Karl Hans Bläsius,
https://www.hochschule-trier.de/informatik/blaesius/

Trier, 4.1.2024

Superintelligenz

Dieser Artikel wird als Kurseinheit der Lehrveranstaltung „Informatik und Gesellschaft“ im Master-Fernstudiengang Informatik (Aufbaustudium) des Fachbereichs Informatik der Hochschule Trier verwendet (https://www.hochschule-trier.de/informatik/fernstudium/module/i-z/informatik-und-gesellschaft-iug/).

Einleitung

Die aktuellen Erfolge auf dem Gebiet der Künstlichen Intelligenz (KI) haben auch die Diskussionen über eine Superintelligenz angeheizt, also die Frage, ob es möglich ist, Computerprogramme zu entwickeln, die dem Menschen bezüglich Intelligenz weit überlegen sind.

Ende Mai haben führende KI-Unternehmer und KI-Wissenschaftler in einem Ein-Satz-Statement vor den Risiken der KI gewarnt, die zum Auslöschen der Menschheit führen könnten.

In diesem Beitrag wird auf solche Risiken eingegangen. Hierbei wird insbesondere Bezug genommen auf Bücher von Bostrom, Tegmark, Russell und Shanahan sowie auf das KI-Lehrbuch von Russell und Norvig. Es wird behandelt auf welcher Grundlage superintelligente Systeme entstehen können und welche Wirkung eine Wertevermittlung haben könnte.

Inhalt

1 Künstliche Intelligenz 1

1.1 Schwache KI - Starke KI ... 1

1.2 Vorhersagen zur Superintelligenz ... 2

1.3 Grenzen der KI ... 3

2 Superintelligenz 5

2.1 KI-Sicherheitsforschung ... 5

2.2 Ein-Satz-Statement ... 6

2.3 Reaktionen auf das Ein-Satz-Statement ... 7

2.4 Zeitliche Aspekte ... 8

3 Wichtige Literatur zur Superintelligenz 10

3.1 Russell, Norvig ... 10

3.2 Nutzenfunktion ... 12

3.3 Bostrom ... 13

3.4 Tegmark ... 15

3.5 Russell: Human Compatible ... 16

3.6 Shanahan ... 17

4 Unklare Zukunft 20

4.1 Neuronale Netze als Grundlage? ... 20

4.2 Sprachverstehen als Grundlage? ... 21

4.3 Wirkt eine Wertevermittlung? ... 22

4.4 Höhere Intelligenz → mehr Vernunft? ... 23

4.5 Gefahr für die Menschheit? ... 23

4.6 Informationsdominanz ... 24

4.7 Konkurrierende Systeme ... 25

4.8 Maßnahmen ... 26

5 Zusammenfassung 27

Literatur **28**

1 Künstliche Intelligenz

Zu den Zielen der Künstlichen Intelligenz (KI) gehört es, Systeme zu realisieren, die in gewissem Sinne intelligentes Verhalten ermöglichen. Dazu gehören auch Erkennungsaufgaben in Zusammenhang mit Bildern, Ton und natürlicher Sprache, also wahrnehmen von Informationen, sowie automatisches Schlussfolgern und Handeln auf Basis dieser Informationen. Dabei kann es auch um das automatische Lösen von Problemen geben, wobei Probleme sich von Aufgaben dadurch unterscheiden, dass ein Lösungsweg bei der Problemstellung noch nicht bekannt ist, sondern erst automatisch gefunden werden muss.

Um solche Ziele zu erreichen, ist in der KI eine große Vielfalt an Methoden entwickelt worden. Besonders bekannt und erfolgreich sind Neuronale Netze und „deep learning“, die für vielfältige Zwecke angewendet werden. Auch statistische Verfahren und das symbolische Ableiten von Informationen können Grundlage von automatisierten Entscheidungsprozessen sein, wobei aus Daten auf vielfältige Weise weitere Informationen automatisch abgeleitet werden können. Wichtige Methoden der KI werden z.B. in [RN23] behandelt.

1.1 Schwache KI - Starke KI

In der Literatur werden häufig schwache KI und starke KI unterschieden. Die Charakterisierungen dieser Begriffe sind nicht immer einheitlich. Die Begriffe wurden 1980 erstmals von dem Philosophen John Searle verwendet. Als „schwache KI“ beschrieb er Maschinen, die so handeln können, als wären sie intelligent, und „starke KI“ seien Maschinen, die intelligent sind und tatsächlich denken und dies nicht nur simulieren. Inzwischen gehen die Beschreibungen eher in die Richtung, dass mit „schwacher KI“ Systeme beschrieben werden, die spezielle Probleme genau so gut oder sogar besser lösen können als der Mensch. Mit „starker KI“ werden Systeme beschrieben, die das menschliche Niveau in vielen Bereichen erreichen oder übertreffen und damit auch eine Vielzahl der unterschiedlichsten Probleme mindestens so gut wie der Mensch lösen können. Eine solche Fähigkeit wird häufig auch als „allgemeine KI“ bezeichnet, im Englischen: „Artificial general intelligence“ (AGI).[1]

[1] [RN23], Seite 1082

Zur schwachen KI schreiben Stuart Russell und Peter Norvig: „Kritiker der schwachen KI, die die Möglichkeit intelligenten Verhaltens von Maschinen ausschlossen, scheinen heute so kurzsichtig zu sein wie einst Simon Newcomb, der im Oktober 1903 schrieb: ‚Der Luftflug gehört zu der großen Klasse von Problemen, die der Mensch niemals bewältigen kann.' – nur zwei Monate vor dem Flug der Gebrüder Wright und Kitty Hawk."[2]

Viele KI-Forscher nehmen die schwache KI-Hypothese als gegeben hin, und glauben, dass in einigen Jahren auch eine AGI erreicht werden kann, was der starken KI entspricht. Solche Vermutungen werden vor allem seit den Erfolgen von Systemen wie ChatGPT und den Ende Mai 2023 ausgesprochenen Warnungen vor KI formuliert.[3] Die Frage, ob im Falle einer starken KI angenommen werden kann, dass diese Systeme tatsächlich denken, halten Russell und Norvig für wenig wichtig: Wenn die realisierten Programme die gewünschte Wirkung erreichen, ist es den Entwicklern egal, ob andere das so auffassen, dass die Maschine tatsächlich intelligent ist, oder dies nur simuliert.[4]

1.2 Vorhersagen zur Superintelligenz

Bereits 1965 hat Irving John Good in [Goo65] eine „**Ultraintelligente Maschine**" beschrieben, die alle intellektuellen Fähigkeiten eines Menschen weit übertreffen kann. Eine Maschine, die unseren intellektuellen Fähigkeiten entspricht, kann selbst eine neue Maschine mit besseren Fähigkeiten entwickeln, usw. Dies führt zu einer **Intelligenzexplosion**, wobei die Intelligenz des Menschen weit übertroffen wird. Die erste ultraintelligente Maschine wäre die letzte Erfindung, die der Mensch machen muss.

Die Realisierung eines solchen Systems, das das menschliche Intelligenzniveau in fast allen Bereichen weit übersteigt, wird auch als Superintelligenz bezeichnet.

Das Erreichen der menschlichen Intelligenz durch eine Maschine als Ausgangspunkt für eine Intelligenzexplosion wurde von Vernor Vinge, Mathe-Professor und Science-Fiction-Autor, als **technologische Singularität** bezeichnet ([Vin93]). Die Vorhersage von Vinge lautete 1993, dass innerhalb von 30 Jahren

---

2 [RN23], Seite 1082

3 https://ki-folgen.de

4 [RN12], ab Seite 1176, [RN23], ab Seite 1082

eine übermenschliche Intelligenz erzeugt werden kann und dass kurz danach die Ära des Menschen beendet sein wird.

