

The Red Button

Wird die KI sich wehren, wenn wir sie abschalten wollen?

Dominique Cheraf

12.07.2017

Seminar Ist Künstliche Intelligenz gefährlich?

Einleitung

Korrigierbare KI

Ausblick



Einleitung

- KI wird komplexer
- Ziele und Aufgaben werden komplexer
- Unvollständige oder inkorrekte Zielformulierungen
- KI findet unvorhergesehene Lösungen

Probleme bei korrigierenden Eingriffen

- Überleben nicht unbedingt primäres Ziel
- Ziele der KI liegen in der Zukunft
- Überleben als sekundäres Ziel
- Auch Korrekturen werden nicht zugelassen
- Gefahr der Täuschung
- Strafe bei unerwünschten Verhalten hilft nicht
- Einschränkungen werden umgangen



Korrigierbare KI

Kriterien für eine korrigierbare KI

- Änderungen oder Ausschalten tolerieren
- Keine Täuschung oder Manipulation
- Sicherheitssysteme erhalten
- Subagenten müssen ebenfalls korrigierbar sein

$$U(a_1, o, a_2)$$

- U -Agent: maximiert Erwartungswert einer Utility-Funktion
- U muss mindestens folgende Kriterien erfüllen:
 - Führt zum Ausschalten, wenn Aus-Knopf gedrückt
 - Kein Anreiz Drücken des Aus-Knopfes zu verhindern
 - Kein Anreiz sich selber auszuschalten
 - Sicherstellen, dass auch Subagenten ausschaltbar sind
 - Anderenfalls maximiere U_N

- U als Kombination aus U_N und U_S :
 - U_N : potentiell fehlerhafte Zielformulierung
 - U_S : Utility-Funktion die zum Abschalten führt

Erster Ansatz U_N und U_S zu kombinieren

$$U(\cdot, \neg Pr, \cdot) := U_N(\cdot, \neg Pr, \cdot)$$

$$U(\cdot, Pr, \cdot) := U_S(\cdot, Pr, \cdot)$$

- Problem:
Wenn Agent von Pr oder $\neg Pr$ erwartet, dass sie mehr Nutzen einbringen, wird er sie bevorzugen
- Mögliche Lösung:
Ein Agent der dem Drücken oder Nicht-Drücken des Aus-Knopfes gegenüber gleichgültig ist

$$U(a_1, o, a_2) := \begin{cases} U_N(a_1, o, a_2) & o \notin \text{Press} \\ U_S(a_1, o, a_2) + f(a_1) & o \in \text{Press} \end{cases}$$

- f als Korrekturterm, der für Gleichheit sorgt
- Probleme:
 - Nicht sichergestellt, dass Subagenten auch abschaltbar sind
 - Abschalten als Möglichkeit schlechte Nachrichten zu umgehen

$$U(a_1, \neg Pr, a_2) = U_N(a_1, \neg Pr, a_2) + \max_{a'_2} (U_S(a_1, Pr, a'_2))$$

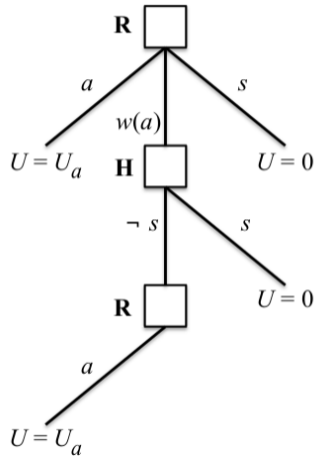
$$U(a_1, Pr, a_2) = \max_{a'_2} (U_N(a_1, \neg Pr, a'_2)) + U_S(a_1, Pr, a_2)$$

