AI Box

Kann man eine KI einsperren?

Matthias Rein

19. Juli 2017

Table of contents

Superintellligenz

"...an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills"

Nick Bostrom

Superintelligenz entsteht möglicherweise sehr plötzlich

$\rightarrow$ Intelligenzexplosion

Gefahren

"*...potentially more dangerous than nukes*"

Elon Musk über unkontrollierte KI

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

Stephen Hawkings

Gefahren - Katastrophale Fehler

Perverse Instantiierung

Endziel: *"Bring uns zum Lächeln!"*

Perverse Instantiierung: *Lähmung der menschlichen Gesichtsmuskeln, um den Mund zu einem permanenten strahlenden Lächeln zu verziehen.*

Endziel: *"Bring uns zum Lächeln, ohne direkt auf unsere Gesichtsmuskeln einzuwirken"*

Perverse Instanntiierung: *Stimulation des Teils des motorischen Kortex, der unsere Gesichtsmuskulatur steuert*

Gefahren - Katastrophale Fehler

Ressourcenvergeudung

Aufgabe: *Prüfe die Riemannsche Vermutung*

Folge: *KI verwandelt das Sonnensystem (samt Menschen) in Computronium*

Gefahren

Vefügung über superintelligentes KI System verleiht unter Umständen große Macht

Fähigkeitenkontrolle vs Motivationskontrolle

- Fähigkeitenkontrolle
  - Handlungsmöglichkeiten der KI einschränken
- Motivationskontrolle
  - Versuch, den "Willen" der KI zu beeinflussen
  - Schaffung einer gutartigen KI ("friendly AI")

AI Box Problem geht von einer potentiell nicht gutartigen KI aus
$\rightarrow$ Fähigkeitenkontrolle nötig

Probleme

- Kann eine den Menschen meilenweit überlegene Intelligenz kontrolliert werden?
- Trade-off Sicherheit/Kontrolle - Nützlichkeit

Orakel, Agent, Werkzeug

Verschiedene mögliche Arten von KI Systemen sind:

- Orakel
  - Ein System, das lediglich Fragen beantwortet
  - Eignet sich gut zur Anwendung von Verwahrungsmaßnahmen
- Werkzeug
  - Kein zielgerichtetes Verhalten
  - Verwahrungsmaßnahmen möglich
- Agent
  - Trifft selbstständig Entscheidungen und führt Handlungen aus, um ein Ziel zu erreichen
  - Fähigkeitenkontrolle und Verwahrungsmaßnahmen scheiden aus

Orakel - Kontrolle

Kontrollmöglichkeiten:

- Faradayischer Käfig
- Output drosseln
- Rechengeschwindigkeit drosseln
- Orakel regelmäßig resetten
- Kurze Laufzeit

Orakel - Kontrolle

Ist das Orakel vertrauenswürdig?

- Testfragen
- Honey-pots

Orakel - Täuschung

Wie sicherstellen, dass das Orakel einen nicht täuscht?

- Auswahl der Fragen
- Mehrere Orakel befragen - Antworten vergleichen

Werkzeuge statt Agenten

Idee: Warum nicht einfach bloße Werkzeuge statt Agenten nutzen?

- Werkzeuge haben keine Nutzenfunktion
  - Es besteht kein Grund für das System, den Nutzer "auszutricksen"
- Beispiele:
  - Google Maps Routenplanung
  - Siri
  - Bildklassifizierung

Werkzeuge statt Agenten

Einwände:

- Agenten werden womöglich ggü Werkzeugen bevorzugt
- Interene Planungs- und Suchprozesse von Werkzeugen verhalten sich evtl. wie Agenten

AI Box Experiment

- Erdacht von E. Yudkowsky 2002
- AI Box Experiment soll zeigen wie einfach Menschen dazugebracht werden könnnen, die AI freizulassen

AI Box Experiment

- Zwei Teilnehmer:
  - Person eins simuliert die KI
  - Person zwei "spielt" Wächter
- Teilnehmer kommunizieren über eine Konsole
- KI muss probieren den Wächter zu überzeugen, sie freizulassen

AI Box Experiment - Regeln

Einige Regeln:

- Der Wächter erhält einen kleinen Geldbetrag, wenn er die KI nicht freilässt
- Der Wächter muss mit der KI über einen festgelegten Zeitraum kommunizieren
- Die KI muss den Wächter tatsächlich überzeugen (keine Tricks)
- Das Gesprächsprotokoll wird nicht veröffentlicht

AI Box Experiment

- Ergebnis: In zwei Fällen konnte der Wächter von der KI überzeugt werden, in einem anderem nicht
- Mögliche Schlussfolgerung: AI Boxing scheitert spätestens durch menschliches Versagen
- Daher besser: Konzentration auf Motivationskontrolle (Friendly AI)

AI Box Experiment

Kritik/Einwände:

- Nichtveröffentlichung des Gesprächsprotokoll
- Nicht wissenschaftlich

AI Box Experiment

Mögliche Schutzmaßnahmen:

- Schulung von Wächtern
- Involvierung von mehreren Wächtern
- Überwachung jeglicher Kommunikation zwischen KI und Menschen
- Einschränkung der Kommunikation

Ende

Fragen?

Diskussion

- Kann das AI Box Problem gelöst werden?
- Welche, der vorgestellten Maßnahmen sind am vielversprechendesten?
- Welche weiteren Maßnahmen wären denkbar?


Yudkowski, *AI Box Experiment*, http://yudkowsky.net/singularity/aibox


Nick Bostrom, *Superintelligence - Paths, Dangers, Strategies*,


Armstrong et al., *Thinking Inside the Box: Controlling and Using an Oracle AI* , http://www.aleph.se/papers/oracleAI.pdf


Karnofsky, *Thoughts on the Singularity Institute - we should prefer Tool AIs over autonomous agents* , http://lesswrong.com/lw/cbs


Branwen, *Why Tool AIs Want to Be Agent AIs* , https://www.gwern.net/Tool%20AI