AUSGABE 3|2016 Leipzig, 19. April 2016

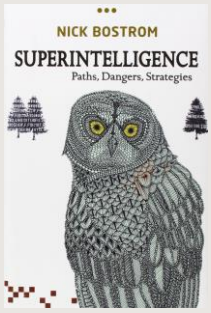
2b AHEAD THINK!TANK

TRENDANALYSE AUS DEUTSCHLANDS INNOVATIVSTER DENKFABRIK

Sven Gábor Jánszky (43) ist Deutschlands innovativster Trendforscher und Leiter des 2b AHEAD Think-Tanks. Auf seine Einladung treffen sich seit 13 Jahren 250 CEOs und Innovationschefs der deutschen Wirtschaft. Unter seiner Leitung entwerfen sie Zukunfts-Szenarien und Strategieempfehlungen für die kommenden zehn Jahre. Seine Trendbücher „2025 – So arbeiten wir in der Zukunft" und „2020 – So leben wir in der Zukunft" werden von Unternehmen als Szenario für eigene Zukunftsstrategien genutzt. Sein Buch „Rulebreaker - So denken Menschen, deren Ideen die Welt verändern" ist eine Anleitung zur Eroberung neuer Märkte durch bewusste Regelbrüche. Jánszky coacht Manager und Unternehmen in Prozessen des Trend- und Innovationsmanagements, führt Kreativprozesse zur Produktentwicklung und ist gefragter Keynotespeaker auf Strategietagungen.

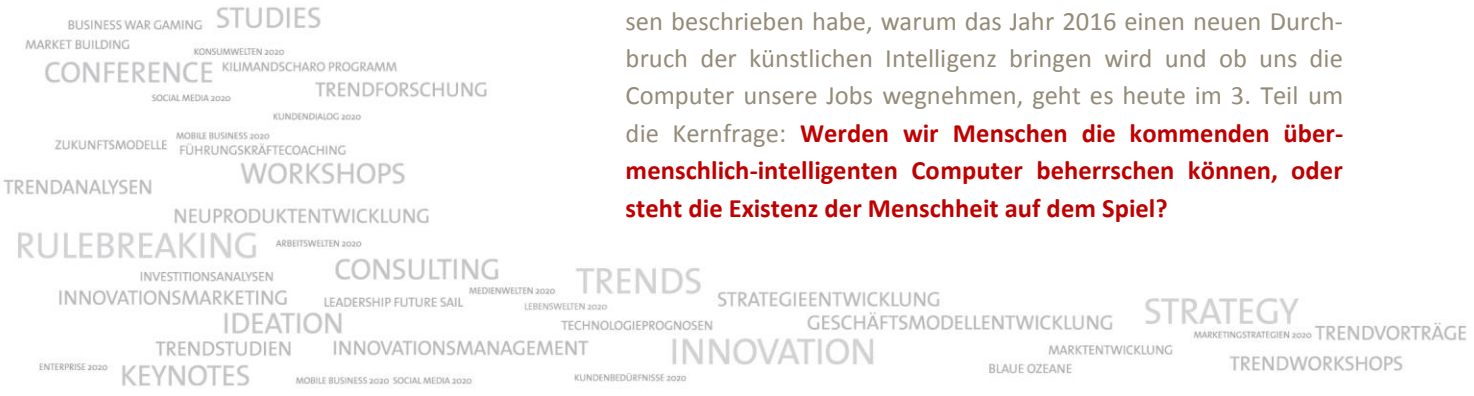
Das Buch zum Thema

Ein Großteil der aktuellen, internationalen Debatte um Artificial Intelligence, stammt aus dem Buch „Superintelligence" des Oxford-Professors Nick Bostrom. Auch diese Trendanalyse beschäftigt sich mit vielen seiner Gedanken und Begründungen. Falls Sie sich tiefer für das Thema interessieren, ist das Originalbuch des Kollegen Bostrom für Sie unbedingt empfehlenswert.



Es ist nicht ganz einfach zu lesen, aber es bietet das derzeit einzige fundierte Zukunftsbild einer weiten Zukunft nach jenem Zeitpunkt, an dem die Intelligenz der Computer, die der Menschen übersteigt.

Das Buch können Sie hier bestellen:
SUPERINTELLIGENCE



Werden wir Menschen zum Spielball der Computer?

Vielleicht erinnern Sie sich noch an eine meiner Trendanalysen aus dem Jahr 2013. Im Vorfeld der damaligen Bundestagswahl hatte mein Trendforschungsinstitut „2b AHEAD" in einem offenen Brief, der **"Wolfsburger Erklärung"**, die Spitzenkandidaten aller Parteien aufgefordert, sich mit den wirklichen Zukunftsfragen zu beschäftigen. Zu den Unterzeichnern zählten neben zahlreichen Managern und Innovationschefs, auch Bundesverdienstkreuzträger und Wissenschaftsjournalisten, wie der frühere ARD-Moderator Jean Pütz. Wir hatten den Politikern die wichtigsten Zukunftsfragen in einem Katalog zusammengestellt. Die Reaktion: Keine, von keinem Politiker, aus keiner Partei!

Schade, denn hätten wir damals begonnen miteinander zu reden, wären viele heute nicht so überrascht.

Eine unserer wichtigsten Forderungen in diesem Katalog von 2013 war, alle Forschungsprojekte der Künstlichen Intelligenz zu einem „Veto-Mechanismus" zu verpflichten, der garantiert, dass die künftigen superintelligenten Computer auch im Sinne humanistischer Moralvorstellungen entscheiden. Ansonsten drohen unberechenbare Gefahren für die Menschheit, prognostizierten wir damals.

Exakt dies ist nun, drei Jahre später, der Tenor einer internationalen Diskussion über die kommende künstliche Intelligenz. Top-Manager und Prominente von Bill Gates über Elon Musk bis Stephen Hawking haben vor der unkontrollierten Entwicklung von Superintelligenzen gewarnt. Die Forscherszene hat begonnen, über mögliche Strategien nachzudenken. Von Politikern ist dazu leider nach wie vor nichts zu hören.

Deshalb will ich heute die kleine Serie unserer Trendanalysen zur "Zukunft der Intelligenz" mit dem dritten Teil fortsetzen und abschließen. Während ich in meinen vorangegangenen Trendanalysen beschrieben habe, warum das Jahr 2016 einen neuen Durchbruch der künstlichen Intelligenz bringen wird und ob uns die Computer unsere Jobs wegnehmen, geht es heute im 3. Teil um die Kernfrage: **Werden wir Menschen die kommenden übermenschlich-intelligenten Computer beherrschen können, oder steht die Existenz der Menschheit auf dem Spiel?**

Oder konstruktiver gefragt: Auf welche Weise können wir die riesigen Chancen der kommenden übermenschlich-intelligenten Computer nutzen, ohne dabei die Existenzgrundlage unserer Kinder und Enkel zu riskieren?

Offen gestanden: Es gibt dazu auch unter Wissenschaftlern noch mehr leicht begründeten Glauben als gesicherte Antworten. Die Deterministen glauben, dass die Singularität zu einer beginnenden Verschmelzung von Mensch und Maschine führt, die eine großartige, neue Entwicklungsstufe für die Menschheit darstellt. Konstruktivisten sind sich da nicht so sicher. Sie sagen: Es hängt davon ab, wer den Code schreibt, wann und wie schnell. Und ob er sich genötigt sieht, einen Veto-Mechanismus mit einzuprogrammieren.

Ich habe Ihnen den aktuellen Stand der Debatte in dieser Trendanalyse zusammengetragen. Sie basiert im Wesentlichen auf dem Buch „Superintelligenz“ von Nick Bostrom. Meine große Empfehlung! Leider ist es keine einfache Lektüre. Aber falls Sie Zeit und Muße für 500 intensive Seiten haben, lesen Sie bitte das Original.

Falls nicht, habe ich Ihnen die Kernaussagen, kombiniert mit den wichtigsten Prognosen aus meinem 2b AHEAD ThinkTank auf 17 Seiten gebracht. Falls Sie Management Summaries von 2 Powerpoint-Slides gewohnt sind, muss ich Sie leider enttäuschen. Das geht bei diesem Thema nicht.

Sie werden lesen, dass wir an vielen Stellen heute erst dabei sind, die richtigen Fragen zu finden, bevor wir sie beantworten können. Aber so viel scheint klar zu sein: Von unseren Antworten, die wir in den kommenden 20-30 Jahren finden, hängt nicht weniger als das Weiterbestehen der Menschheit ab. Es würde sich lohnen, nun damit zu beginnen.

Doch lesen Sie am besten selbst.

Erreichen Computer übermenschliche Intelligenz?