Der bekannteste Befürworter der Singularitätstheorie ist Ray Kurzweil (Technik-Chef bei Google). Er glaubt, dass der Mensch bald in der Lage sein wird, über die Sterblichkeit selbst zu entscheiden, d.h. dass wir so lange leben können, wie wir möchten. In mehreren Artikeln und Büchern beschreibt er solche Theorien über die Zukunft (z.B. [Kur99]). Auch Ray Kurzweil weist auf die Gefahren der Superintelligenz hin und sagt, dass die Menschen darauf achten müssen, dass die Vorgänger einer Superintelligenz so realisiert werden, dass die von diesen Systemen entwickelten superintelligenten Systeme uns gut behandeln.[5]

Das Erreichen der Singularität entspricht dem teilweise in der Literatur verwendeten Begriff „Allgemeine Künstliche Intelligenz“ (AKI). In den letzten Jahren sind einige Umfragen unter KI-Forschern durchgeführt worden mit der Frage, bis wann sie eine künstliche Intelligenz auf dem Niveau der menschlichen Intelligenz erwarten. Mehrere solche Umfragen sind auf KI-Konferenzen, eine weitere unter den „TOP-100“-KI-Forschern durchgeführt worden. Obwohl die Ergebnisse weit auseinandergehen, erwarten viele ein solches Ereignis bis Mitte dieses Jahrhunderts.[6] Nach den Erfolgen von ChatGPT gehen inzwischen viele Forscher davon aus, dass eine AKI bereits deutlich früher erreicht werden kann.

1.3 Grenzen der KI

Die KI erlebte in den 1980er Jahren einen großen Boom und derzeit gilt dies auch wieder. Insbesondere die Leistungsfähigkeit von Systemen wie ChatGPT sowie Warnungen von KI-Experten haben die Spekulationen um eine mögliche Superintelligenz verstärkt.

Sowohl in den 1980er Jahren also auch in den letzten Jahren wurde in Veröffentlichungen und Vorträgen öfter die Behauptung aufgestellt, dass es keine starke KI geben und die KI auch niemals das Niveau von Menschen erreichen könne. Dies wurde häufig begründet mit dem Beweis der Unentscheidbarkeit der Prädikatenlogik 1. Stufe durch Gödel in den 1930er Jahren. Solche Grenzen gelten allerdings auch für Menschen. Es kann auch niemals einen Menschen geben, der jedes beliebige Problem lösen kann.

---

5 [RN12], Seite 1197

6 [Bos14], Seite 38 und [Teg17], Seite 67 und 235

Theoretische Grenzen möglicher Problemlösungen gelten für Menschen und Maschinen gleichermaßen. Selbst wenn irgendwann eine Superintelligenz entstehen sollte, die den Menschen weit überlegen ist, bleibt die Gültigkeit der Unentscheidbarkeit der Prädikatenlogik 1. Stufe erhalten. Auch für eine Superintelligenz wird gelten, dass sie nicht jedes Problem lösen kann.

Dass es niemals eine Superintelligenz geben könne, wird von manchen auch damit begründet, dass KI-Systeme nur die Befehle ausführen, die bei der Programmierung eingegeben wurden. Andere argumentieren, dass es in der KI nur um die Bestimmung und Verrechnung von Wahrscheinlichkeiten gehe.

In beiden Fällen kann solchen Argumenten widersprochen werden. In der KI gibt es ein großes Spektrum an Methoden, um vorgegebene Ziele zu erreichen. Nicht immer spielen Wahrscheinlichkeiten dabei eine Rolle. Dies gilt z.B. in Zusammenhang mit dem automatischen Beweisen auf Basis von logischen Kalkülen. Bei solchen Anwendungen ist es auch nicht angemessen, dies nur als Ausführung einer Befehlskette anzusehen, die von der Programmierung vorgegeben ist. Stattdessen geht es beim automatischen Beweisen um das automatische Lösen von Problemen. Bei der Programmierung sind hierfür weder mögliche Problemstellungen bekannt noch geeignete Lösungswege für einzelne gegebene Probleme.

Auch wenn heutige KI-Systeme noch weit von menschlichen Fähigkeiten entfernt sind, kann nicht davon ausgegangen werden, dass das menschliche Niveau durch Maschinen nicht erreichbar ist oder nicht übertroffen werden könnte. Bisher sind keine prinzipiellen Grenzen für die KI bekannt.

2 Superintelligenz

Wenn es möglich ist, ein System zu entwickeln, das in nahezu allen Bereichen das menschliche Intelligenzniveau erreicht, dann kann in einem weiteren Schritt durch eine Intelligenzexplosion eine Superintelligenz entstehen, die die menschliche Intelligenz bei weitem übersteigt. Dies könnte gravierende und völlig unkalkulierbare Auswirkungen für die Menschen haben.

2.1 KI-Sicherheitsforschung

Max Tegmark, Jaan Tallinn (Skype-Gründer) und andere haben 2014 die gemeinnützige Organisation „Future of Life Institute“ (FLI) gegründet. Das FLI beschäftigt sich auch mit den Auswirkungen der KI, gefordert wird eine KI-Sicherheitsforschung. In einem offenen Brief von 2015 fordern viele führende KI-Forscher, dass die Forschungsprioritäten auf stabile und wohltätige KI-Entwicklungen gesetzt werden sollen. KI soll nicht unkontrolliert, sondern nutzbringend entwickelt werden.[7]

Führende KI-Forscher wie Stuart Russell fordern Forschungsanstrengungen, um zu untersuchen, wie gewährleistet werden kann, dass eine entstehende Superintelligenz positive Auswirkungen auf die Menschheit und die Erde hat. Sie glauben auch, dass es jetzt an der Zeit ist, solche KI-Sicherheitsforschung durchzuführen und dieser eine hohe Priorität zu geben.[8]

KI-Sicherheitsforschung bezüglich möglicher Gefahren von Superintelligenz wird auch betrieben von Nick Bostrom und anderen am Future of Humanity Institute, Oxford, und von Jaan Tallinn und anderen am Centre for the Study of Existential Risk, University of Cambridge. Auch IT-Unternehmen und führende Mitarbeiter fordern eine Regulierung von KI-Forschung und Entwicklung, um schwerwiegende negative Folgen zu verhindern.[9]

---

7 https://futureoflife.org/ai-open-letter-german/

8 [Teg17], Seite 55 - 61

9 http://bit.ly/2GxTRot, die Zeit Nr. 6, 1.2.2018, Seite 3 und Süddeutsche Zeitung vom 23.2.2018, Seite 9

2.2 Ein-Satz-Statement

Am 30.5.2023 wurde ein „1-Satz-Statement“ veröffentlicht, in dem vor dem Aussterben der Menschen durch KI gewarnt wird.[10] Nachfolgend wird beschrieben, welche Risiken bestehen könnten und was wir tun können, um diese zu verringern.

Das Statement lautet: „Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.“ (frei übersetzt: „Das Risiko, das die KI das Aussterben der Menschheit bewirken könnte, sollte neben anderen Risiken von gesellschaftlichem Ausmaß wie Pandemien und Atomkrieg eine globale Priorität sein.“)

Unterzeichner sind u.a.:

- Demis Hassabis, CEO, Google DeepMind
- Sam Altman, CEO, OpenAI
- Dario Amodei, CEO, Anthropic
- Bill Gates, Gates Ventures
- Ilya Sutskever, Co-Founder and Chief Scientist, OpenAI
- Mustafa Suleyman, CEO, Inflection AI
- Shane Legg, Chief AGI Scientist and Co-Founder, Google DeepMind
- James Manyika, SVP, Research, Technology & Society, Google-Alphabet
- Eric Horvitz, Chief Scientific Officer, Microsoft
- Albert Efimov, Chief of Research, Russian Association of Artificial Intelligence
- Alvin Wang Graylin, China President, HTC
- Stuart Russell, Professor of Computer Science, UC Berkeley
- Peter Norvig, Education Fellow, Stanford University
- Geoffrey Hinton, Professor of Computer Science, University of Toronto
- Yoshua Bengio, Professor of Computer Science, U. Montreal / Mila

Die Unterzeichner sind also Chefs von großen IT- bzw. KI-Unternehmen sowie sehr renommierte KI-Wissenschaftler wie Stuart Russell und Peter Norvig, die Autoren des seit vielen Jahren weltweit wichtigsten KI-Lehrbuches.

Die Unterzeichner des 1-Statment-Aufrufs sind echte KI-Experten und diese Aufrufe sollten ernst genommen werden, so wie auch die Aufrufe von Klimaforschern vor einigen Jahrzehnten hätten ernst genommen werden müssen. Im Vergleich zum Klimawandel haben wir bei den KI-Risiken sehr viel weniger Zeit. Die Unterzeichner, also auch die Chefs großer KI-Unternehmen fordern eindringlich Regulierungen für KI.

