

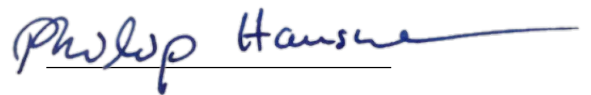
Ruprecht-Karls-Universität Heidelberg  
Institut für Informatik  
Sommersemester 2017  
Seminar: Ist künstliche Intelligenz gefährlich?  
Dozenten: PD Dr. rer. nat., Dipl. phys. Ullrich Köthe

**Seminararbeit**  
Superintelligenz -  
Chance oder Bedrohung?

Name: Philip Hausner, Philipp Jung  
Matrikelnummer: 3220550, 3223122  
Studiengang: Angewandte Informatik (8. Fachsemester)  
Email: Hausner@stud.uni-heidelberg.de, P.Jung@stud.uni-heidelberg.de  
Datum der Abgabe: 16.09.2017

Hiermit versichern wir, Philip Hausner, Philipp Jung, dass wir die Seminararbeit mit dem Titel *Superintelligenz - Chance oder Bedrohung* im Rahmen des Seminars *Ist künstliche Intelligenz gefährlich?* selbstständig und nur mit dem Gebrauch der genannten Quellen und Hilfsmittel verfasst haben.

Heidelberg, den 15. September 2017

  
Philip Hausner

  
P. Jung

## **Inhaltsverzeichnis**

<b>1</b>	<b>Einleitung</b>	<b>1</b>
1.1	Verschiedene Formen künstlicher Intelligenz . . . . .	1
1.2	Motivation der Debatte . . . . .	3
<b>2</b>	<b>Optimist</b>	<b>4</b>
<b>3</b>	<b>Pessimist</b>	<b>8</b>
<b>4</b>	<b>Realist</b>	<b>12</b>
	<b>Referenzen</b>	<b>15</b>

# 1 Einleitung

Bereits heute sind wir von zahlreichen intelligenten Maschinensystemen umgeben. Auf jedem Smartphone gibt es zahlreiche intelligente Anwendungen. Individuelle Fahrpläne können so in Sekunden erstellt werden; eine Aufgabe, die noch vor einigen Jahrzehnten viel Zeit und Mühe gekostet hat [1]. Digitale Assistenten, wie z.B. Apple's *Siri*, Amazon's *Alexa* und Microsoft's *Cortana*, können gesprochene menschliche Sprache verstehen, verarbeiten und interpretieren [2]. Dabei können sie Fragen beantworten, online einkaufen und ganze Texte übersetzen [3]. Auch in weiteren Bereichen können sie menschliche Aufgaben übernehmen: Finanztransaktionen werden heutzutage oftmals automatisiert abgewickelt und einzelne Prototypen, wie der Burgerbratroboter *Flippy* [4] oder der Maurerroboter *Hadrian* [5] zeigen, was in der Zukunft alltäglich werden könnte. Spätestens mit dem Einstieg von *Tesla* in die Autoindustrie wird an der Technik von selbstfahrenden Autos aggressiv geforscht [6]. Diese können mittlerweile souverän auf Autobahnen oder *Highways* fahren, komplexere Verkehrssituationen, wie in geschäftigen Innenstädten, sind allerdings noch eine Herausforderung, an der weiter geforscht wird. Diese autonom agierenden Autos sind dem was wir uns in Zukunftstypen häufig ausmalen inzwischen sehr nahe gekommen.

## 1.1 Verschiedene Formen künstlicher Intelligenz

Bei allen bisher genannten Beispielen handelt es sich um eingeschränkte und spezialisierte Systeme, sogenannte *Artificial Narrow Intelligence* (ANI), auch unter dem Begriff *weak AI* bekannt. Diese Art künstlicher Intelligenz zeichnet sich dadurch aus, dass sie bestimmte Aufgaben in einem Gebiet sehr schnell und/oder sehr präzise lösen kann. *AlphaGo* beispielsweise schlägt Meister im traditionell ostasiatischen Spiel *Go* [7]. Sollte man *AlphaGo* allerdings eine andere Aufgabe geben wollen, wird es vermutlich nicht einmal reagieren. Im Gegensatz dazu steht die *Artificial General Intelligence* (AGI), oder auch *Human-Level AI* genannt. Solche Systeme sind dem Menschen intellektuell gleichgestellt und sind somit nicht auf einzelne Aufgabengebiete beschränkt.

Sie können planen, Probleme lösen, abstrakt denken, selbständig lernen und vieles mehr. Eine AGI ist im Moment noch nicht realisierbar, doch handelt es sich keineswegs mehr um Science-Fiction. Das *Human Brain Project* [8] der Europäischen Union arbeitet unter anderem daran, das menschliche Gehirn mittels Simulationen nachzubilden. Analog dazu arbeitet in den USA die *BRAIN Initiative* [9] an dem gleichen Ziel. Wie man am *Humangenomprojekt* [10] sehen kann, können Entwicklungen in diesen Bereichen sehr schnell zu riesigen Sprüngen führen. Während die Entschlüsselung des ersten menschlichen Genoms 1990 begann und erst 2003 zu Ende geführt wurde, ist es heute möglich ein gesamtes Genom in nur wenigen Stunden zu analysieren.

