zeit.de

34C3: Notwehr against the machine

Von Patrick Beuth, Leipzig

5-6 Minuten

"Irgendwann in den nächsten Jahren wird irgendein Gremium aus Imperiumsbürokraten vor der Entscheidung stehen: Sollen wir jetzt diese selbstlernende, fehlerbeseitigende, uns turmhoch überlegene künstliche Intelligenz anschalten – oder lieber nicht? Und dann wird einer sagen: Jetzt hamwa's bezahlt, jetzt schalten wir's auch an." So stellt sich Marc-Uwe Kling den Anfang vom Ende der Menschheit vor. Beim Autor der *Känguru-Chroniken* und der Science-Fiction-Satire *Qualityland* klingt das lustig, aber auf dem 34. Chaos Communication Congress (34C3) ist seine Geschichte durchaus als Warnung zu verstehen.

Künstliche Intelligenz und deren Teilbereiche Machine Learning und Deep Learning sind in diesem Jahr nicht nur in Forschung, Wirtschaft, Politik und Medien, sondern auch auf dem 34C3 ein zentrales Thema. In mehreren Vorträgen geht es um selbstlernende Maschinen, ihre Auswirkungen auf die Gesellschaft – und ihre Kontrolle.

"Geldwäsche für Vorurteile"

Schon in der Eröffnungsrede warnt der Science-Fiction-Autor Charles Stross: "Wir müssen eine Strategie entwickeln, um Herr über diese Dinge zu werden, bevor sie Herr über uns werden." Stross denkt an künftige Machine-Learning-basierte Systeme, die extrem personalisierte, staatliche Propaganda im Netz verbreiten, Videos und Audioaufnahmen nach Belieben fälschen oder einzelne Internetnutzer derart genau analysieren, dass sie ihnen stets exakt jene Reize bieten können, die sie noch länger am Bildschirm halten.

Die beiden deutschen KI-Forscher Hendrik Heuer und Karen Ullrich warnen tags darauf vor Verzerrungen in Machine-Learning-Systemen, die bereits verwendet werden. "Eine Art Geldwäsche für Vorurteile", nennt Heuer solche Modelle, die einseitige Zielsetzungen haben, oder Trainingsdaten verwenden, die nicht repräsentativ sind. Die einseitige Vorhersage für die Risikoeinschätzung von schwarzen gegenüber weißen Strafgefangenen, die *ProPublica* im Mai 2016 in einem aufsehenerregenden Bericht beschrieben hatte, ist eines seiner Beispiele: Die Diskriminierung von Afroamerikanern, die in den Trainingsdaten des US-Strafverfolgungs- und Justizsystems steckt, wird durch die Entscheidung eines emotionslosen Automaten nicht entfernt – sie ist nur nicht mehr so einfach sichtbar.

Mit Verweis auf Facebooks Algorithmus, der den Newsfeed

der Nutzer individuell, aber nach nicht nachvollziehbaren Maßgaben sortiert, sagt Heuer: "Solche Systeme, die im Schatten agieren, rekonfigurieren menschliche Beziehungen", sie "beeinflussen die Weltsicht".

Nicht nur Facebooks Gesichtserkennung überlisten

Markus Beckedahl, der Gründer von *netzpolitik.org*, fordert in seinem Vortrag deshalb demokratisch legitimierte Kontrollinstanzen "irgendwo zwischen Datenschutzbehörden und Kartellämtern". Sie sollen das *algorithmic decision making* nicht nur bei Facebook, sondern zum Beispiel auch im Gesundheitsbereich oder bei der Kreditvergabe nachvollziehbar machen. Notfalls auf Kosten einiger Geschäftsgeheimnisse, findet er.

Und zur praktischen Gegenwehr ruft die US-Amerikanerin Katharine Jarmul auf. Normalerweise berät sie Unternehmen beim Einsatz von Machine-Learning-Systemen. Auf dem Kongress zeigt sie, "wie man künstliche Intelligenzen hereinlegt". Adversarial (Machine) Learning heißt ihr Ansatz, was letztlich bedeutet, eine KI zu hacken. Mit speziell präparierten Trainingsdaten oder Inputs kann eine KI dazu gebracht werden, unsinnige oder gezielt falsche Ergebnisse zu liefern. Das fängt an bei Spamfiltern und Antivirenprogrammen, die ausgetrickst werden können. Und es endet mit fahrerlosen Autos, die mit unauffällig manipulierten Straßenschildern in die Irre geleitet werden könnten.

Jarmul selbst hat ein Foto von sich Schritt für Schritt so verändert, dass es Facebooks Gesichtserkennung überfordert hat. Ihr Ziel war es, Facebook glauben zu lassen, sie sei eine Katze. Das hat nicht geklappt, aber zumindest hat Facebook bei einem hinreichend veränderten Bild aufgehört, ihr vorzuschlagen, sich selbst namentlich zu markieren. Das System des Unternehmens war sich an dieser Stelle also nicht mehr sicher, ein Bild von Jarmul oder überhaupt das eines menschlichen Gesichts vorgelegt bekommen zu haben.

Es gibt eine Reihe von Open-Source-Werkzeugen, mit denen solche Experimente möglich sind. Sie heißen DeepFool, cleverhans oder deep-pwning. Jarmul ruft die Hacker-Community auf, sie für "gute Zwecke" zu nutzen. Für die Umgehung allgegenwärtiger Überwachungstechnik etwa, oder schlicht für die Untersuchung von Black-Box-Systemen, die das Leben von Menschen beeinflussen können, aber intransparent sind.