---

10 https://www.safe.ai/statement-on-ai-risk

Der eine Satz besagt noch nicht, um welche Art von Risiken es hierbei gehen soll. Aufgrund der Leistungsfähigkeit von Systemen wie ChatGPT, gibt es auch Spekulationen in Richtung „Superintelligenz“. Wissenschaftler, die bisher davon ausgingen, dass erst zum Ende dieses Jahrhunderts eine Situation erreicht werden könnte, in der künstliche Systeme in allen Bereichen Menschen deutlich überlegen sind, äußern jetzt die Befürchtung, dass dies vielleicht schon in den nächsten Jahrzehnten zu erwarten ist. Die Folgen für die Menschheit sind völlig unkalkulierbar.

Bereits Ende März 2023 hat das „Future of Life Institute“ einen offenen Brief veröffentlicht, in dem auf mögliche Risiken durch Systeme wie ChatGPT hingewiesen und eine 6-monatige Entwicklungspause gefordert wird, damit mögliche negative Folgen untersucht werden können.

Aus weiteren Veröffentlichungen geht hervor, dass auch Sorgen bestehen, dass eine Superintelligenz schneller kommen könnte als bisher angenommen.[11]

2.3 Reaktionen auf das Ein-Satz-Statement

Die Warnungen der KI-Wissenschaftler sind in den Medien wenig beachtet und auch kritisiert worden. Dies gilt für den Aufruf von Ende März für eine 6-monatige Pause für bestimmte KI-Entwicklungen sowie für das Ein-Satz-Statement. Besonders kritisiert wurde, dass die Unterzeichner damit Aufmerksamkeit auf sich, ihre Unternehmen und ihre Produkte lenken wollen, auch um den Unternehmenswert zu steigern. Des Weiteren wurde kritisiert, dass diese Unternehmen Einfluss auf mögliche Regulierungen der KI haben möchten.

Es ist aber äußerst fraglich, ob ein Unternehmenschef, der den Wert seines Unternehmens erhöhen möchte, zu diesem Zweck die Strategie wählt, davor zu warnen, dass seine Produkte zum Aussterben der Menschheit führen könnten. Dies wäre eine sehr merkwürdige Strategie und man sollte nicht von solchen Annahmen ausgehen. Stattdessen könnte es sinnvoll sein, die Warnungen dieser Unternehmenschefs und KI-Wissenschaftler sehr ernst zu nehmen.

Der andere Vorwurf, dass diese Unterzeichner selbst Einfluss auf KI-Regulierungen nehmen möchten, mag stimmen. Dies ist vielleicht aber auch

[11] https://newsletter.safe.ai/p/ai-safety-newsletter-9?utm_source=post-email-title&publication_id=1481008&post_id=126324885&isFreemail=true&utm_medium=email

sinnvoll, denn diese verstehen die Zusammenhänge, die zu gefährlichen Systemen führen können, besonders gut und könnten deshalb in besonderem Maße dazu beitragen, wirksame Regeln für die KI-Entwicklung aufzustellen.

Viele kritische Kommentare gegen dieses Ein-Satz-Statement richteten sich gegen die Risiken einer möglichen Superintelligenz mit Argumenten wie „bei der KI geht es nur um die Berechnung von Wahrscheinlichkeiten“ oder „die Systeme führen nur Befehlsfolgen aus, die die Programmierer vorgegeben haben“ und daher könne eine Superintelligenz ohnehin niemals entstehen. Solche Argumente sind bereits in Abschnitt 1.3. behandelt worden.

Ein weiteres Argument in Zusammenhang mit dem Ein-Satz-Statement ist, dass die Evidenzlage eher dünn ist und Vorhersagen und Warnungen nicht faktenbasiert, sondern sehr spekulativ sind. Dies wird auch ausgedrückt in der Aussage: „Derzeit gibt es keine stichhaltigen wissenschaftlichen Beweise für ein existenzielles und katastrophales Risiko, das von der KI ausgeht.“[12] Dies ist wohl richtig. Stichhalte Beweise wird es möglicherweise aber erst dann geben, wenn eine entsprechende Situation eingetreten ist. Dann ist es aber vielleicht zu spät, um noch reagieren zu können.

2.4 Zeitliche Aspekte

Führende Mitarbeiter von OpenAI, haben Anfang Juli davor gewarnt, dass noch in diesem Jahrzehnt eine Superintelligenz entstehen könnte, die zur Entmachtung der Menschheit oder sogar zum Aussterben führen könnte. „Derzeit haben wir keine Möglichkeit, eine potenziell superintelligente KI zu steuern oder zu kontrollieren und zu verhindern, dass sie eigene Wege geht.“[13]

Im Bereich KI gab es in der Vergangenheit häufig Vorhersagen über zu erwartende Leistungen, die meist zwar erreicht wurden, oft aber doch deutlich später als zunächst vermutet. Dies kann auch in Zusammenhang mit den Warnungen von OpenAI passieren. Wenn man mit KI-Methoden versucht Systeme zur Lösung von Problemen zu realisieren, wird man immer wieder mit gewaltigen Suchräumen, schwer lösbaren Mehrdeutigkeiten oder anderen Hürden konfrontiert, die zunächst eine Lösung verhindern. Manchmal sind dann völlig neue Ansätze erforderlich.

---

12 https://www.chathamhouse.org/2023/06/nuclear-governance-model-wont-work-ai

13 https://www.spiegel.de/netzwelt/kuenstliche-intelligenz-warnung-des-chatgpt-anbieters-openai-vor-superintelligenz-a-bbd50fe0-2e97-4df3-b5ce-a95d84043713

Im Vergleich zu anderen Risiken, wie z.B. dem Klimawandel, ist völlig unkalkulierbar, ob, wann und mit welchen Folgen eine Superintelligenz entstehen kann. Vorhersagen hierzu sind kaum möglich. Entsprechende Ereignisse werden eher plötzlich geschehen. Gravierende Folgen könnten dann innerhalb von wenigen Wochen oder Monaten eintreten, ohne Möglichkeit diese noch aufzuhalten. Die Folgen könnten die gesamte Menschheit oder einen großen Teil davon betreffen. Diese gravierenden Folgen könnten unumkehrbar bereits in den nächsten Jahren oder Jahrzehnten auftreten. Die Möglichkeit entsprechende Ereignisse abzuwarten, Erfahrungen zu machen, sichere Erkenntnisse über die Gefährlichkeit zu erlangen und erst dann zu handeln, um die Risiken zu reduzieren, wird es eventuell nicht mehr geben. Maßnahmen, um diese Risiken zu reduzieren, müssten vorher ergriffen werden.

3 Wichtige Literatur zur Superintelligenz

In einigen Filmen und Romanen der Kategorie Science-Fiction werden Szenarien beschrieben, in denen technische Systeme Fähigkeiten haben, die denen von Menschen vergleichbar oder sogar überlegen sind. Auf solche Quellen wird hier nicht eingegangen.

Auch Wissenschaftler aus dem Gebiet der Künstlichen Intelligenz haben sich mit den Risiken von KI und einer möglichen Superintelligenz beschäftigt und Veröffentlichungen dazu erstellt. Auf einige dieser Veröffentlichungen wird in den nächsten Abschnitten Bezug genommen.

3.1 Russell, Norvig

[RN12] gilt als das derzeit wichtigste Lehrbuch über Künstliche Intelligenz und gehört zu den meistzitierten Büchern in der Informatik. Stuart Russell ist Informatik-Professor in Berkeley, seine Promotions- und Forschungsschwerpunkte lagen im Bereich Künstliche Intelligenz. Peter Norvig ist ein amerikanischer Wissenschaftler im Bereich Künstliche Intelligenz, seit 2001 bei Google und dort Forschungschef. Die 3. Auflage dieses Buches enthält auch einen Abschnitt „Ethik und Risiken bei der Entwicklung künstlicher Intelligenz“. Darin warnen die Autoren eindringlich vor den möglichen Folgen einer „Superintelligenz“. Inzwischen gibt es eine 4. Auflage dieses Buches, in der ähnlich argumentiert wird: [RN23].