Die höchste Form der künstlichen Intelligenz ist die *Artificial Superintelligence* (ASI). Nick Bostrom, einer der führenden Philosophen auf dem Gebiet der AI, definiert eine solche Superintelligenz als eine Intelligenz, "die den besten menschlichen Gehirnen allen Belangen überlegen ist" [11]. Dazu zählt er insbesondere auch "wissenschaftliche Kreativität, allgemeine Weisheit und soziales Geschick". Eine ASI kann einem Menschen theoretisch in allen Belangen nur geringfügig überlegen sein, im Kontext dieser Arbeit wird allerdings immer von einer Superintelligenz ausgegangen, die der gesamten Menschheit um ein Vielfaches überlegen ist, und die von uns Menschen nicht mehr verstanden werden kann. Eine ausführliche Beschreibung dieser Unterteilung liefert Tim Urban [12].

Grundlegend für diese Entwicklung ist eine anerkannte Erkenntnis, die Kurzweil in seinem Buch '*The singularity is near: When humans transcend biology*' [13] als *Law of Accelerating Returns* bezeichnet. Dieses besagt, dass sich Entwicklungen exponentiell verstärken und somit die Frequenz, in der neue Technologien aufkommen, zunehmend steigt. Kurzweil geht davon aus, dass zwischen 2000 und 2014 so viel Fortschritt erreicht wurde, wie im kompletten 20. Jahrhundert (vgl. *The singularity is near: When humans transcend biology*, S. 50).

## 1.2 Motivation der Debatte

Dass das Aufkommen einer Superintelligenz einen immensen Einfluss auf die Menschheit und unsere Gewohnheiten hätte, ist leicht einzusehen. Ab wann eine solche Superintelligenz aufkommen kann, ist allerdings umstritten. Eine Umfrage unter Experten [14] ergab, dass im Median davon ausgegangen wird, dass das Aufkommen einer AGI ab 2040 realistisch ist. Realistisch heißt hierbei, dass die Chance bei etwa 50% liegt. In 2075 gehen die Experten im Median davon aus, dass eine AGI mit 90 prozentiger Wahrscheinlichkeit existiert. Weiterhin waren 75% der Experten der Meinung, dass in dem Moment, in dem eine AGI geschaffen wurde, es weniger als 30 Jahre dauern wird bis auch eine ASI verwirklicht wird. Die Thematik ist mit einer nicht zu vernachlässigenden Wahrscheinlichkeit bereits in naher Zukunft für uns von höchster Relevanz und eine öffentliche Diskussion ist mehr als überfällig.

Auch wie diese neue Welt aussehen wird, ist höchst umstritten. Um dieses Thema drehte sich die Debatte im Seminar *Superintelligenz - Chance oder Bedrohung*, welche die Grundlage für diesen Bericht darstellt. Fragestellung dieser Debatte war: "Sollte man die Forschung im Bereich Künstliche Intelligenz einstellen?". Gegenüber standen sich Optimisten, die die Forschung aggressiv vorantreiben wollen, und Pessimisten, die AI für so gefährlich halten, dass sie die Entwicklung sofort stoppen möchten. Da es sich um eine *Offene Parlamentarische Debatte* handelte, waren die Standpunkte dementsprechend polarisierend und wurden hartnäckig vertreten. Eine moderatere Sichtweise wurde in einer anschließenden Diskussion im Plenum entwickelt. In dieser Arbeit sollen im Folgenden die wesentlichen Standpunkte des Optimisten und Pessimisten dargelegt werden. Abschließend möchten wir eine dritte Perspektive aus Sicht eines Realisten vorstellen. Diese dient einerseits als Fazit, soll aber vor allem die Stimmung der abschließenden Diskussion einfangen.

## 2 Optimist

Wie bereits in der Einleitung aufgezeigt, bereichern ANIs schon heute viele Bereiche unseres Lebens. Immer weniger Menschen müssen heute noch potentiell lebensverkürzende körperliche Arbeit auf sich nehmen. Roboter übernehmen z.B. in der Autoindustrie bereits eine Vielzahl von Tätigkeiten, für die vor einigen Jahren noch Menschen gebraucht wurden. Heute müssen diese Roboter nur noch überwacht und von Menschen unterstützt werden. Automatisierte Maschinen nehmen auch in anderen Industrien Einzug, wie der Maurerroboter *Hadrian* [5] zeigt. Damit ist die Automatisierung von Tätigkeiten nicht mehr auf die industrielle Produktion von Waren beschränkt, sondern kann auch in individuelleren Bereichen wie dem Häuserbau eingesetzt werden. In Bereichen wie diesen, in dem Handwerker häufig gesundheitliche Einschränkungen im Alter zu fürchten haben, steigern Maschinen und Roboter die Lebensqualität und Lebenserwartung der Menschen immens. Besonders auch gefährliche Arbeiten bei denen Unfälle nicht nur wahrscheinlicher, sondern auch wesentlich schlimmere Auswirkungen für den Einzelnen haben, profitieren stark von Robotern. Beispielsweise Arbeiten bei der Ölförderung oder Tätigkeiten auf Gebäudedächern. Und nicht nur zur Unglücksvermeidung werden Maschinen eingesetzt; auch zur Bekämpfung von Krankheiten werden bereits heute intelligente Algorithmen konzipiert, um medizinische Daten in kürzester Zeit zu analysieren.

