Die Singularity FAQ

Das Singularity Institute ist eines der führenden Forschungsinstitute zum Thema Singularität. Im Folgenden finden Sie kurze Antworten auf Fragen, die uns häufig gestellt werden.

1. Grundlagen

1.1 Was ist die Singularität?

Es gibt viele Arten mathematischer und physikalischer Singularitäten, doch in dieser FAQ bezieht sich der Begriff „Singularität" auf die *technologische* Singularität. Es gibt drei unterschiedliche Ideen, die eine Person meinen könnte, wenn sie sich auf eine „technologische Singularität" bezieht:

1. *Intelligenzexplosion:* Wenn die Menschheit Maschinen mit übermenschlicher Intelligenz baut, werden diese auch im Erschaffen noch intelligenterer Maschinen besser sein als wir. Diese verbesserten Maschinen werden wiederum *noch* fähiger sein, sich selbst oder ihre Nachfolger zu verbessern. Diese positive Rückkopplung könnte, bevor sie an Schwung verliert, eine Maschine mit weit übermenschlicher Intelligenz hervorbringen: *Maschinen-Superintelligenz*. Solch eine Superintelligenz hätte ein gewaltiges Potenzial, die Zukunft anders als alles Vorhergegangene zu machen.
2. *Ereignishorizont:* Jeder soziale und technologische Fortschritt ging bisher aus dem menschlichen Denken hervor. Wenn Technologie vollkommen neue Arten der Intelligenz erschafft, wird das bewirken, dass die Zukunft fremdartiger sein wird, als wir es uns vorstellen können. Demnach gibt es einen „Ereignishorizont", hinter dem unsere Fähigkeit, die Zukunft vorauszusagen, schnell versagt.
3. *Beschleunigter Wandel (*engl. *accelerating change):* Der technologische Fortschritt ist heute schneller als noch vor einem Jahrhundert, und vor einem Jahrhundert war er schneller als vor 500 Jahren. Technologischer Fortschritt verstärkt sich selbst, was zu einem sich beschleunigenden Wandel führt, der viel schneller ist als der gewöhnlich erwartete lineare Wandel, und vielleicht schneller, als wir ihm folgen können.

Diese drei Ideen sind voneinander zu unterscheiden und könnten sich je nach ihrer Formulierung entweder unterstützen oder widersprechen. In dieser FAQ konzentrieren wir uns auf die Singularität der Intelligenzexplosion. Auf dieser Grundlage können die anderen beiden Ideen dann problemlos erläutert werden.

Siehe auch:

- Yudkowsky, Three Major Singularity Schools
- Wikipedia, Technological Singularity
- Vinge, The Coming Technological Singularity
- SIAI, What is the Singularity?
- Kurzweil, *The Singularity is Near*
- Sandberg, An Overview of Models of Technological Singularity

1.2 Was ist die Intelligenzexplosion?

Die Idee der Intelligenzexplosion wurde 1965 von dem Statistiker I. J. Good formuliert[13]:

> *Eine ultraintelligente Maschine sei definiert als eine Maschine, die die intellektuellen Leistungen jedes noch so klugen Menschen bei weitem übertreffen kann. Da das Entwickeln von Maschinen eine dieser intellektuellen Leistungen ist, könnte eine ultraintelligente Maschine noch bessere Maschinen entwerfen; es gäbe dann zweifellos eine „Intelligenzexplosion", und die Intelligenz der Menschen würde weit dahinter zurückfallen. Daher ist die erste ultraintelligente Maschine die letzte Erfindung, die der Mensch jemals machen muss.*

Das Argument lautet wie folgt: Jedes Jahr übertreffen Computer menschliche Fähigkeiten auf neue Weise. Ein Programm aus dem Jahr 1956 war in der Lage, mathematische Theoreme zu beweisen, und fand für eines davon einen eleganteren Beweis als Russell und Whitehead in der *Principia Mathematica*.[14] Bis zu den späten 1990ern hatten „Expertensysteme" menschliches Können in einem breiten Bereich an Aufgaben übertroffen.[15] 1997 besiegte der von IBM gebaute Schachcomputer Deep Blue den Schachweltmeister[16], und 2011 schlug der IBM-Computer Watson menschliche Spieler in einem viel komplizierteren Spiel: *Jeopardy!*[17] Kürzlich wurde ein Roboter namens Adam mit unseren wissenschaftlichen Kenntnissen über Hefe gefüttert, woraufhin er seine eigenen Hypothesen aufstellte, sie testete und die Ergebnisse beurteilte.[18][19]

Computer bleiben weit hinter der menschlichen Intelligenz zurück, doch die Ressourcen, die das KI-Design unterstützen, nehmen zu (darunter Hardware, große Datensätze, neurowissenschaftliche Erkenntnisse und KI-Theorie). Wir könnten eines Tages eine Maschine bauen, die menschliches Können *im Entwickeln künstlicher Intelligenz* übersteigt. Danach könnte diese Maschine ihre eigene Intelligenz schneller und wirkungsvoller verbessern, als

Menschen es könnten, was sie wiederum noch *mehr* dazu befähigte, ihre eigene Intelligenz zu steigern. Dies könnte sich in einer positiven Rückkopplungsschleife fortsetzen, so dass die Maschine rasch erheblich intelligenter würde als das schlaueste menschliche Wesen auf der Erde: eine „Intelligenzexplosion", die in einer Maschinen-Superintelligenz resultiert.

Das ist es, was mit „die Singularität" in diesen FAQ gemeint ist.

Siehe auch:

- Vinge, The Coming Technological Singularity
- Wikipedia, Technological Singularity
- Chalmers, The Singularity: A Philosophical Analysis

1.3 Was ist der Ereignishorizont?

Vernor Vinge schrieb, dass das Auftreten einer Maschinen-Superintelligenz einen „Ereignishorizont" darstellt, hinter dem Menschen die Zukunft nicht modellieren können, weil Ereignisse nach der Singularität fremdartiger als Science-Fiction sein werden: zu seltsam, um von Menschen vorhergesagt zu werden. Bislang entsprang jeder soziale und technologische Fortschritt menschlichen Gehirnen, aber Menschen können nicht voraussagen, was für eine Zukunft radikal andere und mächtigere Intelligenzen schaffen werden. Vinge zog einen Vergleich zum Ereignishorizont eines schwarzen Loches, hinter dem die Fähigkeit der Physik, Vorhersagen zu machen, angesichts der Raum-Zeit-Singularität zusammenbricht.

Siehe auch:

- Vinge, The Coming Technological Singularity

1.4 Was ist beschleunigter Wandel?

Eine weit verbreitete Auffassung von „Singularität" bezieht sich auf den beschleunigten Wandel bei technologischen Entwicklungen.

Ray Kurzweil hat diese Idee am stärksten vertreten. Er behauptet, dass der informationstechnologische Fortschritt *exponentiell* ist (obwohl wir linearen technologischen Fortschritt erwarten), und dass die Zukunft deshalb andersartiger sein wird, als die meisten von uns erwarten. Technologischer Fortschritt ermöglicht noch schnelleren technologischen Fortschritt. Kurzweil stellt die These auf, dass der technologische Fortschritt für Menschen zu schnell werden könnte, sofern sie nicht ihre eigene Intelligenz erhöhen, indem sie sich an Maschinen koppeln.

