learning machines

data science for sustainability

Wer denkt?

Überlegungen zu starker KI

Jochen Sautter
01.03.2021

Learning Machines GmbH
Gerberau 9a
79098 Freiburg

Wer denkt?

Wahrscheinlich haben Menschen nur sehr selten ein Werkzeug entwickelt, das so universell einsetzbar ist wie die die Algorithmen der KI. Andrew Ng meint, es sei viel schwieriger, Bereiche zu finden, wo man KI nicht einsetzen kann, als solche, wo man sie verwenden kann. Ein schlagendes Beispiel für diese Eigenschaft ist Alpha Go Zero, ein Programm, das sich selbst beigebracht hat, übermenschlich gut Go zu spielen. Wenig später erzielte ein Derivat derselben Technologie, Alpha Fold, einen bedeutenden Durchbruch in der Biotechnologie, indem es die räumliche Struktur von Proteinen auf Basis von DNA Sequenzen vorhersagen kann.

Die Universalität der KI spiegelt die Universalität der Sprache der Mathematik. Einer Mathematik übrigens, die nicht wirklich neu ist: So erfordert z.B. die 1918 veröffentlichte Relativitätstheorie weit modernere mathematische Werkzeuge als heutige state-of-the-art KI-Algorithmen. Insofern ist nicht die KI jetzt in die Welt gekommen, sondern die Welt ist der KI entgegengewachsen, indem sie die technologischen Ökosysteme an Daten und Rechnerkapazitäten entwickelt hat, in der diese Algorithmen zum Leben erwachen.

Schon als Gedanken-experiment und als künftige Möglichkeit sind Künstliche Intelligente Systeme mit menschenähnlichen Fähigkeiten Gegenstand philosophischer Fragen und intensiver Debatten, insbesondere zur Natur des menschlichen Bewusstseins.

Als maschinelle Implementierung von Mathematik kann KI in Domänen agieren, die prinzipiell mathematisch bzw. formallogisch beschreibbar sind. Also allen Naturwissenschaften, den Ingenieurs-, den Wirtschaftswissenschaften, aber auch der Linguistik oder bei Brettspielen wie Go. Schwieriger und kontroverser wird dies bei der Domäne des menschlichen Bewusstseins. Also der Frage, ob Künstliche Intelligenz eines Tages wirklich, im menschlichen Sinne, „intelligent“ werden kann. Die öffentliche Diskussion dieser Frage wird oft von zwei extremen Polen dominiert: Auf der einen Seite stehen Techno-Apokalyptiker wie Elon Musk, die den Atem eines außer Kontrolle geratenen Deus ex Machina schon im Nacken spüren („mark my words: AI is far more dangerous than nukes!“), oder aber, in einer euphorischen Variante, bis etwa 2045 eine Singularität erwarten, eine „Intelligenz-Explosion“, die alle menschlichen Probleme lösen wird, einschließlich der Sterblichkeit (Ray Kurzweil). Auf der anderen, tendenziell eher europäisch und eher geisteswissenschaftlich geprägten Seite, stehen fundamentale Skeptiker, die erklären, schon die Idee einer bewussten Maschine unterliege einem Missverständnis hinsichtlich der Natur des Denkens. „Wer glaubt, dass tote Rechenmaschinen Geist, Bewusstsein oder Gefühle entwickeln, hat nicht mal entfernt verstanden, was Geist, Bewusstsein und Gefühle sind“, so der Bestseller-Philosoph Richard David Precht, leider ohne zu erklären, was denn dieses Unverstandene sei. Ähnlich argumentierte der Medienwissenschaftler Roberto Simanowsky in einem Text über das Sprachprogramm GPT-3: „Aber ‚denken‘, so viel scheint klar, ist ohnehin nicht das richtige Wort. GPT-3 ist eine Art künstliches neuronales Netzwerk, das anders als unser Gehirn Information rein mathematisch verarbeitet.“

Warum ist das „klar“? Und welcher Natur wäre das Denken des menschlichen Gehirns denn dann, wenn kein prinzipiell mathematisch beschreibbarer Prozess? Skeptiker wie Precht, Simanowsky und Andere reproduzieren einen alten Denkreflex: Für wen halten wir uns, wenn wir glauben, wir könnten denkende und fühlende Wesen, also unsereiner, erschaffen? Für Götter? Aber man kann das umdrehen und fragen: Für wen halten wir uns, wenn wir glauben, unsereiner und unser Bewusstsein könne prinzipiell nicht mathematisch

beschrieben und reproduziert werden? Offenbar für übernatürliche Wesen, denn von anderen natürlichen Vorgängen wissen wir, dass sie, zumindest prinzipiell, mathematisch verblüffend gut beschreibbar sind.

