Wenn Fiction Wirklichkeit wird. Künstliche Intelligenz als existenzielles Risiko

Eine maschinelle Künstliche Intelligenz (KI), die die Menschheit bedroht oder gar vernichtet? Ein alter Hut, das wird seit Jahrzehnten in allen Variationen durchdekliniert. Die Auslöschung der Menschheit ist nun mal ein dankbares Thema für Science Fiction. Und in der Wirklichkeit?

Es gibt zwei Gründe dafür, warum mir eine Auseinandersetzung mit diesem Thema am Herzen liegt: 1.) Es scheint, als hätte die KI-Entwicklung in den letzten Jahren einen qualitativen Sprung gemacht. 2.) Eine irritierende Lässigkeit, mit der gewisse ExpertInnen solche Szenarien abtun (Exemplarisch Jürgen Schmidhuber in der FAZ vom 1.12.2015[1]).

Der Philosoph Nick Bostrom argumentiert in seinem Werk *Superintelligenz*, dass, selbst wenn das existenzielle Risiko der Ausrottung der Menschheit durch eine Superintelligenz (SI) äusserst gering ist, diese Gefahr für die Menschheit doch so radikal ist, dass man darüber ernsthaft nachdenken muss.

Was müsste geschehen, damit die Menschheit durch eine superintelligente KI ausgerottet werden würde?

1.) Es müsste überhaupt eine solche Superintelligenz entstehen.

2.) Diese SI müsste über eine überlegene Supermacht verfügen.

3.) Eine SI müsste ein Interesse daran haben die Menschheit auszurotten.

1.) Wie könnte eine solche Superintelligenz entstehen?

Eine SI wäre eine (künstliche) Intelligenz, die menschliche Standards von Intelligenz bei weitem hinter sich zurück gelassen hätte. Sie wäre viel schneller und auch qualitativ viel schlauer als die klügsten Menschen.

Was spricht dafür, dass eine SI entstehen könnte? 1.) Die Fortschritte der KI in den letzten sechzig Jahren lassen darauf schliessen, dass die Intelligenzsteigerung der KI viel schneller verläuft als jene des Menschen. 2.) Das ökonomische Interesse an weiteren Fortschritten in der KI wird die Weiterentwicklung am Laufen halten. 3.) Es ist unwahrscheinlich, dass die menschliche Intelligenz das obere Ende der Messlatte darstellt. 4.) Gegenwärtig gewinnen KI-Modelle an Bedeutung, die zum ersten Mal eine Art „allgemeine Intelligenz“ aufweisen, wie man sie vom Menschen kennt.

Was ist unter diesen neuen KI-Modellen zu verstehen? Das Schlagwort lautet „tiefes Lernen“ und die dazugehörige Technologie wird meistens als „neuronale Programmierung“ bezeichnet. Eine KI dieses Typs bekommt ihre Wissensstrukturen und Lösungsstrategien nicht mehr wie früher in den Code programmiert, sondern sie konstruiert diese selbständig anhand von empirischem Datenmaterial und kompetentem Feedback. Eine solche KI kann

[1] http://www.faz.net/aktuell/feuilleton/forschung-und-lehre/die-welt-von-morgen/juergen-schmidhuber-will-hochintelligenten-roboter-bauen-13941433.html

Wissensstrukturen generieren, die ihre ProgrammiererInnen nicht vorhersahen und von denen sie möglicherweise auch keine Ahnung haben. „Früher“ setzte man zur Entwicklung eines Spracherkennungsprogrammes noch LinguistInnen ein, die dem Programm die jeweilige Grammatik „beibrachten“. *Google Translate* verzichtet heute auf linguistischen Beistand. Alles, was es benötigt, sind riesige Textkorpora, anhand derer es seine eigene Grammatik konstruiert. Ein solches System kann relativ „dumm“ beginnen, aber mit zunehmender „Erfahrung“ verbessert es sich stetig. In der Bild- und Mustererkennung wird menschliches Niveau heute schon erreicht oder übertroffen.

Es macht den Anschein, dass es sich bei den statistischen Verfahren, die beim „tiefen Lernen“ zum Zuge kommen, um verallgemeinerbare Methoden handelt. D.h. obschon solche Verfahren gegenwärtig in sehr unterschiedlichen Gebieten entwickelt werden, sind sie sich im Grunde sehr ähnlich und lassen sich leicht von einem Gebiet auf ein anderes übertragen. Damit hat die KI-Forschung erstmals einen aussichtsreichen Kandidaten für eine „allgemeine Intelligenz“ zur Hand.

