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

Entwurf:
4. Jänner 2020

Abstract

Der Zusammenfassungstext kommt hier her. Abstract ist kein Vorwort und keine Einleitung!

Vorwort

Das Vorwort ist optional: d. h. man muss kein Vorwort schreiben! Wer will, kann das in dieser Form tun. Am Ende sollten Ort, Datum und der Name des Autors des Vorworts angegeben werden.[1]

Wien am 4. Jänner 2020

Franz Srambical

1 Vgl. Weigl, Huberta. *Vorwort.* URL: `http://www.ahs-vwa.at/pluginfile.php/31/mod_data/content/1315/02-VWA-Vorwort.pdf` (besucht am 3. 2. 2017).

Inhaltsverzeichnis

1. Einleitung

Ich möchte diese Arbeit mit einem Gedankenexperiment beginnen.

Es existiere ein System, dass durch ein quantitativ und qualitativ höheres Lernniveau in der Lage ist, Ziele zu erreichen, die die Menschheit ohne eine solches System nicht erreichen könnte. Der Eigentümer einer Büroklammernfabrik ist im Besitz eines solchen Systems und gibt 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]

Es ist durchaus möglich, dass ein solches System mit einer allgemeinen künstlichen Intelligenz beim Erreichen der ihnen vorgegebenen Ziele nebenbei die gesamte Menschheit auslöscht.

Was rechtfertigt diese technovolatile Haltung?

> *„There are all sorts of extreme forces coming onto the game board that were not there before. To expect them to all fail or exactly cancel out for the purpose of making the outcome normal would be one heck of a coincidence.“*[2]

Jede technologische Neuentdeckung bedeutet in erster Linie Veränderung. Die Erfindungen der letzten Jahrhunderte hatten mehrheitlich positive Auswirkungen zur Folge, sonst wäre unser Lebensstandard heute nicht der höchste in der Menschheitsgeschichte.[3] So ermutigend das auch klingt, so dürfen wir nicht einfach nach dem Trend der Vergangenheit in die Zukunft extrapolieren, sondern müssen – so Easterlin – versuchen, die Kräfte zu verstehen, die für den Anstieg der Lebensqualität verantwortlich sind.[4] Was eine allgemeine künstliche Intelligenz betrifft, müssen wir sie nicht nur verstehen, sondern auch lenken können, um das Wohlbefinden der Spezies Mensch nicht zu gefährden, sondern zu stärken.

---

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 *Eliezer Yudkowsky on Intelligence Explosion - YouTube*. url: `https://www.youtube.com/watch?v=D6peN9LiTWA` (besucht am 7. 8. 2019), 30:51–31:07.

3 Vgl. Easterlin, Richard A. „The Worldwide Standard of Living since 1800“. In: *The Journal of Economic Perspectives* 14.1 (2000), S. 7–26. issn: 0895-3309. url: `https://www.jstor.org/stable/2647048`, S. 22–23.

4 Vgl. ebd., S. 23.

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/`, 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. Definition von künstlicher 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 gewonnen 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. Definition von allgemeiner künstlicher 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]

---

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.

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 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. Aber auch Missbrauch in Form einer Machtkonzentration oder Ähnlichem muss unter allen Umständen vermieden werden.

Der Ansatz eine *antropomorphe* Maschine, also ein System mit menschenähnlichen Eigenschaften, zu entwickeln, gilt deshalb als veraltet. 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.

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

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

10 Vgl. Yudkowsky, Eliezer. *What is Friendly AI? | Kurzweil*. 5. März 2001. URL: `https://www.kurzweilai.net/what-is-friendly-ai` (besucht am 1. 10. 2019).

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

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 ist es nicht unwahrscheinlich, dass die KI dies als Glücklichsein 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 KI-Wissenschaftler V. C. Müller und N. 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.

2.6. Die These der Intelligenzexplosion

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

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.

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.

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)[15]

15 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.

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 Antropomorphismus 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 6. 8. 2019).

schon seit jeher in der Geschichte des Menschen zu beobachten. Ich erkäre mir den Antropomorphismus 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. Destruktives Potential

https://80000hours.org/podcast/episodes/allan-dafoe-politics-of-ai/ https://80000hours.org/topic/pri
paths/ai-policy/

3.2.2. Machtverschiebung -und konzentration

3.2.3. Missbrauch

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

4. Maschinelle Werteanpassung

Es ist schwer menschliche Werte in Computersystemen 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* (DRL, dt. *mehrschichtiges bestärkendes Lernen*; siehe Abbildung 4.3). 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, 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, 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).

