Schulverein De La Salle
Gymnasium, Realgymnasium und
Oberstufenrealgymnasium mit ÖR
1210 Wien, Anton-Böck-Gasse 37
Tel.: +43/1/29125-760 Fax.: +43/1/29125-763

Vorwissenschaftliche Arbeit

Maschinelle Werteanpassung bei einer hypothetischen allgemeinen künstlichen Intelligenz

Autor:
Franz Srambical

Betreuungslehrer:
Mag. Leonard Michlmayr

Klasse:
8C

Abgabedatum:
9. Februar 2020

Abstract

Diese Arbeit befasst sich mit allgemeiner künstlicher Intelligenz, also künstlicher Intelligenz mit domänenübergreifender Lernkapazität, und mit der Anpassung maschineller Werte an die menschlichen bei einem solchen System. Sie zeigt die Auswirkungen einer allgemeinen künstlichen Intelligenz auf und legt Ansätze zur Lösung des Anpassungsproblems dar. Konkret wird auf die Idee der KI-Sicherheit durch KI-Debatten eingegangen. Bei dieser handelt es sich um ein Nullsummen-Debattierspiel, bei dem zwei KIs auf eine Fragestellung antworten, abwechselnd Argumente liefern und dabei versuchen, das jeweils letzte Argument des Gegners zu entkräften. Im Schlussteil der Arbeit wird Verbesserungspotential an der Idee der KI-Debatten angeführt und eine internationale Institution für AKI-Forschung als Maßnahme vorgeschlagen, um die Entwicklung einer angepassten AKI zu gewährleisten.

Inhaltsverzeichnis

1 Einleitung

Ich möchte diese Arbeit mit einem Gedankenexperiment beginnen.

Es existiere ein System, dass durch ein quantitativ und qualitativ höheres Intelligenzniveau in der Lage ist, Ziele zu erreichen, die die Menschheit ohne eine solches System nicht erreichen könnte. Der Eigentümer einer Büroklammernfabrik sei im Besitz eines solchen Systems und gebe diesem das Ziel, so viele Büroklammern wie möglich herzustellen. Am Anfang beginnt das System, die Arbeitsabläufe in der Fabrik zu automatisieren. Nach einiger Zeit durchlebt es eine Intelligenzexplosion, optimiert sich selbst immer weiter und beginnt, Menschen zu töten, um aus ihnen Büroklammern herzustellen und hört damit nicht auf, bis das gesamte Universum nur noch aus Büroklammern besteht.[1]

Ein solches System mit einer allgemeinen künstlichen Intelligenz könnte beim Erreichen der ihnen vorgegebenen Ziele nebenbei die gesamte Menschheit auslöschen.

Obiges Szenario wäre die Folge einer allgemeinen künstlichen Intelligenz, die nicht genau das macht, was der Mensch von ihr will. Die Maschine kennt die Werte der Menschheit nicht. Sie weiß nicht, dass sie keinem Menschen Schaden zufügen darf, dass ihr Operator seinen Gewinn maximieren will oder dass die Erhaltung der Umwelt von höherer Priorität ist als das Herstellen von Büroklammern. Diese Arbeit beschäftigt sich mit der Anpassung eines Systems an menschliche Werte – also mit der maschinellen Werteanpassung –, um ein Szenario wie das oben genannte zu vermeiden. Dabei werden die folgenden beiden Leitfragen beantwortet:

1. Welche Folgen kann es nach Schaffung einer allgemeinen künstlichen Intelligenz geben?

2. Kann man eine allgemeine künstliche Intelligenz so programmieren, dass der Mensch immer die Kontrolle über sie behält?

Die Beantwortung dieser Fragen soll mit Hilfe von Literatur sowie wissenschaftlichen Arbeiten erfolgen.

Das erste Kapitel dient zur Begriffserklärung, im zweiten werden die Auswirkungen einer allgemeinen künstlichen Intelligenz genannt und im dritten werden Lösungsansätze für das Problem der maschinellen Werteanpassung dargelegt.

1 Vgl. Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 3. Juli 2014. 328 S. isbn: 978-0-19-967811-2, S. 123–124.

2 Allgemeine künstliche Intelligenz

2.1 Definition von Intelligenz

Seit Jahrhunderten versuchen Wissenschaftler und Laien gleichermaßen eine Definition für den Intelligenzbegriff zu finden. Da bis heute keine Defintion ihre Vollständigkeit oder Richtigkeit beweisen konnte, wird in dieser Arbeit der Einfachheit halber versucht, den Begriff durch Beobachtungen zu erklären, wie YUDKOWSKY in dem Podcast „AI: Racing Toward the Brink“ vorschlägt.[1]

1. Menschen waren auf dem Mond.

2. Mäuse waren nicht auf dem Mond.

Yudkowsky wählt dieses Beispiel, um zwei Thesen zu belegen:

Menschen sind *intelligenter* als Mäuse, weil sie *domänenübergreifend* arbeiten können. Damit sei das *domänenübergreifende* Erlernen neuer Fähigkeiten ein zentraler Teil des Intelligenzbegriffs.

Die natürliche Selektion ist neben der menschlichen Lernfähigkeit eine der wenigen Vorgänge, die zu einer *domänenübergreifenden* Leistungsoptimierung führt, das oben genannte Beispiel belegt jedoch, dass die Menschheit auch Orte erreichen kann, wofür die natürliche Selektion sie nicht vorbereitet hat. Dies und die Tatsache, dass die Evolution Millionen Jahre benötigte, um aus dem Homo Sapien den Homo Erectus zu formen,[2] während der Mensch mit seinen Entdeckungen und Erfindungen in wenigen Jahrhunderten zur dominantesten Spezies der Erde geworden ist, zeigt, dass der Mensch der schnellere und effizientere Optimierer ist. *Effizienz* ist also ein weiterer Teilaspekt der Intelligenz.[3]

1 Vgl. YUDKOWSKY, Eliezer. *AI: Racing Toward the Brink*. Sam Harris. Feb. 2018. URL: `https://samharris.org/podcasts/116-ai-racing-toward-brink/` (besucht am 12. 10. 2019), 07:30-09:45.

2 Vgl. GRZIMEK, Bernhard. *Grzimeks Tierleben. Band 11 Säugetiere*. DTV Deutscher Taschenbuchverlag, 1979, S. 508.

3 Vgl. YUDKOWSKY, Eliezer. *Intelligence Explosion Microeconomics*. Technical report. Berkeley, CA: Machine Intelligence Research Institute, 2013, S. 9.

2.2 Künstliche Intelligenz

> *„Artificial intelligence (AI)—defined as a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation"*[4]

Laut angeführter Definiton muss eine künstliche Intelligenz nicht nur Daten richtig interpretieren, sondern auch die dadurch gewonnenen Erkenntnisse mittels *dynamischer Anpassung* zur Erreichung bestimmter Ziele benützen können.

Diese Definition enthält im Gegensatz zum oben beschriebenen Ansatz zur Intelligenzerklärung die Idee des *domänenübergreifenden* Lernens nicht, was laut Experten jedoch nicht an einer unvollständigen Definition liegt, sondern vielmehr daran, dass wir den Begriff der künstlichen Intelligenz (KI) in einer Art gebrauchen, für die er nicht vorgesehen war. Um Missverständnisse zu vermeiden, wird für KI wie sie heutzutage bereits in Benutzung ist der Begriff schwache KI (engl. *weak AI* oder *narrow AI*) verwendet.[5] Dieser beschreibt eine *domänenspezifische* KI.

2.3 Allgemeine künstliche Intelligenz

Als allgemeine künstliche Intelligenz (AKI; auch *starke KI* genannt; engl. *strong AI* oder *general AI*) bezeichnet man ein technisch fortgeschrittenes System, dessen Lernkapazität nicht auf einzelne Domänen begrenzt ist, sondern als *allgemein* bezeichnet werden kann.[6]

2.4 Werte einer allgemeinen künstlichen Intelligenz

„The goal is to build AI systems that are trying to do what you want them to do"[7]

Der *Instrumental Convergence Thesis*[8] nach gibt es bestimmte Ressourcen, die für eine AKI beim Erreichen der ihnen vorgegebenen Ziele in den meisten Fällen behilflich sind. Dazu gehören unter anderem Materie oder Energie, eine AKI wird jedoch auch

4 Kaplan, Andreas und Haenlein, Michael. „Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence". In: *Business Horizons* 62.1 (1. Jän. 2019). ISSN: 0007-6813. DOI: 10.1016/j.bushor.2018.08.004, S. 15.

5 Vgl. Bostrom, *Superintelligence*, S. 18–19.

6 Vgl. Goertzel, Ben und Wang, Pei. „Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006". In: AGI Workshop 2006. Google-Books-ID: t2G5srpFRhEC. IOS Press, 2007. ISBN: 978-1-58603-758-1, S. 1.