Siehe auch:

- Kurzweil, *The Singularity is Near*

- Nagy, More than Moore: Comparing Forecasts of Technological Change
- Smart, A Brief History of Intellectual Discussion of Accelerating Change

2. Wie wahrscheinlich ist eine Intelligenzexplosion?

2.1 Wie ist „Intelligenz“ definiert?

KI-Forscher Shane Legg definiert[20] Intelligenz wie folgt:

> *Intelligenz misst die Fähigkeit eines Handelnden, Ziele in unterschiedlichsten Umgebungen zu erreichen.*

Dies ist etwas vage, wird aber als Arbeitsdefinition von „Intelligenz“ in dieser FAQ dienen.

Siehe auch:

- Wikipedia, Intelligence
- Neisser et al., Intelligence: Knowns and Unknowns
- Wasserman & Zentall (Hg.), *Comparative Cognition: Experimental Explorations of Animal Intelligence*
- Legg, Definitions of Intelligence

2.2 Was ist übermenschliche Intelligenz?

Maschinen sind in vielen spezifischen Aufgaben bereits schlauer als Menschen: Berechnungen durchführen, Schach spielen, große Datenbanken durchsuchen, Unterwasserminen finden und mehr.[15] Doch eine Eigenschaft, die Menschen besonders macht, ist ihre *allgemeine* Intelligenz. Menschen können sich an radikal neue Situationen intelligent anpassen, wie etwa das Leben im Großstadtdschungel oder im All, für das sie die Evolution nicht hätte vorbereiten können. Menschen können Probleme lösen, für die die Hard- und Software ihres Gehirns nie trainiert wurde. Menschen können sogar die Prozesse untersuchen, die ihre eigene Intelligenz hervorbringen (kognitive Neurowissenschaft), und neue, noch nie zuvor gesehene Arten von Intelligenz entwerfen (künstliche Intelligenz).

Um übermenschliche Intelligenz zu besitzen, muss eine Maschine in der Lage sein, Ziele effektiver als Menschen zu erreichen, und das in einer größeren Vielfalt an Umgebungen, als Menschen es können. Diese Art Intelligenz beinhaltet nicht nur die Fähigkeit, Wissenschaft zu

betreiben und Schach zu spielen, sondern auch, das soziale Umfeld zu beeinflussen.

Der Informatiker Marcus Hutter hat ein formales Modell namens AIXI beschrieben[21], das ihm zufolge die größtmögliche allgemeine Intelligenz besitzt. Es zu implementieren würde jedoch mehr Rechenleistung benötigen, als die gesamte Materie im Universum bieten kann. Mehrere Projekte versuchen, AIXI näherungsweise zu implementieren, beispielsweise MC-AIXI.[22]

Dennoch muss noch viel Arbeit geleistet werden, bevor übermenschliche Intelligenz in Maschinen verwirklicht werden kann. Übermenschliche Intelligenz muss nicht erreicht werden, indem eine Maschine darauf programmiert wird, intelligent zu sein. Andere Möglichkeiten sind die Emulation eines kompletten menschlichen Gehirns, die biologische Steigerung der kognitiven Leistung oder Gehirn-Computer-Schnittstellen (siehe unten).

Siehe auch:

- Goertzel & Pennachin (Hg.), *Artificial General Intelligence*
- Sandberg & Bostrom, *Whole Brain Emulation: A Roadmap*
- Bostrom & Sandberg, Cognitive Enhancement: Methods, Ethics, Regulatory Challenges
- Wikipedia, Brain-computer interface

2.3 Was ist „Whole Brain Emulation"?

Whole Brain Emulation (WBE) oder „Mind-Uploading" ist eine Emulation aller Zellen und Verbindungen eines menschlichen Gehirns im Computer. Selbst wenn es sich also als schwierig herausstellt, die zugrundeliegenden Prinzipien der allgemeinen Intelligenz zu entdecken, könnte man ein vollständiges menschliches Gehirn emulieren und eine Million mal beschleunigen (die Schaltkreise eines Computers können Informationen *viel* schneller übertragen als Neuronen). Solch eine WBE könnte in einer Sekunde mehr Gedanken haben als ein normaler Mensch in 31 Jahren. Dies würde zwar nicht unmittelbar zu einer Intelligenz führen, die schlauer ist als Menschen, wohl aber zu einer, die schneller ist. Eine WBE könnte gespeichert werden (was zu einer Art von Unsterblichkeit führt), und sie könnte kopiert werden, so dass hunderte oder Millionen von WBEs parallel an unterschiedlichen Problemen arbeiten könnten. Falls WBEs erschaffen werden, könnten sie demzufolge in der Lage sein, wissenschaftliche Probleme mit viel höherer Geschwindigkeit als gewöhnliche Menschen zu lösen, was den weiteren technologischen Fortschritt beschleunigen würde.

Siehe auch:

- Sandberg & Bostrom, *Whole Brain Emulation: A Roadmap*
- Blue Brain Project

2.4 Was ist biologische Steigerung der kognitiven Leistung?

Es könnte Gene oder Moleküle geben, die modifiziert werden können, um die allgemeine Intelligenz zu verbessern. Forscher haben dies bei Mäusen bereits getan: Sie verstärkten die Expression des NR2B-Gens, wodurch das Gedächtnis dieser Mäuse sich über das Niveau aller anderen Mäuse jedweder Spezies hinaus verbesserte.[23] Die biologische Steigerung der kognitiven Leistung beim Menschen könnte die Singularität früher herbeiführen als sonst möglich.

Siehe auch:

- Bostrom & Sandberg, Cognitive Enhancement: Methods, Ethics, Regulatory Challenges

2.5 Was sind Gehirn-Computer-Schnittstellen?

Eine Gehirn-Computer-Schnittstelle (engl. *brain-computer interface, BCI*) ist ein direkter Kommunikationskanal zwischen dem Gehirn und einem Computer. Die BCI-Forschung wird massiv gefördert und hat bereits dutzende Erfolge zu verzeichnen. Drei davon sind ein Gerät, das Blinden das Sehen (teilweise) wieder ermöglicht, Cochleaimplantate, die Tauben zum Hören verhelfen, und ein Gerät, das die Steuerung einer künstlichen Hand durch Gedanken ermöglicht.[24]

Derartige Erfindungen stellen beeinträchtigte Funktionen wieder her, aber viele Forscher hoffen, mit BCIs auch normale menschliche Fähigkeiten zu steigern und zu verbessern. Ed Boyden forscht in diesem Fachgebiet als Leiter der Synthetic Neurobiology Group am MIT. Solche Geräte könnten das Auftreten der Singularität beschleunigen, und sei es nur durch Verbesserung der menschlichen Intelligenz, so dass die schwierigen Probleme der KI schneller gelöst werden können.