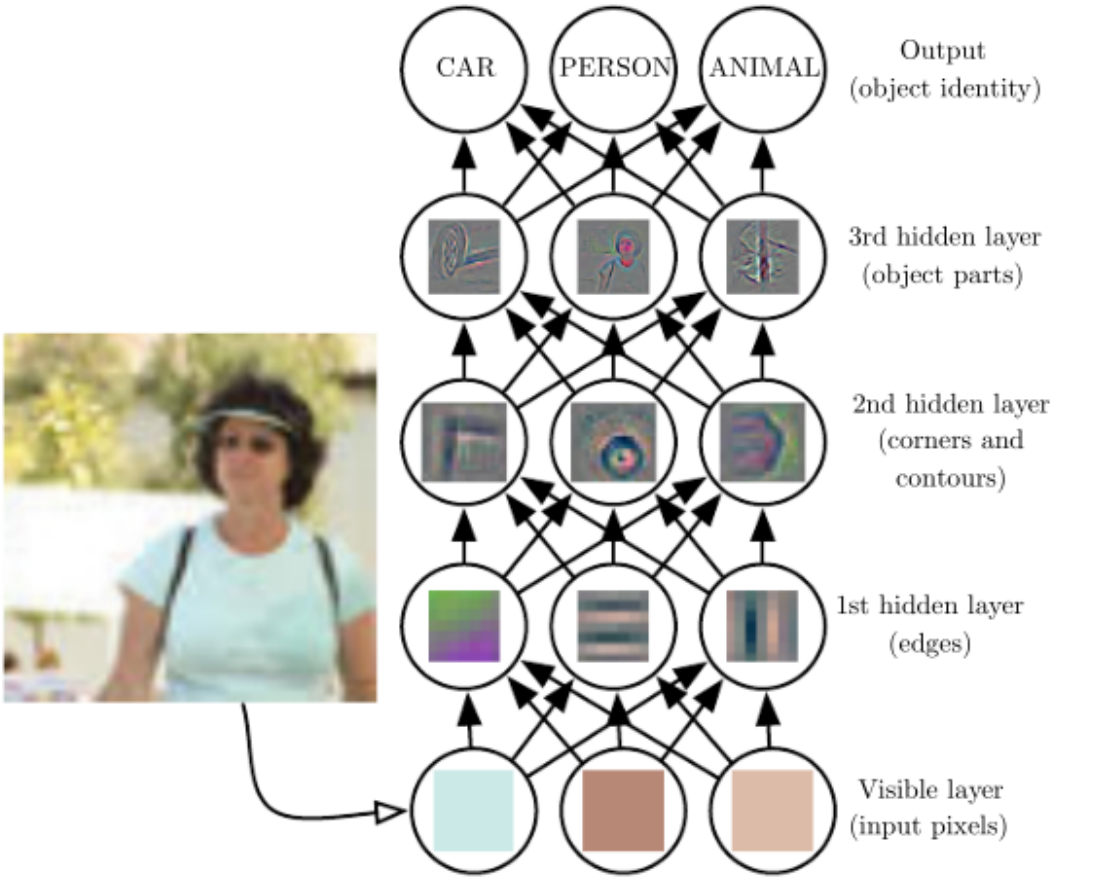

Abbildung 4.1.: Veranschaulichung eines DL-Modells. Die KI bekommt rohe Pixeldaten als Input. Mit jeder Schicht wendet sie ein neues Konzept auf das vorherige an, die Konzepte sind also aufbauend. Durch Analyse der Helligkeit umgebener Pixeln werden Ränder erkannt (1. Schicht). Ansammlungen von Rändern 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]