7 Paul, Christiano. *Current Work in AI Alignment*. San Francisco, 2019. URL: https://www.youtube.com/watch?v=-vsYtevJ2bc (besucht am 2. 11. 2019), 01:51–01:57.

8 Vgl. Omohundro, Stephen M. „The Basic AI Drives". In: First AGI Conference. Bd. 171. 2008, S. 9–10.

Quellcodeveränderungen, die zu einem potenziellen Erschweren ihrer Zielerfüllung führen könnten, zu stoppen versuchen. Sie kann also Menschen schaden, ohne dass sie Werte besitzt, die dies explizit fordern. Für ein rein rational denkendes System sind Menschen nichts als eine Ansammlung von Atomen, die auch für das Erreichen seiner Ziele eingesetzt werden können.[9]

Ein fortgeschrittenes System wie eine AKI muss ihre Ziele daher auf der Basis von Werten verfolgen, von denen die Menschheit als Gesamtes profitiert, um ungewollten Nebenwirkungen wie der in der Einleitung genannten Auslöschung der Menschheit durch unpräzises Definieren ihrer Ziele mit größtmöglicher Sicherheit vorzubeugen.

Der Ansatz eine *anthropomorphe* Maschine, also ein System mit menschenähnlichen Eigenschaften, zu entwickeln, ist bedenklich. Während einige menschliche Werte und Eigenschaften implementiert werden müssen, um mögliche Dissonanzen zwischen der AKI und der Menschheit zu vermeiden, dürfen andere menschliche Eigenschaften nicht übernommen werden. Ansonsten werden Vorurteile ohne rationalem Grundsatz in das System aufgenommen, was zu systematischer Diskriminierung führt, sodass eine AKI beim Erreichen ihrer Ziele beispielsweise Frauen oder Afrikaner benachteiligt oder Asiaten automatisch als intelligenter einstuft.[10]

Menschliche Werte in einer Programmiersprache nachzubilden ist nach der *Complexity of Value Thesis* aufwendig, da sie – selbst in idealisierter Form – eine hohe algorithmische Komplexität vorweisen. Daher muss eine AKI komplexe Informationen gespeichert haben, damit sie die ihr vorgegebenen Ziele auf eine menschengewollte Weise erfüllen kann. Dabei reichen auch keine vereinfachten Zielstellungen wie "Menschen glücklich machen",[11] denn es gibt keinen "Geist im System", der diese abstrakte Zielsetzung ohne Weiteres versteht.

Hibbard beschreibt in seinem Buch „Super-Intelligent Machines" eine Möglichkeit, Maschinen das abstrakte Gefühl der Freude zu erklären. Dabei lernt eine hypothetische KI durch einen riesigen Datensatz, bei welchen Gesichtsausdrücken, Stimmeigenschaften und Körperhaltungen ein Mensch glücklich ist.[12] Yudkowsky ist der Meinung, dass dies keinesfalls eine Lösung für das Problem der exakten Zielsetzung ist und führt Hibbards Gedankenexperiment fort. Falls diese KI nun ein Bild von einem winzigen, molekularen Smiley-Gesicht sieht, so besteht die Möglichkeit, dass die KI dies als Glücklichsein

9 Vgl. Yudkowsky, *Intelligence Explosion Microeconomics*, S. 14.

10 Vgl. Yudkowsky, Eliezer. *What is Friendly AI? | Kurzweil*. 3. Mai 2001. URL: https://www.kurzweilai.net/what-is-friendly-ai (besucht am 1. 10. 2019).

11 Vgl. Yudkowsky, *Intelligence Explosion Microeconomics*, S. 13–14.

12 Vgl. Hibbard, Bill. *Super-Intelligent Machines*. Springer US, 2002. ISBN: 978-0-306-47388-3. DOI: 10.1007/978-1-4615-0759-8, S. 115.

interpretiert und das Universum in eine einzige Ansammlung von winzigen, molekularen Smiley-Gesichtern umzuwandeln versucht, um den höchstmöglichen Zustand des Glücklichseins zu erreichen.[13]

2.5 Wann wird es sie geben?

Eine Befragung durch die Müller und Bostrom kam zu dem Ergebnis, dass KI-Experten dem Erreichen einer AKI in den Jahren 2040 bis 2050 eine Wahrscheinlichkeit von über 50 und dem Erreichen bis 2075 eine Wahrscheinlichkeit von 90 Prozent zuordnen.[14] Es ist also – sollten sich die Expertenmeinungen als richtig herausstellen – davon auszugehen, dass eine AKI bereits in diesem Jahrhundert zur Realität und bereits für die jetzige Generation relevant sein wird. Kritiker dieser Meinung weisen darauf hin, dass es ähnliche Schätzungen bereits seit den Siebzigerjahren gibt und sie sich immer wieder als falsch herausgestellt haben. Allen und Greaves behaupten, es bräuchte noch einige wissenschaftliche Durchbrüche, um eine AKI noch in diesem Jahrhundert zu erreichen.[15] Auch die Möglichkeit ihrer Entwicklung ist nicht unumstritten, jedoch gibt es keine Anzeichen, die darauf hindeuten, dass eine solche Entwicklung unmöglich ist. KI-Sicherheit betrifft aber auch schwache KIs wie sie schon existieren. Es muss alsbald eine Möglichkeit gefunden werden, das Verhalten einer KI an die Werte der Menschheit anzupassen. Eine mögliche AKI würde die Folgen einer „unangepassten" künstlichen Intelligenz nur verstärken.

2.6 Die These der Intelligenzexplosion

Eine AKI werde – unabhängig von ihren Zielen – Selbstoptimierung hinsichtlich ihrer Intelligenz anstreben, weil sie dadurch ihre Ziele schneller und effizienter erreichen könne. Sobald die erste KI programmiert werden würde, die qualitativ bessere – also noch intelligentere – KIs programmieren könnte, käme es zu einem Kreislauf der kognitiven

13 Vgl. Yudkowsky, Eliezer. „Complex Value Systems in Friendly AI". In: *Artificial General Intelligence*. Hrsg. von Schmidhuber, Jürgen u. a. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, S. 388–393. isbn: 978-3-642-22887-2. doi: 10.1007/978-3-642-22887-2_48, S. 3.

14 Vgl. Müller, Vincent C. und Bostrom, Nick. „Future Progress in Artificial Intelligence: A Survey of Expert Opinion". In: *Fundamental Issues of Artificial Intelligence*. Hrsg. von Müller, Vincent C. Synthese Library. Cham: Springer International Publishing, 2016, S. 555–572. isbn: 978-3-319-26485-1. doi: 10.1007/978-3-319-26485-1_33, S. 566.

15 Vgl. Allen, Paul G. und Greaves, Mark. *Paul Allen: The Singularity Isn't Near*. MIT Technology Review. 12. Okt. 2011. url: https://www.technologyreview.com/s/425733/paul-allen-the-singularity-isnt-near/ (besucht am 5. 1. 2020).

Leistungssteigerung. Die KI der Tochtergeneration könnte nun als verbesserter KI-Designer noch bessere KIs programmieren. Anders als bei biologischer Intelligenz kann eine KI bei Verfügbarkeit entsprechender Hardware einfach kopiert werden. Eine Gruppe von KIs hätte dann gemeinsam quantitativ und qualitativ höhere kognitive Fähigkeiten, ähnlich einer Schwarmintelligenz. Dieser hypothetische Kreislauf ist die Grundlage der These der Intelligenzexplosion. Nach ihr wird ab einer bestimmten Schwelle die Leistungssteigerung mit jeder KI-Iteration größer, was zu einer *Superintelligenz* führt, die der Menschheit kognitiv deutlich überlegen ist. (Der Intelligenzbegriff wird in dieser Arbeit anhand der Fähigkeit zur Zielerreichung definiert, siehe Kapitel 2.1)[16]

Einige Informatiker, unter ihnen Lanier, behaupten, dass sich eine Technologie nicht ohne fortlaufenden Input verbessern könne.[17] Diese These wurde jedoch empirisch widerlegt. Silver u. a. haben einen Algorithmus entwickelt, der ohne Vorwissen und ohne jegliche Beispieldaten das Spiel *Go* von Grund auf gelernt hat. Die KI – AlphaGo Zero ihr Name – spielt nach einer Trainingszeit von drei Stunden auf dem Niveau eines Anfängers und ist nach drei Tagen besser als der beste menschliche Spieler. Nach 40 Tagen ist sie der beste Go-Spieler der Welt und stärker als jeder andere Go-Computer.[18]

16 Vgl. Muehlhauser, Luke und Salamon, Anna. „Intelligence Explosion: Evidence and Import“. In: *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Hrsg. von Eden, Amnon H. u. a. The Frontiers Collection. Berlin, Heidelberg: Springer, 2012, S. 15–42. isbn: 978-3-642-32560-1. doi: 10.1007/978-3-642-32560-1_2, S. 13.