Man kann einwenden, dass die enorme Leistung des menschlichen Gehirns noch weitgehend unverstanden ist. Mathematische Beschreibungen der menschlichen Gehirnprozesse oder mentalen Aktivität existieren bislang allenfalls als spekulative Grobskizzen. Projekte wie das großangelegte „Human Brain Project", das 2013 mit dem Ziel startete, ein künstliches Gehirn in einem Supercomputer zu simulieren, haben sich bislang als vermessen herausgestellt. Und man kann Simanowsky nur zustimmen, wenn er sagt, GPT 3 „verstehe" nicht wirklich was er sage. GPT3 kann verblüffend gut über kosmologische Grenzfragen schwadronieren, z.B. was dem Urknall vorausgegangen ist, oder Kurzgeschichten im Stile von Raymond Chandler schreiben. Allerdings beantwortet dasselbe System die Frage „wie viele Augen hat Dein linker Fuß?" ohne mit der Wimper zu zucken mit „zwei". Frei nach Wittgenstein würden wir jemandem, der so etwas sagt, absprechen, an den menschlichen „Sprachspielen" als vollwertiges Mitglied teilzunehmen. Ebendies wäre aber vorausgesetzt, wenn wir „verstehen" oder „denken" sagen. GPT3 ist nicht mehr und nicht weniger als ein verblüffend guter Fake.

Aber ist dieser Unterschied tatsächlich von unüberbrückbarer, fundamentaler Natur? Oder haben wir es nicht eher mit einem Unterschied im Maß der Komplexität zu tun? Die Zahl der Neuronen in einem menschlichen Gehirn (rund 100 Milliarden) und ihren Verbindungen (schätzungsweise 100 Billionen) liegt derzeit noch um einige Zehnerpotenzen über der von GPT3, dem derzeit größten Neuronalen Netz mit 175 Milliarden gewichteten Verbindungen. Vor allem aber sind Hard- und Software eines menschlichen Gehirns ungleich komplexer, „höher entwickelt" und in sehr weiten Teilen unverstanden, es spielen neben elektrischen Signalen chemische Vorgänge und dynamische Prozesse eine Rolle. Dass die Mathematik des menschlichen Gehirns weit davon entfernt ist, entschlüsselt zu sein, ist also kein Argument dagegen, dass es sie gibt. Die künstlichen Systeme befinden sich in einer rasant verlaufenden Evolution, und ich sehe keinen fundamentalen Grund, warum sich diese Linien nicht irgendwann schneiden sollten. 2016 befragte die Oxford University 350 führende KI-Forscher nach ihrer Prognose der künftigen Entwicklung: Im Durchschnitt erwarteten die Befragten, dass KI Systeme in 45 Jahren in allen Aufgaben Menschen mindestens ebenbürtig sind. Hinsichtlich des Zeitplans ist Skepsis erlaubt, immerhin glaubte man sich schon in den 50er Jahren nur 20 Jahren von einer wirklich intelligenten Maschine entfernt.

Ein Einwand lautet, dass maschinelle Systeme gar keine „eigenen" Gedanken hätten, sondern nur reproduzierten, was sie, im Falle von GPT aus menschengemachten Texten gelernt haben. Simanowksy schreibt: „GPT wird nichts verstehen von der Erfahrung, die er repräsentiert, denn er operiert rein statistisch. Seine Antworten repräsentieren dann nicht die Meinung der KI, sondern den statistischen Höchstwert dessen, was die Menschheit zu einem Thema denkt." Aber gilt nicht ähnliches auch für die Sozialisation eines Menschen? Der KI Forscher Francois Chollet (Google) weist in einem Essay von 2017 darauf hin, dass Intelligenz grundsätzlich nicht als hermetische Eigenschaft eines Systems in einem Blechkasten - oder Schädel - gedacht werden könne: „Beyond your brain, your body and senses & are a fundamental part of your mind. Your environment is a fundamental part of your mind. Human

culture is a fundamental part of your mind. These are, after all, where all of your thoughts come from. You cannot dissociate intelligence from the context in which it expresses itself." Analog hierzu kann ein KI-Algorithmus nur funktionieren und verstanden werden im Zusammenhang mit den Daten, mit denen er trainiert wurde. Was also soll „verstehen" heißen, wenn nicht sinnvolles Reagieren eines zuvor durch Interaktion mit einer Umgebung konditionierten Systems auf aktuellen Input?