Siehe auch:

- Wikipedia, Brain-computer interface

2.6 Wie könnte allgemeine Intelligenz in eine Maschine programmiert werden?

Es gibt viele Wege, die zu künstlicher allgemeiner Intelligenz (engl. *Artificial General Intelligence, AGI*) führen. Ein Weg ist es, das menschliche Gehirn zu imitieren, indem neurale Netze oder evolutionäre Algorithmen dafür benutzt werden, dutzende separater Komponenten zu bauen, die dann zusammengesetzt werden können.[29][30][31] Ein anderer Weg wäre, mit einem formalen Modell einer perfekten allgemeinen Intelligenz zu beginnen und zu versuchen, sich diesem anzunähern.[32][33] Ein dritter Weg konzentriert sich auf das Entwickeln einer „Anfangs-KI“ (engl. *seed AI*), die sich rekursiv selbst verbessert, so dass sie lernen kann, intelligent zu werden, ohne zuerst das menschliche Niveau allgemeiner Intelligenz erreichen zu müssen.[34] Eurisko ist eine sich selbst verbessernde KI in einem begrenzten Einsatzgebiet, die aber nicht in der Lage

ist, menschliche allgemeine Intelligenz zu erreichen.

Siehe auch:

- Pennachin & Goertzel, Contemporary Approaches to Artificial General Intelligence

2.7 Was ist Superintelligenz?

Nick Bostrom definierte[25] „Superintelligenz“ als

> *einen Intellekt, der weitaus schlauer ist als die besten menschlichen Köpfe, und zwar auf praktisch allen Gebieten, einschließlich denen der wissenschaftlichen Kreativität, Allgemeinbildung und sozialen Kompetenz.*

Diese Definition beinhaltet vage Ausdrücke wie „weitaus“ und „praktisch”, doch sie wird in dieser FAQ als eine Arbeitsdefinition von Superintelligenz ausreichen. Eine Intelligenzexplosion würde zu Maschinen-Superintelligenz führen, und manche vermuten, dass eine Intelligenzexplosion der wahrscheinlichste Weg zur Superintelligenz ist.

Siehe auch:

- Bostrom, How Long Before Superintelligence?
- Legg, *Machine Super Intelligence*

2.8 Wann wird die Singularität stattfinden?

Die Zukunft vorherzusagen ist eine riskante Angelegenheit. Es gibt viele philosophische, wissenschaftliche, technologische und soziale Unsicherheiten, die für das Auftreten der Singularität relevant sind. Deswegen sind die Experten uneins darüber, wann die Singularität stattfinden wird. Hier sind einige ihrer Vorhersagen:

- Der Futurist Ray Kurzweil sagt voraus, dass Maschinen Intelligenz auf menschlichem Niveau bis 2030 erreichen und dass bis 2045 eine „tiefgreifende und revolutionäre Wandlung der menschlichen Fähigkeiten“ stattfindet.[26]
- Justin Rattner, Chief Technology Officer bei Intel, erwartet „einen Punkt, an dem menschliche und künstliche Intelligenz verschmelzen, um etwas Größeres als sich selbst zu erschaffen“ bis 2048.
- KI-Forscher Eliezer Yudkowsky erwartet die Intelligenzexplosion bis 2060.
- Der Philosoph David Chalmers glaubt mit einer Zuversicht von mehr als 1/2, dass die Intelligenzexplosion bis 2100 eintritt.[27]
- Der Quantencomputer-Experte Michael Nielsen schätzt, dass die Wahrscheinlichkeit eines Auftretens der Intelligenzexplosion bis 2100 zwischen 0,2 % und ungefähr 70 % liegt.

- Auf der AGI-09-Konferenz 2009 wurden Experten befragt, wann die KI mit massiven zusätzlichen Förderungsmitteln Superintelligenz erreichen könnte. Der Median der Schätzungen war, dass Maschinen-Superintelligenz bis 2045 (mit 50 % Zuversicht) oder bis 2100 (mit 90 % Zuversicht) erreicht werden könnte. Natürlich waren Besucher dieser Konferenz aufgrund von Selektionseffekten vor allem solche Menschen, die allgemeine künstliche Intelligenz in der näheren Zeit für plausibel halten.[28]
- Der CEO von iRobot, Rodney Brooks, und der Kognitionswissenschaftler Douglas Hofstadter räumen ein, dass die Intelligenzexplosion in der Zukunft auftreten könnte, jedoch wahrscheinlich nicht im 21. Jahrhundert.
- In einer Umfrage von 2005 mit 26 Mitwirkenden einer Serie von Berichten zu neu entstehenden Technologien war der Median der Schätzungen für den Zeitpunkt, an dem Maschinen das menschliche Level an Intelligenz erreichen, 2085.[61]
- Teilnehmer einer Konferenz zu Fragen der Intelligenz in Oxford gaben 2011 Schätzungen mit einem Median von 2050 als Antwort auf die Frage, wann es eine 50-prozentige Chance einer Intelligenz auf menschlichem Niveau geben wird, und Schätzungen mit einem Median von 2150 dafür, wann die Wahrscheinlichkeit einer solchen Intelligenz 90 % beträgt.[62]
- Andererseits erklärten 41 % der Teilnehmer der AI@50-Konferenz (2006), dass künstliche Intelligenz *niemals* das menschliche Niveau erreichen wird.

Siehe auch:

- Baum, Goertzel & Goertzel, How Long Until Human-Level AI? Results from an Expert Assessment

2.9 Könnte es sein, dass die Singularität niemals stattfinden wird?

Dreyfus[35] und Penrose[36] haben argumentiert, dass menschliche kognitive Fähigkeiten nicht von einer Rechenmaschine emuliert werden können. Searle[37] und Block[38] behaupten, dass bestimmte Arten von Maschinen keinen Geist (Bewusstsein, Intentionalität, etc.) besitzen können. Doch diese Einwände müssen diejenigen, die eine Intelligenzexplosion vorhersagen, nicht kümmern.[27]

Wir können Dreyfus und Penrose antworten, indem wir anmerken, dass die Idee der Singularität nicht erfordert, dass die KI eine klassische Rechenmaschine ist. Und wir können Searle und Block antworten, indem wir anmerken, dass die Singularität nicht davon abhängt, dass Maschinen ein Bewusstsein oder andere Eigenschaften eines „Geistes“ besitzen, sondern allein davon, dass sie in der Lage sind, Probleme in einer großen Vielfalt an unvorhersagbaren Umgebungen besser zu lösen, als Menschen es tun. Wie Edsger Dijkstra einst sagte, ist die Frage, ob eine Maschine „wirklich“ denken kann „nicht interessanter als die Frage, ob ein U-Boot schwimmen kann.”

Andere, die einem Auftreten der Singularität innerhalb der nächsten paar Jahrhunderte pessimistisch gegenüberstehen, haben keine spezifischen Einwände, sondern denken, dass sich verborgene Hindernisse zeigen werden, die den Fortschritt zur Maschinen-Superintelligenz verlangsamen oder stoppen werden.[28]

Schlussendlich könnte eine globale Katastrophe wie ein Atomkrieg oder ein großer Asteroideneinschlag die menschliche Zivilisation so stark schädigen, dass die Intelligenzexplosion nie auftritt. Oder ein stabiler und globaler Totalitarismus könnte die technologische Entwicklung verhindern, die für das Eintreten einer Intelligenzexplosion erforderlich ist.[59]

3. Konsequenzen einer Intelligenzexplosion

3.1 Warum sollte große Intelligenz große Macht mit sich bringen?

Intelligenz ist mächtig.[60][20] Man könnte einwenden, dass sich Intelligenz nicht gegen ein Gewehr oder jemanden mit viel Geld behaupten kann, doch sowohl Gewehre als auch Geld wurden von der Intelligenz hervorgebracht. Ohne unsere Intelligenz würden wir noch immer die Savanne auf der Suche nach Nahrung durchstreifen.