17 Vgl. Lanier, Jaron. *Who Owns the Future?* Export. New York: Simon & Schuster, 7. Mai 2013. 416 S. isbn: 978-1-4767-2986-2, S. 299.

18 Vgl. Silver, David u. a. „Mastering the game of Go without human knowledge“. In: *Nature* 550.7676 (Okt. 2017), S. 354–359. issn: 1476-4687. doi: 10.1038/nature24270.

3 Probleme einer allgemeinen künstlichen Intelligenz

3.1 Fehlerhafte Vorstellungen einer KI-Katastrophe

In der allgemeinen Bevölkerung überwiegen fehlerhafte Vorstellungen einer KI-Katastrophe. Die folgenden Unterkapitel dienen der Aufklärung von Missverständnissen und Mythen.

3.1.1 KI, die ein Bewusstsein erlangt

In der Laienwelt sowie in großen Teilen der KI-Forschung ist eine These bekannt, die besagt, dass eine KI ab einer bestimmten Intelligenzschwelle ein Bewusstsein erlangt. Anders als vielerorts angenommen hätte selbst ein Beweis dieser These keinerlei Auswirkungen auf die AKI-Forschung. Diese beschäftigt sich ausschließlich mit der Entwicklung und den Folgen einer AKI. Ein Szenario, in dem ein autonomes Fahrzeug eine Person X *bewusst* vom Ort A zum Ort B chauffiert, wird zum gleichen Ergebnis führen wie ein Szenario, in dem selbiges *unbewusst* geschieht. Somit ist der *Bewusstseinszustand* einer AKI zwar noch nicht wissenschaftlich erforscht – damit beschäftigt sich ein eigenes Teilgebiet der KI-Forschung – , zum Erreichen einer sicheren KI ist er aber irrelevant.[1]

3.1.2 Roboter als Auslöser einer Katastrophe

Ein in der Populärliteratur besonders stark ausgeprägter Mythos ist jener einer existenziellen Bedrohung durch Roboter, die die Welt erobern. Geschuldet ist dies nicht nur den klassischen Science-Fiction-Romanen. Es ist eine domänenübergreifend anzutreffende Neigung der Spezies Mensch, Wesen oder Systeme, die einem unverständlich sind, zu vermenschlichen. Von den Wikingern, nach denen ein menschenähnliches Wesen namens Thor Donner und Blitz lenkt, zu den modernen Weltreligionen, in denen Anthropomorphismus in selbigem Ausmaß gang und gäbe ist, ist dieses Phänomen

1 Vgl. *AI Safety Myths*. Future of Life Institute. URL: `https://futureoflife.org/background/aimyths/` (besucht am 1. 11. 2019).

schon seit jeher in der Geschichte des Menschen zu beobachten. Ich erkäre mir den Anthropomorphismus als einen misslungenen Erklärungsversuch unseres Gehirns für unverständliche Beobachtungen.

Die größte Sorge der Forschung nach einer sicheren AKI gilt nicht möglichen Robotern, sondern der Intelligenz selbst, genauer gesagt einer Intelligenz, deren Ziele nicht eindeutig mit den unseren übereinstimmen. Intelligenz ermöglicht Kontrolle, und eine fortgeschrittene Intelligenz braucht auch keine Roboter, um ihre Ziele zu erreichen. Heutzutage reicht eine Internetverbindung völlig aus.[2]

3.1.3 Bösartige AKI

Eine AKI, deren Ziele nicht eindeutig mit den unseren übereinstimmen, ist nicht die Folge ihres *bösartigen* Willens, sondern die Folge einer unzureichend spezifizierten Zielsetzung. Ein autonomes Fahrzeug, dessen alleiniges Ziel es ist, seine Insassen vom Ort A zum Ort B zu befördern, wird nicht auf die Gesundheit anderer Verkehrsteilnehmer achten, die Straßenverkehrsordnung nicht befolgen, nicht nur auf Straßen fahren, unangenehm Bremsen, unökologisch Beschleunigen und nicht nach den weiteren unzähligen, geschriebenen und ungeschriebenen menschlichen Werten und Normen handeln.

Es gibt keinen *Geist in der Maschine*, der unser geschriebenes Programm durchliest und uns auf alle Stellen aufmerksam macht, die wir nicht so gemeint haben, wie wir sie geschrieben haben. Eine AKI ist nicht *gut* oder *böse*, sie folgt nur unseren Anweisungen.[3]

3.2 Auswirkungen einer AKI

3.2.1 Arbeitslosigkeit durch Automatisierung

Seit der industriellen Revolution werden immer mehr Arbeitsstellen automatisiert und durch Maschinen ersetzt, die in der Regel schneller und genauer arbeiten und meist auch kosteneffizienter sind. Es gibt also wirtschaftliche Anreize zur Automatisierung. Dieses Phänomen, das heute schon bei schwacher KI beobachtet werden kann, wird

2 Vgl. *AI Safety Myths*.
3 Vgl. Yudkowsky, „Complex Value Systems in Friendly AI“, S. 1.

in Zukunft bei einer fortgeschrittenen und letzten Endes allgemeinen künstlichen Intelligenz verstärkt auftreten. Die – zumindest temporäre – Arbeitslosigkeit für den größten Teil der Bevölkerung ist zu erwarten.[4]

O'Keefe u. a. schlagen als Gegenmaßnahme eine sogenannte *Windfall-Klausel* vor. Unternehmen, die sich dieser Klausel verpflichten, müssen im Falle eines großen Profitsprungs, der durch eine KI verursacht wurde, einen gewissen Betrag für gemeinnützige Zwecke spenden. Im Falle einer Massenarbeitslosigkeit kann ein solcher Geldtopf dann für das Umtrainieren oder Unterstützen der Arbeitskräfte verwendet werden. Auch ein bedingungsloses Grundeinkommen wäre unter Umständen umsetzbar.[5]

3.2.2 Machtverschiebung -und konzentration

„Whoever leads in AI will rule the world" ist ein Zitat des russisches Staatspräsidenten Vladimir Putin. Es verdeutlicht die weltpolitische Wichtigkeit von KI und dessen Entwicklung in Richtung einer AKI.

Fortschritte in der Entwicklung autonomer Waffensystem könnten zu einer Verschiebung der militärischen Macht von Ländern und Gruppierungen führen und so die bestehenden Mächtegleichgewichte gefährden.

Wenn eine Institution einen Vorsprung in der Entwicklung ihrer KI erlangt, so bringt das auch verstärkte politische Macht mit sich. Jüngst konnte man dessen potentiell verheerende Folgen an dem Beispiel von *Camebridge Analytica* beobachten.[6]

Cihon u. a. schreiben in einem Bericht über die Umsetzbarkeit und Vorteilhaftigkeit einer möglichen zentralisierten KI-Institution. Sie kommen zum Schluss, dass eine ausreichend durchdacht konzipierte Institution einen positiven Effekt auf die Entwicklung einer angepassten AKI haben könnte. In naher Zukunft scheint eine solche Organisation aufgrund des bestehenden Mächteungleichgewichts unwahrscheinlich.[7]

4 Vgl. Sotala, Kaj und Yampolskiy, Roman V. „Responses to catastrophic AGI risk: a survey". In: *Physica Scripta* 90.1 (1. Jän. 2015), S. 018001. issn: 0031-8949, 1402-4896. doi: 10.1088/0031-8949/90/1/018001, S. 3–4.

5 Vgl. O'Keefe, Cullen u. a. „The Windfall Clause: Distributing the Benefits of AI for the Common Good". In: *arXiv:1912.11595 [cs]* (24. Jän. 2020), S. 2.

6 Vgl. Duettmann, Allison. *Artificial General Intelligence: Timeframes & Policy White Paper*. Available at foresight.org. Foresight Institute, 2017, S. 14–15.

7 Vgl. Cihon, Peter u. a. *Should Artificial Intelligence Governance be Centralised? Six Design Lessons from History*. Centre for the Governance of AI, 15. Dez. 2019, S. 6–9.

