Ethik in AI

Inhalt

- Warum Ethik in AI?
- Positionen
- SKYNET
- Project Maven
- Probleme in der Sicherheit von KI
- Softwaredemo

Warum Ethik in AI?

- Steigende Verantwortung von KI-Systemen
- Immer weitreichendere Entscheidungen in Bereichen wie
  - Banken
  - Versicherungen
  - Geheimdienste
  - Militär

Warum Ethik in AI?

- Flexibilität erlaubt vielseitige Verwendung der gleichen Technologie
- Selbstfahrendes Fahrzeug
  - Selbstfahrender PKW
  - Selbstfahrendes Waffensystem
- Gesichtserkennung
  - Zählen von Personen für Statistik
  - Identifizierung von Zielen

Warum Ethik in AI?

- Keinerlei verbindliche Regeln für Einsatz von KI
- Verantwortlichkeit liegt bei Entwicklern
- Kontroverses Thema mit verschiedenen Positionen

Elon Musk

- Gegner der aktuellen Entwicklung im Bereich KI
- Befürchtet Pannen mit verheerenden Folgen für die Menschheit
- Fordert Regulierungen für die Forschung und den Einsatz
- Kontrollen durch unabhängige Stelle, die Einblicke erhält

Ronald Arkin

- US-amerikanischer Forscher auf dem Gebiet der Robotik
- Arbeitet an der Entwicklung von autonomen Waffensystemen
- Sieht in Robotern den Vorteil, keine Gefühle wie Rache zu kennen
- KI soll dazu beitragen Menschliche Fehler im Krieg zu vermeiden und so ethischer zu handeln
- Mensch soll vorgeben nach welchen Richtlinien Maschinen handeln
- Verantwortung für einzelne Einsätze muss auch weiterhin bei Menschen liegen

SKYNET

- Programm der NSA
- Ziel: Identifizierung von Mitglieder von Terrororganisationen
- Verwendung von Bewegungsdaten auf Basis von GSM-Funkzellen
- Anwendung in Pakistan
- Eng verbunden mit Drohnenprogramm

SKYNET

- NSA gibt false positive Quote mit 0,008% an
- 200 Mio. Pakistaner => 16.000 falsch klassifiziert
- In der Praxis evtl. höhere Fehlerquoten durch zu kleinen Trainingsdatensatz

Project Maven

- Zusammenarbeit von Unternehmen mit Pentagon
- Ziel: Automatisiertes Auswerten von Drohnen-Bildmaterial
  Klassifizierung von militärischen Zielen
- Bekannt geworden durch Wiederstand von Google Mitarbeitern
  => Google will sich aus dem Projekt zurückziehen

Sicherheitsprobleme

- Unerwartetes Verhalten von KI kann Gefahr darstellen
- Ursachen häufig schwierig zu identifizieren
- Fehlerquellen lassen sich in Kategorien einteilen

Negative Seiteneffekte

- Entstehen während der Erfüllung der eigentlichen Aufgabe
- z.B. Putzroboter wirft Vase um, da er so schneller putzen kann
- Verhinderung über Anpassung der Belohnungsfunktion möglich
- Manuelles Festlegen aller Verbote aufwändig

  => Muster erlernen lassen

Reward Hacking

- Maximierung der Belohnungsfunktion auf unvorhergesehene Weise
- z.B. Putzroboter verursacht selber Schmutz, um diesen wieder beseitigen zu können
- Ursache von logischen Fehlern bis hin zu Implementierungsdetails
- Verhinderung während der Entwicklung und durch Überwachung der Belohnung

Sicheres Experimentieren

- Verbesserung während der Laufzeit durch Versuche
- z.B. Putzroboter sucht Möglichkeit Steckdose zu säubern
  => verwendet nassen Mopp
- Experimentieren muss in vorgegebenem Rahmen stattfinden
- Überwachung durch Mensch
- Auslagerung der Versuche in virtuelle Welt

Umgebungswechsel

- Wechsel der Umgebung kann zu gefährlichem Verhalten führen
- z.B. Putzroboter wird in Büro trainiert und in Fabrik eingesetzt
- KI muss eigene Unsicherheit erkennen können

Softwaredemo

Fragen