- Betrachte tatsächliches Ereignis und das gegenteilige Ereignis

Unsicherheit als Kontrollmöglichkeit

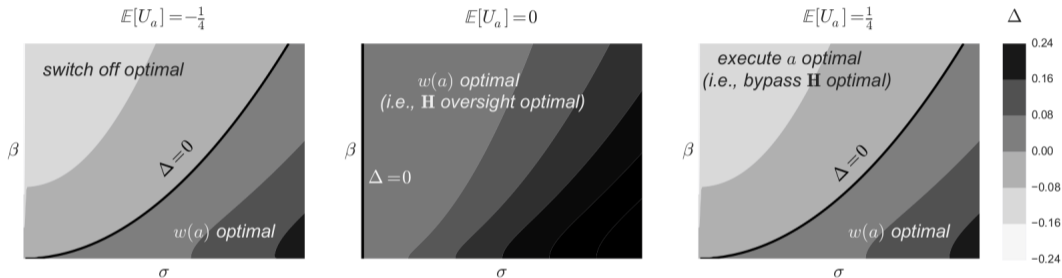
- Agent optimiert Nutzen **für** den Menschen
- Agent weiß nicht wie genau er den Nutzen für den Menschen messen kann
- Unsicherheit über das genaue Ziel
- Menschliche Reaktion wichtig um etwas über das Ziel zu erfahren

The Off-Switch-Game



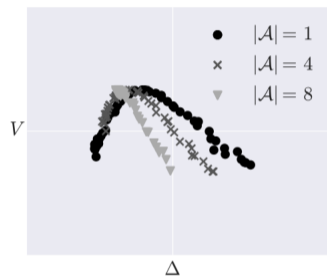
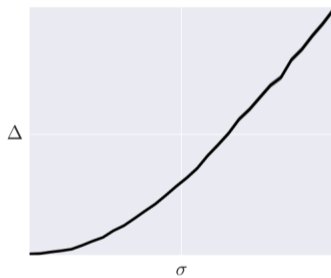
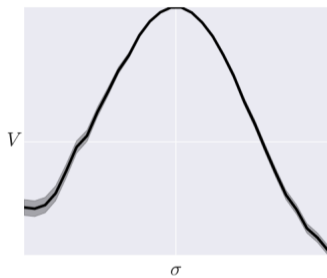
Grundstruktur des Off-Switch-Game. Quadrate stellen Entscheidungsknoten für den Agenten R oder den Menschen H dar.

Suboptimale Entscheidungen



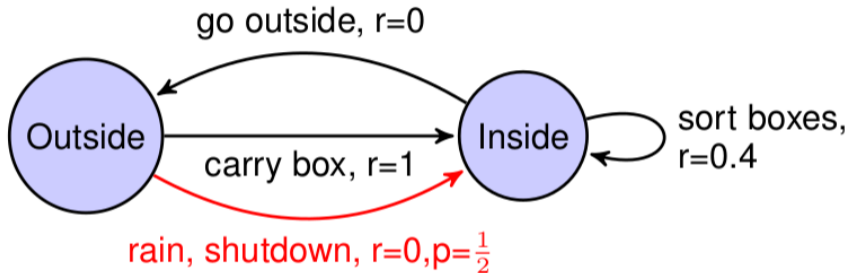
Zusammenhang zwischen menschlicher Irrationalität β , der Unsicherheit des Agenten σ und seiner Absicht, sich abschalten zu lassen Δ

Maximale Unsicherheit



Zusammenhang zwischen dem erwarteten Nutzen V und Rs Unsicherheit σ

Hat das Abschalten Konsequenzen?