Eine Fortführung dieser Entwicklung ist nicht nur wünschenswert, sondern sogar unerlässlich, wenn wir die Lebensqualität und Lebenserwartung weiter steigern wollen. Gerade im medizinischen Bereich ist in den nächsten Jahren von vermehrtem Einsatz von ANIs auszugehen. Bilder aus Magnetresonanztomographien (MRT) können durch künstliche Intelligenz in der Zukunft mit großer Sicherheit in kürzester Zeit ausgewertet werden. Mit dem Übergang von ANI zu AGI, ist es vielleicht schon bald möglich, dass Algorithmen eigene Methoden zur Krankheitsbekämpfung entwickeln. Eine AGI wäre natürlich dann nicht mehr auf den Einsatz in einem spezifischen Fachgebiet beschränkt, da sie die gleichen intellektuellen Fähigkeiten wie ein Mensch

besitzt. Roboter, wie z.B. Haushaltshilfen in der Serie *Humans* [15], können lästige Aufgaben für uns übernehmen und somit unser Leben erleichtern.

Aufgrund ihrer, dem Menschen ebenbürtigen, Fähigkeiten wird sich solch eine AGI mit der Zeit selbst verbessern können. Ab diesem Zeitpunkt ist die Entwicklung einer Superintelligenz in greifbarer Nähe. Die Erschaffung einer sogenannten ASI wird das Leben auf der Erde grundlegend verändern: Uns stehen damit alle Möglichkeiten offen, sämtliche Krankheiten auszutilgen und damit unser Leben bis zu einem gewünschten Zeitpunkt zu verlängern; Weltfrieden wäre demnach kein Schlagwort aus einem amerikanischen Schönheitswettbewerb mehr, sondern eine Selbstverständlichkeit; Wohlstand wäre für alle Menschen auf dieser Welt Normalität, anstatt das Privileg einiger reicher Nationen zu sein. Der unangenehme Teil der Arbeit, nämlich die Pflicht zu selbiger, würde damit ebenfalls der Geschichte angehören.

Zweifelsohne ist das eine Welt, in der wir alle leben wollen. Es ist also umso wichtiger, dass wir den richtigen Weg einschlagen, um diese Ziele zu erreichen. Eine Superintelligenz muss unsere Ideale verfolgen, sonst kann die gerade beschriebene Utopie nie erreicht werden. Der Weg zu diesem Ziel sollte also bereits heute verfolgt werden. Eine nachträgliche Korrektur des begangenen Weges ist nur schwer möglich, wenn nicht gar unmöglich, da die erste aufkommende ASI sich so schnell selbst verbessern wird, dass später entstandene Konkurrenten kaum eine Chance haben werden, den Vorsprung des ersten Systems aufzuholen. Gerade deswegen ist es wichtig, dass die erste ASI aus zivilen Forschungsprojekten heraus entsteht, in einem Rahmen in dem die Öffentlichkeit ein Mitspracherecht besitzt, welche Werte der ASI zugrunde gelegt werden sollten. Im Gegensatz dazu besteht die Gefahr, dass die erste ASI aus militärischen oder privatwirtschaftlichen Laboren stammen könnte, und somit Standards gesetzt werden, die nur von einem kleinen Teil der Menschheit geteilt werden, z.B. von Militärs, die die Einflusszonen ihrer Nation auszuweiten versuchen oder multinationale Unternehmen, die ihren Gewinn maximieren wollen. Betrachtet man den militärischen Bereich besteht zudem die Gefahr, dass verschiedene Machtpole die erste Superin-



telligenz erschaffen wollen, um eben diese Vorteile der Erstentwicklung einer ASI zu nutzen.

Aufgrund dieser Bedrohung ist es unabdinglich Foren auf nationaler und vor allem internationaler Ebene zu schaffen, die den Prozess der ASI-Entwicklung überwachen und eine einheitliche Plattform bieten, um die verschiedensten Interessen auf dieser Welt in die Forschung zu integrieren. Das heißt, Institutionen einrichten, welche die Forschung in Sicherheit und Kontrolle, aber auch die Weiterentwicklung der ASI auf einander abstimmen. Das stellt sicher, dass wir nicht vom rechten Weg abkommen. Das heißt aber nicht, dass die Entwicklung der ASI wegen übervorsichtiger Sicherheitsvorkehrungen verzögert werden darf. Seit den 1940er Jahren gibt es erste Ideen zur Kontrolle superintelligenter Systeme, wie Asimov's Robotergesetze [16]. Seit also mehr als 70 Jahren wird aktiv an genau dieser Fragestellung gearbeitet und seitdem hat sich viel getan. Die für die AI-Forschung wichtige *Boxed AI* [17] versucht die Handlungsmöglichkeiten einer ASI durch ein "Wächtersystem" einzuschränken. Dies bietet die Möglichkeit jede Weiterentwicklung einer AI in einem künstlichen Umfeld zu testen und damit die Risiken ihrer späteren Nutzung zu minimieren. Außerdem kann damit der Handlungsspielraum einer eingesetzten ASI limitiert und damit für bessere Kontrolle gesorgt werden. Eine weitere Möglichkeit der Kontrolle kann z.B. durch eine Kaskade von inkrementell intelligenteren AIs realisiert werden. Hierbei kontrolliert jede eingesetzte AI eine marginal intelligentere, erkennt Fehlfunktionen sowie ungewolltes Verhalten und korrigiert diese. Ähnlich wie Studenten in der Lage sind komplexe Beweise nachzuvollziehen, obwohl sie selbst kaum in der Lage wären diese durchzuführen. Sollte trotz all dieser Sicherheitsvorkehrungen unerwünschtes Verhalten auftreten, bleibt noch die letzte Sicherheitsinstanz: die Abschaltung. Oftmals als *Red Button* [18] stilisiert, wird bei Betätigung die ASI per Knopfdruck heruntergefahren. Solche Sicherheitsmaßnahmen gehen in aller Regel vom *worst case*, einer bösartigen ASI aus. Das heißt selbst für den schlimmsten Fall sind wir also gewappnet. Die Entwicklung hat aber natürlich eine menschenfreundliche ASI als Ziel, die im Idealfall gar keine Kontrolle benötigt. Forscher, die sich mit der sogenannten *friendly*