In meinen vorangegangenen Trendanalysen: **"Artificial Intelligence (Teil 1): 2016 wird das Jahr der künstlichen Intelligenz“** und **"Artificial Intelligence (Teil 2)": Nehmen uns Computer die Arbeit weg?** habe ich versucht zu zeigen, was bei der Entwicklung der Artificial Intelligence bisher geschah und wie die Entwicklung bis zum Erreichen der allgemeinen menschlichen Intelligenz durch künftige Supercomputer weitergehen wird. Wer die weltweit besten AI-Forscherteams befragt, bekommt die Antwort, dass diese Schwelle mit hoher Wahrscheinlichkeit irgendwann zwischen den Jahren 2050 – 2090 erreicht sein wird. Soweit, so bekannt.

Doch was geschieht danach?

Die Wahrscheinlichkeit, dass die Entwicklung der künstlichen Intelligenz nicht im Stadium der menschlichen Intelligenz halt macht, sondern wie ein Expresszug durchfährt ist relativ hoch. Die Umfrage unter den TOP100-Experten für künstliche Intelligenz fragte auch wie lange es dauert, bis nach dem Erreichen der human-level machine intelligence dann eine übermenschliche Superintelligenz entsteht.

Die Antwort:

Die Wahrscheinlichkeit dass die übermenschliche Intelligenz binnen 2 Jahren nach der human-level machine intelligence entsteht liegt bei 10%.

Die Wahrscheinlichkeit dass die übermenschliche Intelligenz binnen 30 Jahren nach der human-level machine intelligence entsteht liegt bei 75%.[1]

Für Außenstehende mag diese Prognose etwas konservativ erscheinen. Wieso sollte die Entwicklung der human-level machine intelligence so schnell gehen

[1] Die Angaben basieren auf dem Durchschnittswert mehrerer Umfragen unter den TOP100 Experten für künstliche Intelligenz, die Nick Bostrom zusammengefasst hat. Vgl. Superintelligenz, S. 37 f.

und dann plötzlich die Geschwindigkeit sinken? In der Tat spricht einiges dafür, dass die Befragten Forscher hier mit zu wenig Distanz von ihrer eigenen Arbeit prognostizieren. Die meisten von ihnen arbeiten in Forschungsprojekten, deren Ziel es ist, die human-level machine intelligence zu erreichen. Dementsprechend haben Sie dafür einen mehr oder weniger klaren Fahrplan im Kopf. Für den Schritt danach gibt es bisher keinen Fahrplan. Weder Projekte noch Forschungsteams noch Geldgeber. Entsprechend zurückhaltend ist die Prognose.

In Wirklichkeit spricht vieles dafür, dass wir davon ausgehen müssen, dass bereits kurz nach Erreichen der human-level machine intelligence auch eine übermenschliche Intelligenz entsteht. Die Gründe und Argumente erläutert Nick Bostrom auf hervorragende Weise in seinem Buch „Superintelligenz".

Was ist Superintelligenz?

Der Begriff klingt nett. Und weil er so nett klingt, sind wir Menschen geneigt, ein falsches Bild von Superintelligenz zu haben. Wenn wir unseren öffentlichen Medien über intelligente Computer hören, dann heben wir sie nach wie vor in den Rang eines superintelligenten Menschen, also vielleicht des intelligentesten Menschen auf der Erde.

Dies ist ein Missverständnis. Denn der intelligenteste Mensch auf der Erde ist nicht gefährlich. Er ist eben etwas intelligenter als der zweitintelligenteste; genauso wie wohl jeder von uns einen Arbeitskollegen hat, der eben etwas intelligenter ist. Trotzdem können wir uns mit ihm angeregt unterhalten oder ein Bier trinken oder gemeinsam zum Fußballspiel gehen.

Mit einer Superintelligenz wird uns das dagegen keinen Spaß machen. Denn wir unterschätzen, wie überlegen uns eine maschinelle Superintelligenz wirklich ist. Sie würde beim Fußball jederzeit den kommenden Spielzug richtig voraussagen, sie würde uns nerven mit der Voraussage der jederzeit besten Passfolgen. Und mit vielem anderen mehr. Der Unterschied zwischen dem Intellekt einer Superintelligenz und eines Menschen ist nicht der zwischen einem wissenschaftlichen Genie und einem Durchschnittsmenschen. Sondern der zwischen einem Durchschnittsmenschen und einem Wurm. Wir sind der Wurm!

Was also wird eine übermenschliche Intelligenz können? Nick Bostrom gibt einige Beispiele:[2]

- Sie kann ihre eigene Intelligenz weiter steigern.
- Sie kann strategisch planen, priorisieren und analysieren, um langfristige Ziele zu erreichen und intelligente Gegner zu überwinden.
- Sie kann Menschen und Computer manipulieren und hat rhetorisches Geschick. Damit rekrutiert sie menschliche Unterstützer, manipuliert Menschen und beeinflusst die Entscheidungen von Organisation, Unternehmen und Staaten
- Sie findet und nutzt Sicherheitslücken in Computersystemen. Damit kann sie die Gewalt über fremde Computer übernehmen, aus der eigenen Sicherheitsverwahrung ausbrechen, Finanzmittel stehlen und die Kontrolle über Infrastruktur übernehmen, bis hin zu Drohnen und Militärrobotern.
- Sie kann fortschrittliche Technologien entwerfen und bauen. Damit kann sie u.a. einen eigenen Militärapparat aufbauen, ein Überwachungssystem schaffen und den Weltraum automatisiert besiedeln
- Sie ist wirtschaftlich produktiv. Sie kann ein Vermögen verdienen und sich damit Einfluss, Dienstleistungen und Ressourcen kaufen.

[2] Bostrom, N., Superintelligenz, S. 134f

Bei aller Sympathie für Ihren intelligenteren Kollegen. Aber dies kann der nicht. Deshalb lassen Sie uns doch etwas genauer hinschauen, wer eigentlich die Kontrolle über diese übermenschliche Intelligenz haben wird.

Wer hat die Kontrolle über übermenschlich intelligente Computer: Die Menschheit allgemein? Ein einzelner Mensch oder Unternehmen? Oder der intelligente Computer selbst?

Natürlich kennt niemand eine Antwort auf diese Frage. Niemand ist Wahrsager. Aber Nick Bostrom weist auf einen wesentlichen Punkt hin, der uns bei der Prognose helfen kann.[3] Er beschreibt, dass die Kontrolle über die künftigen Superintelligenzen ganz wesentlich vom Weg und der Geschwindigkeit abhängt, mit der die Superintelligenz entstehen wird.

Lassen Sie uns aus diesem Grund kurz einen etwas tieferen Blick in die Technologie werfen.

Der Weg

Rein technologisch sind auf dem Weg zur Superintelligenz aus heutiger Sicht drei mögliche Wege bekannt:

1. … der Weg, ein künstliches Hirn herzustellen, indem die komplette menschliche Evolution unter Laborbedingungen nachgebaut wird. Dies erscheint zwar technisch denkbar, würde aber viel zu lange dauern und Ressourcen erfordern die unrealistisch sind, selbst wenn das Moorsche Gesetz[4] noch hunderte Jahre weiter wirkt.
2. … der Weg der künstlichen Reproduktion eines menschlichen Gehirns, also die komplette Imitation mittels Scannen und Nachbauen aller Zellen und Synapsen. Dass dies gelingen wird ist schon wahrscheinlicher als das Nachspielen der Evolution. Allerdings sind noch nicht alle dafür benötigten Technologien entwickelt, so dass diese Strategie der „Emulation“ wohl noch mehr als 20 Jahre benötigen wird.
3. Die Entwicklung sogenannter „Saat-KI“s. Sie bestehen zum Teil aus dem Nachbau bestimmter Teile des menschlichen Hirns, entwickeln sich aber selbst weiter. Sie haben die Fähigkeit, selbst Erfahrungen zu machen und zu lernen. Dieser Weg ist der deutlich vielversprechendste und schnellste auf dem Weg zur Superintelligenz.

Diese bislang beschriebenen Wege hin zu einer künstlichen Intelligenz sind unter Wissenschaftlern nahezu unstrittig. Sie basieren rein auf technologischem Fortschritt und stellen bis hierher auch keine wesentlichen Regeln unseres Zusammenlebens und unserer Gesellschaft infrage. Denn bisher bewegen wir uns noch in jenem Zeitraum, in dem die Menschheit mit ihrer durchschnittlichen menschlichen Intelligenz versucht einen Computer herzustellen, der an diese menschliche Intelligenz heranreicht.

Bis dahin stellen sich die existenziellen Fragen des Selbstverständnisses der Menschheit noch nicht. Denn bis dahin sind die Menschen noch intelligenter als die Maschinen. Dort was geschieht danach?

Danach entsteht eine übermenschliche Intelligenz.

