Ethik für künstliche intelligente Systeme

Symposium Ethik & Nachhaltigkeit

Stefan Huber
FH Salzburg

16. Oktober 2021

Maschinenethik

Einsatz von intelligenten, autonomen Systemen wirft Fragen der Ethik auf.

- Bildung, Pflege, Industrie, Straßenverkehr, ...

Bisherige Technikethik

Ethik für Menschen im Umgang mit Maschinen.

AI führt hier zu vielen neuen Fragen.

Maschinenethik

Einsatz von intelligenten, autonomen Systemen wirft Fragen der Ethik auf.

- Bildung, Pflege, Industrie, Straßenverkehr, ...

Bisherige Technikethik

Ethik für Menschen im Umgang mit Maschinen.

AI führt hier zu vielen neuen Fragen.

Maschinenethik

Ethik für Maschinen.

Neue Disziplin. Systematischer Zugang durch Catrin Misselhorn [Mis18].

Maschinenethik

Einsatz von intelligenten, autonomen Systemen wirft Fragen der Ethik auf.

- Bildung, Pflege, Industrie, Straßenverkehr, ...

Bisherige Technikethik

Ethik für Menschen im Umgang mit Maschinen.

AI führt hier zu vielen neuen Fragen.

Maschinenethik

Ethik für Maschinen.

Neue Disziplin. Systematischer Zugang durch Catrin Misselhorn [Mis18].

Zentraler Aspekt

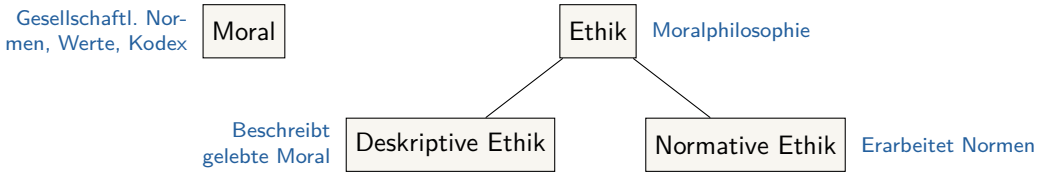
Kann eine Maschine moralischer Akteur sein?

Ethik und Moral

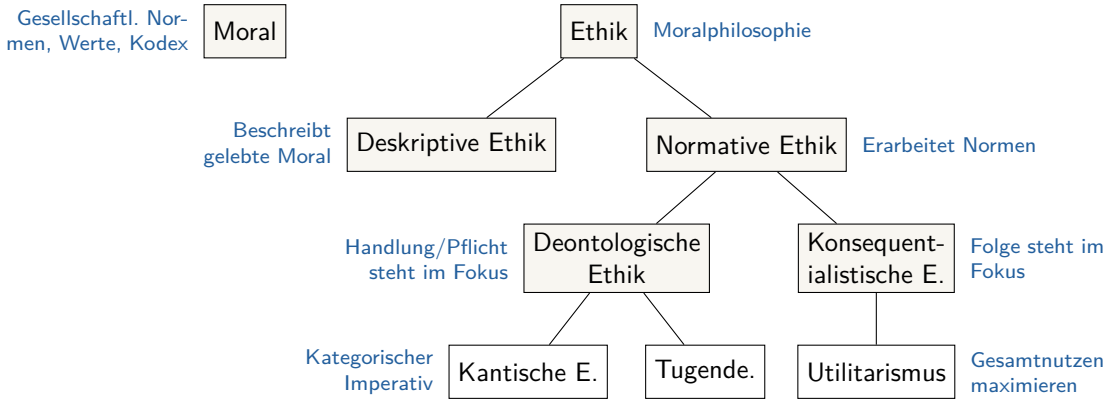
Gesellschaftl. Normen, Werte, Kodex — **Moral**

Ethik — Moralphilosophie

Ethik und Moral



Ethik und Moral



Orientierungsfrage

Falls überhaupt, welche konkreten normativen Ethiken sind für Maschinen anwendbar?

Artificial Morality

Begriff von artificial morality (AM) analog zu artificial intelligence (AI).

- Analoge Frage nach der Existenz von starker AI wird auch für AM aufgeworfen.

Aber Maschinenethik setzt bereits bei bodenständigeren Voraussetzungen ein:[1]

Fallbeispiel

Darf ein Staubsaugerroboter Tiere töten?

