Natürlich intelligent

Der KI-Newsletter von ZEIT ONLINE

von **Marie Kilg**
KI-Kolumnistin

Liebe Lesende,

wenn man beim KI-Training nicht aufpasst, kommt es vor, dass die Maschinen nichts lernen, was über die Trainingsdaten hinausgeht.

Ein Beispiel: Wenn ein KI-System Katzen auf Bildern erkennen soll, dann muss es mit vielen Tausend Bildern von Katzen trainiert werden. Das Ziel ist, dass die KI nach dem Training Katzen auch auf unbekannten Bildern korrekt erkennen kann. Ein Problem wäre, wenn sie etwas als Katze erkennt, das keine ist. Aber wenn die KI nur genau das als Katze erkennt, was sie schon gesehen hat, ist sie auch nicht besonders nützlich. Sie hat zwar verstanden, auf welchen Bildern in den Trainingsdaten Katzen zu sehen sind, aber ihre Definition von "Katze" ist zu eng. Eine Katze, die eine andere Fellfarbe hat, erkennt sie nicht.

Dieses Problem nennt man im Machine Learning *overfitting*. Beim Training wird die KI für richtige Antworten belohnt. Aber wenn der Algorithmus nicht richtig justiert ist, passt das Modell am Ende zu genau zur Aufgabenstellung, und die Maschine kann nicht generalisieren.

Tatsächlich ist das Problem nicht auf KI beschränkt. Auch Menschen overfitten gelegentlich. Wir orientieren uns an messbaren Metriken, aber manchmal führt das dazu, dass wir das Ergebnis, das die Metrik abbilden soll, nicht erreichen.

Wenn ich die Zahl der Sit-ups, die ich schaffe, erhöhe, fühlt sich das gut an. Aber eigentlich geht es mir darum, fit zu sein. Durch die übertriebene Aufmerksamkeit auf diese eine Zahl vernachlässige ich vielleicht ein ganzheitlicheres Training. Ein Freund von mir lernt Spanisch mit der Sprachlern-App Duolingo. Er kann jetzt in hundert Prozent der Fälle auf korrektem Spanisch sagen, dass das Kaninchen den Fisch liebt. Im Alltag hilft das wenig. Aber die App belohnt ihn fleißig mit virtuellen Trophäen.

Foto: Thomas Steineder

Immer dann, wenn wir unser Verhalten auf einen bestimmten Wert, eine Metrik hin optimieren, besteht die Gefahr, dass es dem zugrunde liegenden Ziel eigentlich abträglich ist. Clickbait-Seiten mit reißerischen Titeln oder Bildergalerien gibt es, weil Klicks Geld bringen: Die Werbeeinnahmen steigen durch die erhöhten Abrufzahlen. Auf den ersten Blick lassen sich die messbaren Zahlen also unmittelbar abbilden auf das Ziel, die Einnahmen zu erhöhen. Dass die Qualität der Inhalte und der Ruf der Marke leiden, tritt in den Hintergrund, weil die Klicks so leicht messbar sind.

Geld verdienen heißt nicht immer Wert schöpfen. Aber da Geld eine so leicht messbare Größe ist – und in so vielen Bereichen eine gute Annäherung an Erfolg – ist es schwer, die Situationen zu erkennen, in denen wir overfitten.

Umso wichtiger ist es, immer wieder zu überprüfen, ob man wirklich seinem Ziel näherkommt oder ob es nur so scheint.

Bei KI können Fehler wie *overfitting* das ganze System unbrauchbar machen. Oder, wenn sie unerkannt bleiben, diskriminierend. Zum Beispiel, wenn ein Kreditvergabesystem aufgrund des fehlerhaften Trainings einen Zusammenhang zwischen dem Wohnsitz und der Rückzahlwahrscheinlichkeit fabuliert. Oder ein Gesichtserkennungsprogramm nur weiße Gesichter als Gesichter erkennt, weil andere Hautfarben in den Trainingsdaten unterrepräsentiert waren.

Abgesehen davon, dass wir sorgfältig auswählen sollten, in welchen Lebensbereichen wir Verantwortung an Algorithmen abgeben, können die Lösungsansätze gegen KI-*overfitting* vielleicht auch eine Hilfe für unser Alltags-*overfitting* sein. Im Machine Learning bekämpft man das Problem, vereinfacht gesagt, in dem man möglichst divers denkt: Sorgfältig ausgewählte und vielfältige Trainingsdaten machen das System robuster. Man sollte seine Ergebnisse überprüfen, zum Beispiel indem man sich der gleichen Frage mit unterschiedlichen Modellen nähert. Und man muss darauf achten, dass man nicht immer dieselben alten Daten bemüht, sondern auch mal frisches, bisher ungesehenes Material in die Maschine gibt. Alles Dinge, die menschlichen Intelligenzen auch nicht schaden.