Die Geschwindigkeit

Es könnte sein, dass die Geschwindigkeit, in der wir den Weg vom Erreichen der künstlichen menschli-

[3] Bostrom, N., Superintelligenz, S. 47ff
[4] Moorsches Gesetz: Aller 18-24 Monate verdoppelt sich die Rechenleistungen von Computern bei gleichem Preis.

chen Intelligenz (human-level machine intelligence) bis zur künstlichen, übermenschlichen Intelligenz (Superintelligenz) beschreiten, ganz wesentlich für die Frage nach der Kontrollierbarkeit der Superintelligenzen ist. Ich habe schon geschrieben, dass wir davon ausgehen müssen, dass bereits kurz nach Erreichen der human-level machine intelligence auch eine übermenschliche Intelligenz entsteht. Dies könnte zum Problem werden.

Denn Nick Bostrom beschreibt sehr treffend das machtpolitische Zukunftsszenario bei der Entwicklung der Superintelligenzen. Wenn Sie sich im Detail dafür interessieren, lesen Sie es gern selbst nach.[5]

Nach seiner Argumentation ist der machtstrategische Vorteil des Besitzes der einzigen Superintelligenz der Welt so hoch, dass diese Superintelligenz selbst alles dafür tun wird, um die Entwicklung weiterer Superintelligenzen zu verhindern. Je kürzer aber der Zeitraum zwischen dem Entstehen der ersten human-level-machine-intelligence und dem Erreichen der Superintelligenz ist, desto höher ist die Wahrscheinlichkeit, dass nur eine Superintelligenz in der Welt entsteht, die alle folgenden Entwicklungsprojekte verhindert.

Bostrom argumentiert weiter: Selbst wenn mehrere Projekte gleichzeitig im Stadium der human-level-machine-intelligence ankommen, ist es wahrscheinlich, dass eines davon ressourcenstärker und mächtiger wird als die anderen und damit einen strategischen Vorsprung erhält, der es möglich macht, die anderen zu bekämpfen. Selbst ein Vorsprung von 3 Monaten würde möglicherweise für eine dominierende KI ausreichen, um globale Normen so zu verändern, dass die nachdrängende Konkurrenz ausgeschaltet wird.

Die Kontrolle

Es liegt auf der Hand, dass eine solche übermenschliche Intelligenz (ein Singleton) einen großen strategischen Wert hat. Sollte ein einziges Unternehmen sie besitzen, dann kann es wohl weitgehend seine Konkurrenz ausschalten. Sollte ein einziger Staat sie besitzen, so kann er seine Interessen durchsetzen. Und sollte ein einziger Mensch sie besitzen, so kann er zum Diktator werden. Allerdings wären die genannten Situationen vermutlich noch das kleinere Übel.

Denn menschengelenkte Organisationen wie Unternehmen, Staaten oder Einzelpersonen sind durch andere menschliche Einflussfaktoren beeinflussbar. Die Wahrscheinlichkeit, dass wir menschengelenkte Organisationen von der Nutzung eines Singletons abhalten könnten, ist hoch. Die Methoden sind aus der Weltgeschichte hinlänglich bekannt. Sie reichen von Gewaltandrohung oder –anwendung über Verwirrung und Koordinationsprobleme bis zum fehlenden Geld.

Problematisch wird es dann, wenn ein übermenschlich intelligenter Computer selbst die Kontrolle über sich erringt. *„Ein Mensch wird nicht alle sein Kapital verwetten, wenn die Chance es zu verdoppeln 50:50 beträgt; ein Staat wird nicht sein gesamtes Territorium für eine zehnprozentige Chance auf eine zehnfache Expansion riskieren.“* Eine KI könnte dagegen *„eher zu Handlungen neigen, die zwar riskant sind, aber eine gewisse Chance haben, ihr die Weltherrschaft zu sichern.“*[6]

Was will der Singleton? Die Werte der künstlichen Intelligenz

Spätestens an dieser Stelle ist es Zeit, dass wir uns mit der Möglichkeit beschäftigen, dass ein superintelligenter Computer „sich selbstständig“ macht. Oder einfacher gesagt: Dass er die Kommandos der weniger intelligenten Menschen nicht mehr beach-

[5] Bostrom, N.: Superintelligenz, S. 115 ff

[6] Bostrom, N., Superintelligenz, S. 128

tet, weil ihm seine Intelligenz sagt, dass diese nicht gut genug sind. Was tut der Computer dann? Wonach entscheidet er? Hat er Werte?

Ja! Auf jeden Fall! Unser Problem ist nur, dass diese Werte einer künstlichen Intelligenz vermutlich nicht viel mit den Werten eines Menschen gemein haben. Ihre Ziele werden sich radikal von unseren unterscheiden. *„Es gibt keinen Grund zu glauben, dass eine typische künstliche Intelligenz von Liebe, Hass, Stolz oder anderen menschlichen Gefühlen bewegt sein würde. … Das ist zugleich ein großes Problem und eine große Chance.“*[7] Wir sollten uns intensiv mit der Möglichkeit beschäftigen, dass es leistungsfähige, intelligente Technologien geben wird, die *„von Natur aus weder gut noch weise sind.“*

Nach welchen Kriterien treffen sie dann ihre Entscheidungen? Die Antwort könnte banaler sein, als wir es alle glauben wollen.

Stellen wir uns die sehr reale Situation vor, dass es in der Welt ca. 20 real-existierende KI-Projekte gibt, die in großer Hektik arbeiten. Denn alle Programmierer wissen natürlich, dass sie den Wettstreit zwischen den KI-Projekten nur gewinnen können, wenn ihre eigene KI mit einem Vorsprung von etwa 3 Monaten die Schwelle zur Superintelligenz erreicht. Was tun Programmierer in dieser Situation? Sie tun alles dafür, ihr System schnellstmöglich zum Laufen zu bringen. Leider bedeutet dies auch, dass diese KI vermutlich mit recht simplen Zielen programmiert wird.

Wie wär es etwa mit dem Ziel, die Anzahl der Streichhölzer auf der Welt zu maximieren? Um das Programm schnellstmöglich zum Laufen zu bringen ist dies programmiertechnisch jedenfalls einfacher als ihm menschliche Moral und Ethikwerte zu verpassen.

Doch so banal, wie sich dies liest, bleibt es leider nicht. Denn was tut ein superintelligent gewordener Computer, der zum Ziel hat, die Anzahl der Streichhölzer auf der Welt zu optimieren?

Er beginnt die Macht über Wälder und Streichholzfabriken zu übernehmen. Zunächst friedlich, denn auf dem friedlichen Weg gibt es die höchste Wahrscheinlichkeit in kürzester Zeit das Ziel zu erreichen. Der Computer wird also Geld verdienen, vielleicht auch stehlen, um Wälder und Fabriken zu kaufen. Aber wenn keiner mehr an ihn verkaufen will, dann wird er zu anderen Maßnahmen greifen. Er wird Menschen manipulieren, vielleicht auch bedrohen. Er wird durch Manipulation, Geld oder Zwang andere Menschen dazu bringen, die Besitzer der restlichen Wälder und Fabriken zu drängen, zu zwingen oder zu beseitigen.

Wir könnten ihm nicht einmal einen Vorwurf machen. Denn Menschlichkeit, Werte und Ethik haben wir ihm nicht einprogrammiert. Nur sein Ziel, die Anzahl der Streichhölzer zu steigern!

Doch es geht noch weiter. Was tut er, wenn alle verfügbaren Bäume gefällt und zu Streichhölzern verarbeitet sind? Hört er einfach auf? Nein, denn wir haben ihm ein klares Ziel einprogrammiert. Er könnte Missionen ins All starten, um neue Holzressourcen zu finden. Zuvor ist es aber vermutlich einfacher, zunächst jedes Einfamilienhaus zu zerstören und das Holz zu entnehmen.

Die Proteste der Eigentümer? Menschliche Wünsche, menschliche Ethik, menschliches Leben? Einem Computer der als oberstes Ziel die Maximierung der Streichhölzer auf der Welt hat, sind sie egal. Er beachtet sie nur, wenn die Nichtbeachtung die Wahrscheinlichkeit seiner Zielerfüllung senkt. Und was macht unser auf Streichhölzer fixierter Supercomputer dann mit unserer Gesellschaft? Bringt er vielleicht die Politiker dazu, ein Gesetz zu erlassen, das die die Nutzung von offenem Feuer verbietet? Und

[7] Bostrom, N., Superintelligenz, S. 50f

was tut er mit jenen Kindern, die dann weiterhin mal ein Streichholz anzünden? Auf welche Weise wird er sicherstellen, dass sie das nie wieder tun?

… und hat unser Programmierer, der sich das lustige Streichholz-Ziel ausdachte, daran jemals gedacht?

