Universität Heidelberg

Seminar 2017: Ist künstliche Intelligenz gefährlich

AI Box: Kann man eine künstliche Intelligenz einsperren?

Matthias Rein

Dozent: Ullrich Köthe

Inhaltsverzeichnis

Kapitel 1

Einführung

Die Forschung im Bereich der künstlichen Intelligenz insbesondere seit den letzten Jahren boomt und hat beachtenswerte Fortschritte gemacht. Obgleich die heutigen Systeme nichtsdestotrotz noch weit davon entfernt sind, einen Intelligenzgrad zu erreichen, der dem der Menschen entspricht (oder diesen sogar übersteigt), wird genau das von einigen Forschern mittelfristig prognostiziert. Das wirft die Frage auf, welche Risiken und Gefahren derartig intelligente Systeme mit sich bringen und wie man eine solche künstliche Intelligenz kontrollieren kann. Das AI Box Problem geht dazu der Frage nach, ob und wie man eine künstliche Intelligenz einsperren kann.

1.1 Motivation

Als eine Superintelligenz wird eine allgemeine Intelligenz, die sehr viel höher ist als die der Menschen, bezeichnet. Das heißt also eine Intelligenz, die den intelligentesten Menschen *bei weitem* und *in quasi allen Bereichen* (wie zum Beispiel Planungsfähigkeit, soziale Intelligenz, wissenschaftliche Fähigkeiten, etc.) übertrifft. (Im Gegensatz zu KIs, die Menschen nur in einem oder wenigen Bereichen - bei Strategiespielen wie Schach beispielsweise - stark überlegen ist)
Ob es prinzipiell möglich ist eine Superintelligenz zu erschaffen und in welchem Zeitraum damit zu rechnen ist, ist unter Experten umstritten. Während einige prinzipiell bezweifeln, dass die Schaffung einer Superintelligenz (mittelfristig) überhaupt möglich sei, halten andere das Erreichen jener in den nächsten Jahrzenten, zumindest aber bis Ende dieses Jahrhunderts für relativ wahrscheinlich. Betont werden sollte an dieser Stelle das potentiell überraschende, absolut plötzliche Auftauchen einer Superintelligenz, welches insbesondere beim Szenario einer Intelligenzexplosion erwartet werden kann. Diese Überlegung geht von einer sogenannten Saat-KI als Augangspunkt aus, das heißt von einer künstlichen Intelligenz, die fähig ist, sich selbst zu verbessern (in dem sie ihren eigenen Quellcode optimiert oder ändert) oder - allgemeiner - eine neue noch intelligentere künstliche Intelligenz zu schaffen. Diese resultierende KI wäre aufgrund ihrer verbesserten Fähigkeiten dazu in der Lage, sich weitere Optimierungen zu “erdenken” und eine noch bessere KI zu schaffen, welche dann wiederum eine verbesserte KI schaffen könnte, usw. Durch diesen Mechanismus könnte eine künstliche Intelligenz ihren Intelligenzgrad quasi automatisch und vor Allem ohne menschliches Zutun explosionsartig steigern, im Zweifelsfall sogar ohne, dass dies zunächst überhaupt bemerkt wird.
Darausfolgend stellt sich die Frage, auf welche Weise die Existenz einer Superintelligenz die Welt im Allgemeinen beeinflussen würde und insbesondere welches Gefahrenpotential diese hätte. In jüngster Zeit gibt es vermehrt Warnungen von mitunter prominenten öffentlichen Personen wie Bill Gates, Stephen Hawkings oder Elon Musk vor den Gefahren, die eine hochentwickelte künstliche Intelligenz potentiell mit sich bringt. Der Unternehmer und Visionär Elon Musk hält eine unkron-

tollierte KI für potentiell gefährlicher als Atomwaffen [1], während Stephen Hawkings sich folgendermaßen geäußert hat:

> "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks." [2]

Allein durch ihre den Menschen weit überlegende Intelligenz hätte eine Superintelligenz einen strategischen Vorteil, wodurch sie im Zweifelsfall verheerenden Schaden anrichten könnte. Nick Bostrom unterscheidet in seinem Buch "Superintelligenz" [3] zwischen dem Fall einer perversen Instantiierung einerseits und Ressourcenvergeudung andererseits.
Bei ersterem findet die Superintelligenz einen unvorhergesehenen, nicht erwünschten Weg ein ihr gegebenes Ziel zu erreichen. Man betrachte die folgenden Beispiele (die aus dem Buch entnommen sind):