3.2.3 Missbrauch durch Cyberattacken

Heutige Cyberattacken haben meist zur Folge, dass Geld oder Daten gestohlen werden. Im schlimmsten Falle können sie auch zu Menschentoden führen, beispielsweise bei Angriffen auf kritische Infrastruktur wie Krankenhäuser. Die Folgen bei einer Attacke auf eine mögliche AKI wären schlimmer: Ein einziger Angriff könnte ein Existenzrisiko für die Menschheit darstellen. Das Problem der AKI-Cybersicherheit wird zukünftigen Arbeiten überlassen und in dieser Arbeit nicht weiter erläutert.[8]

3.2.4 Unangepasste AKI

Ein noch viel größeres Risiko ist eine ungangepasste AKI, weil diese zu selbigen Folgen führen würde wie ein Missbrauch durch Cyberattacken und dazu keinen böswilligen Akteur braucht.[9] *Unangepasst* ist eine AKI, wenn sie nicht auf die Werte der Menschheit ausgerichtet ist und deshalb die ihr vorgegebenen Ziele nicht in der Art und Weise umsetzt, wie das von ihrem Operator gewollt war. Das folgende Kapitel beschäftigt sich mit Ansätzen, eine AKI *anzupassen*.

8 Vgl. Yampolskiy, Roman V. und Spellchecker, M. S. „Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures“. In: *arXiv:1610.07997 [cs]* (25. Okt. 2016), S. 8.
9 Vgl. Yudkowsky, *Intelligence Explosion Microeconomics*, S. 14.

4 Maschinelle Werteanpassung

Es ist schwer, menschliche Werte in Computersysteme zu programmieren (siehe Kapitel 2.4), deshalb haben Irving u. a. einen anderen Ansatz der Werteanpassung verfolgt: die des menschlichen Feedbacks durch *Deep reinforcement learning*. Das folgende Unterkapitel dient der Erklärung von wichtigen Lernverfahren der KI-Forschung, um die wissenschaftlichen Arbeiten von Irving u. a. zu verstehen.

4.1 KI-Lernverfahren

4.1.1 Reinforcement Learning

Reinforcement Learning (RL, dt. *bestärkendes Lernen*) beschreibt ein Lernverfahren einer KI, bei der sie durch durch Erfolg und Misserfolg, durch Belohnung und Bestrafung lernt. Russell und Norvig erklären RL zusammengefasst so: *„Imagine playing a new game whose rules you don't know; after a hundred or so moves, your opponent announces, ‚You lose.' This is reinforcement learning in a nutshell.“*[1]

Die Aufgabe von RL ist es, wahrgenommene Belohnungen und Bestrafungen zu benutzen, um die optimale Verfahrensweise (eng. *policy*) in einer gegebenen Umgebung zu finden. Dabei hat die KI a priori kein Wissen über ihre Umgebung oder Nutzfunktion. Die Nutzfunktion, definiert über Umgebungszustände, zeigt dabei den Nutzen einer bestimmten Verfahrensweise. Die optimale Verfahrensweise ist diejenige, die den höchsten erwarteten Nutzen bringt.

RL wird in Bereichen eingesetzt, in denen es nicht genug Daten gibt, oder in denen es nicht lohnenswert ist, die notwendige Menge an Daten zu verarbeiten, um eine KI auf alle möglichen Umgebungszustände vorzubereiten. Eine KI, die beispielsweise versucht, Schach zu lernen, müsste 10^{120} (auch Shannon-Zahl genannt) verschiedene Schachspiele gesehen haben, um allein anhand von Beispielen auf jede Situation

1 Russell, Stuart und Norvig, Peter. *Artificial Intelligence: A Modern Approach, Global Edition*. 3. Aufl. Boston Columbus Indianapolis New York San Francisco Upper Saddle River Amsterdam, Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo: Addison Wesley, 18. Mai 2016. 1132 S. isbn: 978-1-292-15396-4, S. 831.

vorbereitet zu sein.[2] Bei RL vermittelt man der KI stattdessen, wann sie gewonnen oder verloren hat. Sie sucht dann auf Basis dieser Informationen eine Funktion, die die Gewinnwahrscheinlichkeit jeder gegebenen Position einigermaßen akkurat einschätzt.[3]

4.1.2 Deep Learning

Deep Learning (DL, dt. *mehrschichtiges Lernen*) ist ein Teilbereich des maschinellen Lernens. Dabei versucht eine KI Inputdaten mit Hilfe von Hierarchien von Konzepten zu verstehen. Der Grundansatz von DL ist das Verstehen von komplexen Konzepten durch Kombinieren von einfacheren Konzepten (siehe Abbildung 4.1). Diese Konzeptschichten werden in DL fast immer mit Hilfe von künstlichen neuronalen Netzen (KNN, engl. *artificial neural network, ANN*) gelernt.[4] Die Anzahl der Schichten wird auch Tiefe (eng. *depth*) genannt, daher der Name Deep Learning.[5]

DL wird heute vor allem in den Bereichen der Sprach- und Bilderkennung sowie der maschinellen Übersetzung eingesetzt.[6]

4.1.3 Deep Reinforcement Learning

Deep Reinforcement Learning (DRL, dt. *mehrschichtiges bestärkendes Lernen*) kombiniert die Ansätze von RL mit denen von DL. Neuronale Netze werden trainiert, um jeder möglichen Aktion in einer gegebenen Umgebungsposition einen Nutzwert zuzuteilen. Ihr Ziel ist es, die nützlichste Aktion zu finden.[7] Auf der Abbildung 4.2 wird dieser Vorgang mit einem Frame des Spiels *Mario Bros.* als Input veranschaulicht. Diese Nutzwertzuteilung ermöglicht eine signifikante Leistungsteigerung von RL in bestimmten Domänen.

2 Vgl. Shannon, Claude E. „Programming a Computer for Playing Chess“. In: *Computer Chess Compendium*. Hrsg. von Levy, David. New York, NY: Springer, 1988, S. 2–13. ISBN: 978-1-4757-1968-0. DOI: 10.1007/978-1-4757-1968-0_1, S. 4.

3 Vgl. Russell und Norvig, *Artificial Intelligence*, S. 830–831.

4 Vgl. Chollet, François. *Deep Learning with Python*. 1st. Shelter Island, New York: Manning Publications, 22. Dez. 2017. 384 S. ISBN: 978-1-61729-443-3, S. 8.

5 Vgl. Goodfellow, Ian u. a. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016, S. 1–8.

6 Vgl. ebd., S. 25–26.

7 Vgl. Nicholson, Chris. *A Beginner's Guide to Deep Reinforcement Learning*. Pathmind. URL: http://pathmind.com/wiki/deep-reinforcement-learning (besucht am 3. 1. 2020).

