**Projektbericht «Gefahren der AI-Entwicklung»**

**Einleitung**

Die Künstliche Intelligenz (Im Projekt wird die Abkürzung «AI» verwendet) gewinnt zunehmend an Wichtigkeit in unserer Gesellschaft. Obschon die Forschung bezüglich AI schon in den 50er-Jahren begonnen hatte, ist das Thema erst in den letzten Jahren hochaktuell geworden. Gründe dafür sind technologischer Fortschritt, aber auch Notwendigkeit in Industrie, Kommerz, Militär und Gesellschaft. Das positive Ertragspotential solcher Technologie ist immens. Und so sind auch die Gefahren. Es lässt sich nur schwer prognostizieren, wie sich die Menschheit mit der Verwendung von AI entwickeln wird. Klar ist aber, dass sich die AI schon in der heutigen Welt nicht mehr wegdenken lässt. Zukünftige Generationen der AI-Technologien werden wirtschaftlich effizienter und mächtiger werden.

Viele Personen aus der Politik, Industrie, Akademie und Medien[[1]](#footnote-1) haben sich zur Verwendung von AI geäussert. Neben Proponenten der AI gibt es auch sehr kritische Ansichten, die eine oft dystopische Zukunft voraussagen. Im Jahr 2015 verfassten Wissenschaftler des Institutes «Future of Life»[[2]](#footnote-2) einen «Open-Letter»[[3]](#footnote-3); den Bericht «Research Priorities for Robust and Beneficial AI», welches auf die verschiedenen Gefahren und Chancen in der AI-Entwicklung hinweist. Dieses 10-seitige Dokument benutzten wir als Einstiegshilfe und mithilfe von Artikeln und Papiern aus der ausführlichen Quellenangabe, auf die sich dieser "Open Letter" stützt, konnten wir die verschiedenen Argumente rekonstruieren. Wir beschränkten uns dabei auf die Gefahren der AI-Entwicklung.

Wir zielten auf folgende Erkenntnisse ab:

* Wir wollten sehen, wie weit die AI schon heute verwendet wird und wie weit sie in Gebieten wie Wirtschaft, Militär und Gesellschaft eine Rolle spielen. Durch die rapide Erweiterung der Einsatzgebiete der AI nimmt sie immer stärkeren Einfluss und hat Auswirkungen auf unseren Alltag. In welchen Bereichen stellt die wachsende Zunahme der Verwendung von AI eine Gefahr dar und hat Auswirkung auf die Zukunftsgestaltung?
* Durch die Entwicklung einer Superintelligenz[[4]](#footnote-4) besteht die Gefahr einer Singularität[[5]](#footnote-5). Stellt diese Entwicklung ein existenzielles Risiko für die Menschheit dar?

**Themenfindung und Textgrundlage**

Als spannendes und aktuelles Thema ist das Gebiet der AI, dessen Anwendungspotenzial sehr weitgehend ist und die dazugehörenden Herausforderungen fundamentale Veränderungen in der Gesellschaft hervorbringen können, bot sich eine Argumentationsrekonstruktion bezüglich der Gefahren der AI-Entwicklung gut an. Da die Thematik sehr breit ist und auch stark in die Tiefe geht, war eine Einschränkung der Fragestellungen im Rahmen dieses Projektes sehr wichtig. Das Projekt sollte eine gute Übersicht über potenzielle Gefahren aus verschiedenen Thematiken wie Wirtschaft, Militär, Gesellschaft und Ethik liefern. Dabei wurden grundlegende Fragen aus der Philosophie des Geistes bezüglich Bewusstseins, Identität und dem Turing-Test ausgeklammert.

Der erste Versuch war, eine prominente Person zu finden, welche aktiv an der Diskussion um die Gefahren der AI-Entwicklung teilnimmt. Eine solche Person, die eine kritische Position zu AI vertritt und uns auch auf mögliche Gefahren hinweist, ist Elon Musk[[6]](#footnote-6). Durch seine Geschäftstätigkeiten in der autonomen Verkehrsführung mit «Tesla», der Gründung von «OpenAI» (Entwicklung einer Open-Source-KI) und der Firma «Neuralink» (Entwicklung einer Gehirn-Maschine-Schnittstelle), sowie seiner Präsenz in den Medien, dachten wir, er biete sich hervorragend für eine Argumentationsanalyse an. Doch wurde nach eingehender Recherche festgestellt, dass Elon Musk nur sehr wenige Argumente vorträgt. Er beschränkt sich im Rahmen der Interviews und Symposien, welche wir untersucht haben, auf oberflächliche Aussagen und für Medien attraktive «Sound-bites». Deshalb wurde bald klar, dass wir eine viel fundiertere Textgrundlage finden mussten.