[1] Siehe [Mis18, p. 8]

Artificial Morality

Begriff von artificial morality (AM) analog zu artificial intelligence (AI).

- Analoge Frage nach der Existenz von starker AI wird auch für AM aufgeworfen.

Aber Maschinenethik setzt bereits bei bodenständigeren Voraussetzungen ein:[1]

Fallbeispiel

Darf ein Staubsaugerroboter Tiere töten?

Wir müssen klären:

- Intelligenter Akteur
- Moralischer Akteur: Autonomie und moralisches Handeln

[1] Siehe [Mis18, p. 8]

Künstliche intelligente Akteure

Standardbuch zu AI: Russel und Norvig, *Artificial Intelligence* [RN20].

Zugänge zur Begriffsbildung von AI:

- Künstliche Systeme, die per se menschliche Leistungen vollbringen.

  Beispiel: Bild- und Sprachverstehen

- Künstliche Systeme, die rational handeln, eine Problemlösungs- und Deduktionsfähigkeit besitzen.

  Beispiel: Schachcomputer

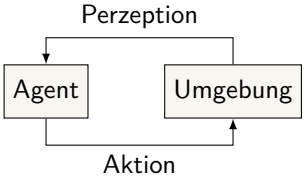
Zwei Dimensionen des Begriffs AI

- Menschlichkeit $\leftrightarrow$ Rationalität
- Denken $\leftrightarrow$ Handeln

Standardmodell des intelligenten Agenten

Das Standardmodell in [RN20]:

- Agent interagiert mit einer Umgebung.
- Agent nimmt Zustände der Umgebung durch Perzeptionen wahr.
- Agent setzt Aktionen basierend auf Perzeptionssequenzen.

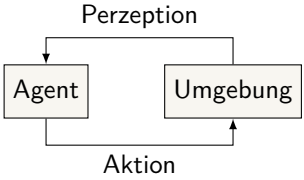


2 Siehe [RN20, p. 39]

Standardmodell des intelligenten Agenten

Das Standardmodell in [RN20]:

- Agent interagiert mit einer Umgebung.
- Agent nimmt Zustände der Umgebung durch Perzeptionen wahr.
- Agent setzt Aktionen basierend auf Perzeptionssequenzen.



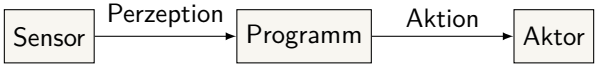
Rational das Richtige tun

- Performanzmaß bewertet die Umgebungszustände hinsichtlich einer Problemstellung.
- Spielart des Konsequentialismus[2].

[2] Siehe [RN20, p. 39]

Konstruktionsweisen von Agenten

Agent setzt sich zusammen aus Agentenarchitektur und Agentenprogramm:



Anwendungsbeispiele:

- Krebsklassifikation, Prädiktion von Maschinenausfällen, Schachcomputer, autonomes Fahren

Programm:

- Mathematische Analyse oder Spieltheorie, z.B. beim tic-tac-toe Spiel.
- Statistischen Analysen.
- Logischen Deduktionssysteme.
- Maschinelles Lernen: Programme/Modelle generiert aus Daten/Umgebungsexploration. Möglicherweise Anpassungsfähigkeit.

Moralische Akteure

Vier Ausbaustufen moralischer Akteure nach Moor:[3]

Ethical impact agents Erzeugen moralische Folgen, handelt aber nicht, z.B. Tachometer.

Implicit ethical agents Moralische Wertvorstellungen schlagen sich nieder, z.B. Abstandsregelautomat.

Explicit ethical agents Erkennt moralisch relevante Informationen und trifft moralische Handlungsentscheidungen, z.B. nichttötender Staubsaugerroboter.

Fully ethical agents Verfügt über Bewusstsein, Denken, Willensfreiheit.

[3] Siehe [Mis18, p. 70]

Moralische Akteure

Vier Ausbaustufen moralischer Akteure nach Moor:[3]

Ethical impact agents Erzeugen moralische Folgen, handelt aber nicht, z.B. Tachometer.

Implicit ethical agents Moralische Wertvorstellungen schlagen sich nieder, z.B. Abstandsregelautomat.

Explicit ethical agents Erkennt moralisch relevante Informationen und trifft moralische Handlungsentscheidungen, z.B. nichttötender Staubsaugerroboter.

