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:
Prof. Mag. Kurt Rauch & Mag. Leonard Michlmayr

Klasse:
8C

Entwurf:
1. Oktober 2019

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 1. Oktober 2019

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 eine so technovolatile Haltung wie diese?

> *„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 Richard A. 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 bestärken.

Kommentare: Werte formulieren die keinen schaden anrichten (Beispiel finden) VERZERRUNG BIASES übersetzung

---

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` (besucht am 9. 8. 2019), 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 Vollständig- oder Richtigkeit beweisen konnte, wird in dieser Arbeit der Einfachheit halber versucht, den Begriff durch Beobachtungen zu erklären, wie Eliezer Yudkowsky in dem Podcast "AI: Racing Toward the Brink" vorschlägt.

1. Menschen waren auf dem Mond.
2. Mäuse waren nicht auf dem Mond.

Yudkowsky wählt dieses Beispiel um zu demonstrieren, dass Menschen auch Orte erreichen, wofür die natürliche Selektion sie nicht vorbereitet hat. Daraus könne geschlossen werden, dass Menschen *intelligenter* als Mäuse sind, weil sie *domänenübergreifend* arbeiten können. Deshalb sei das *domänenübergreifende* Erlernen neuer Fähigkeiten ein zentraler Teil des Intelligenzbegriffs.[1]

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"*[2]

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 die Idee des *domänenübergreifenden* Lernens im Gegensatz zum oben beschriebenen Ansatz zur Intelligenzerklärung nicht, was laut Experten jedoch nicht an einer unvollständigen Definition liegt, sondern vielmehr daran, dass

1 Vgl. Yudkowsky, Eliezer. *AI: Racing Toward the Brink*. Sam Harris. Feb. 2018. URL: https://samharris.org/podcasts/116-ai-racing-toward-brink/, 06:01–09:49.

2 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), S. 15–25. ISSN: 0007-6813. DOI: 10.1016/j.bushor.2018.08.004. URL: http://www.sciencedirect.com/science/article/pii/S0007681318301393 (besucht am 6. 8. 2019).

wir den Begriff der KI in einer Art gebrauchen, wofür 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.[3] Dieser beschreibt eine *domänenspezifische* KI.

2.3. Definition von allgemeiner künstlicher Intelligenz

Als allgemeine künstliche Intelligenz bezeichnet man ein technisch fortgeschrittenes System, dessen Lernkapazität nicht auf einzelne Domänen begrenzt ist, sondern als *allgemein* bezeichnet werden kann.[4]

2.4. Werte einer allgemeinen künstlichen Intelligenz

Es ist essenziell, dass ein so fortgeschrittenes System wie eine AKI die Werte der Menschheit teilt, um ungewollten Nebenwirkungen wie der in der Einleitung genannten Auslöschung der Menschheit durch unpräzises Definieren ihrer Ziele mit größtmöglicher Sicherheit vorzubeugen. Dabei geht es nicht darum, eine - wie Ray Kurzweil es in einem Artikel ausdrückt - antropomorphe Maschine,[5] also ein System mit menschenähnlichen Eigenschaften zu entwickeln, sondern dafür zu sorgen, dass ein solches System einen gleichermaßen positiven Effekt auf die gesamte Menschheit hat.

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 allgemeinen künstlichen Intelligenz in den Jahren 2040 bis 2050 eine Wahrscheinlichkeit von über 50, und dem Erreichen

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

4 Vgl. Goertzel, Ben und Wang, Pei. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms : Proceedings of the AGI Workshop 2006*. Google-Books-ID: t2G5srpFRhEC. IOS Press, 2007. 305 S. isbn: 978-1-58603-758-1, S. 1.

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

bis 2075 eine Wahrscheinlichkeit von 90 Prozent zuordnen.[6] 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 mehr als nur relevant sein wird.

2.6. Die These der Intelligenzexplosion

6 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. url: https://doi.org/10.1007/978-3-319-26485-1_33 (besucht am 5. 9. 2019), S. 566.

3. Probleme einer allgemeinen künstlichen Intelligenz

3.1. Fehlerhafte Vorstellungen einer KI-Katastrophe

3.1.1. Bösartige KI

3.1.2. KI, die ein Bewusstsein erlangt

3.1.3. Roboter als Auslöser einer Katastrophe

3.2. Gesamtmenschheitlicher Konsens über gemeinsame Werte

3.3. "Gute" und "schlechte" menschliche Werte

3.4. Wertekodierung in einer Programmiersprache

3.4.1. Statische Wertekodierung

3.4.2. Dynamisch-maschinelle Werteanpassung

3.5. Biases

3.5.1. Verzerrung in der Risikoeinschätzung

Auch Zeitpunkt einer AKI

3.5.2. Verzerrung in der Werteformulierung

3.5.3. Verzerrung in der Kodierung

Nutzenfunktion (eng. *utility function*)

3.6. Sichere und vertrauenswürdige KI

[1]

3.7. KI-Ethik

1 Vgl. Yudkowsky, Eliezer. „Intelligence Explosion Microeconomics“. In: (), S. 96.

4. Lösungsansätze

4.1. Bestärkendes Lernen

4.2. Reziprok-bestärkendes Lernen

4.3. Mensch-Maschinen-Interface

4.4. Hirnemulation

Literaturverzeichnis

Print-Quellen

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

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 (besucht am 9. 8. 2019).

Goertzel, Ben und Wang, Pei. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms : Proceedings of the AGI Workshop 2006*. Google-Books-ID: t2G5srpFRhEC. IOS Press, 2007. 305 S. isbn: 978-1-58603-758-1.

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), S. 15–25. issn: 0007-6813. doi: 10.1016/j.bushor.2018.08.004. url: http://www.sciencedirect.com/science/article/pii/S0007681318301393 (besucht am 6. 8. 2019).

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. url: https://doi.org/10.1007/978-3-319-26485-1_33 (besucht am 5. 9. 2019).

Yudkowsky, Eliezer. „Intelligence Explosion Microeconomics“. In: (), S. 96.

Audio-Quellen

Yudkowsky, Eliezer. *AI: Racing Toward the Brink*. Sam Harris. Feb. 2018. url: https://samharris.org/podcasts/116-ai-racing-toward-brink/.

Internet-Quellen

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

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*. What is Friendly AI? 3. Mai 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, 1. Oktober 2019

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