Künstliche Intelligenz

Vom Durchbruch zur Regulierung

Damian Borth

Das Jahr 2012 brachte einen Wendepunkt in der Erforschung der Künstlichen Intelligenz (KI). Beim jährlich stattfindenden Forschungswettbewerb für Bilderkennungssysteme, der »ImageNet Large Scale Visual Recognition Challenge«, gewann ausgerechnet eine Technologie, die damals bereits über 20 Jahre alt und fast schon in Vergessenheit geraten war – ein neuronales Netz namens AlexNet.

Die Fachwelt war überrascht, die großen Tech-Konzerne wie Google, Amazon, Facebook, Apple und Microsoft reagierten zunächst verhalten. Heute setzen sie tiefe neuronale Netze in fast jedem ihrer Produkte und digitalen Dienste ein. Denn diese können Bilder verstehen, Sprache erkennen, Texte analysieren, Dialoge führen – oder sogar selbstständig Inhalte erzeugen.

Außerhalb der Tech-Industrie integriert eine Branche nach der anderen die neue KI in ihre Prozesse oder Produkte, angefangen bei der Automobilindustrie – zum Beispiel selbstfahrende Fahrzeuge – bis hin zur Pharmaindustrie und KI-gestützter Medikamentenentwicklung. Künstliche Intelligenz ist heute Teil der globalen Wirtschaft. Allein im Jahr 2021 wurden weltweit rund 94 Milliarden US-Dollar in KI-Startups investiert.

Man kann die heutige Künstliche Intelligenz als eine *neue* Art von Software verstehen, die Daten analysiert und automatisiert Entscheidungen trifft. Nur dass diese Software nicht von Hand programmiert wird, sondern aus Daten lernt und im Labor trainiert werden muss, bevor sie in den Einsatz kommt. Und ähnlich wie zu Beginn der Softwareentwicklung in den 1960er Jahren lernen wir erst langsam zu verstehen, was es bedeutet, diese neue Art von Software einzusetzen, und wie sie sich in der realen Welt verhält. Dabei erleben wir auch, dass es zu unerwünschten Folgen kommen kann.

In den Anfangsjahren der Computer gab es regelrechte Unfälle wegen fehlerhafter Software. So sind damals Raketen explodiert oder haben

Prof. Dr. Damian Borth ist Ordentlicher Professor für Künstliche Intelligenz und Maschinelles Lernen und Direktor am Institut für Informatik an der Universität St. Gallen.

Röntgengeräte Menschen verletzt. Bei tiefen neuronalen Netzen sind Fehler im System nicht im Programmcode zu finden, der dann identifiziert und isoliert werden kann. Vielmehr sind bei einer Fehlfunktion Millionen, wenn nicht Milliarden von Verbindungen eines tiefen neuronalen Netzes betroffen.

»Tiefe neuronale Netze sind keine Blackbox. Wir können nachverfolgen, was passiert. Wir können es nur nicht interpretieren.«

Diese Komplexität wird oft als Blackbox-Verhalten von tiefen neuronalen Netzen bezeichnet. Allerdings ist dieses »Feuern« für uns Menschen leider nicht einfach zu interpretieren und deswegen schwer erklärbar. Das wirft zwangsläufig die Frage nach der Vertrauenswürdigkeit von Künstlicher Intelligenz und KI-gestützten Anwendungen und Produkten auf.

Die moderne KI muss sich neuen Herausforderungen stellen. Zentral sind hier die ethische Fundierung automatischer Entscheidungen und deren Fairness. Die Europäische Kommission hat sich mit diesen Themen auseinandergesetzt und dafür eine »High-Level Expert Group« ins Leben gerufen. Diese Gruppe hat in den letzten Jahren eine Übersicht zu dem Thema unter dem Namen »Trustworthy AI« zusammengestellt. Trustworthy AI – also »vertrauenswürdige KI« – definiert, wie sich KI-Systeme zu verhalten haben und welche Eigenschaften diese nach außen tragen sollen. Das ist wichtig und gut.

Mir kommt es allerdings so vor, dass wir damit jetzt alle zufrieden sein sollen. Was mir bei dieser ganzen Diskussion fehlt und was vernachlässigt wird, ist die Frage, wie die technische Umsetzung dieser Forderungen aussehen soll.

Denn wenige machen sich Gedanken darüber, wie das zu implementieren ist, und noch wichtiger, wie das kontrolliert werden kann. Hier stehen wir noch ganz am Anfang.

Wir müssen unser Augenmerk viel stärker auf die technische Umsetzung richten, daraus Kontrollmechanismen ableiten und eine Art von technischer Überwachung, einen KI-TÜV, aufbauen. Denn die tiefen neuronalen Netze sind keine Blackbox. Wir können alles, was innerhalb dieser Netze passiert, nachverfolgen und korrigieren. Nur wenn wir es schaffen, fehlerhafte KI-Systeme zu identifizieren und zu isolieren, können wir diese Technologie nachhaltig mit uns Menschen und unserer Umwelt interagieren lassen.

Einen ersten Schritt in diese Richtung haben wir mit unserer Forschung bereits getan: Wir können erste tiefe neuronale Netze ohne Testdaten testen und diese eines Tages in ihrem Verhalten zertifizieren. Weitere Schritte müssen folgen, wenn wir Künstliche Intelligenz als Technologie in unserer Mitte behalten wollen.