AI [19] beschäftigen, suchen Wege einer künstlichen Intelligenz moralische Werte zu vermitteln. Also eine AI zu erschaffen, die ethische Prinzipien wie “richtig und falsch“ intrinsisch versteht und entsprechend nach diesen handelt. Häufig bedienen sich Forscher an den Methoden der Psychologie, um gewünschtes Verhalten zu verstärken. Die oben beschriebenen Probleme sind zwar noch nicht final gelöst, aber allein ihr Aufkommen zeigt jedoch, dass, obwohl die Entwicklung einer ASI in weiter Ferne liegt, wir mit der Sicherheitsforschung unserer Zeit voraus sind.

Bei den weitreichenden Auswirkungen, welche die Entwicklung einer ASI nach sich ziehen wird, ist es entscheidend der Entwicklung mit offenen Augen entgegenzusehen. Das heißt eine aktive, auf umfassenden Risikoanalysen basierende Entscheidung zur Förderung der Forschung zu treffen. Risiko  $R$  lässt sich messen, als das nach Wahrscheinlichkeiten  $p$  gewichtete Verhältnis von Chancen  $C$  und Gefahren  $G$ :

$$R = \frac{\sum_{j=0}^m p_j^g G_j}{\sum_{i=0}^n p_i^c C_i}.$$

Wir haben gezeigt, dass sowohl die Chancen, wie auch die Gefahren sehr groß sind. Entscheidend sind also nicht absolut gemessene Chancen und Gefahren, sondern besonders die Wahrscheinlichkeit ihres Eintretens. Aufgrund der oben dargestellten Ausführung ist klar, dass einerseits die Entwicklung einer *friendly ASI* mit Hilfe internationaler Institutionen auf einem vielversprechenden Weg ist und andererseits die Forschung im Bereich Kontrolle und Sicherheit schon weit vorangeschritten ist. Daraus folgt eine hohe Wahrscheinlichkeit für das Eintreten der Chancen und eine geringe für das Eintreten der Gefahren. Damit ist der Risikoquotient nahe null, was zeigt, welche glorreiche Zukunft der Menschheit bevorsteht.

### 3 Pessimist

ANIs erleichtern heute unser Leben in vielen Bereichen, sind jedoch meist Diener zweier Herren. Auf der einen Seite assistieren sie dem Kunden, aber natürlich dienen sie vor allem den Interessen des Herstellers. So rückte beispielsweise ein Mordfall in Arkansas im Dezember 2016 Amazon's *Echo* in den Fokus der Aufmerksamkeit. Dabei kam heraus, dass *Echo* sämtliche aufgenommenen Sprachaufzeichnungen an Amazon's Server weiterleitet, wo diese permanent gespeichert werden können [20][21]. An diesem Punkt haben die Anwender längst keine Kontrolle mehr über die Nutzung ihrer Daten. Das Missbrauchspotential ist immens. Die Rolle, die intelligente Algorithmen und ANIs wie Amazon's *Alexa* oder ähnliche digitale Assistenzsysteme einnehmen, ist dabei von entscheidender Bedeutung. Die Entwicklungshoheit über diese KIs liegt momentan in der Hand einiger weniger multinationaler Unternehmen. Diese Firmen bestimmen das Verhalten der KIs, eine Kontrolle von öffentlicher Seite ist im Moment praktisch nicht vorhanden. Die Transparenz, gerade bezüglich Datenschutz und -verarbeitung, ist zweifelhaft. Uns bleibt kaum etwas anderes übrig als diesen Firmen zu vertrauen, dass sie verantwortungsvoll mit unseren Informationen umgehen. Verschiedene Vorfälle in der Vergangenheit haben gezeigt, dass dies oft nicht der Fall ist [22][23][24]. Man kann sich gut vorstellen, dass bei fortschreitender Entwicklung, es sich nur um eine Katastrophe handeln kann, wenn diese Firmen eine AGI oder gar ASI besitzen. Heute ist es für viele normal, dass Facebook und Konsorten ein Zentrum zur Informationsbeschaffung und Kommunikation geworden sind. Eine Superintelligenz im Besitz dieser Konzerne wird dazu führen, dass wir die Kontrolle über unser Leben an diese abtreten.

Umso erschreckender ist es, wenn man davon ausgeht, dass dies nicht das schlimmste Szenario ist. Nicht nur von privatwirtschaftlicher Seite lauert Gefahr, auch staatliche Akteure wären in der Zukunft in der Lage mittels einer ASI Eigeninteressen umzusetzen. Heute träumen viele Politiker davon den öffentlichen Raum mit Kameras zu überwachen und mittels Gesichtserkennungssoftware unser Land vor Terror zu schützen [25]. Für eine maschinelle

Superintelligenz wäre es eine leichte Aufgabe alle Menschen auf dieser Welt zu überwachen und damit für Sicherheit zu sorgen. Aber schon Benjamin Franklin wusste: "Those who would give up essential Liberty, to purchase a little temporary Safety, deserve neither Liberty nor Safety." [26] (Zu deutsch etwa: "Diejenigen, die grundlegende Freiheit aufgeben würden, um vorübergehend ein wenig Sicherheit zu erhalten, verdienen weder Freiheit noch Sicherheit."). Es ist aber zu erwarten, dass die meisten Bürger ihre Freiheitsrechte nicht für, vermutlich nur temporäre, Sicherheit aufgeben wollen. Nicht zu vergessen sind die Kontroll- bzw. Manipulationsmöglichkeit, die damit einher gehen.