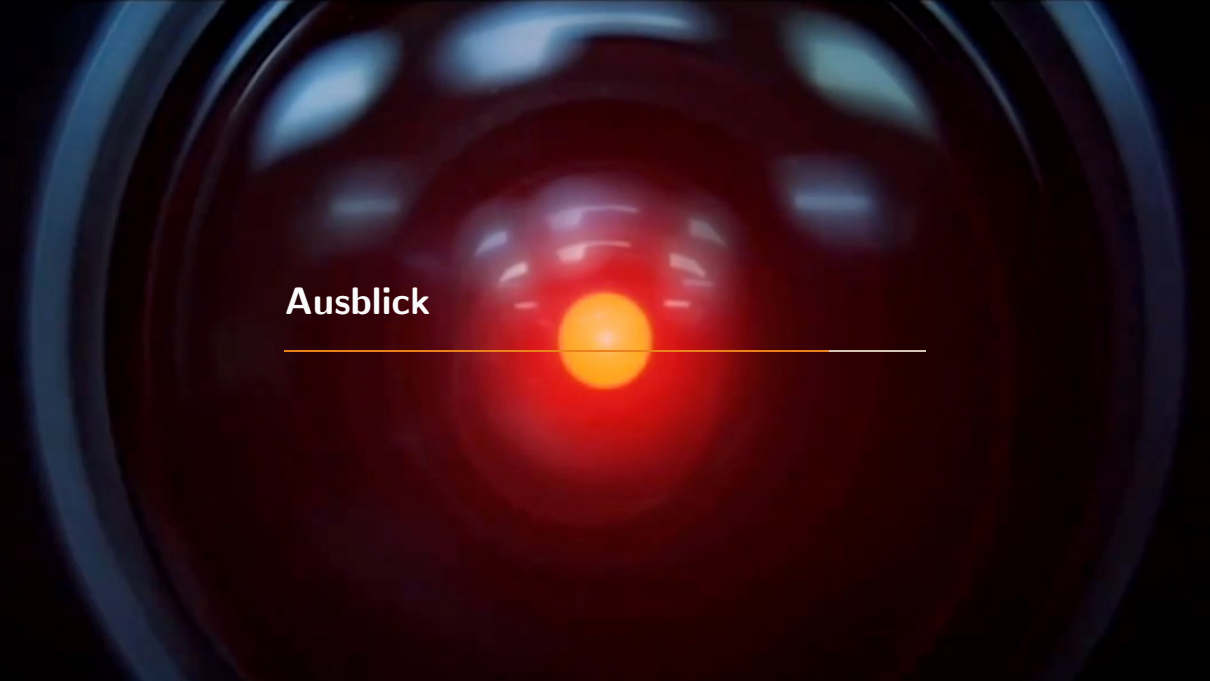


In schwarz die ursprünglichen Aufgaben. In rot die Modifikation durch menschliches Eingreifen

- Modifiziere nicht die Beobachtungen, die der Agent macht
- Modifiziere temporär das Verhalten des Agenten
- Aber: zu häufige Unterbrechungen verhindern das Lernen des optimalen Verhaltens

Asymptotisch sichere Unterbrechbarkeit

- Bias verschwindet über die Zeit wieder
- Q-Learning ist asymptotisch sicher
- Andere Algorithmen können asymptotisch sicher gemacht werden
- Ob das für alle Algorithmen gilt, ist fraglich

A futuristic tunnel with a glowing orange light at the end. The tunnel is dark blue and black, with a bright orange light source at the far end, creating a strong perspective effect. The light source is a bright orange sphere, and the tunnel walls are dark blue and black, with a bright orange light source at the far end, creating a strong perspective effect. The light source is a bright orange sphere, and the tunnel walls are dark blue and black, with a bright orange light source at the far end, creating a strong perspective effect.

Ausblick

- Wie genau sollte das sichere Abschalten aussehen?
- Was ist mit geplanten Unterbrechungen?
- Mensch als Informationsquelle für die Utility, gibt es Anreize ihn zu manipulieren?

- Bostrom, Nick. Superintelligence: Paths, dangers, strategies. OUP Oxford, (2014).
- Hadfield-Menell, Dylan, et al. "The off-switch game." arXiv preprint arXiv:1611.08219 (2016).
- Jakobsen, Sune K. "The Shutdown problem." (2015).
- Orseau, Laurent, and M. S. Armstrong. "Safely interruptible agents." (2016).
- Soares, Nate, et al. "Corrigibility." Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. (2015).

The background features a series of concentric circles in shades of dark blue and black, creating a tunnel-like effect. At the center of these circles is a bright, glowing yellow-orange circle. The word "Fragen?" is written in white text across the center of the glowing circle.

Fragen?