Man kann einwenden, dass Prozesse in biologischen Systemen erstens teilweise zufallsgesteuert und zweitens intrinsisch unscharf sind im Gegensatz zu deterministisch ablaufenden Programmen auf digitalen Computern. Tatsächlich ist es jedoch leicht möglich, sowohl Randomisierung als auch Unschärfe in Maschine Learning Algorithmen einzubauen, und ebendies wird auch praktiziert. So kann man die Leistungsfähigkeit neuronaler Netze oft verbessern, indem während des Trainingsprozesses ein zufällig ausgewählter Teil der beteiligten Neuronen ausgeschaltet wird („Dropout"). Das System wird so gezwungen, etwas Ähnliches auszubilden wie die Plastizität des Gehirns, also Robustheit seiner Funktion gegen den Ausfall von Teilen der Hardware. Interessanterweise kann dieser Eingriff, obwohl er eigentlich eine empfindliche Störung darstellt, helfen, aus Trainingsdaten gelernte Gesetzmäßigkeiten auf neue Daten anzuwenden.

Wenn es also künftig eine Maschine geben sollte, die "ich" sagt, und zwar auf ebenso originelle, überzeugende und kohärente Weise wie ein Mensch, die also einen Turing-Test wirklich besteht, mit welcher Begründung wollte man einem solchen System absprechen, dieses "ich" zu sein, und die geäußerten Gedanken tatsächlich zu denken? Das Gedankenexperiment eines solchen Systems wirft interessante philosophische Fragen auf. Man kann diese entweder barsch vom Tisch fegen, indem man kategorisch ausschließt, dass Bewusstseinsphänomene auf mathematisch beschreibbaren Strukturen beruhen könnten, und jedes derartige System, egal wie überzeugend es sich artikuliert, zum bewusstlosen Zombie erklären. Oder man könnte sich auf die Frage einlassen, was es für unser Selbstverständnis bedeutet, wenn unser eigenes reflektierendes Ich nicht mehr und nicht weniger ist als eben dies: eine in der Interaktion mit der Welt und Anderen geprägte informationsverarbeitende Struktur, die so komplex ist, dass sie die Perspektive und das Verhalten eines menschlichen Ichs ausbilden kann. Die letztere Variante scheint mir nicht nur überzeugender, sondern auch inspirierender.

Skeptiker wie Simanowsky oder Precht scheinen die menschliche Würde verteidigen zu wollen gegen die womöglich kränkende Vorstellung, wir seien nichts grundlegend anderes als eine raffiniert gemachte „tote Rechenmaschine". Sie entwerfen eine dualistische Perspektive, wo der menschliche Geist in einer numinosen Sphäre von „Qualia" residiert, die sich einer mathematischen Beschreibung prinzipiell entzieht und also für maschinelle Verfahren unerreichbar bleiben muss. Aber vielleicht auch einer Sphäre, in der die Sozial- und Kulturwissenschaft sich gegen eine erfolgsbesoffene, übergriffig werdende KI- und Neurowissenschaft zu behaupten sucht, die in einem naiv-reduktionistischen Durchmarsch sich anschicken könnte, diese alten Disziplinen mittels Hirn-Scans und KI-Algorithmen wegzuerklären.

Ich denke, beide Befürchtungen beruhen auf einem Missverständnis:
Ein humaner Begriff menschlicher Würde braucht keine übernatürliche Sphäre (dass die katholische Kirche so etwas vielleicht braucht, sei dahingestellt). Ein Mensch verliert keine der Eigenschaften, die ihn zu einem fühlenden, verantwortungsfähigen, rechtsfähigen Wesen, also zu einer Person machen, wenn uns künftig eine mathematische Beschreibung mentaler Prozesse zunehmend besser gelingen sollte.
Und was die Konkurrenz der Disziplinen angeht, so hat die KI Forschung schon immer begriffen, dass intelligente Systeme, ob menschlich oder künstlich, gar nicht anders gedacht werden können als eingebettet in die sie prägende Umgebung. Also eine Umgebung, die bei fortgeschritten Systemen nicht ohne die Begriffe der Kultur- und Sozialwissenschaften beschrieben werden kann.

Jochen Sautter ist Experte für Deep Learning und Mitgründer des Freiburger Start-Ups Learning Machines, das individuelle KI-Anwendungen entwickelt.