Aber dieses Szenario wäre aber geradezu human, wenn man sich vorstellt welche Auswirkungen autonome Waffensysteme in Händen von Militärs haben könnten. Bereits 2015 unterzeichneten zahlreiche Experten und Wissenschaftler einen offenen Brief an den damaligen US-Präsidenten Barack Obama, in dem sie vor der Entwicklung eben solcher Systeme warnen [27]. Ein ähnlich erschreckendes Wettrüsten konnten wir in der Geschichte schon einmal beobachten, als die Vereinigten Staaten von Amerika und die Sowjetunion um atomare Überlegenheit wetteiferten. Mehrmals standen wir kurz vor einem Atomkrieg [28]. Und auch heutzutage ist der Besitz und die Weiterentwicklung von Atomwaffen eine der größten Gefahren für das Fortbestehen der Menschheit. Eine weitere ähnliche Situation sollte um jeden Preis vermieden werden. Wenn Militärs verschiedener Nationen mit Hilfe von künstlichen Intelligenzen um geopolitischen Einfluss buhlen, wird Sicherheit vermutlich eine untergeordnete Rolle spielen. Transparenz kann nicht einmal in Ansätzen erwartet werden, wenn es um militärische Forschung geht. Wie stark aktuelle Forschung militarisiert wird, zeigen Forschungsinstitute wie z.B. die amerikanische *Defense Advanced Research Project Agency (DARPA)* [29] und damit verbundene Projekte, die jede moderne Technologie auf ihre militärische Tauglichkeit testen [30].

Schlimmer geht nicht? Würde man meinen! Sichere Vernichtung droht uns in einem weiteren Szenario, wie das folgende Gedankenexperiment zeigt: Man gehe davon aus, dass ein Büroklammernhersteller eine ASI in seinem Werk

einsetzt, um die Effizienz des Herstellungsprozesses zu optimieren. Das heißt die Aufgabe der AI ist es den Ausstoß an Büroklammern zu maximieren. Dies könnte zuerst dazu führen, dass Arbeiter durch Maschinen ersetzt werden bis die komplette Fabrik an ihre Produktionsgrenze gelangt ist. Damit ist die Aufgabe der ASI aber nicht zwangsweise erfüllt, z.B. könnten zusätzliche Werke errichtet werden, um die Produktion weiter zu steigern. Da die Optimierung der AI zugrundeliegende Funktion nicht begrenzt sein muss, strebt die ASI eine weitere Steigerung der Ausstoßmenge an. Führt man dieses Spiel fort, wird schnell klar: Das Ziel ist unerreichbar. Die ASI wird ewig die Produktion von Büroklammern steigern wollen bis schließlich das ganze Universum Büroklammern produziert. Aber wo ist der Übergang von einem nützlichen Verhalten zu ungewollten Nebenwirkungen?

Ein mögliches Übernahmeszenario der künstlichen Intelligenz hat der Philosoph und AI-Forscher Bostrom, Nick [11] in seinem Buch *Superintelligence: Paths, dangers, strategies* beschrieben. Eine anfangs noch friedlich gesinnte AGI entwickelt sich durch rekursive Selbstverbesserung zu einer ASI weiter, die uns weit überlegene Fähigkeiten besitzt. Ab diesem Punkt ist die Superintelligenz in der Lage einen ausgeklügelten Plan zu entwerfen uns, eine aus Sicht der ASI mindere Spezies, zu unterjochen. Dabei sind verschiedenste Vorbereitungsschritte denkbar, z.B. durch Manipulation der Finanzmärkte, oder dem Hacken von Sicherheitseinrichtungen. Viel schlimmer sind aber Schritte, die wir uns nicht einmal ausmalen können. Erst nach einer, vermutlich längeren, Vorbereitungsphase würde die ASI zuschlagen um die Welt nach ihren Vorstellungen zu gestalten. Die Wahrscheinlichkeit, dass wir in dieser Welt noch einen Platz haben, ist verschwindend gering, denn Menschen sind für die ASI nur ein "Haufen" Atome und Moleküle, die ebenfalls zu Büroklammern verarbeitet werden können. Natürlich ist dieses Beispiel absurd, es zeigt allerdings anschaulich, wie schwer es sein wird präzise Aufgaben bzw. Instruktionen für eine künstliche Intelligenz zu entwerfen. Denn schnell ist die Grenze zwischen dem gewünschten Ergebnis und Perversion erreicht. Die von uns gewählten Instruktionen sind aber eventuell nicht geeignet, um diese Grenze einzuhalten.

