Kann künstliche Intelligenz moralisch denken?

Wenn ein selbstfahrendes Auto einen Unfall verursacht, können wir die Maschine nicht zur Verantwortung ziehen – zumindest noch nicht

CHRISTIAN HUGO HOFFMANN

Die Entwicklung künstlicher Intelligenz (KI) ist einer der wichtigsten Trends der Welt – und sie wirft eine Reihe von interessanten ethischen Fragen auf. Die entsprechenden Debatten werden in der breiteren Öffentlichkeit jedoch häufig engstirnig und schlecht informiert geführt. Ein Beispiel dafür ist die Versteifung auf moralische Dilemmata – man denke etwa an das gerne besprochene, aber sehr unwahrscheinliche Szenario, in dem ein autonomes Fahrzeug entweder plötzlich auf die Fahrbahn springende Kinder erwischt oder mit den Insassen in ein parkiertes Auto kracht.

Zwar mag die Frage, wer in einer unvermeidbaren Unfallsituation wie viel Verantwortung trägt – die Nutzer des Autos, der Programmierer der KI, die KI-Firma? –, durchaus interessant sein, doch viel wichtiger ist in diesem Zusammenhang der Hinweis auf einen eklatanten Mangel: Wir Menschen verfügen über sehr strikte Regeln für Medikamententests oder Flugreisen, haben aber keine oder kaum Regeln für selbstfahrende Autos. Das ist erstaunlich, denn klar müsste zumindest so viel sein: Das autonome Fahrzeug selber trägt keine (Mit-) Verantwortung für den Unfall. Dies zunächst aus einem einfachen Grund: KI ist heutzutage ziemlich dumm, und wir wiederum sind schlecht beraten, wenn wir ihre Fähigkeiten überschätzen, sie mit moralischen Akteuren verwechseln oder ihr Autonomie einräumen.

Simulieren statt wissen

In einer detaillierteren Betrachtung sollten wir zwischen mittelfristig und längerfristig realisierbarer KI unterscheiden. Auf mittlere Sicht sind KI nicht als «intelligent», sondern als Tool oder meistens als Datenauswertungsmaschinen zu begreifen. Beim maschinellen Lernen werden Trainingsdaten (etwa von Autos) sowie Metadaten («Dieses Bild zeigt ein Auto») in ein bestimmtes Computerprogramm eingegeben. Daraufhin versucht das Programm durch stetiges Anpassen in den Daten Muster und Gesetzmässigkeiten zu erkennen – um schliesslich auch bei Bildern, die es noch nicht «gesehen» hat, mit einer statistischen Wahrscheinlichkeit vorauszusagen, dass es sich um ein Auto handelt. Wie sollte eine solche KI einen moralischen Status besitzen?

In der Tradition Immanuel Kants ist der moralische Status oder die Würde an der Autonomie festzumachen, die Menschen eigen ist. Demnach ist es die menschliche Fähigkeit, (Handlungs-) Gründe abzuwägen und Entscheidungen zu treffen, die den besten dieser Gründe folgen, die Menschen zu autonomen und willensfreien Akteuren macht. Demgegenüber handeln unsere heutigen KI nicht nach eigenen Gründen, weil sie keine solchen besitzen.

Über eine allfällige Simulation hinaus haben sie kein moralisches Empfinden, keine Intentionen (das Gerichtetsein des Geistes auf etwas), und sie können diese Dinge auch keinen Personen zuschreiben. Sie «wissen» nichts, weil sie kein Wissen besitzen können, und selbst wenn dem anders wäre, wüssten sie nicht, wie es sich anfühlt, etwa wohlwollend zu agieren. Ohne diese Fertigkeiten aber ist eine angemessene moralische Praxis nicht möglich. Trotz vielleicht perfekter Simulation von Denken oder Empathie liegt dem «Agieren» des Computers kein eigenes verständiges Erfassen, kein Problembewusstsein, keine Einsicht zugrunde.

Folgt daraus schon, dass es unmöglich ist, eine moralische Maschine zu bauen? Wie könnte eine KI künftig doch moralischen Status erlangen? Einerseits könnte man sofort einen Perspektivwechsel vornehmen und sich von Kant verabschieden. Sogenannte Behavioristen würden behaupten, dass das Verhalten von Menschen (und Maschinen) ohne Introspektion oder Einfühlung zu untersuchen und zu erklären ist. Nach dieser Lesart, die zum Beispiel Ludwig Wittgenstein und Alan Turing verbindet, heisst etwa wohlwollend sein nichts anderes, als ein bestimmtes Verhalten zu zeigen. Dennoch bleibt das Problem bestehen, dass funktionale Äquivalenz zwischen Mensch und Maschine (die mit heutiger KI noch nicht zu erreichen ist) aufseiten Letzterer noch kein Verständnis gewährleistet. Und Verantwortungsübernahme oder -zuschreibung ohne Verständnisgabe bleibt eine Chimäre.

Andererseits sollten wir anerkennen, dass Begriffe wie «Verantwortung», «Schuld» oder «Bedauern» die Währung eines kausal denkenden Geistes darstellen. Um sie zu verstehen, müssen wir beziehungsweise der Computer in der Lage sein, das, was passiert ist, mit dem zu vergleichen, was unter einer alternativen Hypothese passiert wäre. Somit würde die vielleicht erste Anforderung an eine moralische Maschine in der Fähigkeit bestehen, über ihr eigenes Handeln nachzudenken. Sie müsste also kausale beziehungsweise kontrafaktische Analysen vornehmen können: Was wäre der Fall gewesen, wenn ich anders gehandelt und das Auto in das parkierte statt in das fahrende gelenkt hätte?

> KI ist heutzutage ziemlich dumm, und wir wiederum sind schlecht beraten, wenn wir ihre Fähigkeiten überschätzen.

Fakten sind nicht alles

Zwar ist KI auf mittlere Sicht in der Lage, Regelmässigkeiten zu erkennen und anhand von Datenmustern zu beobachten, wie sich kausale Fragen der einfachsten Art entwickeln: Was passiert, wenn ein Auto mit 50 Kilometern pro Stunde in ein parkiertes kracht? Doch eine zur Moral befähigte KI müsste durch kausale Modelle bereichert werden und insbesondere das Kontrafaktische erfassen können – das aber allein in der Vorstellungskraft und gerade nicht in Daten liegt. Denn Daten sind per Definition: Fakten.

Bis dahin müssen wir uns in den philosophischen Diskussionen damit abfinden, dass die KI im autonomen Fahrzeug nicht zur Rechenschaft für Unfälle gezogen werden kann. Ganz praktisch stehen wir schon jetzt vor der nicht zuletzt juristischen Schwierigkeit, wie eine Person für eine Maschine geradestehen kann, wenn diese eigene Verhaltensweisen erlernt. Frei nach Stanley Kubrick ist die Hölle damit vielleicht ein Ort, an dem der Mensch konsequenzialistisch programmierten, aber denkunfähigen Computern die Macht gegeben hat, über Leben und Tod zu entscheiden.