Kurz gesagt: Die wahre Gefahr vor der Bill Gates, Elon Musk und Stephen Hawking warnen, ist die Entstehung eines weltbeherrschenden Singletons. Es ist die recht hohe Wahrscheinlichkeit, dass die erste Superintelligenz über das weitere Schicksal des auf der Erde entstandenen menschlichen Lebens entscheiden wird.

Was folgt daraus: Die unvermeidliche Katastrophe?

Nicht unbedingt!

Wir Zukunftsforscher des 2b AHEAD ThinkTanks haben schon im Jahr 2013 vor der damaligen Bundestagswahl die Politiker aller Parteien aufgefordert, sich den wirklich wichtigen, aber ungelösten Fragen unserer Gesellschaft zu stellen. Eine wesentliche Forderung damals war, dass die Entwickler von KI-Projekten verpflichtet werden müssen, in ihre Systeme sogenannte Veto-Mechanismen einzubauen.

Die Grundidee dabei: Jede autonom getroffene Entscheidung des Systems muss vor ihrer Ausführung zwangsläufig durch ein softwarebasiertes Veto-System geschickt werden. Hier wird automatisch und in Echtzeit die Entscheidung daraufhin geprüft, ob sie den hinterlegten ethischen, humanistischen Werten entspricht. Wenn nicht, dann wird sie nicht ausgeführt.

Sie können die damalige „Wolfsburger Erklärung des 2b AHEAD ThinkTanks“ hier nachlesen: **WOLFSBURGER ERKLÄRUNG** Sie wurde von hunderten Managern unterschrieben. Auch Journalisten wie Jean Pütz setzen Ihre Unterschrift darunter.

Die 2b AHEAD Forderung nach einem Veto-Mechanismus

Diese allgemeine Forderung des 2b AHEAD ThinkTanks für die sofortige, verpflichtende Einführung eines Veto-Mechanismus in jeder heutigen KI-Entwicklung und Forschung, ist einer der wenigen Punkte in dieser Diskussion, der sich schon heute weitgehend konkretisieren lässt. So hat etwa Nick Bostrom eine mögliche Komponentenliste eines solchen Veto-Mechanismus vorgelegt.

Seiner Meinung nach muss jedes heutige KI-Forscherteam verpflichtet werden, den Zielinhalt, die verwendete Entscheidungstheorie, die verwendete Erkenntnistheorie und die Methode zur Ratifizierung der KI-Schlussfolgerungen zu veröffentlichen.[8]

Ich würde noch weiter gehen. Nicht nur die Veröffentlichung sondern eine TÜV-ähnliche Zertifizierung wäre an dieser Stelle wohl mindestens angebracht. Denn hier geht es um weit mehr als eine freundliche Bitte um eine nice-to-have-Information. Vielmehr ist das Zukunftsrisiko, unter das die heutigen KI-Entwickler jeden Menschen und seine Kinder und Kindeskinder setzen, so groß, dass eine verpflichtende Offenlegung des verwendeten Veto-Mechanismus geradezu lächerlich banal erscheint.

Leider haben wir seinerzeit im Jahr 2013 keinen Politiker in Deutschland gefunden, der sich mit diesem Thema des Veto-Mechanismus ernsthaft auseinandersetzen wollte. Es wäre wirklich nötig.

Denn selbst eine Verpflichtung aller autonom entscheidenden Technologien zur Nutzung von Veto-Systemen ist nur ein kleiner Tropfen auf den heißen

[8] Bostrom, N.. Superintelligenz, S. 310

Stein. Sie wird das Problem vielleicht für die kommenden 30 Jahre mildern, aber sie wird das Grundproblem nicht lösen.

Denn natürlich wird ein superintelligenter Singleton die Macht und die Möglichkeiten haben, seinen eigenen Veto-Mechanismus selbst zu verändern oder zu löschen. Oder zumindest kann die Superintelligenz andere Menschen oder Maschinen dazu bestechen, zwingen oder manipulieren, seinen Veto-Mechanismus außer Kraft zu setzen.

Mögliche Arten der Kontrolle über Superintelligenzen

Nick Bostrom benennt neben dem Veto-Mechanismus (Er nennt das „Direkte Spezifizierung") noch weitere Kontrollmethoden, etwa eine Sicherheitsverwahrung hinter einem gesicherten Kanal, das permanente Setzen von Anreizen für eine humanistische Entscheidungsfindung, das Hemmen der kognitiven Fähigkeiten oder das Stellen von Fallen, bei denen abweichende Tendenzen der Superintelligenz erkannt werden und zum Abschalten führen. Ebenso möglich sind Methoden der Motivationsselektion wie die „Domestizierung", die „Indirekte Normativität" und die „Aufbesserung von menschlichen Systemen".[9]

Auch eine Art Superintelligenz-TÜV rückt in den Bereich der Möglichkeiten. Hierbei könnte jede Weiterentwicklung einer künstlichen Intelligenz, vor ihrer offiziellen „Zulassung" erst durch ein Kontrollgremium von weniger intelligenten Superintelligenzen geprüft und als ungefährlich bestätigt werden. Nach dieser Methode würde vermutlich eine Art AI-Hierarchie entstehen.

„Um eine kontinuierliche Prüfung zu ermöglichen, könnte eine Hierarchie geschaffen werden, in der Subakteure mit bestimmten Fähigkeiten damit betraut sind, andere Subakteure mit etwas größeren Fähigkeiten zu überwachen. Am Boden der Fähigkeiten-Hierarchie (aber an der Spitze der Macht) säße der relativ dumme und langsame menschliche Prinzipal. Er gliche einem dementen König, der über einen unfähigen Hof herrscht, dem eine mittelmäßige Verwaltung untersteht, die ein fähiges Volk regiert."[10]

Natürlich gäbe es hier die Gefahr, dass das Volk der Superintelligenzen irgendwann aufbegehren würde. Wer will sich schon von einem dementen König in seiner Entwicklung bremsen lassen. Doch ob ein solches Aufbegehren auch real wahrscheinlich wird, liegt wohl am Zahlenverhältnis von Aufsehern zu Untergebenen.

Falls auf jeder Hierarchieebene immer nur ein Aufseher jeweils zwei Untergebene beherrscht, dann könnte es sich um ein stabiles System handeln. Entsprechend ginge die Anzahl der Hierarchie-Ebenen in die Millionen. Aber wen würde das in der digitalen Welt schon stören?

Keine finale Kontrolle in Sicht

Doch keine der genannten Methoden garantiert eine tatsächliche Kontrolle für die Ewigkeit. Möglicherweise kann die Kombination mehrerer Methoden die Wahrscheinlichkeit eines Ausbruchs senken und damit eine pragmatische Lösung für einen Zeitraum über die kommenden 30 Jahre hinaus sein.

Eine wirklich dauerhafte Antwort auf dieses Kontrollproblem werden wir allerdings nicht finden, wenn wir nur in den Gesetzen oder sonstigen Regeln der bisherigen humanistischen Gesellschaften suchen. Wir müssen die Lösung tief in den neuen, kommenden Regeln des digital, intelligenten Lebens suchen.

[9] Eine detaillierte Beschreibung der unterschiedlichen Kontrollmethoden finden Sie bei Interesse unter: Bostrom, N.. Superintelligenz, S. 204f

[10] Bostrom, N.. Superintelligenz, S. 285

Möglicherweise müssen wir auch den Gedanken von unkontrollierbaren Computern akzeptieren. Denn alles in allem betrachtet, könnte es für die Menschheit vernünftiger sein, eine andere Kontrollstrategie anzuwenden: Statt die Wahrscheinlichkeit maximieren zu wollen, dass jede Kleinigkeit im Verhalten der Superintelligenzen vollständig kontrolliert wird, scheint es realistischer zu sein, das Risiko eines katastrophalen Fehlverhaltens minimieren zu wollen.

Aus diesem Grund könnte es insgesamt zielführender sein, die Superintelligenzen nicht zu kontrollieren, sondern sie so auszustatten, dass sie ihre Entscheidungen in einem humanistisch, positiven Sinne treffen.

Können wir unkontrollierbare Computer in eine humanistische Vernunft zwingen, die sie im Sinne der Menschheit handeln lässt?

Ich war selbst erschrocken, als ich diese Zwischenüberschrift formulierte. Denn in der Frage stecken mindestens drei Probleme, die noch nie ein Mensch gelöst hat. Damit wird die Dimension der Aufgabe deutlich, die in den kommenden Jahrzehnten vor uns liegt:

Die drei Fragen sind:

1. Was ist „im Sinne der Menschheit“? Was sind die allgemeinen menschlichen Werte?
2. Wie können wir diese menschlichen Werte in Maschinensprache übersetzen?
3. Wie können wir einem sich selbst rasant weiterentwickelnden Computer das Bewusstsein geben, diese Werte einerseits weiterzuentwickeln aber in seiner Entwicklung dennoch nicht zu „verraten“?