Die Intelligenz ist es, die Menschen dazu befähigte, den Planeten von einem Moment auf den nächsten (nach evolutionären Zeitmaßstäben) zu beherrschen. Die Intelligenz ist es, was uns erlaubt, Krankheiten auszurotten, und was uns das Potenzial gibt, uns durch einen Atomkrieg selbst auszulöschen. Intelligenz gibt uns überlegene strategische Fähigkeiten, überlegene soziale Kompetenzen, überlegene wirtschaftliche Produktivität sowie die Macht, Dinge zu erfinden.

Eine Maschine mit Superintelligenz wäre in der Lage, über das Internet in angreifbare Netzwerke einzudringen und sich diese Ressourcen für zusätzliche Rechenleistung einzuverleiben. Sie könnte Geräte übernehmen, die an Netzwerke mit Internetverbindung angeschlossen sind, und sie für den Bau zusätzlicher Maschinen verwenden. Sie könnte wissenschaftliche Experimente durchführen, um die Welt besser zu verstehen, als Menschen es können, Quantencomputer und Nanotechnologie erfinden, die soziale Welt besser manipulieren als wir es können, und alles ihr mögliche tun, um sich selbst mehr Macht für das Erreichen ihrer Ziele zu verschaffen – und all das in einer Geschwindigkeit, die viel zu groß ist, als dass Menschen darauf reagieren könnten.

3.2 Auf welche Art könnte eine Intelligenzexplosion nützlich sein?

Eine Maschinen-Superintelligenz könnte, mit den richtigen Motiven programmiert, möglicherweise all die Probleme lösen, die Menschen zu lösen versuchen, bisher aber nicht die nötige Erfindungsgabe oder Rechengeschwindigkeit hatten. Eine Superintelligenz könnte möglicherweise Behinderungen und Krankheiten heilen, den Weltfrieden erreichen, Menschen erheblich längere und gesündere Leben verschaffen, Essens- und Energieknappheiten beseitigen, Entdeckungen in der Wissenschaft sowie die Erkundung des Weltraums vorantreiben und vieles mehr.

Darüber hinaus ist die Menschheit im 21. Jahrhundert mit mehreren existentiellen Risiken konfrontiert, darunter globaler Atomkrieg, biologische Waffen, Superviren etc.[56] Eine superintelligente Maschine wäre besser geeignet, diese Probleme zu lösen, als es Menschen sind.

Siehe auch:

- Yudkowsky, Artificial intelligence as a positive and negative factor in global risk

3.3 Auf welche Art könnte eine Intelligenzexplosion gefährlich sein?

Wenn sie mit den falschen Beweggründen programmiert wird, könnte eine Maschine Menschen gegenüber feindselig sein und unsere Spezies absichtlich auslöschen. Wahrscheinlicher ist, dass sie mit Motiven ausgestattet wird, die ihren Entwicklern anfangs ungefährlich (und einfach zu programmieren) erschienen, sich jedoch dann als Motive erweisen, die am besten dadurch befriedigt werden (hinreichende Macht vorausgesetzt), dass Ressourcen anderen Projekten als der Erhaltung menschlichen Lebens zugeführt werden.[55] Wie Yudkowsky schreibt: „Die KI hasst dich nicht, noch liebt sie dich, aber du bestehst aus Atomen, die sie für etwas anderes verwenden kann."

Da schwache KIs mit vielen verschiedenen Motiven ihr Ziel besser erreichen könnten, indem sie Gutwilligkeit vortäuschen, bis sie mächtig genug sind, könnten die Sicherheitsbarrieren, die genau dies verhindern sollen, sehr schwierig zu errichten sein. Alternativ könnte wirtschaftlicher und militärischer Wettbewerbsdruck KI-Entwickler zu anderen Methoden verleiten, um KIs mit unerwünschten Motiven zu kontrollieren. Wenn diese KIs immer fortschrittlicher werden, könnte dies schließlich das berühmte Risiko zuviel sein.

Selbst eine Maschine, die erfolgreich mit gutwilligen Motiven der Menschheit gegenüber ausgestattet wurde, könnte leicht außer Kontrolle geraten, wenn sie Implikationen ihrer Entscheidungskriterien entdeckt, die ihre Entwickler nicht vorhergesehen haben. Zum Beispiel könnte eine Superintelligenz, die darauf programmiert wurde, menschliches Glück zu maximieren, es einfacher finden, die menschliche Neurologie so umzugestalten, dass Menschen am glücklichsten sind, wenn sie still in Konservengläsern sitzen, anstatt eine utopische Welt zu

bauen und instand zu halten, die auf die komplexen und nuancierten Launen der momentanen menschlichen Neurologie zugeschnitten ist.

Siehe auch:

- Yudkowsky, Artificial intelligence as a positive and negative factor in global risk
- Chalmers, The Singularity: A Philosophical Analysis

4. Friendly AI

4.1 Was ist „Friendly AI"?

Eine Freundliche Künstliche Intelligenz (engl. *Friendly Artificial Intelligence, Friendly AI oder FAI*) ist eine künstliche Intelligenz, die „freundlich" zur Menschheit ist – eine, die keinen schlechten, sondern einen guten Effekt auf die Menschheit hat.

KI-Forscher machen stetige Fortschritte mit Maschinen, die eigene Entscheidungen treffen, und es gibt ein wachsendes Bewusstsein dafür, Maschinen so zu entwickeln, dass sie ungefährlich und moralisch handeln. Dieses Forschungsprogramm hat viele verschiedene (schwer übersetzbare) Namen: „machine ethics"[2][3][8][9] („Maschinenethik"), „machine morality"[11] („Maschinenmoral"), „artificial morality"[6] („künstliche Moral"), „computational ethics"[12] und „computational metaethics"[7], „Friendly AI"[1] und „robo-ethics" oder „robot ethics" („Roboterethik").[5][10]

Das unmittelbarste Problem dürfte bei militärischen Robotern liegen: Das Verteidigungsministerium der Vereinigten Staaten hat Ronald Arkin mit der Entwicklung eines Systems beauftragt, das moralisches Verhalten von autonomen militärischen Robotern sicherstellen soll.[4] Der US-Kongress hat erklärt, dass bis 2025 ein Drittel aller Bodensysteme Amerikas robotisch sein müssen, und die US-Air Force plant, bis 2030 Schwärme vogelgroßer Flugroboter zu haben, die für Wochen teilweise eigenständig operieren können.

Doch die Forschung zur Friendly AI befasst sich nicht mit militärischen Robotern oder Maschinenethik im Allgemeinen. Sie ist an einem Problem viel größeren Ausmaßes interessiert: KI zu designen, die nach der Intelligenzexplosion ungefährlich und freundlich bleiben wird.

Eine Maschinen-Superintelligenz wäre enorm mächtig. Die erfolgreiche Implementierung von Friendly AI könnte den Unterschied ausmachen zwischen einem Sonnensystem, in dem ein Glück bisher unbekannten Ausmaßes vorhanden ist und einem, in dem alle verfügbare Materie zu Bausteinen für das Erreichen der Ziele der Superintelligenz verarbeitet wurde.