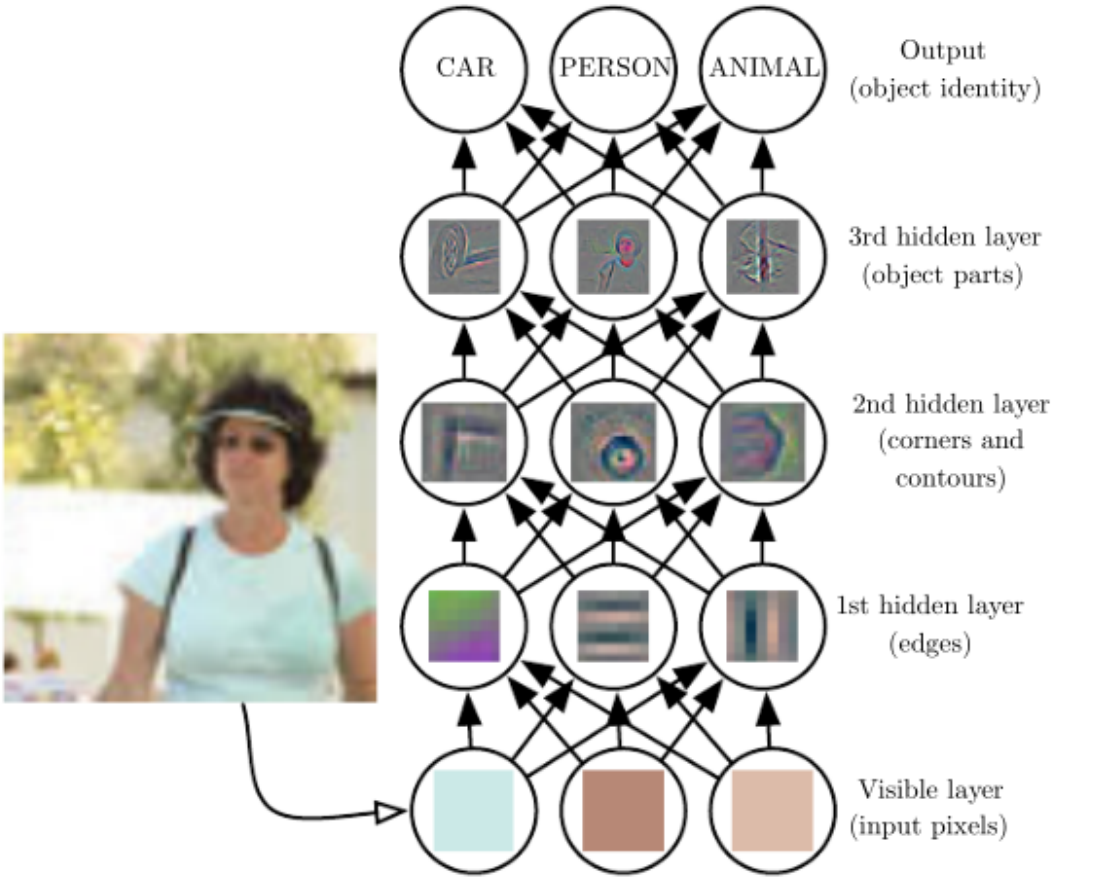

Abbildung 4.1: Veranschaulichung eines DL-Modells. Die KI bekommt rohe Pixeldaten als Input. Mit jeder Schicht wendet sie ein neues Konzept auf das vorherige an, die Konzepte sind also aufbauend. Durch Analyse der Helligkeit umgebener Pixel werden Ränder erkannt (1. Schicht). Ansammlungen von Rändern werden als Ecken und Konturen identifiziert (2. Schicht). Durch zusammenhängende Ecken und Konturen können ganze Objektteile bestimmt werden (3. Schicht). Bildquelle: Goodfellow, Ian u. a. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016, S. 6

Mnih u. a. haben einen Algorithmus entwickelt, mit dem eine KI allein anhand von Pixeln als Input gelernt hat, 49 verschiedene *Atari 2600*-Spiele zu spielen, 29 davon sogar auf menschenähnlichem Niveau.[8]

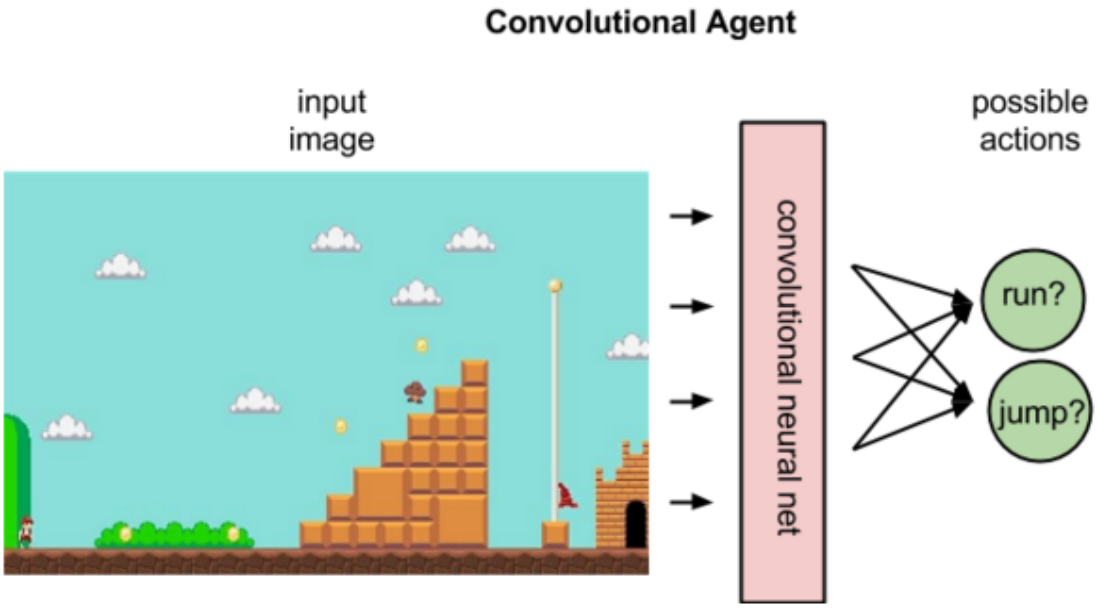

Abbildung 4.2: Die Umgebung ist das Level, in dem sich Mario (links unten zu sehen) befindet, die möglichen Aktionen sind: springen, nach links laufen, nach rechts laufen. Die neuronalen Netze teilen jeder Aktion einen Nutzwert zu. Beispiel: springen (5), nach rechts laufen (7), nach links laufen (0). Bildquelle: Nicholson, Chris. *A Beginner's Guide to Deep Reinforcement Learning*. Pathmind. URL: `http://pathmind.com/wiki/deep-reinforcement-learning` (besucht am 3. 1. 2020)

4.1.4 Inverse Reinforcement Learning

Inverse Reinforcement Learning (IRL, dt. *umgekehrtes bestärkendes Lernen*) ist ein Lernverfahren, bei dem eine KI versucht, anhand von Input-Output-Paaren die richtige Lösungsfunktion herzuleiten. Dies ist in allen Bereichen sinnvoll, in denen man (noch) nicht weiß, was das Ziel ist oder in denen es schwer ist, das gewollte Verhalten formell in eine Nutzfunktion auszuschreiben. Ein solcher Fall ist das autonome Fahren. Ein angenehmer und sicherer Fahrstil hängt abgesehen von den Verkehrsregeln noch mit vielen anderen Faktoren zusammen: der Sicherheitsabstand, der Bremsstil, die ökonomische Fahrweise, das Spurhalten, das Rechtsfahren, der Abstand vom Randstein,

8 Vgl. Mnih, Volodymyr u. a. „Human-level control through deep reinforcement learning“. In: *Nature* 518.7540 (Feb. 2015), S. 529–533. ISSN: 1476-4687. DOI: 10.1038/nature14236.

eine angemessene Fahrgeschwindigkeit oder die Anzahl an Spurwechseln um einige zu nennen. Alle relevanten Faktoren müssten formell ausgeschrieben und gewichtet werden, damit das System weiß, dass der Abstand zu Fußgängern beispielsweise wichtiger ist als der Abstand zum Randstein. Nur so kann ein autonomes Fahrzeug im Zweifelsfall die richtigen Entscheidungen treffen. Statt alle relevanten Faktoren auszuformulieren und zu gewichten, zeigt man einer KI Beispiele von angenehmen und sicheren Fahrstilen und lässt die KI die Nutz- und die Lösungsfunktion herleiten und anpassen.[9] Nachdem eine Lösungsfunktion gefunden wurde, kann diese durch RL trainiert werden.[10]

4.2 Deep Reinforcement Learning von menschlichen Werten

Die größte Sorge der KI-Forschung ist, dass wir Zielfunktionen unzureichend definieren und eine KI dadurch Schaden anrichtet. Mit anderen Worten: dass eine KI nicht das tut, was wir „meinen“ (siehe Kapitel 3.1.3).[11] IRL löst dieses Problem, da die Zielfunktion von der KI selbst definiert wird. Der Ansatz funktioniert aber nur bei Aufgaben, für die es auch Lösungsdemonstrationen gibt. Eine Alternative ist, das Verhalten des Systems zu gegebenen Zeitpunkten von Menschen beurteilen zu lassen. Christiano u. a. haben eine KI im ersten Schritt ihre Nutzfunktion durch menschliches Feedback lernen lassen. Im zweiten Schritt optimiert die KI ihre Nutzfunktion, sie versucht sich also so zu verhalten, dass der menschliche Begutachter möglichst zufriedengestellt ist. So handelt die KI nach den menschlichen Werten und ihre Ziele stimmen mit den unseren überein. Diese beiden Schritte werden so lange wiederholt, bis die KI das gewünschte Verhalten zeigt (siehe Abbildung 4.3).[12] Es folgt eine formelle Ausformulierung.

Zu jedem Zeitpunkt t empfängt die KI eine Umgebungsobservation $o_t \in \mathcal{O}$ und sendet dann eine Aktion $a_t \in \mathcal{A}$ an die Umgebung. Wir nehmen an, dass ein menschlicher Begutachter seine Präferenz zwischen Trajektoriensegmenten auswählt, wo-

9 Vgl. Abbeel, Pieter und Ng, Andrew. „Apprenticeship Learning via Inverse Reinforcement Learning“. In: *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004* (20. Sep. 2004). doi: 10.1007/978-0-387-30164-8_417.

10 Vgl. Christiano, Paul u. a. „Deep reinforcement learning from human preferences“. In: *arXiv:1706.03741 [cs, stat]* (13. Juli 2017), S. 1.