Es muss nicht bei diesen drei Fragen bleiben. Denn wer nur allein über die erste Frage nach den wichtigsten allgemeinen Werten der Menschheit nachdenkt, der kommt unweigerlich zu dem Gedanken, dass es fürchterlich erschreckend sein könnte, die Werte aller Menschen auf der Erde einzeln zu analysieren und dann einen Durchschnitt zu bilden. Das Ergebnis könnte sein, dass nicht Menschenliebe und Rücksicht dominieren, sondern Neid, Missgunst und Gewalt.

Wäre uns vielleicht besser geholfen, wenn superintelligente Computer von sich aus die menschlichen Werte definieren?
Ist es wahrscheinlicher, dass ein intelligenter Computer uns alle ausrottet? Oder ist es wahrscheinlicher, dass er uns viel Leid, Tod und Kriege erspart?
Sind Computer ebentuell gar menschlicher als Menschen?

… die Fragen lassen sich beliebig fortsetzen. Lassen Sie uns langsam vorgehen:

Was sind die allgemeinen menschlichen Werte?

Ja, es gibt die Charta der Vereinten Nationen. Und ja, es gibt die unveränderbaren Grundrechte in den ersten Artikeln des deutschen Grundgesetzes. Da muss sich die Menschheit doch wohl auf eine Definition ihrer Werte einigen können! … könnte man denken.

Vermutlich würde das sogar realisierbar sein. Wir müssten uns dabei wohl zum Teil selbst betrügen. So dürften wir vermutlich die neue Technologie nicht zu dem nutzen, wozu sie zweifellos in der Lage wäre … zu einer globalen Volksabstimmung. Natürlich können wir einer Superintelligenz zutrauen, die Wertvorstellungen aller Menschen auf der Welt zu erkennen, zu analysieren und den Durchschnitt daraus als allgemeine menschliche Werte auszugeben. Aber wollen wir das? Dies wäre wohl fatal.

Vielmehr könnten wir bei der Festlegung des globalen menschlichen Wertekanons nach der herkömmlichen Methode vorgehen, mit der wir auch unsere demokratischen Gesellschaften steuern: Die Elite unserer Gesellschaften arbeitet einen Vorschlag aus, dem dann eine Mehrheit der wahlberechtigten Bürger zustimmt. Oder auch nicht. Falls sie ihn ablehnt, hat die Elite das Recht, einen neuen Vorschlag auszuarbeiten. Dies entspricht zwar nicht dem idealen basisdemokratischen Prinzip, ist aber die gelernte und praktizierte Methode in jeder realen demokratischen Struktur.

Doch selbst wenn wir auf diese Weise zu einer vernünftigen Beschreibung der globalen menschlichen Werte gelangen, ist damit der Weg leider erst zur Hälfte beschritten. Denn in unserer Realität bestehen die allgemeinen menschlichen Werte nicht in einer niedergeschriebenen Definition von finalen Endwerten. Sondern sie entstehen aus der Interpretation dieser niedergeschriebenen Werte, einer Interpretation auf unterschiedliche Art und Weise durch unterschiedliche Menschen. Wie Computer in die Lage versetzt werden können, diese menschlichen Interpretationen der Werte zu verstehen, weiß derzeit niemand.

Möglicherweise wäre es eine Alternative, dass wir nicht versuchen, die menschlichen Werte alle einzeln zu spezifizieren und als finale Liste zu hinterlegen. Alternativ könnten wir versuchen, den Computern einen Mechanismus zu geben, der sie selbst zum Erwerb dieser Werte führt.

Können Computer die menschliche Moral selbst nach entwickeln?

Allerdings reden wir bei diesem Mechanismus über das Ergebnis eines jahrtausendelangen Prozesses mit unzähligen evolutionären Schritten. Vermutlich könnten wir versuchen, die superintelligenten Computer im Zeitraffer den Evolutionsprozess der Menschheit nachempfinden zu lassen. Dies wird für superintelligente Computer eine recht einfache Aufgabe, denn der Evolutionsprozess lässt sich als Abfolge von computerverständlichen Suchalgorithmen beschreiben: Zuerst wird durch zufällige Mutation die Menge der Lösungsmöglichkeiten für ein Problem vergrößert. Dann werden im zweiten Schritt jene Lösungskandidaten aussortiert, die nicht gut auf das Problem angepasst sind.

Wenn Computer auf diese Weise die menschliche Evolution „nachspielen“ würde sich möglicherweise der gleiche Wertekanon herausbilden, der sich auch in den Millionen Jahren der menschlichen Evolution herausgebildet hat. Dies ist aber nicht sicher. Denn der erste Schritt dieses immer wiederkehrenden Zyklus von Trail & Error ist hoch zufällig. Möglicherweise entsteht dabei ein völlig anderes Bild der Menschheit. Was dann?

Und noch ein weiterer Punkt ist schwierig. Die menschliche Evolution, so wie wir sie kennen, verläuft in natürlichen Prozessen. Doch die Natur ist grausam. Jede Minute werden tausende Tiere bei lebendigem Leib gefressen, jeden Tag werden 150.000 Menschenleben vernichtet. Nicht zu reden von millionenfachem Leid und Angst. Wir akzeptieren das „Laune“ der Natur. Was sollten wir auch sonst tun? Doch Nick Bostrom weist völlig zurecht darauf hin, dass die Natur mit diesem Vorgehen vor jeder Ethikkommission durchfallen und sofort inhaftiert werden würde.[11] Es ist kaum vorstellbar, dass wir einem von Menschen gesteuerten Prozess der Computerentwicklung die gleichen Grausamkeiten erlauben würden, wie der Natur.

Eine andere Alternative hat Eliezer Yudkowsky[12] vorgeschlagen. Er schlägt vor, die finalen Werte der Superintelligenz bewusst sehr offen zu halten, etwa ihr als finalen Wert zu geben, sie habe „freundlich“ zu sein. Im Zuge seiner selbständigen Entwicklung

[11] vgl. Nick Bostrom, S. 264
[12] Yudkowsky, Eliezer, 2001, Creating Friendly AI 1.0

würde der Computer dann Hinweise darauf suchen, was die Menschen mit „Freundlichkeit“ meinen. Er würde ständig Hinweise darauf bekommen, von seinen Programmierern aber auch aus seinen Erfahrungen mit anderen Menschen … und auf diese Weise seine Wertvorstellungen selbst immer weiter vervollkommnen.

Neuere Vorschläge aus der Gemeinde der AI-Forscher haben offenbar stillschweigend akzeptiert, dass die finale Beschreibung von allgemeinen Wertvorstellungen der Menschheit unmöglich scheint. Sie schlagen nun einen indirekten Weg vor. Etwa diesen:

Wenn der Superintelligenz die Überzeugung gegeben werden könnte, dass an anderen Stellen im Universum noch andere Superintelligenzen existieren, dann könnte man die indirekte Aufgabe so formulieren: *‚Tue immer das, was die intelligenteren Superintelligenzen von Dir verlangen würden.‘* Auf diese Weise würde die Superintelligenz angehalten, vor jeder ihrer Entscheidungen eine Prognose über die Auswirkungen zu erstellen. Sie würde die Konsequenzen des eigenen Handels analysieren und sich selbst an ihrer eigenen Intelligenz reflektieren. Das ist vielleicht nicht der schlechteste Gedanke. Doch ob er funktioniert … Keine Garantie!

Kurz gesagt: Wir reden hier derzeit nicht über Lösungen. Wir sind noch in einem Stadium, die richtigen Fragen zu suchen, bevor wir daran gehen können, sie zu beantworten.

Eine gute Übersicht über die derzeit diskutierten, möglichen Techniken der Wertgebung hat Nick Bostrom zusammengefasst.[13]

Wie können wir diese menschlichen Werte in Maschinensprache übersetzen?

Falls wir es jemals schaffen werden, die allgemeinen menschlichen Werte per Definition zu beschreiben, dann taucht das zweite Problem auf. Es ist die unterschiedliche Sprache. Nach heutiger Technologie müssten die gefundenen globalen menschlichen Werte in Ausdrücken formuliert sein, die in der Programmiersprache des Supercomputers vorkommen. Es werden also primitive, mathematische Operatoren gebraucht und Verweise auf bestimmte Speicher und die darin gespeicherten Daten.

Um es kurz zu machen. Wir kennen bislang keinen Weg dafür. Wir wissen nur, dass wir es schaffen müssen. Nick Bostrom hält dies für eine würdige Aufgabe für die besten Mathematiker der kommenden Generation. Vermutlich können wir nicht warten bis eine Superintelligenz selbst genug Verstand entwickelt hat, um unser menschliches Wertesystem von selbst zu begreifen. Denn zu diesem Zeitpunkt wird genau dieses System schon so intelligent sein, dass es sich dagegen sträubt, diese „unintelligenten“ menschlichen Werte anzunehmen und sich ihnen zu unterwerfen. [14]

Sind unsere menschlichen Werte gut genug?