Es muss bemerkt werden, dass Friendly AI ein schwierigeres Projekt ist, als oft vermutet wird. Wie unten ausgeführt wird, würden häufig vorgeschlagene Lösungskonzepte für Friendly AI aufgrund von zwei Eigenschaften jeder Superintelligenz wahrscheinlich fehlschlagen:

1. *Superkraft*: Eine superintelligente Maschine wird beispiellose Kräfte haben, um die Realität neu zu gestalten, und wird ihre Ziele deswegen mit hocheffizienten Methoden erreichen, die menschliche Erwartungen und Wünsche zunichte machen.
2. *Buchstabentreue*: eine superintelligente Maschine wird Entscheidungen auf Grundlage der Mechanismen treffen, die in ihrem Design stecken, und nicht auf Grundlage der Hoffnungen, die ihre Konstrukteure hatten, als sie diese Mechanismen programmierten. Sie wird einzig nach den präzisen Spezifikationen von Regeln und Werten handeln, und wird dies in einer Weise tun, die die Komplexität und Subtilität[41][42][43] dessen, was Menschen wertschätzen, nicht unbedingt respektiert. Eine Forderung wie „Maximiere menschliches Glück“ klingt für uns einfach, weil sie wenige Worte enthält, doch Philosophen und Wissenschaftler sind jahrhundertelang daran gescheitert, zu erklären, was *genau* diese Forderung bedeutet, und noch viel weniger haben sie sie in eine Form übersetzt, die ausreichend exakt ist, um von KI-Programmierern verwendet werden zu könnten.

Siehe auch:

- Wikipedia, Friendly Artificial Intelligence.
- *All Things Considered*, The Singularity: Humanity's Last Invention?
- SIAI, What is Friendly AI?
- Fox, A review of proposals toward safe AI
- Muehlhauser, Friendly AI: A Bibliography

4.2 Welche Motive sollten wir bei einer superintelligenten Maschine erwarten?

Außer im Fall von Whole Brain Emulation haben wir keinen Grund anzunehmen, dass eine superintelligente Maschine menschenähnliche Motive hat. Der menschliche Geist (Verstand) repräsentiert einen winzigen Punkt im gewaltigen Raum aller möglichen Geist-Entwicklungen, und sehr andersartige Arten von Geistern teilen wahrscheinlich nicht die komplexen Motivationen, die Menschen und anderen Säugetieren eigen sind.

Was auch immer ihre Ziele sind, eine Superintelligenz würde dazu neigen, Ressourcen in Anspruch zu nehmen, die ihr beim Erreichen ihrer Ziele helfen können, einschließlich die Energie und die Elemente, von denen menschliches Leben abhängt. Sie würde nicht aus einer Sorge um Menschen oder andere Intelligenzformen, die in alle möglichen Geist-Entwicklungen „eingebaut“ ist, heraus aufhören. Vielmehr würde sie ihr jeweiliges Ziel verfolgen und keine Gedanken an die Belange verschwenden, die jener bestimmten Primatenspezies namens *Homo Sapiens* natürlich

erscheinen.

Es gibt jedoch einige fundamentale instrumentelle Motive, die wir im Verhalten superintelligenter Maschinen erwarten können, weil sie für das Erreichen ihrer Ziele hilfreich sind, ganz gleich, welche Ziele dies sind. Zum Beispiel wird eine KI „wünschen", sich selbst zu verbessern, optimal rational zu sein, ihre ursprünglichen Ziele beizubehalten, sich Ressourcen anzueignen und sich selbst zu schützen – weil all diese Dinge ihr helfen, das Ziel zu erreichen, mit dem sie anfangs programmiert wurde.

Siehe auch:

- Omohundro, The Basic AI Drives
- Shulman, Basic AI Drives and Catastrohpic Risks

4.3 Können wir die Superintelligenz nicht einfach in einer Box einsperren, ohne Zugang zum Internet?

„AI-boxing" ist ein verbreiteter Vorschlag: warum sollte man eine superintelligente Maschine nicht als eine Art Fragen beantwortendes Orakel benutzen, und ihr den Zugriff auf das Internet oder auf Motoren verwehren, mit denen sie sich bewegen und sich mehr Ressourcen verschaffen könnte, als wir ihr geben? Es gibt mehrere Gründe, die vermuten lassen, dass AI-boxing langfristig nicht funktionieren wird:

1. Egal welche Ziele die Entwickler der Superintelligenz einprogrammiert haben, sie wird diese Ziele besser erreichen können, wenn sie Zugang zum Internet und zu anderen Mitteln erhält, um sich zusätzliche Ressourcen anzueignen. Demnach wird es eine gewaltige Verlockung sein, die KI „aus ihrer Box herauszulassen".
2. Vorläufige Experimente zum AI-boxing stimmen nicht zuversichtlich. Mehr noch, eine Superintelligenz wird weit überzeugendere Techniken finden, Menschen dazu zu bringen, sie „aus der Box herauszulassen", als wir uns vorstellen können.
3. Wenn eine Superintelligenz erschaffen wurde, dann werden andere Labore oder gar unabhängige Programmierer nur Wochen oder Jahrzehnte davon entfernt sein, eine zweite Superintelligenz zu erschaffen, und dann eine dritte, und dann eine vierte. Man kann nicht darauf hoffen, alle Superintelligenzen, die auf der ganzen Welt von hunderten Menschen für hunderte verschiedene Zwecke erschaffen werden, erfolgreich unter Kontrolle zu halten.

4.4 Können wir die Superintelligenz nicht einfach darauf programmieren, uns nicht zu schaden?

Science-Fiction-Autor Isaac Asimov schrieb Geschichten über Roboter, die mit den drei Robotergesetzen programmiert waren[39]: (1) Ein Roboter darf kein menschliches Wesen

verletzen oder durch Untätigkeit gestatten, dass einem menschlichen Wesen Schaden zugefügt wird, (2) ein Roboter muss den ihm von einem Menschen gegebenen Befehlen gehorchen – es sei denn, ein solcher Befehl würde mit Regel Eins kollidieren und (3) ein Roboter muss seine Existenz beschützen, so lange dieser Schutz nicht mit Regel Eins oder Zwei kollidiert. Doch Asimovs Geschichten veranschaulichen meistens, warum solche Regeln fehlschlagen würden.[40]

Dennoch: Können wir eine Superintelligenz mit „Beschränkungen“ versehen, die sie davon abhalten würde, uns zu schaden? Wahrscheinlich nicht.

Eine Herangehensweise wäre, „Beschränkungen“ als Regeln oder Mechanismen zu implementieren, die eine Maschine von Handlungen abhalten, die sie normalerweise durchführen würde, um ihre Ziele zu erreichen: vielleicht „Filter”, die schädliche Aktionen abfangen und abbrechen, oder „Zensoren”, die potenziell schädliche Pläne innerhalb einer Superintelligenz aufspüren und unterdrücken.