Durch weitere Recherche stiessen wir auf den «Open Letter»[[7]](#footnote-7) des «Future of Life»-Institutes. Der 10-seitige Bericht mit dem Titel «Research Priorities for Robust and Beneficial AI» bietet eine hervorragende Übersicht zur Thematik und enthält sehr viele Quellen zu wissenschaftlichen Artikeln, Bücher und Papieren und fand grosse Unterstützung unter Wissenschaftlern, Politikern, den Medien und Technologieschaffenden. Wir haben diesen Bericht genau durchgearbeitet und die Quellen gesammelt, auf welche sich der Bericht stützt. Diese Quellen haben wir nach Thesen untersucht und nach Argumenten geprüft, welche sich für eine Rekonstruktion eignen. Der «Open Letter» und die Quellen ergaben unsere Textgrundlage. Nach genauem Durchlesen und Strukturieren des «Open Letters» hat sich aber gezeigt, dass der Umfang dieses Berichtes den Rahmen des Projektes sprengt. Deshalb musste eine Auswahl getroffen werden, die einen guten Überblick zur Thematik lieferte, eine spannende Argumentationsanalyse ermöglichte und im Zeitplan durchführbar war.

**Organisation**

Um die interne Kommunikation zu vereinfachen, das Material übersichtlich geordnet zu halten und unsere Zusammenarbeit und Versionierungen effizient zu gestalten, richteten wir auf «Github»[[8]](#footnote-8) ein Repository ein. So blieben wir immer auf dem neusten Stand und konnten unseren Arbeitsverlauf kontrollieren. Auf Zotero[[9]](#footnote-9) haben wir eine Bibliothek angelegt, um die Quellenverweise dynamisch verwalten zu können

**Arbeitsteilung und erste Schritte**

Im Team wurde die zu untersuchenden Artikel grob nach kurzfristigen und längerfristigen Gefahren aufgeteilt. Während Claude zuerst das Buch «Superintelligence»[[10]](#footnote-10) des Philosophen und Futurologen Nick Bostrom untersuchte und sich somit in die längerfristigen Gefahren der AI-Entwicklung einarbeitete, recherchierte Sebastian das Buch «The Second Machine Age»[[11]](#footnote-11) von Erik Brynjolfsson und Andrew McAfee, welches den Fokus auf den wirtschaftlichen Auswirkungen der AI legt.

**Resultate der ersten Phase**

Die ersten Resultate waren vielversprechend.

…

(erste karten)

(erste erkenntnisse)

(erste probleme)

(weitere quellen)

(strukturierungsphase)

(endprodukt)

**Fazit**

**Bibliografie:**

Open Letter, «Research Priorities for Robust and Beneficial AI», 2015.

Bostrom, Nick, «Superintelligence: Paths, Dangers, Strategies», 2014.

Brynjolfsson, Erik and McAfee, Andrew, «The Second Machine Age», 2014.

Calo, Ryan, «The Case for a Federal Robotics Commission», 2014.

Docherty, Bonnie, «Losing humanity: the case against killer robots», 2012.

Mokyr, Joel, «Secular stagnation? Not in your life», 2014.

Wallach, Wendell & Allen, Colin, «Moral Machines: Teaching Robots Right from Wrong», 2008.

Future of Life-Institute, online unter: <https://futureoflife.org/> (Zugriff: 10.07.2020).

Github repository, online unter: <https://github.com/flicksolutions/musk> (Zugriff: 10.07.2020).

Zotero, online unter: <https://www.zotero.org/groups/2463181/musk_argumentationsanalyse/collections/2VWWS9ZF> (Zugriff: 10.07.2020).

Alle Texte sind auf Github verfügbar:

<https://github.com/flicksolutions/musk/tree/master/quellen> (Zugriff: 10.07.2020).