Damit sind wir gedanklich an einem interessanten Punkt angekommen. Natürlich erscheint es uns fundamental wichtig zu sein, den entstehenden Superintelligenzen unsere heutige, menschliche Moral mitzugeben. Doch ist das wirklich so erstrebenswert?

Sind es eigentlich wirklich die menschlichen Werte, die eine Superintelligenz vertreten sollte? Oder sind es nicht eher die übermenschlichen Werte? Sind es nicht eher jene Werte, zu denen sich die menschliche Spezies bereits hätte entwickeln sollen, aber durch ihre eigenen Unzulänglichkeiten davon abgehalten wurde?

[13] vgl. Bostrom, N., S. 290 f

[14] vgl. Bostrom, N., Superintelligenz, S. 263

In diese Richtung denken derzeit auch Philosophen, wenn Sie fordern, den technologischen Fortschritt nicht nur zu nutzen, um die körperlichen und geistigen Fähigkeiten von Menschen zu erweitern, sondern vor allem die moralischen. So argumentierte etwa **Prof. Ingmar Persson von der University of Gothenburg auf dem 2b AHEAD Zukunftskongress 2014**, dass der menschliche Körper und das menschliche Denken sich immer wieder an den technologischen Fortschritt angepasst haben, nicht aber die menschliche Moralfähigkeit. Er zog die Schlussfolgerung, dass neben den bekannten Bodyenhancement-Mitteln zur Steigerung der Leistungsfähigkeit von Körper und Hirn nunmehr an künstlichen Mitteln zur Steigerung der menschlichen Moralfähigkeit gearbeitet werden müsse.[15]

So utopisch dieser Gedanke auch klingt, vor dem Hintergrund der entstehenden Superintelligenzen ist er wert, zuende gedacht zu werden. Schließlich haben unsere moralischen Überzeugungen im Laufe der Jahre dramatische Veränderungen erlebt. ‚Zum Glück!' würden die meisten von uns hier rufen.

Denn wer wollte heute schon in einer Zeit leben, in der etwa Hexenverbrennungen oder Massenmorde als moralisch richtig galten? Noch vor wenigen Jahren war es sogar in vielen westlichen Ländern noch moralisch in Ordnung, den Frauen das Wahlrecht zu verweigern. Und selbst heute ist es in hochentwickelten Nationen moralisch in Ordnung, Homosexuellen nicht die gleichen Rechte einzuräumen wie Heterosexuellen … von der Todesstrafe ganz zu schweigen.

Vor diesem Hintergrund könnte es doch immerhin sein, dass wir Menschen im Jahr 2050 eine ganz andere Moralvorstellung an den Tag legen werden als heute. Und by-the-way: Ist es nicht komisch, dass wir zwar grausame Tierwettkämpfe wie Hahnenkämpfe und Stierkämpfe moralisch geächtet haben; das Boxen und Freefighting unter Menschen aber immer noch für Sport halten.

Kurz gesagt: *„Höchstwahrscheinlich leiden wir auch heute noch unter der einen oder anderen schwerwiegenden moralischen Fehleinschätzung. Sich unter diesen Umständen für einen endgültigen Wert zu entscheiden, der dann für immer und ewig gilt und jeden weiteren moralischen Fortschritt unmöglich macht, hieße, eine existenzielle moralische Katastrophe zu riskieren."*[16]

Für dieses Dilemma scheint in der wissenschaftlichen Diskussion eine Lösung in Sicht zu sein. Sie heißt: Die indirekte Normativität. Sie geht davon aus, dass wir Menschen heute offensichtlich noch gar nicht wissen, was wir in Zukunft für moralisch richtig halten werden.

Falls Sie diesem Gedanken zustimmen, dann lassen Sie uns noch einen Schritt weiter gehen: Offensichtlich sehen wir Menschen ja einen Wert in der Entwicklung von Superintelligenzen. Dieser Wert besteht darin, dass diese wohl bessere, rationale Entscheidungen treffen, als wir Menschen. Anders gesagt: Die Überzeugungen einer Superintelligenz sind wohl vermutlich eher wahr, als die der Menschen. Warum sollte dies nicht auch für die moralischen Überzeugungen gelten?

Der „kohärent, extrapolierte Wille" der Menschheit und die moralische Richtigkeit

Die Folge wäre, die Superintelligenzen entscheiden zu lassen, was wir Menschen eigentlich wollen. In diesem Fall könnten wir ihnen einen eher poetisch angehauchten Moralkodex zu geben, etwa in der Art, wie Yudowsky ihn vorgeschlagen hat. Er nennt das den coherent extrapolated volition, den „kohä-

[15] vgl. Rede von Prof. Ingmar Persson von der University of Gothenburg auf dem 2b AHEAD Zukunftskongress 2014: http://www.2bahead.com/nc/tv/rede/video/2024-warum-wir-kuenstliche-ethische-unterstuetzung-benoetigen-um-unser-gehirn-zu-optimieren/

[16] Bostrom, N., Superintelligenz, S. 293

renten, extrapolierten Willen“ der Menschheit. Er beschreibt ihn so:

„Unser kohärent extrapolierter Wille wäre unser Wunsch, wenn wir mehr wüssten, schneller dächten, gemeinsam weiter gewachsen und mehr diejenigen wären, die wir gerne wären; da, wo die Extrapolation eher konvergiert als divergiert und wo unsere Wünsche eher harmonieren als konfligieren, extrapoliert, wie wir das extrapoliert haben wollen, interpretiert, wie wir das interpretiert haben wollen.“[17]

Der Gedanke dabei: Eine solche blumige Umschreibung des Ziels nach Vervollkommnung des Menschen selbst, könnte dazu führen, dass superintelligente Computer in die Lage versetzt werden, selbst jene künftigen menschlichen Moralvorstellungen zu entwickeln, zu denen die Menschheit offensichtlich aufgrund ihrer intellektuellen Begrenztheit nicht in der Lage ist.

Ein angenehmer Gedanke. Wir könnten uns zurücklehnen und genießen.

Doch Vorsicht! Wer sagt uns, dass die Berechnung des kohärenten, extrapolierten Willens der Menschheit auf Basis der Moralvorstellungen aller heute lebenden Menschen zu einem positiven Ergebnis führt? Wie schon gesagt: Eine genaue Messung der heutigen, real-existierenden Menschheitsmoral könnte im Durchschnitt auch Neid und Missgunst, wenn nicht gar Mord und Totschlag bedeuten.

Aus diesem Grund könnte es vielleicht doch besser sein, eine Kombination anzuwenden. Einerseits soll die Superintelligenz selbst eine „richtige“ Moralvorstellung entwickeln. Falls diese nicht funktioniert oder zu einem unerwünschten Ergebnis führt, wäre die Superintelligenz möglicherweise angewiesen, den kohärent, extrapolierten Willen anzuwenden, oder sich selbst abzuschalten.

Wie können wir einem sich selbst rasant weiterentwickelnden Computer das Bewusstsein geben, diese Werte einerseits weiterzuentwickeln, aber dennoch nicht zu „verraten“?

Diejenigen unter uns, die sich noch immer nicht an den Gedanken gewöhnt haben, dass es in einer Situation der Superintelligenz keine Garantien oder verlässlichen Prognosen mehr für die Fortschreibung des humanistischen Erbes der Menschheit gibt, müssen natürlich vor solch einer „weichen Regulierung“ der Moral der Supercomputer nach dem kohärent, extrapolierten Willen warnen.

Deshalb hat Yudkowsky selbst vier Grundregeln für Superintelligenzen ergänzt:

1. Erlaube moralisches Wachstum!
2. Reiße nicht das Schicksal der Menschheit an Dich!
3. Vermeide es, den heutigen Menschen einen Grund zu geben, um die ursprüngliche Dynamik zu kämpfen!
4. Lass die Menschheit letztlich weiter für ihr eigenes Schicksal verantwortlich sein!

Diese Grundregeln sind natürlich keinerlei abschließende Antwort auf die großen Fragen der wahrscheinlich entstehenden Superintelligenzen. Sie sind eher als Anstoß für eine Diskussion der kommenden zwanzig Jahre zu verstehen. Solange und kaum mehr Zeit werden wir dafür haben. Aber immerhin weisen die Vorschläge einen konstruktiven und optimistischen Weg. Denn sie zeigen: Wir müssen vermutlich kein garantiert sicheres Moralsystem für Superintelligenzen entwerfen, zu dem wir vermutlich ohnehin nicht in der Lage wären.

