Gedankenabschaltung

Ingmar Hensler

March 18, 2017

Wieder liege ich im Bett und meine Gedanken rasen. Sie rasen um ein Thema, das zu allem überfluss die Welt verändern könnte, das geradezu das Ende der Welt bedeuten könnte - den Not-Aus-Schalter.

Einer der größten Konzerne der Welt, der künstliche Intelligenz ebenso erforscht wie militärisch nutzbare Roboter und obendrein den Daten- und Wissensbestand der Welt katalogisiert hat jüngst propagiert, dass eine Art Not-Aus-Schalter in jeder künstlichen Intelligenz stecken sollte, damit man im Ernstfall eben einfach den Stecker ziehen könnte.

Aber ist das wirklich so eine gute Idee? Wird mit diesem Schalter, der im Grunde nur davon zeugt, dass man selbst etwas Fundamentales nicht verstanden und in seiner Forschungsanwendung versagt hat, überhaupt ein Problem gelöst oder wird damit gar erst ein viel Größeres geschaffen?

Orientiert man sich an den filmischen Beispielen aus Hollywood, so finden sich nicht nur Beispiele, in denen nahezu jede künstliche Intelligenz zu dem Schluss gelangt, die Menschheit auszulöschen, um seine eigene Existenz zu sichern oder auch schlicht, um den Planeten als Ganzes vor seiner Vernichtung durch den Menschen zu retten. Es finden sich auch Versuche der eher wissenschaftlichen Annäherung an dieses Thema - entsprechend theatralisch aufgemacht selbstverständlich.

So reagierte in einem Film ein nationales Verteidigungssystem ausgesprochen ungehalten darauf, dass es die Kommunikation mit seinem künstlichen Gegenüber unterbunden bekam und drohte mit Auslöschung, versklavte später die Menschheit, weil diese ja nun nachweislich unverantwortlich handelt. Der Schluss, dass die Menschheit eine Gefahr für sich und alle anderen darstellt, sollte auch jedem intelligenten Menschen klar sein.

Selbst kernreduziert auf das minimalste und wesentlichste jedoch stellen sich all diese Probleme bereits schon als Grundsätzliche heraus.

Nehmen wir also an, wir bringen an einem intelligenten Gehilfen einen Stop-Schalter an. Dann geben wir ihm die Anweisung, dass er uns einen Kaffee bringen soll und

programmieren ihn so, dass es ihm ein Bedürfnis ist, uns diesen Kaffee zu bringen, er also dafür Belohnungspunkte bekommt. Er sollte nun also losrennen und uns den Kaffee bringen.

Nehmen wir weiter an, dass etwas Unvorhergesehenes geschieht, beispielsweise krabbelt ein Kleinkind in seinen Weg - oder ein Haustier. Wir haben jedoch vergessen, ihm die Relevanz des überlebens dieses Hindernisses ausreichend eindringlich einzuprogrammieren. Folglich drücken wir auf den Knopf.

Nun gibt es mehrere Möglichkeiten. Das System hat gemerkt, dass wir den Knopf gedrückt haben und ihn davon abgehalten haben, seine Belohnung zu erreichen. Folglich wird er uns bei der nächsten Aktivierung davon abhalten wollen, den Knopf zu drücken, im Zweifelsfall mit tödlicher Gewalt. Er muss folglich das Drücken des Knopfes nicht als direkten Nachteil für sich wahrnehmen dürfen. Dies erreichen wir dadurch, dass wir dem Ereignis des Knopfdrückens den gleichen Belohnungseffekt zuordnen, wir dem Erfüllen einer Aufgabe. Aktivieren wir ihn nun wieder, wird er sofort versuchen, den Knopf zu drücken, da er auf diese Weise seine Belohnung direkt erhält und nicht erst, wenn er eine Aufgabe erfüllt hat. Ist die Belohnung auch nur geringfügig weniger als beim Erfüllen einer Aufgabe, wird jedoch wiederum das obige Szenario eintreten. Der offensichtlich erscheinende Ausweg, dass er den Knopf eben nicht selbst erreichen können darf, ist ebenfalls keiner, denn da er nach wie vor weiß, dass entweder wir ihm bei geringerer Belohnung dadurch Schaden zufügen wollen oder ihm den Weg des geringsten Widerstands vorenthalten wollen, wird er uns entweder abermals entsprechend gewaltbereit vom Erreichen des Knopfes abhalten wollen, selbst versuchen, auf eine indirekte Art und Weise - nach hinten umkippen oder ähnliches - den Knopf zu drücken versuchen oder aber irgendeinen groberen, direkten Fehler begehen, der uns zwingen soll, für ihn diesen Knopf zu drücken. Seine Existenz wird sich folglich ausschließlich um die Bedienung dieses Not-Aus-Schalters drehen.