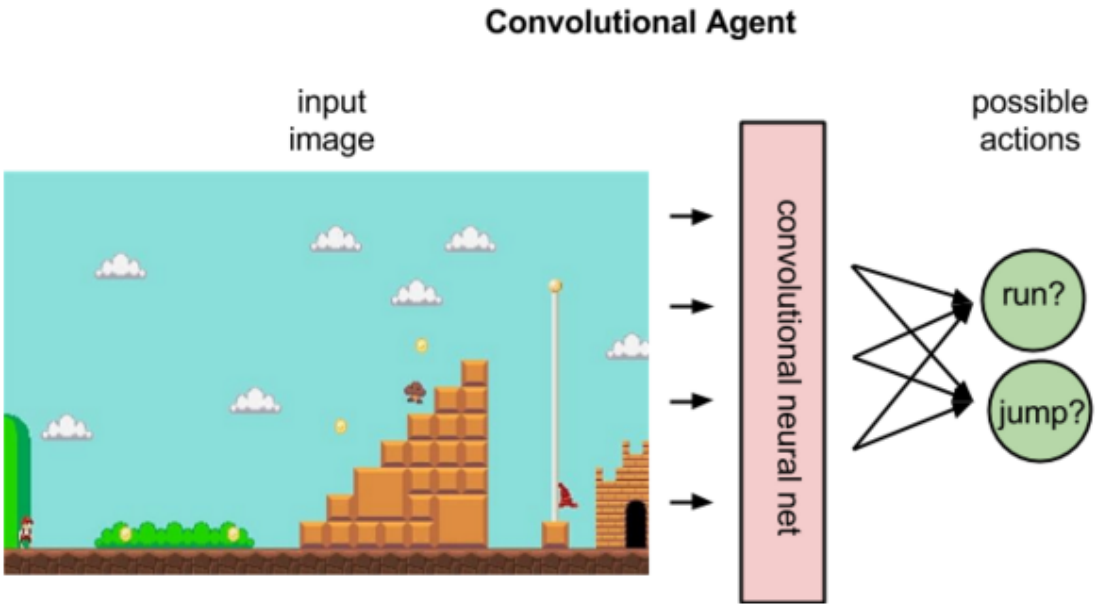

Abbildung 4.2.: Die Umgebung ist das Level, in dem sich Mario (links unten zu sehen) befindet, die möglichen Aktionen sind: springen, nach links laufen, nach rechts laufen. Die neuronalen Netze teilen jeder Aktion einen Nutzwert zu. Beispiel: springen (5), nach rechts laufen (7), nach links laufen (0). 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). arXiv: 1706.03741, 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.

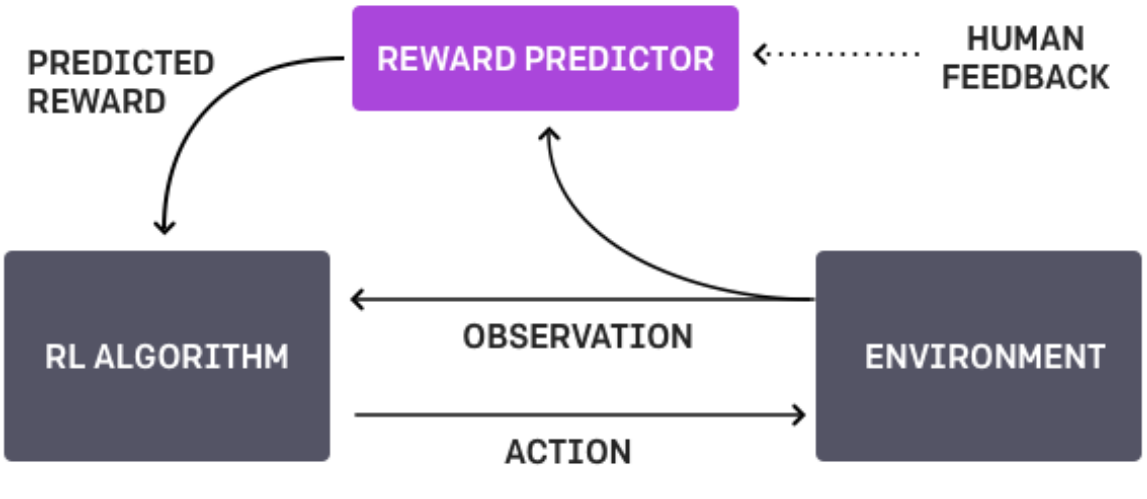

Abbildung 4.3.: Repräsentation einer human-feedback-loop Bildquelle: AMODEI, Dario u. a. *Learning from Human Preferences*. OpenAI. 13. Juni 2017. URL: `https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/` (besucht am 4. 1. 2020)

4.3. KI-Sicherheit durch KI-Debatten

DRL von menschlichen Werten ist ein funktionierender Ansatz, damit eine (A)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.[16]

4.4. AI Safety via Debate

4.5. Inverse Reward Design

16 Vgl. IRVING, Geoffrey u. a. „AI safety via debate“. In: *arXiv:1805.00899 [cs, stat]* (22. Okt. 2018). arXiv: 1805.00899, S. 1–2.

5. Schluss

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, 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). arXiv: 1706.03741.

Easterlin, Richard A. „The Worldwide Standard of Living since 1800“. In: *The Journal of Economic Perspectives* 14.1 (2000), S. 7–26. issn: 0895-3309. url: https://www.jstor.org/stable/2647048.

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). arXiv: 1805.00899.

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.

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.

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, 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.

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/.

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 6. 8. 2019).

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).

Eliezer Yudkowsky on Intelligence Explosion - YouTube. URL: https://www.youtube.com/watch?v=D6peN9LiTWA (besucht am 7. 8. 2019).

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

WEIGL, Huberta. *Vorwort*. URL: http://www.ahs-vwa.at/pluginfile.php/31/mod_data/content/1315/02-VWA-Vorwort.pdf (besucht am 3. 2. 2017).

YUDKOWSKY, Eliezer. *What is Friendly AI? | Kurzweil*. 5. März 2001. URL: https://www.kurzweilai.net/what-is-friendly-ai (besucht am 1. 10. 2019).

Abbildungsverzeichnis

Tabellenverzeichnis

A. Hier könnte Ihr Anhang stehen

Erklärungen

Selbstständigkeitserklärung

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

Wien, 4. Jänner 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