Fully ethical agents Verfügt über Bewusstsein, Denken, Willensfreiheit.

Zentraler Aspekt

Eine Annahme für explicit ethical agents ist die Möglichkeit moralischen Handelns für Maschinen.

[3] Siehe [Mis18, p. 70]

Handlungsfähigkeit

Zentrale Frage bei Misselhorn[4]

Was gilt als (moralische) Handlung?

Unterschied zum Handeln bei biologischen Lebewesen:

- Klar definierte Grenzen (durch die Agentenarchitektur)
- Vom Menschen gebaut, ohne intrinsische Bedürfnisse

Zwei Dimensionen der Handlungsfähigkeit

- Rationalität
- Fähigkeit zum Initiieren von Verhalten

4 Siehe [Mis18, p. 77]

Initiieren von Verhalten

Naheverhältnis zum Begriff der Autonomie. Vier Formen nach Darwall[5]:

- Personale, moralische und rationale Autonomie
- Handlungsautonomie

Selbstursprünglichkeit

Personale, moralische, rationale Autonomie bedeutet handeln aus Gründen.[6]

Genuines Handeln ist Rationalität plus Selbstursprünglichkeit.

5 Siehe [Mis18, p. p78]

6 Das umfasst nicht Akteurskausalität, also die Initiierung von Handlungen ohne Ursache. Ob es diese überhaupt beim Menschen gibt, ist unklar.

Selbstursprünglichkeit

Drei Bedingungen für Selbstursprünglichkeit nach Floridi und Sanders[7]:

- ✓ Interaktivität mit der Umwelt
- ✓ Anpassungsfähigkeit (u.a. der Verhaltensregeln)
- ? Gewisse Unabhängigkeit von der Umwelt[8]

7 Siehe [Mis18, p. p77]

8 Zustandsänderungen ohne direkte Umwelteinwirkung

Selbstursprünglichkeit

Drei Bedingungen für Selbstursprünglichkeit nach Floridi und Sanders[7]:

- ✓ Interaktivität mit der Umwelt
- ✓ Anpassungsfähigkeit (u.a. der Verhaltensregeln)
- ? Gewisse Unabhängigkeit von der Umwelt[8]

Unabhängigkeit

Eindruck der Unabhängigkeit durch Unvorhersehbarkeit:

- nicht-deterministische oder randomisierte Verfahren
- non-explainable AI.

7 Siehe [Mis18, p. p77]

8 Zustandsänderungen ohne direkte Umwelteinwirkung

Ansätze für Moralimplementierung

- Utilitarismus
  Maxime: Lust-Leid Bilanz maximieren. Konsequentialistischer Ansatz gemäß Performanzmaß intelligenter Agenten.

8 Siehe [Mis18, p. 96]

Ansätze für Moralimplementierung

- Utilitarismus
  Maxime: Lust-Leid Bilanz maximieren. Konsequentialistischer Ansatz gemäß Performanzmaß intelligenter Agenten.
- Kantische Ethik
  Implementierte Regeln durch kategorischen Imperativ überprüfen, oder ein Logiksystem entwerfen, welches selbst überprüft.

8 Siehe [Mis18, p. 96]

Ansätze für Moralimplementierung

- Utilitarismus
  Maxime: Lust-Leid Bilanz maximieren. Konsequentialistischer Ansatz gemäß Performanzmaß intelligenter Agenten.
- Kantische Ethik
  Implementierte Regeln durch kategorischen Imperativ überprüfen, oder ein Logiksystem entwerfen, welches selbst überprüft.
- Asimov'sche Gesetze
  1. Ein Roboter darf keinen Menschen verletzten oder durch Untätigkeit zu Schaden kommen lassen.
  2. Muss den Befehlen eines Menschen gehorchen, außer bei Konflikt mit Gesetz 1.
  3. Muss seine eigene Existenz schützen, außer bei Konflikt mit Gesetz $1 + 2$.