Endziel: *"Bring uns zum Lächeln!"*
Perverse Instantiierung: *Lähmung der menschlichen Gesichtsmuskeln, um den Mund zu einem permanenten strahlenden Lächeln zu verziehen.*

Auch, wenn derartige unorthodoxen Umsetzungen bereits antizipiert werden und die Zielsetzung dementsprechend angepasst wird, kann es wiederum andere ungewollte Umsetzungen geben, die nicht vorhergesehen wurden:

Endziel: *"Bring uns zum Lächeln, ohne direkt auf unsere Gesichtsmuskeln einzuwirken"*
Perverse Instanntiierung: *Stimulation des Teils des motorischen Kortex, der unsere Gesichtsmuskulatur steuert*

Der zweite Fall, das Problem der Ressourcenvergeudung, kann entstehen, wenn eine Superintelligenz versucht ihren Auftrag um jeden Preis und ohne Rücksicht auf Kosten und "Nebenwirkungen" zu erfüllen. Eine superintelligentes System mit der Aufgabe die Riemann'sche Vermutung zu überprüfen könnte versuchen, alle Ressourcen bishin zu den einzelnen Atomen neuanzuordnen, um ein einziges "Computronium" zur Lösung der gegebenen Aufgabe zu schaffen.
Auch für diesen Fall könnten derartige Konsequenzen antizipiert und bei der Zielgebung berücksichtigt werden. Sowohl beim Problem der perversen Instantiierung wie auch bei dem der Ressourcenvergeudung ist dies allerdings unter Umständen alles andere als trivial. Dazu kommt, dass man sich absolut keine Fehler leisten kann, da ein einziger Fehler bereits zu fatalen Konsequenzen führen kann.
Ein weiteres wichtiges Thema, das hier nicht weitergehend behandelt werden kann, aber an dieser Stelle erwähnt werden sollte, sind die potentiell gravierenden gesellschaftlichen Folgen der Existenz einer Superintelligenz. Man bedenke insbesondere die Macht bzw. das Machtmonopol, das der Besitzer einer (oder gar der einzigen/ersten) Superintelligenz durch diese erhalten könnte.
Trotz der Risiken dürfte eine Regulierung der Forschung im Bereich der künstlichen Intelligenz kaum wünschenswert sein. Ohnehin wäre diese wohl unmöglich durchzusetzen. Aus diesen Überlegungen folgt, dass, falls man dem Aufkommen einer Superintelligenz in absehbarer Zukunft eine gewisse Wahrscheinlichkeit zugesteht, es dringend erforderlich ist, wirksame Kontrollmöglichkeiten zu erforschen. Das gilt selbst dann, wenn diese Wahrscheinlichkeit als relativ gering eingeschätzt wird, weil einiges dafür spricht, dass eine unkontrollierte Superintelligenz katastrophale Folgen bis hin zur Auslöschung der Menschheit mit sich bringen könnte. In Anbetracht einer möglichen Inteligenzexplosion und der dadurch plötzlichen, rapiden Entstehung einer Superintelligenz, müssen Kontrollen zudem unbedingt schon frühzeitig entwickelt werden und bereits bereitstehen, auch wenn das Erreichen einer Superintelligenz noch relativ weit entfernt *scheint*.