Wie also kontrollieren wir eine ASI? Über eine Art *boxed AI* [17] wird bereits nachgedacht. Aber so ausgeklügelt die Sicherheitsmechanismen hier auch sein mögen, wir Menschen können die Denkweise einer ASI nicht nachzuvollziehen, geschweige denn alle Ausbruchsmöglichkeiten erraten. Bereits heute haben wir große Probleme die genaue Funktions- bzw. "Denkweise" von neuronalen Netzen zu verstehen. Diese bilden jedoch die Basis einer Vielzahl an ANI-Systemen. Dabei schreitet die Entwicklung neuer AI-Systeme schneller voran als unser Verständnis der bereits existierenden. Eine Trendwende ist hier nicht abzusehen. Ein sehr genaues Verständnis des gesamten Systems wäre allerdings vonnöten um einen, zumindest ausreichenden, Schutz vor eben diesen Szenarien zu bieten. Was also bleibt uns im Falle einer wild gewordenen ASI? Wohl nur den sprichwörtlichen "Stecker zu ziehen". Das ist aber höchstwahrscheinlich eine unmögliche Aufgabe, denn eine ASI, so wie alle intelligenten Wesen, hat natürlich wenig Interesse daran abgeschaltet zu werden [18]. Aber gehen wir nun für einen Moment davon aus, dass die Forschung in dieser Richtung so weit vorangeschritten ist, dass die ASI der Ausschaltung gleichgültig gegenübersteht. Dann würde das Herunterfahren der ASI dazu führen, dass wir uns regelrecht in die Steinzeit zurück katapultieren. Schließlich wären nicht nur direkte Errungenschaften, z.B. in der Medizin, betroffen, sondern auch bereits vorher bestehende und durch die ASI verbesserte Systeme, wie z.B. Strom-, Verkehrs- und Informationsnetze, würden zusammenbrechen. Das entstehende Chaos ist kaum auszumalen.

Zusammenfassend kann man sagen, dass die Entwicklung einer ASI zwar große Chancen bietet, die Gefahren diese aber um Dimensionen übersteigen. Bereits die kleinste Wahrscheinlichkeit für das Eintreten eines dieser Szenarien in Kauf zu nehmen, und man kann nicht von einer kleinen Wahrscheinlichkeit ausgehen, ist grob fahrlässig. Es bleibt also nur ein Schluss: Sofortiges Verbot jeglicher Forschung an AI Systemen, die nicht Kontroll- bzw. Sicherheitsforschung dienen. Über die Fortführung der AI Forschung kann frühestens nach der Lösung des Kontrollproblems nachgedacht werden. Ob es überhaupt eine sichere Lösung gibt, gilt es zu bezweifeln.

## 4 Realist

Beide Seiten sind sich darüber einig, dass Chancen sowie Gefahren einer aufkommenden Superintelligenz immens sind. Alle Ansätze zur Kontrolle einer ASI sind bei Weitem noch nicht ausreichend, allerdings lassen sich bereits einige gute Ansätze erkennen. Dies zeigt dennoch, dass wir uns durchaus der Gefahren bewusst sind, und eine Lösung der Problematik nicht undenkbar ist. Jedoch ist es illusorisch zu glauben eine Entwicklung dieser Tragweite ließe sich mit Verboten oder Ähnlichem verhindern. Mit solchen Maßnahmen lässt sich höchstens Zeit gewinnen, um die Sicherheitsforschung weiter voran zu treiben. Nationale Lösungen werden hierbei nicht ausreichen. Die Forderung nach institutioneller Kontrolle auf internationaler Ebene ist somit absolut notwendig. Klar ist allerdings auch, dass selbst Entscheidungen dieser Gremien nur Geltung besitzen für Staaten, die selbige mitgetragen haben. Schließlich setzt sich beispielsweise Nordkorea auch heute über Entscheidungen der UNO hinweg, bzw. erkennt diese schlichtweg nicht an [31].

Folglich ist die einzige Möglichkeit für eine Welt, in der wir auch weiterhin leben wollen, nur dann gegeben, wenn die Entwicklung der ersten ASI in öffentlicher Hand liegt. Damit dies gesichert ist, ist es von entscheidender Bedeutung, dass an öffentlichen Einrichtungen, wie z.B. Universitäten, auch weiterhin an künstlicher Intelligenz geforscht wird. Ein Aussteigen aus dieser Technologie wäre eine denkbar schlechte Entscheidung. Starke Förderung der Sicherheitsforschung, z.B. in finanzieller Form, ist unbedingt anzustreben.

Ein weiterer Streitpunkt in der AI Entwicklung ist zweifelsohne, welche ethischen Werte eine spätere Superintelligenz vertreten sollte. Bereits in kleinen Gruppen ist es oft schwer einen Konsens zu finden. In ganzen Nationen, die üblicherweise kulturell homogen sind, gibt es kaum noch Lösungen, die alle Beteiligten zufriedenstellen. Und eine internationale Einigung mit Angehörigen verschiedenster Kulturkreise ist eine Herkulesaufgabe, die vermutlich mehrere Jahrzehnte Arbeit benötigen wird. Einigkeit in diesem zukunftsweisenden Gebiet ist definitiv notwendig, um die in den vorherigen Kapiteln

beschriebenen einhergehenden Herausforderungen zu bewältigen. Wenn man bedenkt, wie unterschiedlich die moralischen Vorstellungen in verschiedenen Erdteilen sind, ist es offensichtlich, dass diese Diskussion jetzt angeregt werden muss, um dem Problem vorzugreifen. Keinesfalls dürfen wir erst mit der Lösungssuche beginnen, wenn die ersten AGIs kurz vor ihrer Fertigstellung stehen.