Zotero: Auf Zotero haben wir eine Bibliothek angelegt, um die Quellenverweise dynamisch verwalten zu können:

<https://www.zotero.org/groups/2463181/musk_argumentationsanalyse/collections/2VWWS9ZF>

**Argdown-Karte:**

Die Argdown-Karte ist auf Github verfügbar:

Als HTML: <https://flicksolutions.github.io/musk/output/research-priorities.html>

Einzelne Argdown-Dateien:

<https://github.com/flicksolutions/musk/blob/master/argdown/_research-priorities.argdown>

<https://github.com/flicksolutions/musk/blob/master/argdown/_argumente_claude_200510.ad>

<https://github.com/flicksolutions/musk/blob/master/argdown/research-priorities_hirarchy.argdown>

Ein PDF der Argumentationskarte ist auf Github verfügbar:

<https://github.com/flicksolutions/musk/blob/master/output/research-priorities.pdf>

**Projektstatus 14.05.2020:**

Wir haben durch die Auswahl der Texte, welche wir bearbeitet und Argumente daraus rekonstruiert haben, einen guten Überblick der Gefahren der AI-Entwicklung gewonnen. Die Kernaussagen der einzelnen Texte sind analysiert und als Argumente soweit rekonstruiert worden, dass die meisten Argumente an der Sitzung vom 19.05.2020 präsentiert werden können. Durch die breite Natur der Thematik lassen sich eine Vielzahl an weiteren Artikeln und Papern finden, doch mussten wir die Recherche dem Zeitaufwand im Rahmen dieses Projektes anzupassen. So erachten wir das Spektrum der Texte als nahezu komplett. Die Argumente im Einzelnen werden iterativ weiter verbessert und ergänzt. Die Erkenntnisse aus der Sitzung werden zu weiteren Anpassungen führen.

Als momentane Schwierigkeiten im Projekt haben sich drei Punkte herauskristallisiert:

* Übersetzung aus dem Englischen: Wir haben uns entschieden, alle Argumente zu übersetzen. Durch die teilweise komplexe Formulierung mit vielen spezifischen Fachausdrücken in den Texten hat sich die Übersetzung als teilweise schwierig herausgestellt. Gewisse (Fach-) Ausdrücke lassen sich entweder gar nicht oder nur schwer übersetzen, ohne den Inhalt geringfügig zu verändern.
* Konjunktiv und Stärke der Argumentation: Teilweise werden Argumente aufgestellt, welche auf Auswirkungen in der Zukunft abzielen. Dies führt dazu, dass viele Texte auf «es könnte sein», «es kann Gefahren beinhalten», «die Zukunft kann so aussehen», etc. aufgebaut sind.
* Verknüpfung der einzelnen Argumente und Darstellung als Argdown-Karte. Es gibt verschiedene Möglichkeiten, die Argumente zu gruppieren und miteinander zu verknüpfen. Auch hier wird kontinuierlich daran gearbeitet, um die beste Balance zwischen Information und Übersichtlichkeit in der Darstellungsweise zu erlangen.
* Manipulation einzelner Nodes, z.B. des Hauptargumentes.

**Schritte nach dem 19.05.2020:**

* Die Formulierung weiter verbessern.
* Die Argumente finalisieren, dass alle deduktiv gültig sind.
* Fehlende oder ergänzende Argumente implementieren.
* Die Struktur der Karte festigen.
* Projektbericht erstellen, und somit
* Projekt abschliessen.

1. Ein paar prominente Namen sind: Mark Zuckerberg, Elon Musk, Stephen Hawking, Steve Wozniak, Bill Gates und viele weitere. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)
3. Future of Life-Institute: Online unter: <https://futureoflife.org/> (Zugang: 10.07.2020). [↑](#footnote-ref-3)
4. Eine «Superintelligenz» ist ein hypothetischer Agent, welcher eine Intelligenz besitzt, die menschliche Intelligenz bei weitem überschreitet. [↑](#footnote-ref-4)
5. Ein Zeitpunkt, ab dem Maschinen sich selbst verbessern können, den technischen Fortschritt massiv beschleunigen und so unkontrollierbare und irreversible Folgen mit sich bringt. [↑](#footnote-ref-5)
6. [↑](#footnote-ref-6)
7. [↑](#footnote-ref-7)
8. Github repository, online unter: <https://github.com/flicksolutions/musk> (Zugriff: 10.07.2020). [↑](#footnote-ref-8)
9. Zotero, online unter: https://www.zotero.org/groups/2463181/musk\_argumentationsanalyse/collections/2VWWS9ZF (Zugriff: 10.07.2020). [↑](#footnote-ref-9)
10. [↑](#footnote-ref-10)
11. [↑](#footnote-ref-11)