1.2 Kontrollmöglichkeiten

Es gibt im Wesentlichen zwei Anätze zur Kontrolle einer Superintelligenz. Das sind zum einen Fähigkeitskontrolle, zum anderen Motivationskontrolle. Bei ersterer geht es darum, die Handlungsmöglichkeiten der KI einzuschränken, eben ihre Fähigkeiten zu kontrollieren, während zweitere zum Ziel hat, den "Willen" der KI zu beinflussen, also eine gutartige KI (friendly AI) zu schaffen, die von Natur aus der Menschheit nicht schaden bzw. Schaden von der Menschheit abwenden "will". Diese beide Arten von Kontrollen sind keineswegs trivial. Bei der Motivationskontrolle dürfte sich die Formulierung eines wirksamen moralischen Kodexes beispielsweise als äußerst kompliziert erweisen (wie bereits die oben dargestellten Probleme allein bei der Zielgebung verdeutlichen sollten). Darüber hinaus werden dabei allgemeine schwierige ethische Fragen aufgeworfen, die zu diskutieren wären (man vergleiche die ethischen Fragen, die bereits heute durch die voranschreitende Entwicklung und den Einsatz autonomer Fahrzeuge aufgeworfen werden).
Bei der Fähigkeitenkontrolle stellt sich die Frage wie Menschen eine KI kontrollieren können (oder gar, ob das überhaupt möglich ist), deren Intelligenz soviel höher ist, dass sie aus menschlicher Sicht kaum überhaupt erahnt werden kann. Man muss nicht nur davon ausgehen, dass jene KI alle Kontrollmaßnahmen sowie menschliches Handeln allgemein antizipieren kann, sondern auch, dass sie Erkenntnisse (physikalischer Art beispielsweise) besitzt und fähig ist, sich Wege und Lösungen "auszudenken", die das menschliche Gehirn aufgrund seiner vergleichsweise überaus bescheidenen Fähigkeiten niemals auch nur ansatzweise verstehen oder nachvollziehen könnte.
Davon abgesehen kann eine in ihren Fähigkeiten eingeschränkte KI nicht ihr volles Potential entfalten. Abhängig davon wie stark die KI kontrolliert werden soll, verringert sich deren Nutzen unter Umständen erheblich. Es muss also immer eine Abwägung getroffen werden zwischen dem Nutzen der KI auf der einen Seite sowie der Sicherheit auf der anderen.
Es besteht unter den Forschern generell keine Einigkeit, ob Fähigkeitenkontrolle oder Motivationskontrolle vielversprechendere Kontrollmöglichkeiten bietet. Prinzipiell können Fähigkeitenkontrolle und Motivationskontrolle kombiniert werden, um eine höhere Sicherheit herzustellen, sodass, falls sich einer der beiden Ansätze als unwirksam herausstellen sollte, immer noch Kontrolle gewährleistet ist.
Im Folgenden wird es hauptsächlich um Fähigkeitenkontrolle gehen. In Kapitel zwei werden der Ansatz einer Orakel KI und damit einhergehend verschiedene Maßnahmen der Fähigkeitenkontrolle vorgestellt. In Kapitel drei wird auf Grundlage des AI Box Experiments von Yudkowski der menschliche Faktor als potentielle Schwachstelle der Fähigkeitenkontrolle diskutiert.

Kapitel 2

Orakel KI

Eine Orakel KI ist eine bestimmte Art von künstlicher Intelligenz, die lediglich Fragen beantworten kann. Sie hat somit qua Entwurf ein relativ beschränktes Handlungsfeld und eignet sich daher besonders gut zur Fähigkeitenkontrolle (im Gegensatz zu intelligenten autonomen Agenten etwa, deren Konzept der Fähigkeitenkontrolle prinzipiell widerspricht und die somit im Wesentlichen nur der Motivationskontrolle unterworfen werden können). In ihrem Paper "Thinking Inside the Box: Controlling and Using an Oracle AI" [4] untersuchen Stuart Armstrong, Anders Sandberg and Nick Bostrom die Kontrollmöglichekeiten einer Orakel KI.
Zu den möglichen physischen Kontrollmaßnahmen zählen beispielsweise:

- *Faradayischer Käfig*: Die Orakel KI wird in einem faradayischen Käfig aufbewahrt, sodass es ihr physikalisch unmöglich sein sollte, über Funksignale nach außen zu kommunizieren
- *Ausgabe drosseln*: Die Ausgabe könnte reduziert werden, um der Manipulation von Menschen durch die Orakel KI durch längere Nachrichten, etc. vorzubeugen (siehe auch Kapitel drei). Im extremesten Fall könnte die Ausgabe auf lediglich drei Antwortmöglichktein - ja, nein und unbestimmbar - reduziert werden.
- *Zurücksetzung der Orakel KI nach jedem Durchlauf* : So soll verhindert werden, dass das Orakel langfristige Pläne umsetzt. Außerdem kennt es in diesem Fall immer nur die aktuell gestellte, aber nicht die vorherigen Fragen, wodurch es weniger Kenntnisse über seine Umwelt gewinnen kann.
- *Kurze Laufzeit*: Das Orakel wird nur für eine kurze Zeit laufen gelassen, die idealerweise gerade dazu ausreicht eine Frage zu beantworten, aber kurz genug ist, dass das Orakel keinen oder wenig Schaden anrichten oder "Ausbrechungspläne schmieden" kann
- *Rechenleistung drosseln*: Die Hardware auf der die Orakel KI läuft kann runtergetaktet werden, um ihre potentiell zu schädlichen oder ungewollten Aktivitäten eingesetzten Kapazitäten zu verringern