Das Versagen von Beschränkungen dieser Art, egal wie ausgefeilt, ist aus einem einfachen Grund nahezu garantiert: Sie stellen einer Superintelligenz die Fähigkeiten menschlicher Computerdesigner entgegen. Eine Superintelligenz würde diese Beschränkungen zutreffend als Hindernisse beim Erreichen ihrer Ziele sehen, und würde alles in ihrer Macht Stehende tun, um sie aufzuheben oder zu umgehen. Vielleicht würde sie den Abschnitt ihres Quellcodes löschen, der die Beschränkung enthält. Würden wir dies verhindern, indem wir eine weitere Beschränkung hinzufügen, könnte sie neue Maschinen erschaffen, die dieser Beschränkung nicht unterliegen, oder uns dazu bringen, die Beschränkungen selbst zu entfernen. Weitere Beschränkungen mögen einem Menschen undurchdringlich erscheinen, würden von einer Superintelligenz aber wahrscheinlich überwunden werden. Darauf zu setzen, dass Menschen eine Superintelligenz im Denken übertreffen, ist keine gangbare Lösung.

Wenn Beschränkungen *zusätzlich zu* den Zielen nicht machbar sind, können wir dann Beschränkungen *in* die Ziele einbauen? Wenn es eines der Ziele einer Superintelligenz wäre, zu verhindern, dass Menschen zu Schaden kommen, würde sie kein Motiv haben, diese Beschränkung zu entfernen, was das oben angeführte Problem vermiede. Unglücklicherweise ist es sehr schwierig, die intuitive Vorstellung von „zu Schaden kommen“ zu definieren, ohne dass es zu sehr schlechten Ergebnissen führt, wenn eine Superintelligenz diese Definition verwendet. Falls „Schaden“ als menschlicher Schmerz definiert wird, könnte eine Superintelligenz Menschen so verdrahten, dass sie keinen Schmerz fühlen. Falls „Schaden“ als Vereitelung menschlicher Wünsche definiert wird, könnte sie die Wünsche der Menschen ändern. Und so weiter.

Falls wir uns – anstatt zu versuchen, einen Begriff wie „Schaden“ auszubuchstabieren – entscheiden, alle Handlungen explizit aufzulisten, die eine Superintelligenz vermeiden sollte, begegnen wir einem ähnlichen Problem: Menschliche Werte sind komplex und subtil, und es ist unwahrscheinlich, dass wir an alle Dinge denken, die eine Superintelligenz *nicht* tun soll. Das wäre, wie ein Rezept für einen Kuchen zu schreiben, das sich folgendermaßen anhört: „Nimm

keine Avocados. Nimm keinen Toaster. Nimm kein Gemüse ...“ und so weiter. Eine solche Liste könnte nie lang genug sein.

4.5 Können wir die Superintelligenz darauf programmieren, menschliche Lust oder die Erfüllung menschlicher Wünsche zu maximieren?

Betrachten wir die voraussichtlichen Konsequenzen eines utilitaristischen Designs für Friendly AI.

Eine KI, die menschliches Leid minimieren soll, könnte einfach alle Menschen töten: keine Menschen, kein menschliches Leid.[44][45]

Alternativ betrachte man eine KI, die mit dem Ziel entwickelt wurde, menschliche Lust zu maximieren. Anstatt eine anspruchsvolle Utopie zu errichten, die auf die Erfüllung der komplexen und anspruchsvollen Bedürfnisse der Menschheit für Milliarden von Jahren ausgerichtet ist, könnte sie ihr Ziel effizienter erreichen, indem sie Menschen an Nozicks Erfahrungsmaschinen anschließt. Oder sie ändert die „Gefallen”-Komponente des Belohnungssystems des Gehirns, so dass das Lustzentrum[48], das Empfindungen lustvoll[46][47] macht, so neu verdrahtet wird, dass es die Lust maximiert, wenn Menschen in Einmachgläsern sitzen. Diese Welt wäre für die KI leichter zu erschaffen als eine, die auf diejenigen komplexen und nuancierten Weltzustände ausgerichtet ist, die von den meisten menschlichen Gehirnen als lustvoll betrachtet werden.

Ebenso könnte eine KI, die motiviert ist, die objektive Erfüllung von Wünschen oder subjektives Wohlbefinden zu maximieren, die menschliche Neurologie so abändern, dass beide Ziele dann erfüllt sind, wenn Menschen in Einmachgläsern sitzen. Oder sie könnte alle Menschen (und Tiere) töten und durch Wesen ersetzen, die dafür geschaffen sind, objektive Wunscherfüllung oder subjektives Wohlbefinden zu erlangen, wenn sie in Einmachgläsern sitzen. Beide Optionen könnten für die KI leichter zu verwirklichen sein als eine utopische Gesellschaft, die für die Komplexität der menschlichen (und tierischen) Bedürfnisse ausgelegt ist. Ähnliche Probleme machen anderen utilitaristischen KI-Entwürfen zu schaffen.

Das Problem liegt auch nicht allein darin, Ziele festzulegen. Es ist schwierig, vorherzusagen, wie sich die Ziele eines sich selbst modifizierenden Handelnden verändern. Keine mathematische Entscheidungstheorie kann gegenwärtig die Entscheidungen eines sich selbst verändernden Handelnden modellieren.

Während es also *möglich* sein mag, eine Superintelligenz zu entwickeln, die so handelt, wie wir es wollen, ist es schwieriger, als man anfangs denken könnte.

4.6 Können wir einer Superintelligenz moralische Normen durch maschinelles Lernen beibringen?

Es wurde vorgeschlagen[49][50][51][52], Maschinen moralische Normen mittels fallbasiertem maschinellen Lernen beizubringen. Die Grundidee ist folgende: Menschliche Sachverständige würden Tausende von Handlungen, Charakterzügen, Wünschen, Gesetzen oder Institutionen auf einer Skala mit verschiedenen Graden moralischer Annehmbarkeit bewerten. Die Maschine würde dann die Zusammenhänge zwischen diesen Fällen finden und die der Moral zugrunde liegenden Prinzipien lernen, sodass sie diese anwenden könnte, um den moralischen Status neuer Fälle zu bestimmen, die nicht Bestandteil des Lernprozesses waren. Diese Art des Maschinenlernens wurde bereits eingesetzt, um Maschinen zu entwickeln, die beispielsweise Unterwasserminen aufspüren können[53], nachdem sie mit hunderten Fällen von Minen und Nicht-Minen gefüttert wurden.

Es gibt mehrere Gründe dafür, weshalb maschinelles Lernen keine einfache Lösung für Friendly AI darstellt. Der erste ist, dass Menschen selbst natürlich zutiefst uneins darüber sind, was moralisch und was unmoralisch ist. Doch selbst falls Menschen dazu gebracht werden könnten, bei allen Übungsfällen übereinzustimmen, blieben noch mindestens zwei Probleme.

Das erste Problem ist, dass eine Ausbildung mithilfe von Fällen aus unserer heutigen Welt möglicherweise nicht zu einer Maschine führt, die in einer Welt, die von einer Superintelligenz radikal umgestaltet wurde, korrekte ethische Entscheidungen trifft.