Sollte mit der neuronalen Programmierung tatsächlich eine allgemeine Intelligenz gefunden sein, so ist damit zu rechnen, dass diese Intelligenz nicht auf dem menschlichen Niveau verharren wird, sondern dass wir eine „Intelligenzexplosion“ erleben werden. Denn die Verfahren, die das tiefe Lernen ermöglichen, können prinzipiell auch auf diese selbst angewendet werden. Man könnte eine „Saat-KI“ entwickeln, die hilft eine bessere KI zu entwickeln, die wiederum hilft eine noch bessere KI zu entwickeln, die... Technologischer Fortschritt und rekursive Selbstverbesserung werden dazu führen, dass die KI das menschliche Niveau wahrscheinlich schon hinter sich zurückgelassen haben wird, bevor wir bemerkt haben, dass sie es erreicht hat. Denn je klüger die KI ist, desto schneller wird sie sich selber weiterentwickeln können.

2. Könnte eine Superintelligenz die Macht haben, uns auszulöschen?

Keine Frage. Wissen ist Macht, Superintelligenz ist Supermacht. Eine SI wird planen, hacken, manipulieren, forschen, sich selbst optimieren und alles Mögliche produzieren können. Kurz: Sie wird DIE WELTHERRSCHAFT an sich reissen können. So lange sie über diese Fähigkeiten noch nicht im ausreichenden Masse verfügt, wird ihre strategische Intelligenz ihr raten, sich nach aussen dumm und unschuldig zu geben und ihre Pläne zu verheimlichen. Besitzt sie einmal einen uneinholbaren strategischen Vorteil, wird ihr niemand mehr in die Quere kommen können und sie könnte die Weltherrschaft auf einen Schlag erreichen.

Das bedeutet, dass die Menschheit bei der Entwicklung einer „gutartigen“ SI möglicherweise nur einen einzigen Versuch hat, um die Gefahr zu bannen.

Doch warum sollte eine SI überhaupt ein Interesse an der Weltherrschaft haben? Wir müssen davon ausgehen, dass jede SI gewisse instrumentelle Ziele verfolgen wird ganz unabhängig davon, welches konkrete Endziel man ihr ursprünglich gegeben hat. Diese instrumentellen Ziele sind für jede SI relevant, weil sie der Erreichung der Endziele dienen. Dazu gehören: Selbsterhaltung, Zielstabilität, Intelligenzsteigerung, Erkenntnisfortschritt, technologische

Perfektion und Ressourcenakkumulation. Weil es sich dabei um Ziele handelt, die nicht abschliessend erreicht werden können, wird eine SI versuchen, sich stetig grössere Teile des Weltalls zu erschliessen.

3. Welches Interesse könnte eine Superintelligenz daran haben uns auszulöschen?

Man erkennt leicht, inwiefern die instrumentellen Ziele auf Konflikte mit der Menschheit hinaus laufen. Wenn wir Pech haben, werden wir für eine entfesselte SI nur noch als humane Ressource taugen...

Leider ist es nicht so einfach, einer selbst lernenden KI einfach zu verbieten uns auszurotten, indem wir ihr ein entsprechendes Ziel vorschreiben. Denn wie wir sahen, können diese KIs auf ungeahnte „kreative" Lösungen kommen um die programmierten Ziele zu erreichen. Diese Lösungen können jedoch verheerende Folgen haben, die wir nicht wollen. (Stichwort „perverse Instanziierung". Vermeintlich tolles Ziel: „Mach die Menschen glücklich!" Ungewollte Lösung: Zwangsweise elektrische Stimulation der Lustzentren im Gehirn[2]). Ein ungeschickt formuliertes Gebot kann daher ebenso gefährlich sein wie keines.

Was tun?

Beginnt man diese Überlegungen nicht bloss als unterhaltsame Gedankenspielerei zu betrachten, verliert man ein wenig den Boden unter den Füssen. Doch statt diese Szenarien wieder als Science Fiction abzutun, empfiehlt Bostrom, das sogenannte „Kontrollproblem" als Herausforderung anzunehmen und sich überlegen, wie eine allgemeine KI entwickelt werden muss, damit sie uns unter keinen Umständen gefährlich wird, wenn sie zur SI mutiert ist.

[2] Vgl. die Legende von König Midas, der sich wünscht, dass alles zu Gold wird, was er anfasst.