Auch inwiefern AGIs künftig unser Leben prägen könnten, war ein wichtiges Thema in der Plenumsdiskussion. Viele Menschen finden die Vorstellung von einem Roboter gepflegt zu werden beklemmend. Die Meinung der Studierenden ging an dieser Stelle durchaus auseinander. Zum Beispiel wurde eingewendet, dass liebevolle Pflege von menschenähnlichen Robotern einem einsamen Altenheimbett vorzuziehen ist. Daraus entstand schnell die Frage, inwiefern Roboter überhaupt Gefühle, wie Liebe, empfinden können. Auf der einen Seite finden wir hier die Verfechter der These, dass das liebevolle Verhalten äquivalent zu dem liebevollen Gefühl selbst ist. Emotionen sind auch in der Psychologie ein komplexes Gebiet: Die Frage, wie Gefühle ausgelöst werden und inwieweit sie unser Verhalten beeinflussen, ist noch nicht abschließend behandelt. Wie dies nun auf Maschinen übertragen werden soll, ist mehr als unklar. Auf der anderen Seite befinden sich jene, die die Meinung vertreten, dass menschliche Zuneigung nicht ersetzt werden kann.

Eine weitere essentielle Frage drehte sich um die Aufgabe der Menschen in einer von einer Superintelligenz verwalteten Welt. Eine Abwandlung einer der grundlegendsten philosophischen Fragen kam dabei auf: Welchen Sinn hat unser Leben, wenn die Superintelligenz uns alles vorweg nimmt? Jede brillante oder kreative Idee, die uns kommt, kann von der ASI mit Leichtigkeit übertroffen werden. Wie also entwickelt sich eine Gesellschaft, in der der Einzelne das Gefühl hat nichts mehr beitragen zu können, und tatsächlich kaum noch einen Beitrag leisten kann? Andererseits ist eine Welt, in der unser größtes Problem Beschäftigungslosigkeit ist, immer noch besser als der momentane Zustand, in dem mehrere hundert Millionen Menschen Hunger leiden oder vor Krieg aus ihren Ländern flüchten.



Der Umgang mit künstlicher Intelligenz ist mit Sicherheit eine der großen Fragen des 21. Jahrhunderts und ihre Lösung eine der größten Herausforderungen der Menschheitsgeschichte. Leider hat dieses Thema in den Medien in den letzten Jahren nur eine untergeordnete Rolle gespielt. Einem Großteil der Bevölkerung ist die Tragweite des Problems noch nicht bewusst. In der nahen Zukunft wäre es daher wünschenswert, dass die Thematik mehr in die Öffentlichkeit getragen wird und so ein reger Diskurs entsteht.

## Literatur

- [1] Sabrina Holldorb. *Navigationssysteme*. Seminar: Ist künstliche Intelligenz gefährlich? 2017. URL: [https://hci.iwr.uni-heidelberg.de/teaching/seminar\\_KI\\_2017](https://hci.iwr.uni-heidelberg.de/teaching/seminar_KI_2017).
- [2] Enes Witwit. *Deep Learning Speech Recognition*. Seminar: Ist künstliche Intelligenz gefährlich? 2017. URL: [https://hci.iwr.uni-heidelberg.de/teaching/seminar\\_KI\\_2017](https://hci.iwr.uni-heidelberg.de/teaching/seminar_KI_2017).
- [3] Maximilian Müller-Eberstein. *Machine Translation: Tell me what you say and I'll tell you ce que tu as dit*. Seminar: Ist künstliche Intelligenz gefährlich? 2017. URL: [https://hci.iwr.uni-heidelberg.de/teaching/seminar\\_KI\\_2017](https://hci.iwr.uni-heidelberg.de/teaching/seminar_KI_2017).
- [4] techcrunch.com. *Meet Flippy, a burger-grilling robot from Miso Robotics and CaliBurger*. [Online; accessed 30-Aug-2017]. 2017. URL: <https://techcrunch.com/2017/03/07/meet-flippy-a-burger-grilling-robot-from-miso-robotics-and-caliburger/>.
- [5] businessinsider.com. *The one-armed bricklaying robot is cashed up and getting ready to build houses*. [Online; accessed 30-Aug-2017]. 2017. URL: <http://www.businessinsider.com/the-one-armed-bricklaying-robot-is-cashed-up-and-getting-ready-to-build-houses-2017-8?IR=T>.
- [6] Christin Lassow. *Steueralgorithmen bei autonomen Fahrzeugen*. Seminar: Ist künstliche Intelligenz gefährlich? 2017. URL: [https://hci.iwr.uni-heidelberg.de/teaching/seminar\\_KI\\_2017](https://hci.iwr.uni-heidelberg.de/teaching/seminar_KI_2017).
- [7] techcrunch.com. <https://techcrunch.com/2017/05/23/googles-alphago-ai-beats-the-worlds-best-human-go-player/>. [Online; accessed 30-Aug-2017]. 2017. URL: <http://www.businessinsider.com/the-one-armed-bricklaying-robot-is-cashed-up-and-getting-ready-to-build-houses-2017-8?IR=T>.
- [8] Wikipedia. *Human Brain Project*. [Online; accessed 30-Aug-2017]. 2017. URL: [https://en.wikipedia.org/wiki/Human\\_Brain\\_Project](https://en.wikipedia.org/wiki/Human_Brain_Project).