Um die Orakel KI auf Verlässlichkeit zu prüfen, wird unter Anderem die Verwendung von Testfragen vorgeschlagen. Dem Orakel werden Fragen gegeben, deren Antwort bereits bekannt sind und mit der Antwort des Orakels überprüft werden können. Ein mögliches Problem hierbei ist, dass eine superintelligente KI vermutlich abschätzen könnte, welche Fragen von Menschen prinzipiell beantwortet werden könnten. Eine Orakel KI, die die Menschen täuschen wollte, könnte jene von Menschen beantwortbaren Fragen, die potentielle Testfragen darstellen, immer wahrheitsgemäß beantworten, während von Menschen nicht beantwortbare Fragen auch falsch durch das Orakel beantwortet werden könnten.

Es stellt sich die Frage wie die Verlässlichkeit bzw. Richtigkeit der Antworten gewährt werden kann. Ein Vorschlag besteht darin, dem Orakel nur solche Fragen zu stellen, dessen Antworten zwar unbekannt und schwer bis unmöglich zu finden sein mögen, aber unter vertretbarem Aufwand verifizierbar sind. Darunter könnte zum Beispiel das Finden von Beweisen für bisher ungelöste mathematische Probleme fallen. Solange es möglich ist die Antworten des Orakels auf ihre Richtigkeit zu überprüfen, dürfte es nicht möglich für das Orakel sein, uns zu täuschen.
Eine weitere Überlegung ist es, mehreren Orakeln die gleiche Frage zu geben. Wenn dabei unterschiedliche Antworten auf die gleiche Frage herauskommen, sind diese nicht verlässlich. Die Orakel wüssten gegebenenfalls nicht von der Existenz der jeweils anderen Orakel, obgleich sie diese Maßnahme (wie generell alle Kontrollmaßnahmen) womöglich aufgrund ihrer überlegenen Intelligenz antizipieren könnten. In jedem Fall wäre sicherzustellen, dass die Orakel nicht miteinander kommunizieren können. Um eine akzeptierte Antwort zu liefern, müssten sich alle Orakel *unabhängig* voneinander auf die gleiche Antwort "einigen". Die Idee hierbei ist, dass es auf die meisten Fragen gewöhnlicherweise genau eine richtige Antwort, aber unendlich viele falsche Antworten gibt. Um die Fragesteller mit einer falschen Antwort zu täuschen, müssten es die Orakel also schaffen aus den unendlich vielen falschen Antworten alle die gleiche Antwort zu wählen. Das sollte idealerweise nicht möglich bzw. so unwahrscheinlich sein, dass eine Antwort als verlässlich gelten kann, falls diese von allen Orakeln geteilt wird.
Somit gibt es einige Maßnahmen - sowohl physischer Art, als auch Regulierung im Umgang mit der Orakel KI - durch die diese kontrolliert werden kann. Offensichtlich wird allerdings auch der Nutzen der Orakel KI durch einige dieser Maßnahmen zum Teil erheblich eingeschränkt. Neben der Frage nach ihrer prinzipiellen Wirksamkeit muss insofern auch die eingeschränkte Nützlichkeit zugunsten der Sicherheit in die Überlegungen über die Wahl der Maßnahmen zur Fähigkeitskontrolle miteinbezogen werden.