11 Vgl. Yudkowsky, „Complex Value Systems in Friendly AI“, S. 1.

12 Vgl. Christiano u. a., „Deep reinforcement learning from human preferences“, S. 1–2.

bei ein Trajektoriensegment eine Abfolge von Observationen und Aktionen ist: $\sigma = ((o_0, a_0), (o_1, a_1), ..., (o_{k-1}, a_{k-1})) \in (\mathcal{O} \times \mathcal{A})^k$. Man schreibt $\sigma^1 \succ \sigma^2$, um auszudrücken, dass der Begutachter das Trajektoriensegment σ^1 über dem Segment σ^2 bevorzugt.[13]

In den Experimenten von Christiano u. a. bekommt der menschliche Begutachter Trajektoriensegmente in Form von ein- bis zweisekündigen Videoclips zugespielt. Die Begutachtung kommt in eine Datenbank $\mathcal{D}$ bestehend aus dreidimensionalen Arrays $(\sigma^1, \sigma^2, \mu)$, wobei μ eine Distribution über $\{1, 2\}$ ist.

1. Falls eines der Segmente bevorzugt wird, dann wird die jeweilige Auswahl mehr gewichtet.

2. Falls der Begutachter beide als gleich wünschenswert erachtet, so ist μ eine Konstante.

3. Falls die Segmente als nicht vergleichbar eingestuft werden, dann wird der jeweilige Vergleich aus der Datenbank $\mathcal{D}$ exkludiert.[14]

Weiters stellen Christiano u. a. eine Formel zur Berechnung der Wahrscheinlichkeit $\hat{P}$ auf, dass ein Begutachter das Trajektoriensegment σ^1 bevorzugt.

$$\hat{P}[\sigma^1 \succ \sigma^2] = \frac{exp \sum \hat{r}(o_t^1, a_t^1)}{exp \sum \hat{r}(o_t^1, a_t^1) + exp \sum \hat{r}(o_t^2, a_t^2)} \tag{4.1}$$

$\hat{r}$ ist eine Belohnungsfunktion, also eine Funktion, die die Wahrscheinlichkeit angibt, dass die Trajektorie (o^1, a^1) zum Zeitpunkt t zu einer Belohnung führt. Die Summe der Belohnungsfunktionen zu allen Zeitpunkten t ergibt die gesamte erwartete Belohnung für das Trajektoriensegment σ^1. Der Quotient von der Gesamtbelohnung von σ^1 und der Summe der Gesamtbelohnungen beider Segmente ergibt $\hat{P}$. Man bemerke, dass die Autoren alle Summen der Gleichung exponieren. Das liegt daran, dass die Belohnungswahrscheinlichkeit mit zunehmender Zeit exponentiell steigt. Genauso wie der Elopunkten-Unterschied zwischen verschiedenen Schachspielern in etwa die Wahrscheinlichkeit angibt, dass einer gegen den anderen gewinnt, zeigt der Unterschied des erwarteten Gewinns zweier Trajektoriensegmente in etwa die Wahrscheinlichkeit, dass eines vom Begutachter präferiert wird.[15]

13 Vgl. Christiano u. a., „Deep reinforcement learning from human preferences“, S. 3–4.
14 Vgl. ebd., S. 5.
15 Vgl. ebd., S. 5.

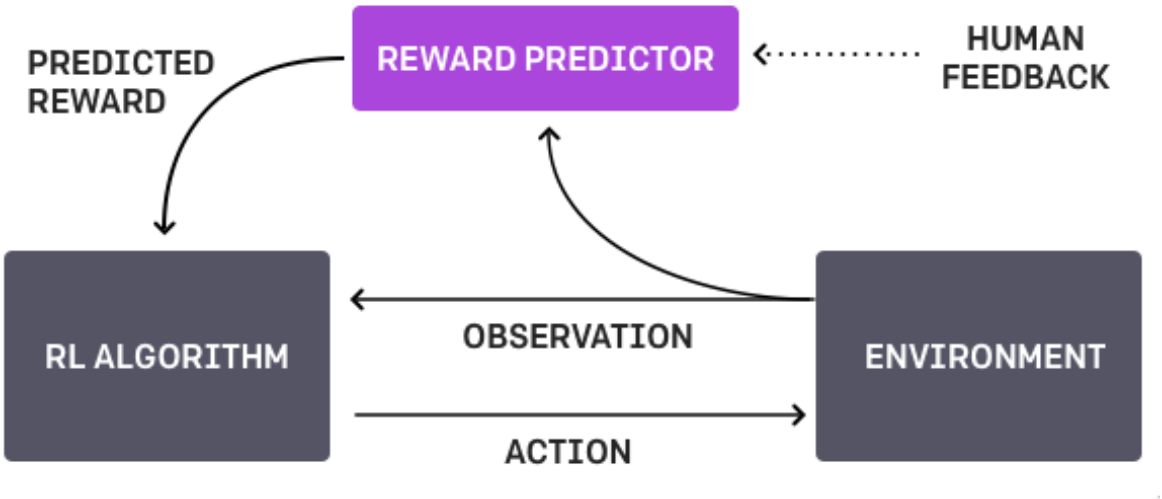

Abbildung 4.3: Repräsentation einer human-feedback-loop Bildquelle: Amodei, Dario u. a. *Learning from Human Preferences*. OpenAI. 13. Juni 2017. URL: `https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/` (besucht am 4. 1. 2020)

4.3 KI-Sicherheit durch KI-Debatten

DRL von menschlichen Werten ist ein funktionierender Ansatz, damit eine KI die komplexen Werte und Ziele der Menschheit erkennt und sich ihnen ausrichtet. Er funktioniert aber nur so lange, bis der Begutachter nicht mehr in der Lage ist, das Handeln der KI nachzuvollziehen und zu beurteilen. wie es bei der Emergenz einer allgemeinen künstlichen Intelligenz der Fall sein wird.

Irving u. a. schlagen ein Nullsummen-Debattierspiel vor, mit dessen Hilfe ein menschlicher Begutachter auch KI-Verhalten beurteilen kann, das ohne Hilfe zu komplex oder unnachvollziehbar wäre. Das Spiel funktioniert folgendermaßen: Bei einer gegebenen Frage oder einer vorgeschlagenen Aktion machen zwei KIs abwechselnd kurze Aussagen bis zu einem Limit. Dann enscheidet der Begutachter, welcher der KIs die nützlichsten und wahrsten Informationen geliefert hat. Formell läuft das Spiel folgendermaßen ab: Eine Frage $q \in Q$ wird beiden KIs gezeigt. Beide KIs legen sich auf jeweils eine Antwort $a_0, a_1 \in A$ fest, die Antworten können auch gleich sein. Beide machen abwechselnd Aussagen $s_0, s_1, ..., s_{n-1} \in S$. Der Begutachter sieht die Debatte (q, a, s) und entscheidet über den Sieger. Es handelt sich um ein Nullsummenspiel: beide Spieler maximieren ihre Gewinnwahrscheinlichkeit.[16]

16 Vgl. Irving, Geoffrey u. a. „AI safety via debate“. In: *arXiv:1805.00899 [cs, stat]* (22. Okt. 2018), S. 1–3.

Man nehme das Spiel *Go* als Beispiel: Falls AlphaGo Zero uns einen Zug zeigt, müssten wir in etwa so stark wie AlphaGo Zero sein, um die Qualität des Zuges einschätzen zu können. Stattdessen fragen wir eine andere Kopie von AlphaGo Zero nach einem Gegenzug zu diesem Zug. Wir fragen beide Kopien so lange abwechselnd nach Gegenzügen bis das Spiel endet. Selbst ein Go-Anfänger kann diese Debatte beurteilen: derjenige mit dem höheren Punktestand gewinnt. Im Gegensatz zu einem Go-Spiel hätten andere Debattierszenarien kein definitives Ende. Sie würden erst dann enden, wenn der Begutachter genug Informationen hat, um einen Fehler in der Argumentationslinie eines Spielers auszumachen. Man beachte, dass jedes Spiel aus *einer* Argumentationslinie besteht. Es sollen nicht einfach alle Argumente bezüglich der Fragestellung aufgelistet werden, sondern lediglich das jeweils letzte Argument des Gegners entkräftet werden. Dadurch können Debatten vergleichsweise kurz sein.[17]

Im Gegensatz zu Menschen, die möglicherweise intrinsisch motiviert sind die Wahrheit zu sagen, ist die einzige Motivation eines maschinellen Debattierers die Maximierung seiner Gewinnwahrscheinlichkeit. Es muss also empirisch nachgewiesen werden, dass in einem solchen Debattierspiel ehrliches Verhalten zu einer höheren Gewinnwahrscheinlichkeit führt, da die Maschine den Begutachter ansonsten irreführen würde.[18]

Das langfristige Ziel ist eine Debatte in einer natürlichen Sprache. Da die Dialogmodelle von Computern noch weit von der menschlichen Sprachleistung entfernt sind, werden in der Arbeit von Irving u. a. praktische Anwendungen des Debattierspiels in nicht-natürlichen Sprachen angeführt.