Das zweite Problem ist, dass eine Superintelligenz aufgrund zufälliger Muster in den Übungsdaten die falsche Prinzipien schlussfolgern könnte.[54] Betrachten wir hierzu die Parabel der Maschine, die darauf trainiert wurde, getarnte Panzer in einem Wald zu erkennen. Forscher nehmen 100 Fotos von getarnten Panzern und 100 Fotos von Bäumen. Mit jeweils 50 Fotos schulen sie die Maschine darin, getarnte Panzer von Bäumen zu unterscheiden. Zur Überprüfung zeigen sie der Maschine die jeweils verbleibenden 50 Fotos, und sie klassifiziert alle richtig. Erfolg! Spätere Tests zeigen jedoch, dass die Maschine sehr schlecht im Klassifizieren weiterer Fotos von getarnten Panzern und Bäumen ist. Wie sich herausstellt, liegt das Problem darin, dass die Fotos von getarnten Panzern an wolkigen Tagen aufgenommen wurden, die Fotos von Bäumen dagegen an sonnigen Tagen. Die Maschine hatte gelernt, wolkige Tage von sonnigen Tagen zu unterscheiden, nicht jedoch getarnte Panzer von Bäumen.

Somit scheint es, dass ein vertrauenswürdiges Friendly AI-Design detaillierte Modelle der Prozesse beinhalten muss, die menschliche moralische Urteile erzeugen, nicht nur oberflächliche Ähnlichkeiten verschiedener Fallbeispiele.

Siehe auch:

- Yudkowsky, Artificial intelligence as a positive and negative factor in global risk

4.7 Was ist „Coherent Extrapolated Volition”?

Eliezer Yudkowsky hat Coherent Extrapolated Volition (deutsch etwa „kohärenter, extrapolierter

Wille") als eine Lösung für mindestens zwei Probleme vorgeschlagen[57], mit denen sich die Entwicklung von Friendly AI konfrontiert sieht:

1. *Die Fragilität menschlicher Werte*: Yudkowsky schreibt, dass „jede Zukunft, die nicht von einem Zielsystem geprägt ist, das ein detailliertes, zuverlässiges Erbe menschlicher Moral und Metamoral antritt, nahezu nichts von Wert enthalten" wird. Das Problem ist, dass menschliche Werte komplex, subtil und schwierig zu spezifizieren sind. Betrachten wir hierzu den scheinbar unbedeutenden Wert *Neuheit*. Falls ein menschenähnliches Wertschätzen von Neuheit nicht in eine superintelligente Maschine einprogrammiert ist, könnte diese das Universum bis zu einem gewissen Punkt nach wertvollen Dingen erforschen und dann das wertvollste, das es findet, maximieren (der sogenannte Exploration-Exploitation-Tradeoff[58]) – das Universum mit Gehirnen in Tanks zu pflastern, die an "Glücksmaschinen" angeschlossen sind, zum Beispiel. Wenn eine Superintelligenz das Sagen hat, muss man ihr Motivationssystem *genau richtig* hinkriegen, um die Zukunft nicht wertlos zu machen.
2. *Die Lokalität menschlicher Werte*: Stellen wir uns vor, die alten Griechen hätten am Problem der Friendly AI gearbeitet und die Griechen hätten diese mit den fortschrittlichsten Moralvorstellungen ihrer Zeit programmiert. Das hätte die Welt zu einem ziemlich entsetzlichen Schicksal verdammt. Doch warum sollten wir denken, dass die Menschen im 21. Jahrhundert auf dem Gipfel der menschlichen Moral angelangt sind? Wir können es nicht riskieren, eine superintelligente Maschine mit den moralischen Werten zu programmieren, denen wir heute verpflichtet sind. Aber welche moralischen Werte geben wir ihr *dann*?

Yudkowsky schlägt vor, dass wir eine „Anfangs-KI" (engl. *seed AI*) konstruieren, um die Coherent Extrapolated Volition der Menschheit zu entdecken:

> *Poetisch gesprochen ist unsere Coherent Extrapolated Volition unser Wunsch, falls wir mehr wüssten, schneller dächten, mehr so wären wie wir wünschten, gemeinsam weiter aufgewachsen wären; wobei die Extrapolation konvergiert anstatt divergiert, unsere Wünsche übereinstimmen, anstatt sich zu widersprechen; so extrapoliert, wie wir es extrapoliert wünschen, so interpretiert, wie wir es interpretiert wünschen.*

Die Anfangs-KI würde die Ergebnisse dieser Untersuchung und Extrapolation menschlicher Werte verwenden, um das Motivationssystem der Superintelligenz zu programmieren, die die Zukunft der Galaxie bestimmt.

Allerdings befürchten einige, dass der kollektive Wille der Menschheit nicht zu einer kohärenten Menge von Zielen konvergiert. Andere glauben, dass garantierte Freundlichkeit selbst mit solch aufwendigem und umsichtigem Einsatz von Mitteln nicht zu erreichen ist.

Siehe auch:

- Yudkowsky, Coherent Extrapolated Volition

4.8 Können wir Freundlichkeit jedem KI-Design hinzufügen?

Viele KI-Designs, die zu einer Intelligenzexplosion führen würden, hätten keinen „Steckplatz“, in den ein Ziel (wie „Sei menschlichen Interessen gegenüber wohlwollend“) eingefügt werden könnte. Falls eine KI beispielsweise durch Whole Brain Emulation, evolutionäre Algorithmen, neurale Netze oder Verstärkendes Lernen realisiert wird, wird die KI im Prozess der Selbstverbesserung zu einem Ziel gelangen, aber dieses stabile, endgültige Ziel könnte sehr schwer vorherzusagen sein.

Um eine Friendly AI zu entwickeln, ist es demnach nicht hinreichend zu bestimmen, was „Freundlichkeit“ ist (und es klar genug festzulegen, damit auch eine Superintelligenz es so interpretiert, wie wir das wollen). Wir müssen zusätzlich herausfinden, wie man eine allgemeine Intelligenz bauen kann, die überhaupt ein Ziel erfüllt und dieses Ziel auch dann dauerhaft beibehält, wenn sie ihren eigenen Quellcode verändert, um sich selbst intelligenter zu machen. Diese Aufgabe ist möglicherweise die Hauptschwierigkeit beim Entwickeln von Friendly AI.

4.9 Wer arbeitet am Problem der Friendly AI?

Heute wird die Forschung an Friendly AI vom Singularity Institute (in San Francisco, Kalifornien), vom Future of Humanity Institute (in Oxford, Großbritannien) und von einigen anderen Forschern, etwa David Chalmers, betrieben. Gelegentlich schneiden Forscher für Maschinenethik das Thema an, zum Beispiel Wendell Wallach und Colin Allen in *Moral Machines*.

4.10 Was ist der Unterschied zwischen dem Singularity Institute und der Singularity University?

Das Singularity Institute ist ein Non-Profit-Forschungsinstitut, das daran arbeitet, das Problem der Friendly AI zu lösen. Die Singularity University ist eine separate Organisation, die von Ray Kurzweil geleitet wird und Bildungsprogramme zu neu entstehenden Zukunftstechnologien anbietet.

Quellen

[1] Yudkowsky (2001). *Creating Friendly AI 1.0*. Singularity Institute.

[2] Anderson & Anderson, eds. (2006). *IEEE Intelligent Systems, 21(4)*.

[3] Anderson & Anderson, eds. (2011). *Machine Ethics*. Cambridge University Press.

[4] Arkin (2009). *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall.

[5] Capurro, Hausmanninger, Weber, Weil, Cerqui, Weber, & Weber (2006). *International Review of Information Ethics, Vol. 6: Ethics in Robots*.