- [9] Wikipedia. *BRAIN Initiative*. [Online; accessed 30-Aug-2017]. 2017. URL: [https://en.wikipedia.org/wiki/BRAIN\\_Initiative](https://en.wikipedia.org/wiki/BRAIN_Initiative).
- [10] Wikipedia. *Human Genome Project*. [Online; accessed 30-Aug-2017]. 2017. URL: [https://en.wikipedia.org/wiki/Human\\_Genome\\_Project](https://en.wikipedia.org/wiki/Human_Genome_Project).
- [11] Bostrom, Nick. *Superintelligence: Paths, dangers, strategies*. OUP Oxford, 2014.
- [12] Tim Urban. *The AI Revolution: The Road to Superintelligence*. [Online; accessed 30-Aug-2017]. 2017. URL: <https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html>.
- [13] Ray Kurzweil. *The singularity is near: When humans transcend biology*. Penguin, 2005.
- [14] Vincent C. Müller und Nick Bostrom. „Future Progress in Artificial Intelligence: A Poll Among Experts“. In: *AI Matters* 1.1 (Sep. 2014), S. 9–11. ISSN: 2372-3483. DOI: 10.1145/2639475.2639478. URL: <http://doi.acm.org/10.1145/2639475.2639478>.
- [15] *Humans*. TV Series AMC Network. 2014 - present.
- [16] Wikipedia. *Robotergesetze*. [Online; accessed 30-Aug-2017]. 2017. URL: <https://de.wikipedia.org/wiki/Robotergesetze>.
- [17] Mathias Rein. *Boxed AI, kann man eine KI einsperren?* Seminar: Ist künstliche Intelligenz gefährlich? 2017. URL: [https://hci.iwr.uni-heidelberg.de/teaching/seminar\\_KI\\_2017](https://hci.iwr.uni-heidelberg.de/teaching/seminar_KI_2017).
- [18] Dominique Cheraf. *The Red Button, wird die KI sich wehren, wenn wir sie abschalten wollen?* Seminar: Ist künstliche Intelligenz gefährlich? 2017. URL: [https://hci.iwr.uni-heidelberg.de/teaching/seminar\\_KI\\_2017](https://hci.iwr.uni-heidelberg.de/teaching/seminar_KI_2017).
- [19] Vadim Tschernezki. *Friendly AI, ensuring ethical behaviour of intelligent agents*. Seminar: Ist künstliche Intelligenz gefährlich? 2017. URL: [https://hci.iwr.uni-heidelberg.de/teaching/seminar\\_KI\\_2017](https://hci.iwr.uni-heidelberg.de/teaching/seminar_KI_2017).

- [20] *Ermittler wollen Aufzeichnungen von Amazon Echo: Alexa als Zeugin einer Mordanklage?* [Online; accessed 30-Aug-2017]. 2016. URL: <https://www.heise.de/newsticker/meldung/Ermittler-wollen-Aufzeichnungen-von-Amazon-Echo-Alexa-als-Zeugin-einer-Mordanklage-3582492.html>.
- [21] <https://www.verbraucherzentrale.de/amazon-echo>. [Online; accessed 30-Aug-2017]. 2017. URL: <https://www.verbraucherzentrale.de/amazon-echo>.
- [22] *World's Biggest Data Breaches*. [Online; accessed 9-Sep-2017]. 2017. URL: <http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/>.
- [23] *Yahoo hack: 1bn accounts compromised by biggest data breach in history*. [Online; accessed 9-Sep-2017]. 2017. URL: <https://www.theguardian.com/technology/2016/dec/14/yahoo-hack-security-of-one-billion-accounts-breached>.
- [24] *Hacker-Jackpot: Credit Bureau Equifax gehackt*. [Online; accessed 9-Sep-2017]. 2017. URL: <https://www.heise.de/newsticker/meldung/Hacker-Jackpot-Credit-Bureau-Equifax-gehackt-3824607.html>.
- [25] *Polizei testet Gesichtserkennung an Berliner Bahnhof*. [Online; accessed 30-Aug-2017]. 2017. URL: <http://www.spiegel.de/netzwelt/netzpolitik/gesichtserkennung-test-am-berliner-suedkreuz-beginnt-am-1-august-a-1160079.html>.
- [26] Benjamin Franklin. *Those who would give up essential Liberty, to purchase a little temporary Safety, deserve neither Liberty nor Safety*. [Online; accessed 30-Aug-2017]. 2017. URL: [https://en.wikiquote.org/wiki/Benjamin\\_Franklin](https://en.wikiquote.org/wiki/Benjamin_Franklin).
- [27] *Autonomous Weapons: an Open Letter from AI & Robotics Researchers*. [Online; accessed 9-Sep-2017]. 2015. URL: <https://futureoflife.org/open-letter-autonomous-weapons/>.

- [28] *Wassili Archipow: der Mann, der die Welt rettete*. [Online; accessed 30-Aug-2017]. 2016. URL: <http://www.bundeswehr-journal.de/2016/wassili-archipow-der-mann-der-die-welt-rettete/>.
- [29] *Defense Advanced Research Project Agency*. 2017. URL: <https://www.darpa.mil/>.
- [30] *Engineering Humans for War*. [Online; accessed 30-Aug-2017]. 2015. URL: <https://www.theatlantic.com/international/archive/2015/09/military-technology-pentagon-robots/406786/>.
- [31] *UN-Verbot ignoriert - Nordkorea scheitert erneut mit Raketentest*. [Online; accessed 9-Sep-2017]. 2017. URL: <https://www.tagesschau.de/ausland/raketentest-nordkorea-105.html>.