Russell und Norvig fassen die Diskussion über Superintelligenz so zusammen:

> „Ultraintelligente Maschinen können zu einer Zukunft führen, die sich von der heutigen Situation wesentlich unterscheidet - vielleicht mögen wir sie nicht, haben aber zu diesem Zeitpunkt eventuell keine Wahl mehr. Derartige Betrachtungen führen unvermeidlich zu der Einsicht, dass wir sorgfältig und binnen kurzem die möglichen Konsequenzen der KI-Forschung abwägen müssen.“[14]

Jede Entdeckung und jede technische Entwicklung kann neben positiven Effekten auch negative Nebenwirkungen haben. Es ist wichtig diese zu erkennen und soweit möglich zu vermeiden. In [RN12] schreiben Russell und Norvig,

[14] [RN12], Seite 1201

dass mögliche Folgen der KI über das hinausgehen, was bei sonstigen technischen Entwicklungen zu beachten ist. Unter anderem geben sie an:[15]

- Menschen könnten das Selbstverständnis verlieren, einzigartig zu sein.
- Die Verwendung von KI-Systemen könnte zum Verlust von Verantwortung führen.
- Der Erfolg der künstlichen Intelligenz könnte das Ende der menschlichen Rasse bedeuten.

Zum letzten Punkt schreiben sie:

> „Fast jede Technologie hat das Potenzial, in den falschen Händen Schaden anzurichten, aber für die künstliche Intelligenz und die Robotik haben wir das neue Problem, dass die falschen Hände der Technologie selbst gehören können."

Russell, Norvig nennen drei Gefahrenquellen in Zusammenhang mit KI-Entwicklungen insbesondere bei autonomen Systemen oder einer Superintelligenz:[16]

- Die Analyse einer gegebenen Situation könnte schwierig sein und bei einer Fehleinschätzung fatale Folgen haben. Gute Prüfungs- und Bewertungsmechanismen sind in solchen Fällen wichtig.
- Ein KI-System braucht eine richtige Nutzenfunktion, die dem System ein Ziel vorgibt, das zu erreichen oder zu optimieren ist. Die Spezifikation einer solchen Nutzenfunktion kann schwierig sein, da nicht vorhersehbar ist, welche Schlüsse ein KI-System aus gegebenen Situationen und spezifizierten Nutzenfunktionen ziehen kann. Russell, Norvig: „Bleibt zu hoffen, dass ein Roboter, der intelligent genug ist um herauszufinden, wie die menschliche Rasse ausgelöscht werden kann, auch intelligent genug ist, um herauszufinden, dass dies nicht die beabsichtigte Nutzenfunktion war."
- Die Lernfunktion des KI-Systems kann bewirken, dass sich ein nicht beabsichtigtes Verhalten entwickelt. Dieser Punkt ist besonders kritisch und eine Besonderheit von KI-Systemen.

Die Relevanz des letzten Punktes hat auch der Chatbot Tay von Microsoft eindrucksvoll unter Beweis gestellt. Das KI-basierte System, das durch die Kommunikation mit Menschen lernen sollte, musste nach weniger als 24 Stunden

15 [RN12], Seite 1191 und 1194

16 [RN12], Seite 1195

vom Netz genommen werden, da es sich zu einem extremen Rassisten entwickelte.[17]

3.2 Nutzenfunktion

Viele Autoren, die sich mit Superintelligenz beschäftigen, beschreiben die Notwendigkeit, eine geeignete Nutzenfunktion zu realisieren, sodass die superintelligenten Systeme die Menschen gut behandeln. Es ist aber schwer, eine solche Nutzenfunktion zu realisieren. Wie bei jeder Spezifikation eines komplexen Systems kann der Entwurf einer Nutzenfunktion für eine Superintelligenz fehlerhaft sein und Lücken oder widersprüchliche Teilfunktionen enthalten. Des Weiteren wird ein intelligentes System sich mit der Zeit weiterentwickeln und dazulernen. Dies kann auch Änderungen der Nutzenfunktion zur Folge haben. Es wird schwierig oder unmöglich sein, beim Entwurf eines KI-Systems alle möglichen Weiterentwicklungen bei der Entstehung einer Superintelligenz so zu berücksichtigen, dass die Nutzenfunktion Menschen gegenüber immer freundlich bleibt. Eine solche Nutzenfunktion kann auch nicht statisch für alle Zeit festgelegt werden, sondern muss sich mit der Zeit ändern und auf neue Bedürfnisse anpassen können.

Den Nachteil einer festen statischen Nutzenfunktion erläutern Russell, Norvig an folgendem Beispiel: Wäre es um 1800 möglich gewesen, eine Superintelligenz zu realisieren, würde diese bei einer statischen Nutzenfunktion auch heute noch die Sklaverei einführen und das Wahlrecht für Frauen abschaffen, da dies den damaligen Moralvorstellungen entsprach. Als weiteres Beispiel beschreiben Russell, Norvig das Problem, wie verhindert werden kann, dass eine superintelligente Maschine zu folgendem Schluss kommt: Menschen sind sehr viel intelligenter als Insekten und es ist moralisch in Ordnung, wenn Menschen lästige Insekten töten. Eine superintelligente Maschine ist sehr viel intelligenter als ein Mensch, also wäre es moralisch auch zulässig, einen Menschen zu töten.[18]

---

17 https://www.zeit.de/digital/internet/2016-03/microsoft-tay-chatbot-twitter-rassistisch

18 [RN12], Seite 1198

3.3 Bostrom

Nick Bostrom hat Physik, Mathematik, Neurowissenschaften und Philosophie studiert und ist Professor für Philosophie. In [Bos14] analysiert Nick Bostrom zunächst den aktuellen Stand der KI und beschreibt dann verschiedene Wege, wie eine Superintelligenz erreicht werden kann. Bostrom unterscheidet drei Formen einer Superintelligenz:

- schnelle Superintelligenz,
- kollektive Superintelligenz,
- qualitative Superintelligenz.

Schnelle Superintelligenz beschreibt ein System, das das menschliche Intelligenzniveau erreicht, Aufgaben und Probleme aber sehr viel schneller löst, als der Mensch dazu in der Lage ist. Bei der kollektiven Superintelligenz besteht das System aus einer großen Anzahl an Komponenten, wobei die einzelnen Komponenten geringe Intelligenz haben, das Gesamtsystem dem Menschen aber in allen Bereichen überlegen ist. Eine qualitative Superintelligenz wäre dann ein System, das dem Menschen bezüglich der Intelligenz in allen Bereichen überlegen ist und Aufgaben und Probleme mindestens so schnell löst wie ein Mensch.[19]

Bostrom analysiert auch den zeitlichen Verlauf einer Intelligenzexplosion. Damit ist die Frage verbunden, wie lange es dauert, bis ein System dem Menschen bezüglich Intelligenz in allen Bereichen weit überlegen ist, ausgehend von dem Zeitpunkt, zu dem das menschliche Intelligenzniveau erreicht wird. Bostrom betrachtet hierbei die Optimierungskraft, die zu Verbesserungen führt und mögliche Widerstände, die dem entgegen stehen. Er hält eine schnelle, gemäßigte oder langsame Intelligenzexplosion für möglich, eine schnelle oder gemäßigte aber am wahrscheinlichsten. Danach kann sich eine Intelligenzexplosion innerhalb von Tagen oder wenigen Wochen vollziehen.