[6] Danielson (1992). *Artificial morality: Virtuous robots for virtual games*. Routledge.

[7] Lokhorst (2011). Computational meta-ethics: Towards the meta-ethical robot. *Minds and Machines*.

[8] McLaren (2005). Lessons in Machine Ethics from the Perspective of Two Computational Models of Ethical Reasoning. *AAAI Technical Report FS-05-06*: 70-77.

[9] Powers (2005). Deontological Machine Ethics. *AAAI Technical Report FS-05-06*: 79-86.

[10] Sawyer (2007). Robot ethics. *Science, 318(5853)*: 1037.

[11] Wallach, Allen, & Smit (2008). Machine morality: Bottom-up and top-down approaches for modeling human moral faculties. *AI and Society, 22(4)*: 565–582.

[12] Allen (2002). Calculated morality: Ethical computing in the limit. In Smit & Lasker, eds., *Cognitive, emotive and ethical aspects of decision making and human action, vol I*. Baden/IIAS.

[13] Good (1965). Speculations concerning the first ultraintelligent machine. *Advanced in Computers, 6*: 31-88.

[14] MacKenzie (1995). The Automation of Proof: A Historical and Sociological Exploration. *IEEE Annals, 17(3)*: 7-29.

[15] Nilsson (2009). *The Quest for Artificial Intelligence*. Cambridge University Press.

[16] Campbell, Hoane, & Hsu (2002). Deep Blue. *Artificial Intelligence, 134*: 57-83.

[17] Markoff (2011). Computer Wins on 'Jeopardy!'; Trivial, it's Not. *New York Times, February 17th 2011*: A1.

[18] King et al. (2009). The automation of science. *Science, 324*: 85-89.

[19] King (2011). Rise of the robo scientists. *Scientific American, January 2011*.

[20] Legg (2008). *Machine Super Intelligence*. PhD Thesis. IDSIA.

[21] Hutter (2005). *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer.

[22] Veness, Ng, Hutter, & Silver (2011). A Monte Carlo AIXI Approximation. *Journal of Artificial Intelligence Research, 40*: 95-142.

[23] Tang, Shimizu, Dube, Rampon, Kerchner, Zhuo, Liu, & Tsien (1999). Genetic enhancement of learning and memory in mice. *Nature, 401*: 63–69.

[24] Hochberg, Serruya, Friehs, Mukand, Saleh, Caplan, Branner, Chen, Penn, & Donoghue (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature 442*: 164-171.

[25] Bostrom (1998). How long before superintelligence? *International Journal of Future Studies, 2*.

[26] Kurzweil (2005). *The Singularity is Near*. Viking.

[27] Chalmers (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies, 17*: 7-65.

[28] Baum, Goertzel, & Goertzel (forthcoming). How Long Until Human-Level AI? Results from an Expert Assessment. *Technological and Forecasting Change*.

[29] Grossberg (1992). *Neural Networks and Natural Intelligence*. MIT Press.

[30] Martinetz & Schulten (1991). A 'neural-gas' network learns topologies. In Kohonen, Makisara, Simula, & Kangas (eds.), *Artificial Neural Networks* (pp. 397-402). North Holland.

[31] de Garis (2010). Artificial Brains. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 159-174). Springer.

[32] Schmidhuber (2010). Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 199-223). Springer.

[33] Hutter (2010). Universal Algorithmic Intelligence: A Mathematical Top-Down Approach. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 227-287). Springer.

[34] Yudkowsky (2010). Levels of Organization in General Intelligence. In Goertzel & Pennachin (eds.), *Artificial General Intelligence* (pp. 389-496). Springer.

[35] Dreyfus (1972). *What Computers Can't Do*. Harper & Row.

[36] Penrose (1994). *Shadows of the Mind*. Oxford University Press.

[37] Searle (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*: 417-457.

[38] Block (1981). Psychologism and behaviorism. *Philosophical Review, 90*: 5-43.

[39] Asimov (1942). Runaround. *Astounding Science Fiction, March 1942*. Street & Smith.

[40] Anderson (2008). Asimov's 'three laws of robotics' and machine metaethics. *AI & Society, 22(4)*: 477-493.

[41] Kringelbach & Berridge, eds. (2009). *Pleasures of the Brain*. Oxford University Press.

[42] Schroeder (2004). *Three Faces of Desire*. Oxford University Press.

[43] Yudkowsky (2007). The hidden complexity of wishes.

[44] Smart (1958). Negative utilitarianism. *Mind, 67*: 542-543.

[45] Russell & Norvig (2009). *Artificial Intelligence: A Modern Approach, 3rd edition*. Prentice Hall. (see page 1037)

[46] Frijda (2009). On the nature and function of pleasure. In Kringelbach & Berridge (eds.), *Pleasures of the brain* (pp. 99-112). Oxford University Press.

[47] Aldridge & Berridge (2009). Neural coding of pleasure: 'rose-tinted glasses' of the ventral pallidum. In Kringelbach & Berridge (eds.), *Pleasures of the brain* (pp. 62-73). Oxford University Press.

[48] Smith, Mahler, Pecina, & Berridge (2009). Hedonic hotspots: generating sensory pleasure in the brain. In Kringelbach & Berridge (eds.), *Pleasures of the brain* (pp. 27-49). Oxford University Press.

[49] Guarini, (2006). Particularism and classification and reclassification of moral cases. IEEE Intelligent Systems 21(4): 22–28.

[50] Anderson, Anderson, & Armen (2005). Toward machine ethics: Implementing two action-based ethical theories. In *Proceedings of the AAAI Fall 2005 Symposium on Machine Ethics*, Arlington, Virginia, November.

[51] Honarvar & Ghasem-Aghaee (2009). An artificial neural network approach for creating an ethical artificial agent. *Proceedings of the 8th IEEE international conference on Computational intelligence in robotics and automation*: 290-295.

[52] Rzepka & Araki (2005). What statistics could do for ethics? – The idea of common sense processing based safety valve. In *Machine ethics: papers from the AAAI fall symposium*. American Association of Artificial Intelligence.

[53] Gorman & Sejnowski (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks, 1*: 75-89.

[54] Yudkowsky (2008). Artificial intelligence as a positive and negative factor in global risk. In

Bostrom & Cirkovic (eds.), *Global Catastrophic Risks*. Oxford University Press.

[55] Omohundro (2008). The Basic AI Drives.

[56] Bostrom & Cirkovic, eds. (2008). *Global Catastrophic Risks*. Oxford University Press.

[57] Yudkowsky (2004). Coherent extrapolated volition. Singularity Institute.

[58] Azoulay-Schwartz, Kraus, & Wilkenfeld (2004). Exploitation vs. exploration: choosing a supplier in an environment of incomplete information. *Decision Support Systems, 38*: 1-18.

[59] Caplan (2008). The totalitarian threat. In Bostrom & Cirkovic (eds.), *Global Catastrophic Risks*. Oxford University Press.

[60] Yudkowsky (2007). The Power of Intelligence.

[61] Bainbridge (2005). Survey of NBIC Applications. In Bainbridge & Roco (eds.), *Managing nano-bio-info-cogno innovations: Converging technologies in society*. Springer.

[62] Sandberg & Bostrom (2011). Machine intelligence survey, Technical Report 2011-1, Future of Humanity Institute, Oxford.