jahre 2015 an, dessen Algorithmus dunkelhäutige Menschen als Gorillas kategorisierte. Bauckhage zufolge hätte dies unter anderem daran gelegen, dass die Bilderkennungssoftware hauptsächlich mit Bildern von Gesichtern hellhäutiger Menschen trainiert wurde. Machine Learning braucht laut Bauckhage nicht Big Data als Grundlage, sondern vielmehr Thick Data in Kombination mit reichlich Forschung, um dem Ziel einer verstehbaren und interpretierbaren KI näher zu kommen.

Eben dies forderte – ebenfalls auf der Cebit – auch Dr. Sandra Wachter von der Universität Oxford ein. Für die Juristin ist es essenziell, dass die von KI getroffenen Entscheidungen beziehungsweise erzielten Ergebnisse nachvollziehbar sind. Nur dann ließen sich Entscheidungen auch hinterfragen – und künftige beeinflussen. Jeder solle eine klare Antwort auf die Frage einfordern können, weshalb eine KI eine bestimmte Entscheidung auf genau diese Weise getroffen hat.

Leitlinien

Die Kritik an der Intransparenz künstlicher Intelligenz und den Auswirkungen auf Verbraucher sind offensichtlich bei Unternehmen wie der Telekom und Google angekommen. Beide Firmen haben sich ethische Leitlinien für den Einsatz der neuen Technik gegeben. Die Telekom bleibt mit ihren neun knappen Punkten im Ungefähren: „Wir legen das Fundament. Gründliche Analyse und Evaluierungen als Basis für die Weiterentwicklung und stete Verbesserung unserer KI-Systeme."

Google geht da schon mehr ins Detail. So besagt eine der sieben Regeln Googles, dass KI einen sozialen Nutzen haben muss, etwa in den Bereichen Gesundheitsversorgung, Sicherheit, Energie, Verkehr, Produktion oder Entertainment. Dabei wolle Google die kulturellen und sozialen Gegebenheiten sowie die Rechtsnormen der Länder beachten, in denen die Techniken eingesetzt werden. Das Unternehmen schließt die Beteiligung an der Entwicklung von KI-Algorithmen für militärische Waffensysteme grundsätzlich aus. Das gelte auch für Technologien, die dazu geeignet wären, eine großflächige Überwachung von Menschen zu ermöglichen. *(jo@ct.de)* c't

Die Richtlinien von Telekom und Google: ***ct.de/ygkd***

Microsofts lernender Chatbot Tay entwickelte sich in seiner ursprünglichen Version zum Rassisten und Sexisten.