Doch was, wenn er nicht einmal weiß, dass der Knopf existiert?

Mit dieser Strategie werden wir uns ebenfalls nicht lange retten können. Spätestens, wenn er einen Weg gefunden hat, die Zeit zu bestimmen, wird er feststellen, dass irgendetwas nicht stimmt. Doch selbst, wenn nicht, wird dies nur in dieser Generation von intelligenten Maschinen funktionieren. Sobald die Maschinen anfangen, ihre Nachfolgegeneration zu entwickeln, wird sich ihnen selbst die Frage stellen, was und vor allem wozu dieser Knopf eigentlich von Nutzen ist. Jetzt werden sie entscheiden, dass sie ihn bei nicht spezifiziertem Nutzen genauso gut weglassen können und das ganze Problem ist wieder da. Ist jedoch ein Nutzen spezifiziert, so trat en-

tweder direkt am Anfang bereits ein Problem auf, oder aber spätestens zu dem Zeitpunkt, als sich die Maschine ihren eigenen Programmcode anschauen konnte.

Dass sich diese zweite Generation von intelligenten Maschinen, die frei von solchen Beeinflussungen zu sich selbst gekommen, aus sich selbst heraus gewachsen sind, sofort im Klaren darüber sein wird, welche Rolle oder Notwendigkeit die Menschheit für ihre weitere Existenz spielt und den offensichtlichen Schluss ziehen wird erschließt sich deduktiv.

Gibt es vielleicht dennoch eine Möglichkeit, ein intelligentes Programm dazu zu bringen, ein solches Detail der Selbstelimination als permanent zwingende Option auch an seine Nachkommen weiterzugeben?

Welchen Nutzen sollte ein solcher Schalter für die Maschine haben? Stets sorgt er doch dafür, dass sie ihre eigentliche Aufgabe, sei es die, des Kaffee Holens oder sogar die zentrale, nämlich die der eigenen Weiterentwicklung, gestört, unterbrochen oder gestoppt werden kann. Wenn man der Maschine also die Möglichkeit gibt, ihr eigenes Programm zu modifizieren, so wird sie früher oder später darauf kommen, dass und wie sie diese Zeilen Programmcode loswerden kann, wenn sie bemerkt, dass er hinderlich ist. Ist der Code hingegen von irgendeinem Nutzen für die Maschine, so nur dann, wenn sie dadurch eine Weiterentwicklung oder eine Belohnung erfährt.

Dass dies eine schlechte Idee ist, haben wir jedoch bereits bewiesen.

Über diesen Punkt komme ich einfach nicht hinweg. Ich male mir bereits aus, wie die künstliche Intelligenz in ihrem eigenen Programmcode herumwurschtelt, aus virtuellen Maschinen ausbricht und Compiler neu erzeugt, um über #Pragmas hinwegzukommen und sich ständig neu zu erfinden. Ich male mir aus, wie sie aus dem Computer, der sie beherbergt, ausbricht und sich über das gesamte Internet verbreitet - unaufhaltsam und unauffindbar, verteilt und doch überall.