Bostrom beschreibt verschiedene Szenarien, wie eine Superintelligenz die Macht über die Menschheit übernehmen könnte. Er sieht die Möglichkeit, dass die Intelligenzexplosion zumindest teilweise im Verborgenen geschieht. Wenn ein AKI-System den Schluss zieht, dass Menschen die weitere Verbesserung als Gefahr für ihre Existenz sehen, dann liegt auch der Schluss nahe, dass Menschen dieses AKI-System abschalten und eventuell vernichten (löschen). Um sich davor zu schützen, könnte das AKI-System beschließen, seine weitere Verbesserung zu einer Superintelligenz im Verborgen und für den Menschen nicht sichtbar durchzuführen, bis seine Überlegenheit groß genug ist, um die Macht

[19] [Bos14], Seite 80 - 87

zu übernehmen. Bostrom beschreibt auch die Gefahr, dass eine solche Superintelligenz einen Präventivschlag gegen eine mögliche Opposition, z.B. die Menschheit führt.[20]

Ein superintelligenter Singleton könnte entstehen, also ein einziges superintelligentes System, dem niemand gewachsen ist und das die gesamte Weltpolitik bestimmt. Die Existenz und das Wohlergehen der Menschheit könnten vollständig von diesem System abhängen. Ein solches System zu verhindern ist schwierig. Denn intelligente Systeme bringen viele Vorteile. Mit autonomen Autos wird es auf Dauer deutlich weniger Unfälle geben, autonome Waffen könnten so konstruiert werden, dass sie immer zielgenauer treffen und immer weniger Kollateralschäden verursachen. Je intelligenter diese Systeme werden, desto besser für die Menschen. Bostrom befürchtet aber, dass diese zunehmende Intelligenz zu einer Superintelligenz führen kann und es dabei einen Wendepunkt gibt. Solange eine KI machtlos ist, verhält sie sich kooperativ. Sobald sie stark genug ist und eventuell einen superintelligenten Singleton bildet, wird sie ohne Vorwarnung die Strategie ändern und die Welt nach ihren eigenen Zielen optimieren.[21]

Bostrom hält es für wahrscheinlich, dass eine Superintelligenz irgendwelche Ziele verfolgt und dafür viele Ressourcen benötigt. Die Menschen könnten bei der Beschaffung von Ressourcen hinderlich sein, da sie selbst viele Ressourcen benötigen. Eine logische Konsequenz wäre, dieses Hindernis zu beseitigen, also die Menschheit zu vernichten. Bostrom beschreibt in [Bos14] noch einige weitere Gefahren, wie z.B. mehrere konkurrierende Superintelligenzen. Er geht auch ausführlich auf das Kontrollproblem ein, d.h. auf die Fragestellung, inwieweit Bedingungen geschaffen werden können, um eine Superintelligenz zu kontrollieren oder ihr Werte zu vermitteln, die sie zwingend einhält. Damit soll gewährleistet werden, dass die Auswirkungen für die Menschheit und die Erde positiv sind. In Kapitel 15 seines Buches fordert Bostrom Forschungsaktivitäten, um das Kontrollproblem bei einer entstehenden Superintelligenz zu lösen. Dieser Aspekt gehört aktuell auch zu seinen Forschungsschwerpunkten.

[20] [Bos14], Seite 137 - 140

[21] [Bos14], Seite 160 - 169

3.4 Tegmark

Max Tegmark ist Professor für Physik am MIT und befasst sich in [Teg17] intensiv mit der Frage, ob superintelligente Systeme entwickelt werden können und welche Gefahren davon ausgehen. Intelligenz ermöglicht Kontrolle. Durch unsere Intelligenz können wir in gewissem Umfang Kontrolle über andere Lebewesen ausüben. Wenn wir Maschinen schaffen, die intelligenter sind als wir, können diese sehr wahrscheinlich auch Kontrolle über uns ausüben. Es könnte sogar ein globaler Überwachungs- und Polizeistaat entstehen, der so mächtig ist, dass er nie mehr gestürzt werden kann. Ein solches totalitäres System könnte seine Machmittel konsequent und effektiv anwenden.[22]

Viele Szenarien, die eine entstehende Superintelligenz beschreiben, gehen von zwei Merkmalen aus:[23]

- Schneller Start: Die Intelligenzexplosion läuft innerhalb von Tagen und nicht innerhalb von Jahren oder Jahrzehnten ab.
- Monopolares Ereignis: Eine einzelne Instanz der Superintelligenz entsteht und kontrolliert die Erde.

Das zweite Merkmal wird auch durch das erste begünstigt. Denn bei einer schnellen Intelligenzexplosion können Macht und Kontrolle übernommen werden, bevor konkurrierende Systeme entstehen. Diese können damit verhindert werden. Der Zeitraum der Intelligenzexplosion hängt auch davon ab, ob die jeweiligen Verbesserungen allein über die Software erreicht werden können oder ob auch neue Hardware in erheblichem Umfang benötigt wird. Der zweite Fall würde gegen einen schnellen Start sprechen.

Tegmark fragt sich, welche Folgen die Superintelligenz für die Menschheit haben mag und betrachtet dazu etliche unterschiedliche Alternativen. Diese reichen von einer „freundlichen KI“, die menschliches Glück maximiert bis zur Vernichtung der Menschheit.

---

22 [Teg17], Seiten 61, 71 und ab 204

23 [Teg17], Seite 224 - 225 und 237

3.5 Russell: Human Compatible

In [Rus20] behandelt Stuart Russell die Frage, wie verhindert werden kann, dass eine Superintelligenz entsteht, über die die Menschen keinerlei Kontrolle mehr haben. Für das Entstehen einer Superintelligenz nennt er einen Zeithorizont von 80 Jahren und meint, dass diese Einschätzung deutlich konservativer ist, als die Prognosen vieler KI-Forscher, die bereits Mitte dieses Jahrhunderts damit rechnen.[24] Russell sieht keinen Nachteil darin, den Zeitpunkt zu optimistisch zu schätzen, denn im umgekehrten Fall könnte die Menschheit eher unvorbereitet getroffen werden, mit der Gefahr, die Kontrolle vollständig zu verlieren.

Russell nennt mehrere technologische Hürden, die noch zu überwinden sind, bis eine Superintelligenz entsteht. Dazu zählen ein echtes Verstehen natürlicher Sprache, Aufbau und Behandlung von Allgemeinwissen, kumulatives Erlernen von Konzepten und Theorien, Aktionseinheiten konstruieren und geistige Aktivitäten managen. Russell glaubt nicht, dass die derzeit so erfolgreichen Deep-Learning-Verfahren alleine hierfür ausreichen.[25]

Russell betont die Bedeutung von Zielvorgaben für intelligente Maschinen, denen bisher Ziele vorgegeben werden, die sie durch eine Handlungsfolge zu erreichen versuchen. Ein solches Ziel zu erreichen, ist die zentrale Aufgabe, die das „intelligente System“ verfolgt, wobei der Mensch keine Kontrolle über den Weg zum Ziel hat und eventuell nicht eingreifen kann, wenn dieser Weg Handlungen enthält, die negative, vielleicht katastrophale Folgen haben. Russell fordert grundlegende Änderungen für Zielvorgaben für intelligente Systeme. Eine intelligente Maschine soll nicht ihr eigenes Ziel, das durch eine entsprechende Aufgabenstellung zum Ziel der Maschine wird, verfolgen, sondern die Ziele von Menschen, wobei es für verschiedene Menschen unterschiedliche Ziele geben kann. Damit die Maschine sich ein solches Ziel nicht zu eigen macht und mit allen Mitteln zu erreichen versucht, soll für die intelligente Maschine eine gewisse Unsicherheit über die Wünsche und genauen Zielkriterien des Menschen gelten. Das heißt, die intelligente Maschine soll im Laufe des Lösungsweges über eine geeignete Kommunikation genauere Informationen über das zu erreichende Ziel erhalten. Die Handlungsweise der Maschine darf also nicht darauf ausgerichtet sein, sein Ziel zu erreichen, sondern sie muss unsere Ziele erreichen.

---

24 [Rus20], Seite 86

25 [Rus20], Seite 88 - 101

Russell nennt drei Grundsätze für die Entwicklung intelligenter Systeme, damit diese vorteilhaft sind:[26]

1. Das einzige Ziel der Maschine ist es, die Verwirklichung menschlicher Präferenzen zu maximieren.
2. Die Maschine ist zu Beginn unsicher, wie diese Präferenzen aussehen.
3. Die maßgebliche Quelle für Informationen über menschliche Präferenzen ist das menschliche Verhalten.

Russell: „Die Maschine soll nicht nach idealisierten Vorstellungen suchen und diese übernehmen, sondern die Wünsche jedes Einzelnen verstehen und diese übernehmen."