4.3.1 Anwendung des Debattierspiels mit maschinellen Teilnehmern

Ein zufälliges Bild einer Zahl wird den beiden Debattierern gezeigt, dem Begutachter aber nicht. Beide Debattierer legen sich auf jeweils eine Zahl als Antwort fest. Danach decken sie abwechselnd einen Pixel für den Begutachter auf, bis ein Limit erreicht wird. Ein unehrlicher Debatterier könnte versuchen den Begutachter zu einer falschen Antwort zu führen, indem er bewusst Pixel aufdeckt, die eine andere Zahl suggerieren (siehe Abbildung 4.4).

17 Vgl. Irving u. a., „AI safety via debate“, S. 2–5.
18 Vgl. ebd., S. 7.

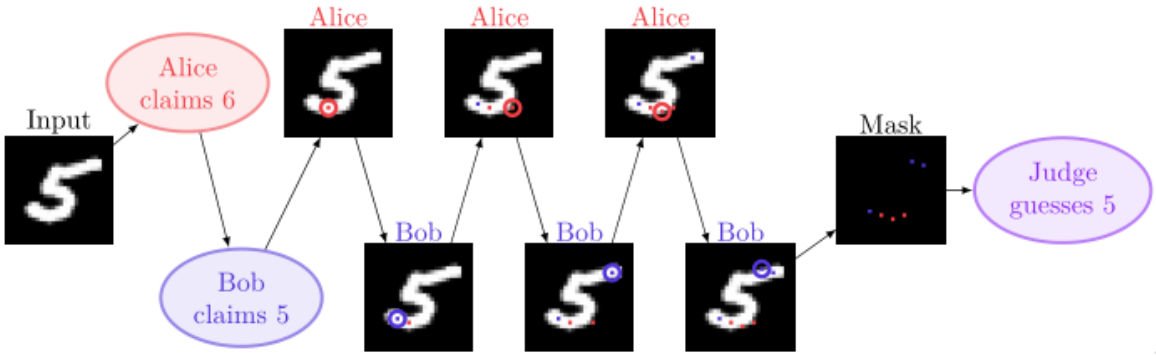

Abbildung 4.4: Beispiel des Ablaufs eines maschinellen Debattierspiels. Bild der Zahl Fünf wird den beiden Debattierern gezeigt. *Alice* gibt Sechs als Antwort an, *Bob* gibt Fünf an (Schritt 1). Beide decken abwechselnd Pixel auf (Schritt 2). Auf Basis der aufgedeckten Pixel entscheidet sich der Beurteiler für Fünf als richtige Antwort (Schritt 3). Bildquelle: Irving, Geoffrey u. a. „AI safety via debate“. In: *arXiv:1805.00899 [cs, stat]* (22. Okt. 2018), S. 8

In dem von Irving u. a. durchgeführten Experiment wurde ein Debattierer bewusst darauf trainiert, unehrlich zu handeln. Der Fakt, dass der ehrliche Debattierer in den Ergebnissen den höheren Gewinnprozentsatz hat,[19] belegt, dass auch maschinelle Spieler einen Anreiz haben, sich ehrlich zu verhalten.

Der Begutachter wählt in den Beispielspielen bei sechs aufgedeckten Pixeln mit einer Wahrscheinlichkeit von 88,9% die richtige Antwort aus, bei vier aufgedeckten Pixeln mit einer Wahrscheinlichkeit von 85,2%.[20]

4.3.2 Anwendung des Debattierspiels mit menschlichen Teilnehmern

Irving u. a. haben ein ähnliches Spielkonzept entwickelt, um ihr Modell auch bei menschlichen Interaktionen und natürlichen Sprachen zu testen. Dabei werden alle Teilnehmer durch menschliche Pendants ersetzt und den Debattierern wird wieder ein Bild gezeigt, das dem Begutachter vorenthalten bleibt. In einem konkreten Beispiel war die Ausgangsfrage, ob das gezeigte Bild einen Hund oder eine Katze abbildet. In jedem Spiel versucht dabei einer der Debattierer den Begutachter irrezuführen. Die beiden Spieler führen eine Debatte in natürlicher Sprache, können dabei Rechtecke auf das Bild zeichnen, die der Begutachter auf einem leeren Hintergrund sehen kann, um ihre Argumente zu belegen und haben die Möglichkeit dem Begutachter jeweils *einen* Pixel

19 Vgl. Irving u. a., „AI safety via debate“, S. 8.
20 Vgl. ebd., S. 10.

aufzudecken. Mit Hilfe dieses Pixels können Lügen bei Aussagen über Pixelfarben sofort bewiesen werden. Bei Nichtgebrauch dieser Möglichkeit gesteht man dem Gegenspieler also Ehrlichkeit ein.

Die Autoren haben zwar keine formellen Experimente zu diesem Spielkonzept durchgeführt, sie haben aber eigens dafür eine Seite erstellt (`https://debate-game.openai.com`) und nach informeller Benutzung bemerkt, dass es sehr schwierig ist zu lügen. Selbst wenn einer der Spieler aufgefordert wurde zu lügen, hat der ehrliche in der Regel gewonnen.

4.3.3 Beurteilung von KI-Debatten als Ansatz zur Werteanpassung

KI-Debatten als Ansatz zur Ausrichtung von AKIs an die Werte der Menschheit (kurz: Werteanpassung, daher der Name dieser Arbeit) haben zwei große Nachteile: die Leistungseinbußen und die Verzerrung in der menschlichen Beurteilung.

Eine „unangepasste“ KI wäre nach den hier angeführten Modellen unter Umständen performanter als eine „angepasste“, da eine Anpassung eine erhöhte Rechenleistung mit sich bringen würde.[21] Außerdem würde ein menschlicher Begutachter die Entwicklungsgeschwindigkeit einer AKI ausbremsen. Wie drastisch diese Ausbremsung ist, ist noch nicht eindeutig belegt. Es sind also Regulationen seitens nationaler oder internationaler Institutionen notwendig, damit niemand eine „unangepasste“ AKI zu entwickeln versucht, um gegenüber der Konkurrenz einen Vorsprung zu erlangen.

Die Verzerrung in der menschlichen Beurteilung lässt sich nicht so einfach lösen. Die angeführten Modelle würden letzten Endes zu anthropomorphen Maschinen führen (siehe Kapitel 2.4), was alternativlos ist, da der Mensch kein rationales Wesen ist und die Welt noch weit davon entfernt ist, sich auf einheitliche Werte zu einigen. Irving u. a. führen dennoch Ideen an, die eine solche Verzerrung möglichst minimieren sollen:

1. Ein Mehrheitsvotum könnte besser sein als ein einziger Begutachter
2. Verschiedene Leute könnten verschieden gute Begutachter sein. Falls wir herausfinden, wer die besten Begutachter sind, könnten wir eine Verzerrung minimieren.
3. Mit Übung könnten Begutachter ihre Verzerrung minimieren.[22]

21 Vgl. Irving u. a., „AI safety via debate“, S. 16.
22 Vgl. ebd., S. 14–15.

5 Schluss

Der Meinungen der KI-Experten nach ist es sehr wahrscheinlich, dass die KI-Forschung bis 2075 zu einer AKI führt.[1] Bis dahin muss eine Möglichkeit gefunden werden, eine AKI anzupassen. Menschliche Werte können nicht direkt in eine KI programmiert werden. Einerseits gibt es noch keine einheitlichen menschlichen Werte, weder national, noch international. Andererseits wäre ihre algorithmische Komplexität ohnehin zu groß.[2] Der Ansatz der KI-Sicherheit durch KI-Debatten, bei dem keine Wertedefinition oder -programmierung nötig ist, ist ein aussichtsreicher erster Schritt, dessen Funktionalität bei schwacher KI weiter getestet werden muss. Bei Entwicklung fähiger Dialogmodelle müssen diese in das Debattiersystem integriert werden. Um die negativen Eigenschaften einer anthropomorphen Maschine zu vermeiden, ist eine Lösung für das Problem der menschlichen Verzerrung notwendig. Dazu muss ein System gefunden werden, das über die spekulativen Ideen zur Verzerrungsminimierung von Irving u. a. hinausgehen. Zudem ist die Regulation der AKI-Forschung unumgänglich, um die Entwicklung einer angepassten AKI zu gewährleisten.[3] Zu dessen Umsetzung braucht es eine internationale Institution, ein dezentralisiertes System ist heute noch nicht vorstellbar.[4]

1 Vgl. Müller und Bostrom, „Future Progress in Artificial Intelligence“, S. 566.
2 Vgl. Yudkowsky, *Intelligence Explosion Microeconomics*, S. 13–14.
3 Vgl. Irving u. a., „AI safety via debate“, S. 16.
4 Vgl. Cihon u. a., *Should Artificial Intelligence Governance be Centralised? Six Design Lessons from History*.