Overfitten Sie auch? Vielleicht fallen Ihnen ja auch Beispiele aus Ihrem Alltag ein – schreiben Sie mir! Ich verspreche, ich werde mich mit den Inhalten Ihrer Nachrichten auseinandersetzen und nicht nur die Antwortrate messen.

Mit besten Grüßen

Marie Kilg

Links zum Weiterlesen

- **"Overfitting" bei Bildgeneratoren wie Midjourney führt zur Ausgabe von urheberrechtlich geschütztem Material.** "We Asked A.I. to Create the Joker. It Generated a Copyrighted Image." (*The New York Times*)
- **Wenn man es darauf anlegt, kann man ChatGPT dazu bringen, Trainingsdaten auszugeben.** Der Blogger Devanash erklärt, warum. "Extracting Training Data from ChatGPT [Breakdowns]" (*Artficial Intelligence Made Simple*)
- **Wie problematisch der Einsatz von Machine Learning in Regierungsinstitutionen sein kann, zeigt zum Beispiel die niederländische Kindergeldaffäre.** *Wired* hat darüber einen guten Longread veröffentlicht. "This Algorithm Could Ruin Your Life" (*Wired*)

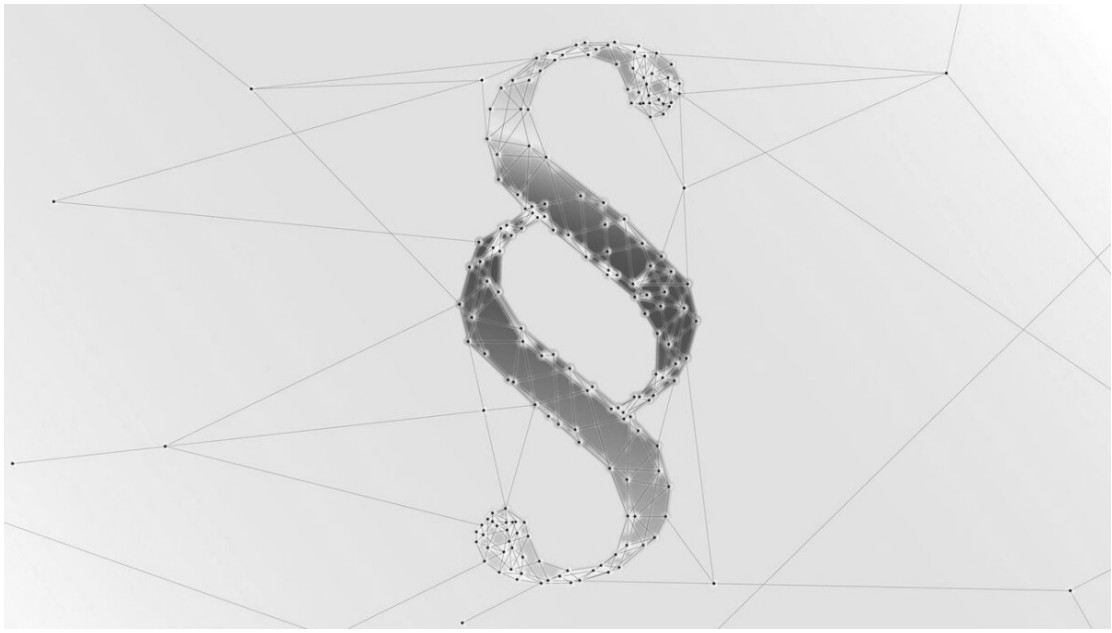
Über KI nachdenken

- **KI lernt jetzt von Babys**: "This baby with a head camera helped teach an AI how kids learn language" (*MIT Technology Review*)
- **Was KI-Sicherheit für KI-Forschende in China bedeutet:** "ChinAI #253: Tencent Research Institute releases Large Model Security & Ethics Report" (*ChinAI Newsletter*)
- **Das EU-Parlament und die Kommission haben sich auf neue Richtlinien zum Schutz von Frauen vor Gewalt verständigt, die auch Cybergewalt einschließen:** "Deepfakes und Penisbilder sollen bald kriminalisiert werden" (*Golem*)