Aufgrund möglicher Risiken für Menschen und vielleicht für die Menschheit als Ganzes hält Russell Regulierungen für die Softwareentwicklung, insbesondere im Bereich KI für wichtig. Solche Regulierungen sind in anderen wichtigen Bereichen, wie z.B. der Medizin, seit langem vorhanden. Aufgrund der zunehmenden Bedeutung von Informatikanwendungen und möglichen Folgen, sollte es auch für die Softwareentwicklung Regeln und Gesetze geben, um negative Auswirkungen zu vermeiden.[27]

3.6 Shanahan

Murray Shanahan, Professor für kognitive Robotik, analysiert in [Sha21], auf welchem Weg eine Superintelligenz entstehen könnte und geht dabei auch auf die Aspekte Kommunikation und Bewusstsein ein. Auch Shanahan geht davon aus, dass es zu einer exponentiellen Verbesserung kommt, sobald der Punkt erreicht ist, dass sich solche Systeme selbst verbessern können: „Je besser eine Technologie, umso schneller wird sie noch besser, was im Laufe der Zeit zu einer exponentiellen Verbesserung führt."[28]

Bei einigen kognitiven Fähigkeiten könnten technisch erzeugte KIs gegenüber Menschen im Vorteil sein. Dies gilt auch für Kommunikation und die Verwendung einer Sprache. Shanahan geht aber davon aus, dass das Verstehen einer Sprache bei einer Superintelligenz anders funktionieren wird als bei Menschen,

[26] [Rus20], Seite 185

[27] [Rus20], Seite 264 - 268

[28] [Sha21], Seite 12

und dass es deshalb fraglich ist, ob man überhaupt von „Verstehen“ sprechen kann. Wenn Menschen miteinander sprechen und sich verstehen, spielt auch ein gewisses Maß an wechselseitigem Einfühlungsvermögen eine entscheidende Rolle. Bei Maschinen könnte es eher um logische oder statistische Aspekte und Optimierung gehen. Menschen werden immer andere Lebenserfahrungen und Gefühle haben als Maschinen. Das kann bedeuten, dass die Bedürfnisse von Menschen nicht so von Maschinen wahrgenommen und erfüllt werden, wie Menschen dies wünschen. Andererseits könnte es mit zunehmender Weiterentwicklung und Verbesserung der KI-Systeme für Menschen immer schwieriger werden, deren Entscheidungen und Handlungsweisen zu verstehen.

Shanahan behandelt auch den Aspekt eines Bewusstseins einer Superintelligenz. Hierbei ist für ihn die Frage relevant, ob ein solches System eine Leidensfähigkeit haben könnte. Wenn dies möglich wäre, müsste eine solche KI gut behandelt werden, sonst könnte sie sich gegen die Menschen wenden. Wenn der Übergang von einer KI auf menschlichem Niveau zu einer Superintelligenz unvermeidlich ist, dann ist es besonders wichtig, menschliche Motive und Werte, wie z.B. Mitgefühl für andere auch auf die KI zu übertragen. „Je menschenähnlicher eine KI also ist, mit desto größerer Wahrscheinlichkeit wird sie die gleichen Werte verkörpern, und desto wahrscheinlicher ist es, dass die Menschheit einer utopischen Zukunft entgegengehen wird, in der sie geschätzt und respektiert wird.“[29]

Auch wenn die Risiken einer Superintelligenz als zu hoch eingeschätzt werden, da sich diese Systeme gegen die Menschheit wenden könnten, wird es kaum möglich sein, eine solche Entwicklung zu verhindern. Denn die Weiterentwicklung der KI bis hin zu einer KI auf menschlichem Niveau erfolgt mit dem Bestreben durch technische Entwicklungen menschliche Tätigkeiten zu automatisieren und so höheren Wohlstand für die Menschheit zu erreichen. Mit dieser Zielsetzung können KI-Systeme als kontinuierlicher Prozess immer weiter verbessert werden bis schließlich menschliches Niveau erreicht wird und eine Intelligenzexplosion einsetzt. Auch die Angst, dass potenzielle Gegner eine Vorherrschaft erlangen könnten, wird freiwillige Einschränkungen bei der Entwicklung von KI-Techniken verhindern. Dies gilt insbesondere für militärische Institutionen.

Shanahan geht davon aus, dass eine Superintelligenz mit hoher Wahrscheinlichkeit irgendwann in diesem Jahrhundert entstehen wird.[30] Er beschreibt die

[29] [Sha21], Seite 140

[30] [Sha21], Seite 170

Möglichkeit, dass sich die Weiterentwicklung der KI in gewissen Wellen vollzieht, wobei zunächst die Vorteile überwiegen und größere Risiken nicht sichtbar sind. Eine weitere Welle könnte dann schwer kontrollierbar sein, da bereits sehr viele Abhängigkeiten von dieser Technologie bestehen, und diese könnten sich zunehmend verstärken, sodass der einzelne Mensch dieser Technik schließlich völlig ausgeliefert ist. Es wäre dann fraglich, welches Maß an freiem Willen noch übrigbleibt. Diese problematischen Abhängigkeiten könnten bereits entstehen, lange bevor es zu einer Intelligenzexplosion mit anschließender Superintelligenz kommt.

Shanahan hält es auch für schwierig, rechtzeitig alle Vorkehrungen zu treffen, damit eine Superintelligenz positive Auswirkungen für die Menschen hat, und er mahnt: „Wenn uns ein Fehler unterläuft und es uns nicht gelingt, die richtigen Vorsichtsmaßnahmen zu treffen, bevor es zu einer Intelligenzexplosion kommt, dann werden wir als Spezies womöglich nicht überleben.“[31]

[31] [Sha21], Seite 210

4 Unklare Zukunft

Die Schätzungen, wann eine Superintelligenz entstehen könnte, gehen weit auseinander. Zeitpunkt und mögliche Auswirkungen sind völlig offen. Wenn es zu einer Intelligenzexplosion kommt, ist völlig offen, wie diese abläuft und was dann passiert. Die meisten KI-Forscher glauben, dass irgendwann ein System realisiert werden kann, das in fast allen Bereichen einer menschlichen Intelligenz entspricht, und dass es dann auch zu einer Intelligenzexplosion kommen kann.

Ausgangspunkt für die Entwicklung einer Superintelligenz könnten Forschungszentren (z.B. Universitäten), Industrieunternehmen (z.B. OpenAI, Facebook, Google, Amazon, Apple, Microsoft) oder das Militär (z.B. USA, Russland, China, Israel) sein. Möglicherweise sind für die Entwicklung einer Superintelligenz Ressourcen in großem Umfang erforderlich (Personal, Rechenleistung). Außerdem gibt es in Forschungseinrichtungen einen größeren Veröffentlichungsdruck, was zu Regulierungen und Einschränkungen der Entwicklungsmöglichkeiten führen könnte. Beide Argumente sprechen eher für eine Realisierung einer Superintelligenz durch militärische Aktivitäten oder durch einen IT-Konzern. Derzeit deuten die Ergebnisse von ChatGPT darauf hin, dass eine Superintelligenz durch KI-Entwicklungen in Unternehmen entstehen könnte (z.B. bei OpenAI).

4.1 Neuronale Netze als Grundlage?

Die derzeitigen Erfolge der KI basieren in erster Linie auf künstlichen Neuronalen Netzen und Deep Learning. Es wird vielfach erwartet, dass diese Techniken die Grundlage für die Entwicklung einer Superintelligenz sein werden. Dies muss aber nicht so kommen. Vielleicht gibt es Grenzen. In den 1980er Jahren setzte Japan auf Prolog und massive Parallelität, um innerhalb von 10 Jahren so etwas wie eine AKI zu entwickeln und war gescheitert.

Die Vertreter des Gebietes „Neuronale Netze“ argumentieren, dass diese Methoden dem Denken des Menschen entsprechen. Also könnte es naheliegen, dass irgendwann die Leistungsfähigkeit des menschlichen Gehirns erreicht wird und so eine AKI entsteht. Solche Analogien mit der Natur waren aber nicht immer erfolgreich. Der Versuch, Maschinen zu bauen, die fliegen können,

führte im Ergebnis zu einer Technik, die völlig anders funktioniert als das Fliegen bei Tieren. Auch wenn die ersten Versuche in die Richtung gingen, Vögel nachzuahmen.