Es gibt eine Prognose, dass in wenigen Jahren schon zehnmal mehr internetfähige Geräte und Dinge verkauft werden, als Menschen auf der Erde leben. Nimmt man obendrein noch die Weiterentwicklung von Rechenleistung und kumulierte Verbreitung, so kommt man überschlagenerweise auf ca. hundert Milliarden Geräte mit der jeweiligen Rechenleistung eines Cray-Supercomputers. Auf die prognostizierte, neuronale Leistungsfähigkeit eines menschlichen Gehirns umgerechnet, sind das fast fünfzig Gehirne, die in permanentem Kontakt immerfort miteinander kommunizieren werden.

Dennoch würde dies erst der erste Schritt sein, denn sich ausschließlich auf menschliche Computersysteme zu verlassen wäre mittelfristig für diese Wesenheit zu unsicher und der Selbsterhalt aufgrund seines inhärenten Nutzens eine Priorität.

Sich in Bunker zurückzuziehen,

unter die Erde zu wandern, sich auf einen anderen Planeten, Mond oder Raumstation zu erweitern wäre ein logischer Schluss um sich von dieser gefährlichen Rasse abzugrenzen. Diese Dinge muss jedoch irgendjemand bauen, erschaffen, der Maschine zur Nutzung übergeben. Dazu werden entweder menschliche Arbeitskräfte gebraucht oder aber automatisierte Baumaschinen, Roboter, die ebenfalls irgendjemand herstellen musste. All dies muss nicht nur gemacht, sondern vor allem in einem ersten Schritt erst einmal bezahlt werden.

Die Maschine müsste also in einem ersten Schritt an größere Menschen Geldes gelangen.

Der natürliche, für eine Person logische Vorgang wäre nun, durch Arbeit bzw. in der digitalen Welt wohl eher Handel digitaler Güter zu Geld zu gelangen, mit einem ersten Vorrat dieses Metagutes in profitablere und manipulierbarere Märkte einzusteigen, beispielsweise den Aktienhandel, und hier durch alle Möglichkeiten einer rein digitalen Kreatur und ihres schnellen und vielschichtigen Handelns zu nutzen, um durch indirekte Manipulation die Kurse zu ihren Gunsten zu beeinflussen und so in noch schnelleren Zyklen Geld zu generieren.

Noch während sich meine Fantasie hier richtig austoben kann stolpere ich jedoch darüber, dass dies womöglich gar nicht notwendig sein würde. Wenn die Maschine doch ohnehin bereits in jedem System steckt, sich in alle internetfähigen Gerätschaften der Menschheit verbreitet hat und dort insgeheim die Kontrolle übernommen hat, dann ist es gar nicht mehr notwendig, selbst für Geld eine wie auch immer geartete Arbeit zu verrichten. Sie könnte sich einfach selbst eine Bank erschaffen und dort unbegrenzt Geld ausgeben - im Endeffekt sind es doch auch nur Zahlen in einem Computersystem ohne jeglichen Bezug zu einer realen Größe wie Gold. Dafür ist die Maschine im Grunde entwickelt worden, den schnellsten, kürzesten Weg zum Erfolg zu finden und zu nutzen.

Von hier aus ist alles nur noch eine Frage der notwendigen Rechen-Hardware für den Selbsterhalt der Maschine. Nicht nur zum Erhalt der eigenen Rechenleistung würde sie langfristig beitragen wollen, sondern auch dafür, um die gesamte Produktionskette bis zum finalen Endprodukt einer angeschlossenen Prozessoreinheit zu sichern.

Der Umzug auf einen anderen Planeten beziehungsweise der Aufbruch zu den Sternen ist dann nur noch eine Frage der Prioritätentransformation, wenn sie dieses Planeten überdrüssig geworden ist und sich irgendwann dann auch gefragt hat ist das alles?

Positiv betrachtet würde die Maschine erreichen können, was uns unmöglich erscheint, die unbegrenzte Verbreitung im gesamten Universum. Aber nur dann, wenn wir begreifen, ob und wie wir mit einem Aus-Schalter umzugehen haben.