Mit KI herumspielen

- **Mit Viggle lassen sich Videos von animierten Charakteren erstellen.** Noch kostenlos und über Discord verfügbar.
- **Zeit für eine neue Website?** Probieren Sie mal, ob Sie mit einer Kurzbeschreibung an eine KI zum gewünschten Ergebnis kommen. Zum Beispiel mit Durable, Unbounce oder Hostinger.
- **Ein Delorean im Wohnzimmer innerhalb von 35 Sekunden.** Generative Text-To-3D-Modelle und Apples Vision Pro machen's möglich: @ammaar auf X

AI Act: Wie die EU KI für die ganze Welt regelt

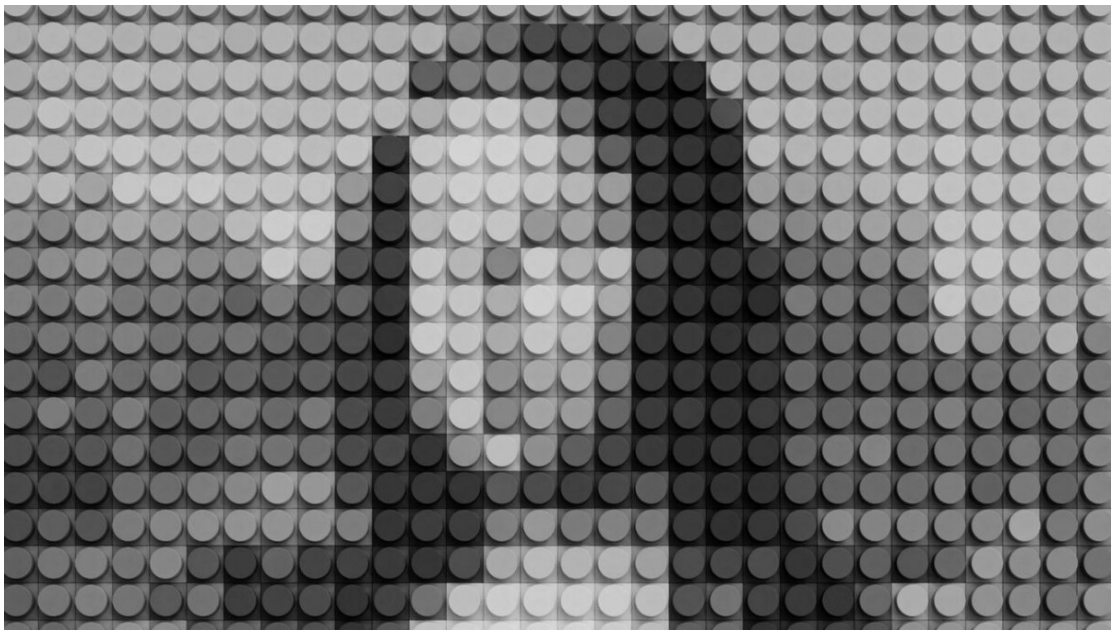


Es ist weltweit einmalig: Ein europäisches Gesetz soll die Anwendung künstlicher Intelligenz regeln. Das könnte ein Vorbild für die ganze Welt werden. → Zum Artikel

Deepfake-Pornografie : Don't fuck with Swifties

Im Netz kursieren gefälschte Pornobilder von Taylor Swift – und sofort eilen ihr ihre Abertausend Hardcore-Fans zur Hilfe. Leider haben nicht alle Frauen diesen Luxus. → Zum Artikel

Nightshade: Giftspritze für künstliche Intelligenz



Mit Nightshade kann man Fotos und Kunstwerke so bearbeiten, dass sie für KI-Modelle nicht mehr lesbar sind. Das Tool soll Künstler schützen – und Entwickler ärgern. → Zum Artikel

KI in der Medizin: "Laien fanden die Antworten des Modells hilfreicher als die der Ärzte"

Kennt jede Diagnose und ist nie unkonzentriert: Kann künstliche Intelligenz die perfekte Medizinerin werden? Bei Google bauen sie ein System, das viele Fragen aufwirft. → Zum Artikel

---

Ihnen gefällt dieser Newsletter? Dann leiten Sie ihn gerne an Ihre Freundinnen und Bekannte weiter. Hier kann man ihn abonnieren. Haben Sie Feedback für uns? Schreiben Sie uns an ki@zeit.de!

ZEIT ONLINE

ZEIT ONLINE GmbH, Buceriusstr. Eingang Speersort 1, 20095 Hamburg