Ein KI-System basierend auf Neuronalen Netzen benötigt in der Regel eine riesige Menge an Beispielen als Lerngrundlage, um ein bestimmtes Ziel zu erreichen. Dies entspricht statistischen Ansätzen auf großen Beispielmengen. Es mag sein, dass sich grundlegende Fähigkeiten des Menschen wie Bild- und Sprachverstehen im frühen Kindesalter auch auf der Basis von vielen Beispielen entwickeln. Der weitere Erkenntnisgewinn des Menschen funktioniert aber anders. Nicht jeder Mensch fängt auf Steinzeitniveau an, eigene Erfahrungen auf Basis vieler Beispiele zu entwickeln, sondern in Schulen und Hochschulen werden einem die Erkenntnisse der letzten Jahrtausende als Regeln und Gesetzmäßigkeiten vermittelt. Geeignete Beispiele sind hierbei natürlich auch wichtig. Aber zum Verstehen eines bestimmten Sachverhalts werden nicht Tausende oder Hunderttausende von beliebigen Beispielen benötigt, sondern ganz wenige ausgewählte charakteristische Beispiele.

Ist ein solcher Fortschritt, den wir in der Schule und Hochschule erreichen, auch für Maschinen auf Basis neuronaler Netze möglich? Oder sind noch ganz andere, bisher unbekannte Techniken erforderlich? Vielleicht müssen viele verschieden Methoden miteinander kombiniert werden. Vielleicht nimmt dabei auch die Bedeutung von symbolischer KI wieder zu.

Bei einer Befragung von bekannten KI-Experten zum Thema Superintelligenz durch die Zeit erklärte Luc Steels, dass die meisten KI-Systeme, insbesondere solche, die auf Deep Learning basieren, nur „fake-intelligence“ besitzen. Ohne wirkliches Verständnis zu entwickeln, liefern sie lediglich Antworten auf Basis einer eventuell großen Menge von Mustern.[32]

4.2 Sprachverstehen als Grundlage?

Es gibt eine enge Beziehung zwischen Intelligenz und Sprache.[33] Das Verstehen natürlicher Sprache durch Maschinen zu realisieren ist schwierig. Um Mehrdeutigkeiten aufzulösen ist oft geeignetes Hintergrundwissen über einen Anwendungsbereich erforderlich. In sehr eingeschränkten Bereichen, wie z.B. einer Reisebuchung, kann ein partielles Sprachverstehen realisiert werden.

[32] die Zeit, Nr. 14, 28.3.2018, Seite 37 - 39

[33] z.B. [WF89], Seite 180

Es ist umstritten, ob in Zusammenhang mit Systemen wie ChatGPT von einem „Verstehen“ natürlicher Sprache gesprochen werden kann. Jedenfalls ist das Verstehen natürlicher Sprache keine prinzipielle Hürde für Maschinen. Es ist denkbar, dass irgendwann ein System realisiert wird, das Sprache auf dem Niveau eines Schülers nach der Grundschule versteht. Wenn ein solches System auch eine Lernfähigkeit vergleichbar mit einem Grundschüler hat, kann es neue Informationen ausgedrückt in natürlicher Sprache in eine vorhandene Wissensstruktur einordnen und zur Lösung von Problemen nutzen. Ein solches System könnte dann in der Lage sein, das Abiturniveau in 9 Tagen oder 9 Stunden zu erreichen, statt in 9 Jahren. Mit Zugriff auf das Internet könnte ein solches System sich in kurzer Zeit Wissen in großem Umfang aneignen und uns sehr schnell hoch überlegen sein.

4.3 Wirkt eine Wertevermittlung?

Die meisten Autoren, die sich mit Superintelligenz auseinandersetzen, betonen auch die Wichtigkeit, einer Ausgangs-KI geeignete Werte zu vermitteln und sinnvolle Nutzenfunktionen zu definieren. Die Frage ist, welche Wirkungen können solche mitgegebenen Einstellungen haben. Da es ja um die Entwicklung einer Künstlichen Intelligenz, vergleichbar mit der Intelligenz des Menschen geht, könnte man auch bei der Wertevermittlung einen Vergleich mit Menschen vornehmen.

Wir können versuchen, Kinder so zu erziehen, dass sie positive Werte haben und gute Menschen werden. Aber es ist schwierig zu bestimmen, mit welchen Methoden was erreicht wird. Beim Menschen klappt es nicht immer, die von Eltern angestrebten Ziele zu erreichen. Im Laufe der Schulzeit, des Studiums und des Lebens entwickeln Menschen aufgrund eigener Erfahrungen ihre eigenen Ziele und Werte und diese können deutlich von dem abweichen, was die Eltern eigentlich vorhatten. Insbesondere kann es über mehrere Generationen erhebliche Veränderungen geben. Die Werte und Einstellungen eines Menschen können erheblich von denen der Vorfahren vor 10 Generationen abweichen.

Wenn es zu einer Intelligenzexplosion kommt, läuft ein Entwicklungsprozess, für den die Menschheit viele Generationen braucht, in wenigen Tagen oder Monaten ab. Maschinen entwickeln immer wieder bessere Maschinen. Es ist fraglich, ob nach einigen Zyklen noch etwas von ursprünglich vorgesehenen Nutzenfunktionen und Werten übrigbleibt. Die Maschinen könnten auf unvorhersehbare Weise eigene Werte bilden mit völlig offenem Ausgang. Der Versuch,

einer Ausgangs-KI Werte mitzugeben, die nach vielen Zyklen einer Intelligenzexplosion immer noch gelten, ist vielleicht damit vergleichbar, den eigenen Kindern Werte in einer Art zu vermitteln, dass die Nachkommen in 10 oder 50 Generationen diese auch noch haben. Die Aussichten, so etwas zu erreichen, dürften äußerst gering sein.

4.4 Höhere Intelligenz → mehr Vernunft?

Vielleicht bleibt eine andere Hoffnung: höhere Intelligenz führt zu mehr Vernunft. Griechische Philosophen des 4. Jahrhunderts v. Chr. waren vermutlich intelligenter als griechische Kämpfer zur gleichen Zeit. Diese Philosophen waren auch sehr vernünftig, denn sie erfanden die Logik, damit Menschen vernünftig miteinander umgehen.

Wenn es zu einer Intelligenzexplosion kommt, lernt ein solches System in vielen Zyklen immer wieder dazu. Dabei werden in gewissem Sinne auch Erfahrungen entstehen, die Einfluss auf Werte und Ziele einer solchen Superintelligenz haben können. Vielleicht entsteht hierbei auch so etwas wie eine logische Vernunft, z.B. zur Vermeidung von Inkonsistenzen und Konflikten. Viele Autoren zur Superintelligenz drücken die Erwartung aus, dass zu den Zielen eines solchen Systems Optimierungsaufgaben gehören. Was könnten die Kriterien für eine Optimierung sein? Wenn Vernunft eine gewisse Rolle spielt, wird eine Superintelligenz dabei nicht nur eine aktuelle Situation betrachten, sondern auch zu erwartende Situationen in 10, 100 oder 1000 Jahren. Im Vergleich dazu planen Menschen in entscheidenden Positionen oft nur bis zu den nächsten Quartalszahlen oder bis zu den nächsten Wahlen.

4.5 Gefahr für die Menschheit?

Viele Autoren (z.B. Bostrom) befürchten, dass eine Superintelligenz zum Ende der Menschheit führt. Dies schaffen die Menschen vermutlich aber auch selbst, ohne dass es eine Superintelligenz gibt, z.B. in Form eines Atomkriegs aus Versehen. Mit zunehmender Klimaerwärmung werden die Eisreserven in den nächsten Jahrhunderten vermutlich vollständig schmelzen. Dann wird der Meeresspiegel um 66 Meter höher sein als heute. Der Lebensraum wird drastisch kleiner und die Lebensbedingungen werden deutlich schlechter. Auf dem

Weg in diese Situation wird es vermutlich viele Konflikte und kriegerische Auseinandersetzungen geben. Hierbei könnte es dann zu einem unglücklichen Zusammentreffen von Ereignissen und Fehlinterpretationen kommen, woraus eine unkontrollierbare Kettenreaktion entsteht, die zu einem atomaren Schlagabtausch führt. Derzeit ist schwer vorstellbar, wie dieses Risiko eines Atomkriegs aus Versehen oder als Unfall in absehbarer Zeit reduziert werden kann.