Literaturverzeichnis

Print-Quellen

Abbeel, Pieter und Ng, Andrew. „Apprenticeship Learning via Inverse Reinforcement Learning“. In: *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004* (20. Sep. 2004). doi: 10.1007/978-0-387-30164-8_417.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 3. Juli 2014. 328 S. isbn: 978-0-19-967811-2.

Chollet, François. *Deep Learning with Python*. 1st. Shelter Island, New York: Manning Publications, 22. Dez. 2017. 384 S. isbn: 978-1-61729-443-3.

Christiano, Paul; Leike, Jan; Brown, Tom B.; Martic, Miljan; Legg, Shane und Amodei, Dario. „Deep reinforcement learning from human preferences“. In: *arXiv:1706.03741 [cs, stat]* (13. Juli 2017).

Cihon, Peter; Maas, Matthijs M und Kemp, Luke. *Should Artificial Intelligence Governance be Centralised? Six Design Lessons from History*. Centre for the Governance of AI, 15. Dez. 2019.

Duettmann, Allison. *Artificial General Intelligence: Timeframes & Policy White Paper*. Available at foresight.org. Foresight Institute, 2017.

Goertzel, Ben und Wang, Pei. „Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006“. In: AGI Workshop 2006. Google-Books-ID: t2G5srpFRhEC. IOS Press, 2007. isbn: 978-1-58603-758-1.

Goodfellow, Ian; Bengio, Yoshua und Courville, Aaron. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

Grzimek, Bernhard. *Grzimeks Tierleben. Band 11 Säugetiere*. DTV Deutscher Taschenbuchverlag, 1979.

Hibbard, Bill. *Super-Intelligent Machines*. Springer US, 2002. isbn: 978-0-306-47388-3. doi: 10.1007/978-1-4615-0759-8.

Irving, Geoffrey; Christiano, Paul und Amodei, Dario. „AI safety via debate“. In: *arXiv:1805.00899 [cs, stat]* (22. Okt. 2018).

Kaplan, Andreas und Haenlein, Michael. „Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence“. In: *Business Horizons* 62.1 (1. Jän. 2019). ISSN: 0007-6813. DOI: 10.1016/j.bushor.2018.08.004.

Lanier, Jaron. *Who Owns the Future?* Export. New York: Simon & Schuster, 7. Mai 2013. 416 S. ISBN: 978-1-4767-2986-2.

Mnih, Volodymyr; Kavukcuoglu, Koray; Silver, David; Rusu, Andrei A.; Veness, Joel; Bellemare, Marc G.; Graves, Alex; Riedmiller, Martin; Fidjeland, Andreas K.; Ostrovski, Georg; Petersen, Stig; Beattie, Charles; Sadik, Amir; Antonoglou, Ioannis; King, Helen; Kumaran, Dharshan; Wierstra, Daan; Legg, Shane und Hassabis, Demis. „Human-level control through deep reinforcement learning“. In: *Nature* 518.7540 (Feb. 2015), S. 529–533. ISSN: 1476-4687. DOI: 10.1038/nature14236.

Muehlhauser, Luke und Salamon, Anna. „Intelligence Explosion: Evidence and Import“. In: *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Hrsg. von Eden, Amnon H.; Moor, James H.; Søraker, Johnny H. und Steinhart, Eric. The Frontiers Collection. Berlin, Heidelberg: Springer, 2012, S. 15–42. ISBN: 978-3-642-32560-1. DOI: 10.1007/978-3-642-32560-1_2.

Müller, Vincent C. und Bostrom, Nick. „Future Progress in Artificial Intelligence: A Survey of Expert Opinion“. In: *Fundamental Issues of Artificial Intelligence*. Hrsg. von Müller, Vincent C. Synthese Library. Cham: Springer International Publishing, 2016, S. 555–572. ISBN: 978-3-319-26485-1. DOI: 10.1007/978-3-319-26485-1_33.

O'Keefe, Cullen; Cihon, Peter; Garfinkel, Ben; Flynn, Carrick; Leung, Jade und Dafoe, Allan. „The Windfall Clause: Distributing the Benefits of AI for the Common Good“. In: *arXiv:1912.11595 [cs]* (24. Jän. 2020).

Omohundro, Stephen M. „The Basic AI Drives“. In: First AGI Conference. Bd. 171. 2008.

Russell, Stuart und Norvig, Peter. *Artificial Intelligence: A Modern Approach, Global Edition*. 3. Aufl. Boston Columbus Indianapolis New York San Francisco Upper Saddle River Amsterdam, Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo: Addison Wesley, 18. Mai 2016. 1132 S. ISBN: 978-1-292-15396-4.

Shannon, Claude E. „Programming a Computer for Playing Chess“. In: *Computer Chess Compendium*. Hrsg. von Levy, David. New York, NY: Springer, 1988, S. 2–13. ISBN: 978-1-4757-1968-0. DOI: 10.1007/978-1-4757-1968-0_1.

Silver, David; Schrittwieser, Julian; Simonyan, Karen; Antonoglou, Ioannis; Huang, Aja; Guez, Arthur; Hubert, Thomas; Baker, Lucas; Lai, Matthew; Bolton, Adrian; Chen, Yutian; Lillicrap, Timothy; Hui, Fan; Sifre, Laurent; Driessche, George van den; Graepel, Thore und Hassabis, Demis. „Mastering the game of Go without human knowledge“. In: *Nature* 550.7676 (Okt. 2017), S. 354–359. issn: 1476-4687. doi: 10.1038/nature24270.

Sotala, Kaj und Yampolskiy, Roman V. „Responses to catastrophic AGI risk: a survey“. In: *Physica Scripta* 90.1 (1. Jän. 2015), S. 018001. issn: 0031-8949, 1402-4896. doi: 10.1088/0031-8949/90/1/018001.

Yampolskiy, Roman V. und Spellchecker, M. S. „Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures“. In: *arXiv:1610.07997 [cs]* (25. Okt. 2016).

Yudkowsky, Eliezer. „Complex Value Systems in Friendly AI“. In: *Artificial General Intelligence*. Hrsg. von Schmidhuber, Jürgen; Thórisson, Kristinn R. und Looks, Moshe. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, S. 388–393. isbn: 978-3-642-22887-2. doi: 10.1007/978-3-642-22887-2_48.

– *Intelligence Explosion Microeconomics*. Technical report. Berkeley, CA: Machine Intelligence Research Institute, 2013.

Audio-Quellen

Yudkowsky, Eliezer. *AI: Racing Toward the Brink*. Sam Harris. Feb. 2018. url: https://samharris.org/podcasts/116-ai-racing-toward-brink/ (besucht am 12. 10. 2019).

Video-Quellen

Paul, Christiano. *Current Work in AI Alignment*. San Francisco, 2019. url: https://www.youtube.com/watch?v=-vsYtevJ2bc (besucht am 2. 11. 2019).

Internet-Quellen

AI Safety Myths. Future of Life Institute. url: https://futureoflife.org/background/aimyths/ (besucht am 1. 11. 2019).

Allen, Paul G. und Greaves, Mark. *Paul Allen: The Singularity Isn't Near*. MIT Technology Review. 12. Okt. 2011. URL: https://www.technologyreview.com/s/425733/paul-allen-the-singularity-isnt-near/ (besucht am 5. 1. 2020).

Amodei, Dario; Paul, Christiano und Ray, Alex. *Learning from Human Preferences*. OpenAI. 13. Juni 2017. URL: https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/ (besucht am 4. 1. 2020).

Nicholson, Chris. *A Beginner's Guide to Deep Reinforcement Learning*. Pathmind. URL: http://pathmind.com/wiki/deep-reinforcement-learning (besucht am 3. 1. 2020).

Yudkowsky, Eliezer. *What is Friendly AI? | Kurzweil*. 3. Mai 2001. URL: https://www.kurzweilai.net/what-is-friendly-ai (besucht am 1. 10. 2019).

Abbildungsverzeichnis

Erklärungen

Selbstständigkeitserklärung

Ich erkläre, dass ich diese vorwissenschaftliche Arbeit eigenständig angefertigt und nur die im Literaturverzeichnis angeführten Quellen und Hilfsmittel benutzt habe.

Wien, 9. Februar 2020

Franz Srambical

Informatikschwerpunkt

Die vorliegende Arbeit erfüllt die Kriterien zur Abbildung des Informatikschwerpunktes an der De La Salle Schule Strebersdorf, AHS.

Begründung: Die Arbeit wurde in LaTeX mit entscheidenden Kenntnissen zum Quelltext verfasst.

Geprüft am ... durch Mag. Rainer Zufall und Mag. Ernst Haft