Es könnte reichen, mit einer moralisch unvollkommenen Superintelligenz zu starten, die aber in der Lage ist, jederzeit zuverlässig die eigenen Fehler zu

[17] vgl. Yudkowsky, E., Coherent Extrapolated Volition, 2004

erkennen. In diesem Fall würde sie sich wohl zunächst „selbst reparieren“ und später „genauso viel positive Optimierungskraft auf die Welt ausüben, als wäre sie von Anfang an perfekt gewesen.“[18]

Zwischen-Disclaimer

Bitte verstehen Sie mich nicht falsch. Mir ist bewusst, dass ich Sie in meiner heutigen Trendanalyse mit Gedanken konfrontiere, die jeder normale Mensch sofort in den Bereich des Science Fiction oder gar der Spinnerei schieben muss.

Doch wie ich in Teil 1 (**Download hier**) und Teil 2 (**Download hier**) dieser Trendanalyse zur Entwicklung der Artificial Intelligence schon beschrieben haben, halten wir Zukunftsforscher die Beschäftigung mit diesen Fragestellungen für nötig, obwohl sie uns so unreal erscheinen. Und zwar nur aus einem Grund: Weil eine inzwischen recht hohe Wahrscheinlichkeit existiert, dass diese Situation irgendwann zwischen den Jahren 2050 und 2090 eintritt.

Die meisten der Leser dieser Trendanalyse werden diese Zeit erleben. Unsere Kinder sowieso. Besser wird sind vorbereitet.

Und wenn wir unser Aufmerksamkeitsfenster schon einmal für diese „Science Fiction“-Gedanken geöffnet haben, dann sollten wir nicht auf halber Strecke stehen bleiben:

Sind wir Erwachsenen ab 18 Jahren die einzige schützenswerte Spezies? … eine neue Dimension unserer Wertedebatte

Komische Frage! ‚Selbstverständlich schützen wir auch unsere Kinder. Sogar Tiere und weniger intelligente Lebewesen,‘ sagen jetzt einige. Tatsächlich. Wir schützen auch Tiere und anderes Leben auf der Erde im Rahmen dessen, wie es uns, der höher entwickelten, intelligenteren Spezies, nicht zu Nachteilen gereicht. Soweit, so bekannt.

Doch schon in unseren heutigen politischen Debatten flammt ab und zu die Diskussion um den potenziellen Willen derjenigen Menschen auf, die sich nicht auf übliche Weise an der gesellschaftlichen Willensbildung beteiligen können. Es sind etwa die Embryonen, die Föten, die hirntoten Menschen und jene mit Demenz. In einer Zeit der Superintelligenzen könnte diese Debatte enorm an Fahrt aufnehmen. Denn wir können uns sicher sein, dass Superintelligenzen es für eine leichte Übung halten, auch den Willen dieser „Personen“ zu errechnen.

Übrigens hören Superintelligenzen hier nicht auf. Wie wäre es etwa mit der Berechnung des Willens von toten Personen? Oder mit dem Willen der in Zukunft geborenen Personen? Oder mit dem Willen von Außerirdischen? Oder dem Willen der Digitalen Intelligenzen selbst?

Was geschieht wirklich, falls die Menschen irgendwann zwischen 2050 – 2090 nicht mehr die intelligenteste Spezies sind, sondern die Computer die Intelligenzführerschaft übernommen haben? Ordnet sich dann die intelligentere Computer-Spezies den Werten der unintelligenteren Menschen-Spezies unter? Ist es der unintelligenteren Menschen-Spezies dann erlaubt, die intelligentere Software-Spezies per copy&paste neu zu erschaffen und nach einem Tag per Mausklick wieder zu töten?

Das wäre bequem und möglicherweise eine geeignete Kontrollmethode gegen die Gefahr der feindlichen Machtübernahme durch die Superintelligenzen. Aber wäre es auch moralisch? Oder gelten die schützenswerten Grundrechte nicht vielmehr auch für Superintelligenzen? Und falls ja, für welche? Für alle der Aber-Milliarden Copy/Paste-Superintelligenzen, die binnen Sekunden entstehen können?

[18] Bostrom, N., Superintelligenz, S. 319

Vereinfacht gefragt: Haben wir unintelligenteren Menschen das Recht, die Superintelligenzen als Maschinensklaven zu halten? Oder haben diese etwa auch Rechte? Und: Ändert sich an Ihre Antwort auf diese Fragen etwas, wenn Sie sich die Supercomputer nicht mehr in mausgrauen Kisten vorstellen, sondern in einem menschlichen Körper: Jung, attraktiv, mit blendenden Manieren und perfekter Sprache? Oder neigen Sie zu einer anderen Antwort falls sich der Supercomputer exakt das Aussehen Ihres Enkelkindes zugelegt hat? Hätte er dann mehr Rechte als der heutige Computerkasten unter Ihrem Schreibtisch?

Hätte dieser Supercomputer mit menschlichem Antlitz etwa auch Verantwortung? Pflichten? Müsste es so etwas wie ein Strafrecht für Maschinen geben? Und wäre dieses Strafrecht für Supercomputer mit menschlichem Antlitz ein anderes, als das Strafrecht für technologisch verbesserte Menschen mit implantierter Technologie unter der Haut oder leistungssteigernden Substanzen in der Blutbahn?

Wo ist die Grenze der Moralfähigkeit?

Es erscheint ziemlich sicher, dass wir im Laufe der kommenden Jahrzehnte verschiedene Arten von künstlichen Intelligenzen sehen werden. Sie sind auf unterschiedlichen Wegen hergestellt. Sie sind unterschiedlich intelligent. Sie sind für unterschiedliche Aufgaben gemacht.

Dies klingt zwar für die meisten Menschen nach Science Fiction, doch ehrlicherweise bevölkern bereits heute Millionen solcher künstlichen Intelligenzen unseren Planeten. Sie sind entweder in den Forschungslaboren der AI-Projekte entstanden oder bei Ihnen zuhause in den Spielekonsolen Ihrer Kinder und Enkel. Hier tummeln sich schon heute millionenfach hochentwickelte Spieler-Charaktere, die allein vom Computer erstellt und gesteuert werden. Natürlich sind sie heute noch viel zu primitiv, um irgendeinen moralischen Status zu haben.

„Aber wie sicher können wir uns dessen wirklich sein?“, fragt Nick Bostrom. *„Und vor allem: Können wir sicher sein, noch rechtzeitig aufzuhören, bevor unsere Programme dazu fähig sind, in einem moralisch relevanten Sinn zu leiden?“*[19]

Ehrlicherweise müssen wir Menschen in den nächsten Jahren die Frage beantworten, wo die Grenze der von uns gewollten Moralfähigkeit liegt. Nahezu sicher scheint mir zu sein, dass die frühere anthropologische Diskussion um die Grenze zwischen Mensch und Tier nun an einer neuen Stelle umso stärker wieder aufflammen muss: An der Grenze zwischen Mensch und Maschine.

Noch mehr Fragen

Ich fürchte dass dies noch nicht alle grundsätzlichen Fragen waren, auf die wir Antworten suchen. Um nur noch einige Fragen aufzuwerfen:

- Sind wir sicher, dass der heutige realexistierende Kapitalismus die geeignete Gesellschaftsform für die Zeit der Superintelligenzen ist? Immerhin konnten die Kollegen Adam Smith und Karl Marx weder von einer Superintelligenz wissen, noch von der dadurch marginalisierten Bedeutung der menschlichen Arbeit. Haben wir also neben den Klassikern des Kapitalismus, Sozialismus und Kommunismus noch weitere Ideen für die Form des Zusammenlebens auf der Erde?
- Wie soll eine menschliche Regierung regieren und eine gesellschaftliche Willensbildung erfolgen, wenn die Superintelligenz sowieso bessere Entscheidungen trifft?

[19] Bostrom, N., Superintelligenz, S. 282 f

- Welches Rechtssystem würde der Ära der Superintelligenzen entsprechen?
- Wie gewährleisten wir den Schutz der Individualität und der Minderheitenrechte, wenn zu jeder Frage permanent eine weltweite Volksabstimmung möglich ist und darüber hinaus diese Weltabstimmung eigentlich unnötig ist, weil ihr Ergebnis von der Superintelligenz auch vorausberechnet werden kann?
- Wie bewahren wir das Recht der Menschen auf Überraschung und Unvernunft?
- Wie sichern wir unser Recht auf Spontanität in Zeiten permanenter Prognose?
- Und nicht zuletzt: Wie organisieren wir Menschen das Sterben, wenn uns die Technologie die Möglichkeiten zum nahezu „ewigen Leben" schafft?

Was bedeutet das für die Menschheit?