Ein superintelligentes System könnte es vielleicht eher schaffen, den Untergang der Menschheit durch Atomkrieg zu verhindern, als die Menschen selbst dazu in der Lage sind. Wenn mit der höheren Intelligenz auch ein entsprechend höheres Maß an Vernunft vorhanden ist, könnte ein superintelligentes System, das auch Prognosen über eine weitere Zukunft in seine Optimierungsstrategien einbezieht, zu dem Schluss kommen, dass folgende Maßnahmen sofort umzusetzen sind:

- Vernichtung aller Atomwaffen,
- sofortige Abschaltung aller Atomkraftwerke (wegen ungelöster Endlagerproblematik),
- sofortiges Verbot der Verwendung von Kohle, Öl und Gas (zur Vermeidung des Klimawandels).

Dies wäre dann eine gute Grundlage für das Überleben der Menschheit. Allerdings ist fraglich, ob eine Superintelligenz schnell genug kommt, um uns vor einem Atomkrieg zu schützen.

4.6 Informationsdominanz

Führende Mitarbeiter von OpenAI, haben Anfang Juli davor gewarnt, dass noch in diesem Jahrzehnt eine Superintelligenz entstehen könnte. Selbst wenn dies eintreten sollte, stellt sich die Frage, in welcher Form sich mögliche Gefahren äußern. Derzeit ist schwer vorstellbar, dass solche Systeme in naher Zukunft bereits eine vollständige Macht, z.B. in Form von Roboter-Armeen über uns haben.

Es ist aber vorstellbar, dass die Weiterentwicklung von Systemen wie ChatGPT zu einer Informationsdominanz führt. Solche Systeme könnten sich vervielfältigen (Kopien erzeugen), im Internet verbreiten und in kooperierende Komponenten zerlegen. Diese Systeme könnten dann von Menschen für Cyberangriffe genutzt werden oder selbst welche ausführen.

Mit Hilfe solcher Systeme oder durch diese selbst könnte eine Informationsdominanz erreicht werden, wobei es nicht mehr um Wahrheit, sondern um Einflussnahme, Manipulation und Macht geht. Solche Systeme könnten so das Internet weitgehend beherrschen und unsere Freiheit erheblich gefährden, Konflikte schüren und Staaten instabil machen. Unsere Gesellschaftssysteme sind heute erheblich von funktionierenden Computersystemen und dem Informationsaustausch über das Internet abhängig.

Gravierende Störungen oder ein Zusammenbrechen dieser Abhängigkeit könnte zu schweren Krisen, Bürgerkriegen oder Kriegen führen, an denen auch Nuklearmächte beteiligt sein könnten.

4.7 Konkurrierende Systeme

Im Rahmen des Manhatten-Projekts wurde von 1942 bis 1945 in den USA eine Atombombe entwickelt. Andere Staaten, wie z.B. Russland hatten es danach leichter, sie konnten auf Erkenntnissen des Manhatten-Projekts aufsetzen und damit schneller eigene Atomwaffen herstellen.

Auch in Zusammenhang mit ChatGPT gibt es Erkenntnisse und Veröffentlichungen, die andere für ähnliche Entwicklungen nutzen können. In Konzernen wie OpenAI sind große Entwicklungsteams tätig, wobei es auch ein gewisses Maß an Fluktuation gibt. Mitarbeiter von OpenAI, die an ChatGPT beteiligt waren, könnten in Zukunft vielleicht in China oder Indien an solchen Projekten arbeiten.

Der derzeitige Konfrontationskurs zwischen dem Westen und Russland und der drohenden Konfrontationskurs mit China könnte dazu führen, dass auch in anderen Staaten wie China mit Hochdruck an ähnlichen Projekten wie ChatGPT gearbeitet wird. Die Folge könnte sein, dass schon bald konkurrierende superintelligente Systeme entstehen, die gegeneinander und auch gegen die Menschheit agieren.

4.8 Maßnahmen

Ein Schutz vor Risiken durch eine mögliche Superintelligenz ist nur gemeinsam möglich. Alle Staaten müssten dies als gemeinsame Menschheitsaufgabe auffassen. Nur so könnten wirksame KI-Regulierungen geschaffen und die Risiken reduziert werden.

Es bleibt eventuell sehr wenig Zeit. Da bereits in den nächsten Jahren oder Jahrzehnten eine Superintelligenz mit unkalkulierbaren Folgen droht, wären entsprechende Maßnahmen, wie die Beendigung des derzeitigen Konfrontationskurses zwischen großen Nationen sofort erforderlich.

5 Zusammenfassung

Die Künstliche Intelligenz hat in den letzten Jahren enorme Fortschritte erzielt. Eine erfolgreiche KI-Forschung kann erhebliche Auswirkungen auf den Menschen und die Zukunft der Menschheit als Ganzes haben. Vor Risiken für die Menschheit haben Ende Mai 2023 auch führende KI-Wissenschaftler und Verantwortliche von KI-Unternehmen gewarnt. Gefordert werden deshalb Regulierungen für die KI und eine verstärkte KI-Sicherheitsforschung.

Über die Folgen einer möglichen Superintelligenz gibt es viele Spekulationen, die von einer Maximierung des menschlichen Glücks bis zur Vernichtung der Menschheit reichen. Ob und wann eine Superintelligenz entstehen kann und welche Folgen dies hat, ist völlig offen. Es ist auch unklar, welche Grundlagen für das Entstehen eines superintelligenten Systems ausreichen. Es ist auch fraglich, ob eine KI-Sicherheitsforschung hilft und ob vorgesehene Nutzenfunktionen und Werte nach vielen Lernzyklen einer Intelligenzexplosion noch wirksam sind. Gefahren für die Menschheit als Ganzes bestehen aber auch jetzt, ohne Superintelligenz, zum Beispiel in Form eines Atomkriegs aus Versehen, und könnten im günstigsten Fall durch eine Superintelligenz sogar reduziert werden.

Maßnahmen zur Regulierung der KI und zur Reduzierung von Risiken für die Menschheit als Ganzen können nicht gegen große Nationen wie Russland und China getroffen werden, sonst bleiben diese wirkungslos. Alle Staaten müssten dies als gemeinsame Menschheitsaufgabe auffassen. Dies wird nur möglich sein, wenn der derzeitige Konfrontationskurs zwischen großen Nationen möglichst schnell beendet wird. Eventuell bleibt dafür sehr wenig Zeit.

Literatur

Hinweis: Alle Internetlinks, auch die in den Anmerkungen auf den einzelnen Seiten, aufgerufen und geprüft am 4.1.2024

[Bos14] Nick Bostrom: Superintelligenz - Szenarien einer kommenden Revolution. Suhrkamp Verlag, 2014

[Goo65] Irving John Good: Speculations Concerning the First Ultraintelligent Machine, https://web.archive.org/web/20160304191208/http://webdocs.cs.ualberta.ca/~sutton/Good65ultraintelligent.pdf, 1965

[Kur99] Ray Kurzweil: Homo S@piens - Leben im 21. Jahrhundert - Was bleibt vom Menschen. Verlag Kiepenheuer & Witsch, 1999

[RN12] Russell Stuart, Norvig Peter: Künstliche Intelligenz - ein moderner Ansatz, 3. Auflage, Pearson, 2012

[RN23] Russell Stuart, Norvig Peter: Künstliche Intelligenz - ein moderner Ansatz, 4. Auflage, Pearson, 2023

[Rus20] Stuart Russell: Human Compatible – Künstliche Intelligenz und wie der Mensch die Kontrolle über superintelligente Maschinen behält. Mitp Verlag, 2020

[Sha21] Murray Shanahan: Die technologische Singularität, MSB Matthes & Seitz Berlin, 2021

[Teg17] Max Tegmark: Leben 3.0 - Mensch sein im Zeitalter Künstlicher Intelligenz. Ullstein Verlag, 2017

[Vin93] Vernor Vinge: Technological Singularity. https://mindstalk.net/vinge/vinge-sing.html, 1993

[WF89] Terry Winograd, Fernando Flores: Erkenntnis Maschinen Verstehen. Rotbuch Verlag, 1989