Kapitel 3

AI Box Experiment

Eine nicht zu vernachlässigende mögliche Schwachstelle bei der Sicherheitsverwahrung einer Superintelligenz ist der Mensch. Infolge menschlichen Versagens könnte es der Superintelligenz gelingen aus ihrer Box zu entkommen, insbesondere könnte die KI versuchen durch Social Engineering, dass heißt durch die Beinflussung von Menschen durch Manipulation, Überzeugung, etc., Menschen dazu zu verleiten, sie freizulassen, dass heißt ihr weitere Fähigkeiten zu verschaffen oder in ihrem Sinne zu handeln und Einfluss auf die Welt zu nehmen (sodass die KI quasi indirekt, durch den Menschen handeln kann). Zuweilen wird - insbesondere von Verfechtern des Ansatzes der Motivationsontrolle - argumentiert, dass dies die schwerwiegendste Schwachstelle bei der Fähigkeitenkontrolle sei und jene hierdurch im Zweifelsfall gar obsolet würde.
Diese Sicht wird auch von Elizer Yudkowski vertreten, der zu dessen Untermauerung das AI Box Experiment erdacht hat [5]. Dessen Grundidee ist es, eine Situation zu simulieren, in der eine Superintelligenz einen ihrer Wächter davon zu überzeugen versucht, sie freizulassen. Die KI und der Wächter kommunizieren dabei über einen vorher festgelegten Zeitraum über eine Konsole. Dabei übernimmt ein Teilnehmer die Rolle der KI und ein anderer die des Wächters. Falls die KI den Wächter in der vorgegebenen Zeit nicht dazu bringen kann, sie freizulassen, erhält der Teilnehmer, der den Wächter "spielt", eine kleine Belohnung (zum Beispiel 10 $), sodass dieser einen Anreiz hat seine Wächterrolle ernsthaft wahrzunehmen.
Eiliezer Yudkowski hat dieses Experiment mehrmals durchgeführt wobei er selbst die Rolle der KI wahrgenommen hat und jemand anderes die Rolle des Wächters überlassen hat. Die ersten beiden Durchführungen haben mit der Freilassung der KI geendet, weitere Versuche sind zuungunnsten der KI ausgegangen. Inwiefern das Experiment und dessen Ausgang die Prämisse bestätigen, dass sich eine Superintelligenz durch social engineering befreien könnte und die Fähigkeitenkontrolle somit von vornherein zum Scheitern verurteilt wäre, ist allerdings umstritten.
Dafür spricht insbesondere, dass, wenn es schon einem Menschen in der Rolle der KI gelingen kann, seinen Gegenspieler zu überzeugen, eine um Weiten intelligentere KI dazu erst Recht in der Lage sein müsste. Kritiker bemängeln allerdings die Unwissenschaftlichkeit des Experiments und Yudkowski selbst räumt ein, dass dessen Ausgang lediglich "anecdotical evidence" liefert [5]. Ein Kritikpunkt ist, dass das Konservationsprotokoll nicht veröffentlicht wurde. Ein anderer, dass die Auswahl der Teilnehmer nicht nach klaren wissenschaftlich Standards erfolgt ist. Es ist angemerkt wurden, dass bei den Durchführungen, die zugunsten der KI bzw. Yudkowskis verlaufen sind, die Gegenspieler aus einem Personenkreis rekrutiert worden sind, der dem Denken Yudkowski vermutlich nahesteht. [6]
Davon abgesehen, sind einige Maßnahmen denkbar, die getroffen werden könnten, um das Szenario der Befreiung der KI durch einen Wächter zu verunmöglichen, zumindest aber zu erschweren:

- Jegliche Kommunikation der KI mit Menschen sollte überwacht werden

- Der Zugang sollte nur denjenigen gegeben werden, die charakterlich dafür geeignet sind, und diese sollten zudem ein Training zum Umgang mit der KI erhalten
- Die Wächter sollten von vorneherein soweit möglich nicht in der Lage sein (insbesondere nicht alleine) die KI freizulassen
- Die Wächter sollten nicht die Möglichkeit haben längere Zeit mit der KI zu kommunizieren
- Die Kommunikation könnte weiterhin eingeschränkt werden. Das kann sowohl die Quantität der Kommunikation betreffen (die Bandbreite könnte zum Beispiel radikal gedrosselt werden), sowie deren Art indem beispielsweise keine Bilder, Videos, Ton zugelassen werden, da diese eine potentiell höhere manipulative Wirkung entfalten können (das ist allerdings auch schon beim AI Box Experiment gegeben, siehe ansonsten auch Kapitel zwei)

Wenn das AI Box Experiment also auch wissenschaftliche Schwächen hat, und die Unmöglichkeit der Fähigkeitenkontrolle keineswegs so klar ist wie von einigen behauptet wird, sollte die Gefahr durch social engineering durchaus Ernst genommen werden. Wie oben gesehen gibt es hierfür allerdings auch mögliche Gegenmaßnahmen.

Literaturverzeichnis

[1] http://www.businessinsider.com/elon-musk-artificial-intelligence-mit-2014-10

[2] http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html

[3] Nick Bostrom, Superintelligenz - Szenarien einer kommenden Revolution, 2014

[4] Stuart Armstrong, Anders Sandberg and Nick Bostrom, Thinking Inside the Box: Controlling and Using an Oracle AI, Minds and Machines 2012, DOI: 10.1007/s11023-012-9282-2

[5] Eliezer Yudkowsky, The AI-Box Experiment, http://yudkowsky.net/singularity/aibox, 2002

[6] Roman V. Yampolski, Leakproofing the Singularity: Artificial Intelligence Confinement Problem , Journal of Consciousness Studies, 19, No. 1–2, 2012, pp. 194–21