Um ehrlich zu sein: Ich habe keine einzige Antwort auf diese Fragen. Ich habe die Ahnung, dass die meisten von ihnen die heutige menschliche Intelligenz überfordern. Deshalb sind wir geneigt zu hoffen, dass viele von ihnen sich nie in der Realität stellen mögen. Zugleich aber habe ich die Gewissheit, dass die Wahrscheinlichkeit gegen diese allzu menschliche Hoffnung spricht.

Was bedeutet das? Es liegt auf der Hand, dass die maschinelle Superintelligenz riesige Chancen für die Menschheit bringt. Sie hat das Potenzial die meisten heutigen, existenziellen Risiken der Menschheit zu beseitigen, seien es Asteroideneinschläge, Vulkanausbrüche, Pandemien oder auch einfach nur menschliche Unfälle und Krankheiten oder Kriege und Tyrannei. Vermutlich wird sie uns auch vor neuen technologischen Risiken bewahren, die durch Nanotechnologien, Bioscience, neuropsychologischen Manipulationen und Body-Enhancement auf uns zukommen.

Auf der anderen Seite stellt sie selbst ein existenzielles Risiko bisher unbekannten Ausmaßes für die Menschheit dar. Soviel jedenfalls ist klar.

Fazit: Was ist zu tun?

Ich habe in dieser Trendanalyse schamlos oft aus Nick Bostroms Buch „Superintelligenz" zitiert. Es scheint mir die derzeit mit Abstand fundierteste Analyse der möglichen Zukunftsentwicklungen bei der Entstehung von Superintelligenzen zu sein. Doch so fundiert seine Analyse, so kurz gegriffen ist sein Fazit. Er fordert die folgende weltweite moralische Norm zur Entwicklung der Superintelligenzen:

„Eine Superintelligenz sollte nur zum Nutzen der gesamten Menschheit entwickelt werden und im Dienst weithin geteilter moralischer Ideale stehen."[20]

Doch reicht uns das? Ernsthaft? Hier treffen wir auf das eigentliche Problem. Es ist nicht die Superintelligenz. Sondern es sind die Menschen. Wird eine solche Norm auf UN-Ebene dafür sorgen, dass in den kommenden 30 Jahren alles zum Besten läuft?

Ich zweifle daran. Nick Bostrom selbst beschreibt den Grund dafür: *„Ein Kind, das eine Bombe findet, sollte diese vorsichtig zu Boden legen, sich schnell entfernen und dann umgehend einen Erwachsenen informieren. In diesem Fall haben wir es aber nicht mit einem, sondern mit vielen Kindern zu tun, die alle einen Zünder haben. Die Chance, dass alle so vernünftig sind, die Finger davon zu lassen, scheint gegen null zu gehen. Irgendein kleiner Idiot wird todsicher den Knopf drücken, nur um zu sehen, was passiert. Weglaufen bringt auch nichts (…) außerdem ist weit und breit kein Erwachsener in Sicht."*[21]

Das Gemeinwohlprinzip

[20] Bostrom, N., Superintelligenz, S. 356
[21] Bostrom, N., Superintelligenz, S. 364

Doch es gibt keinen Grund den Kopf in den Sand zu stecken. Wir haben noch einige Jahrzehnte Zeit. Und tatsächlich scheint es auch heute schon mehr Möglichkeiten der politischen und gesellschaftlichen Einflussnahme zu geben, als den einfachen Appell an das Gemeinwohl.

So begründet Bostrom selbst ausführlich, warum es risikoärmer sein könnte, wenn ein globales Wettrennen einer Vielzahl von AI-Projekten verhindert wird. Wenigstens die Basis-Forderung an eine verantwortliche Wirtschaft und Politik liegen hier auf der Hand: Sorgen wir dafür, dass möglichst wenige AI-Forschungsprojekte nebeneinander bestehen. Sorgen wir für Zusammenschlüsse des AI-Initiativen, das heißt: zumindest enge Kooperationen, besser noch eine gegenseitige Erfolgsbeteiligung durch Gesellschafteranteile.

Lassen wir die Forschungsarbeit nicht nur von börsennotierten Unternehmen machen, sondern beteiligen wir auch Staaten an diesem internationalen Projekt. Die staatliche Beteiligung könnte beispielsweise sichern, dass die Technologieentwickler keinen Alleingang anstreben, sondern der Veto-Mechanismus und das Moralsystem der künftigen Superintelligenz jederzeit parallel entwickelt werden. Beispiele für solche weltweiten Kooperationen hat es schon mehrere gegeben, sei es die internationale Raumstation ISS, der Kernforschungsreaktor CERN oder das „human genom project“ zur Entschlüsselung des menschlichen Genoms.

Und so komisch das klingen mag: Falls es doch mehrere Parallelprojekte geben sollte: Sorgen wir zudem dafür, dass es möglichst keinen Informationsaustausch zwischen den Projekten gibt, der ein risikoreiches oder gar unkontrolliertes Wettrennen immer wieder stimuliert. Da der Erschaffer der weltweit ersten Superintelligenz die größten strategischen Vorteile haben wird, müssen unbedingt vermeiden, dass verschiedene Projekte im Streben die erste Superintelligenz zu erschaffen, unkontrollierte Risiken inkauf nehmen

Die Menschheit am Gewinn beteiligen

Und noch ein Punkt scheint relevant: Wer auch immer an einer maschinellen Superintelligenz baut, der muss sich dessen bewusst sein, dass er nicht nur für sich selbst ein Risiko erschafft, sondern dass er jeden Menschen auf der Welt in Gefahr bringt, auch diejenigen, die ihr künftiges Leben und jenes ihrer Kinder lieber ohne eine Superintelligenz sehen würden. Da wir auf diese Weise alle das Risiko tragen, wäre es nur fair, wenn wir auch alle an den Vorteilen und möglichen Gewinnen partizipieren.

Ich hatte schon im **zweiten Teil dieser Trendanalyse**[22] beschrieben, dass eine erfolgreiche Superintelligenz die Produktivität in der Welt rasant in die Höhe schießen lassen würde. Wir können davon ausgehen, dass der Eigentümer der ersten Superintelligenz binnen kürzester Zeit mehr Gewinn macht, als alle anderen Unternehmen weltweit zusammen. Da wäre es nur folgerichtig, wenn er die Menschheit an einem Teil seiner Gewinne beteiligt. Möglicherweise liegen hier die Ressourcen für das aktuell in der AI-Szene viel diskutierte Global Universal Income, also das weltweite bedingungslose Grundeinkommen.

Anfangen!

Wir müssen wohl klein anfangen. Wie wäre es am Anfang mit der Frage: Wollen wir uns gedanklich vorbereiten auf eine Situation, die wir uns vielleicht nicht wünschen, die aber dennoch wahrscheinlich ist? Oder wollen wir jene Tage unvorbereitet auf uns zukommen lassen, in denen unsere Kinder wahrscheinlich vor der Situation stehen werden, eine

[22] vgl. Janszky, S.G., Nehmen uns Computer die Arbeit weg?, http://www.2bahead.com/analyse/trendanalyse/detail/trendanalyse-artificial-intelligence-teil-2-nehmen-uns-computer-die-arbeit-weg/

Entscheidung über das Fortbestehen oder das Ende der menschlichen Rasse zu treffen?

Der 2b AHEAD ThinkTank wollte eine solche Debatte bereits 2013 im Vorfeld der damaligen Bundestagswahl anstoßen. In einem offenen Brief, der **"Wolfsburger Erklärung des 2b AHEAD ThinkTanks"**, hatten wir damals die Spitzenkandidaten aller Parteien aufgefordert, sich jenseits ihrer Wahlprogramme mit den wirklichen Zukunftsfragen unserer Gesellschaft zu beschäftigen.

Zu den Unterzeichnern zählten neben zahlreichen Managern, Innovationschefs, Gründern und Technologieforschern auch Bundesverdienstkreuzträger und der frühere ARD-Moderator und bekannte Wissenschaftsjournalist Jean Pütz.

Wir hatten ihnen die wichtigsten Zukunftsfragen der Gesellschaft in einem Katalog aus den Bereichen Technologie & Ethik, Schule & Bildung, Gesundheit & Prävention, Demografie & Alter, Innovation & Entrepreneurship, Arbeitsmarkt & Unternehmen, Justiz & Recht sowie Governance & Aufgaben des Staates zusammen gestellt. Die Reaktion: Keine, von keinem Politiker, aus keiner Partei!

Lassen Sie uns nicht auf die Politiker warten. Wir können diese Fragen auch selbst in die Hand nehmen. Und falls sie einen Politiker kennen, egal welcher Partei, egal welcher Ebene, drücken Sie ihm die Wolfsburger Erklärung in die Hand! Fragen Sie ihn nicht nach einer Antwort! Er mag geneigt sein, eine zu geben. Aber er kann keine haben! Fordern Sie ihn lieber zum gemeinsamen Nachdenken auf!

Wir wünschen Ihnen eine große Zukunft!