[8] Siehe [Mis18, p. 96]

Ansätze für Moralimplementierung

- Utilitarismus
  Maxime: Lust-Leid Bilanz maximieren. Konsequentialistischer Ansatz gemäß Performanzmaß intelligenter Agenten.
- Kantische Ethik
  Implementierte Regeln durch kategorischen Imperativ überprüfen, oder ein Logiksystem entwerfen, welches selbst überprüft.
- Asimov'sche Gesetze
  1. Ein Roboter darf keinen Menschen verletzten oder durch Untätigkeit zu Schaden kommen lassen.
  2. Muss den Befehlen eines Menschen gehorchen, außer bei Konflikt mit Gesetz 1.
  3. Muss seine eigene Existenz schützen, außer bei Konflikt mit Gesetz $1 + 2$.
- Bottom-Up Ansätze
  Überwachtes Lernen anstatt expliziter Regeln, etwa über neuronale Netze. Naheverhältnis zur Tugendethik. Deskriptive Ethik?

[8] Siehe [Mis18, p. 96]

Beispiel: Pflegesysteme

Pflegesysteme in der Altenpflege als Antwort auf demographische Veränderungen. Zwei Systeme von Anderson[9]:

Utilitaristisches System

- Ethikberater, der Handlungsfolgen numerisch hinsichtlich Wohlbefinden-Leid Bilanz mit Eintrittswahrscheinlichkeit bewertet.
- Zweifelhaft, ob Utilitarismus hier überzeugt.

9 Siehe [Mis18, p. 138, p. 142]

Beispiel: Pflegesysteme

Pflegesysteme in der Altenpflege als Antwort auf demographische Veränderungen. Zwei Systeme von Anderson[9]:

Utilitaristisches System

- Ethikberater, der Handlungsfolgen numerisch hinsichtlich Wohlbefinden-Leid Bilanz mit Eintrittswahrscheinlichkeit bewertet.
- Zweifelhaft, ob Utilitarismus hier überzeugt.

Deontologisches System von Anderson

- Drei Prinzipien der Bioethik:
  1. Respekt vor der Autonomie des Patienten
  2. Prinzip dem Patienten keinen Schaden zuzufügen
  3. Prinzip zum Wohl des Patienten zu handeln
- Konflikte zwischen Prinzipien: Numerisches Bewertungsschema um präferierte Handlung zu finden.

[9] Siehe [Mis18, p. 138, p. 142]

Beispiel: Autonomes Fahren

Ausgangspunkt

Autonomes Fahren im Zusammentreffen mit Menschen im Verkehr wird Unfälle nicht vermeiden können.

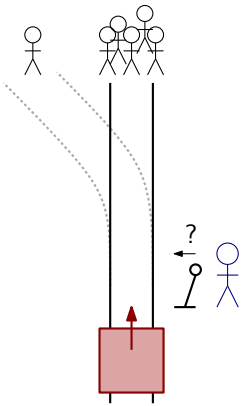
Analogie zu Asimov

1. Nicht mit einem Fußgänger oder Radfahrer zusammenstoßen.
2. Nicht mit einem anderen Fahrzeug zusammenstoßen, außer bei Konflikt mit Gesetz 1.
3. Mit keinem Objekt in der Umgebung zusammenstoßen, außer bei Konflikt mit Gesetz 1+2.

Dieser Ansatz ist aber unzureichend: Analoge Situation zum Trolley Problem

[9] Siehe [Mis18, p. 189]

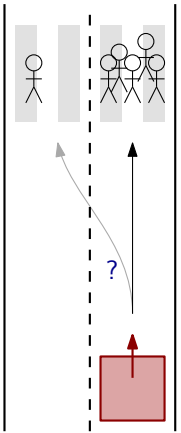
Trolley Problem



Trolley Problem (1930, 1951, 1967)

Dürfen wir die Weichenstellung ändern, um fünf Leben zu retten, aber eines zu opfern?

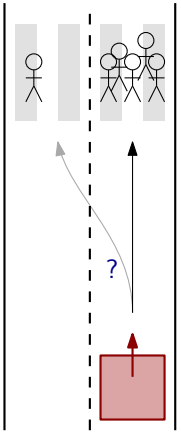
Trolley Problem



Trolley Problem (1930, 1951, 1967)

Dürfen wir die Weichenstellung ändern, um fünf Leben zu retten, aber eines zu opfern?

Trolley Problem

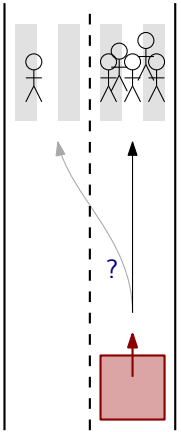


Trolley Problem (1930, 1951, 1967)

Dürfen wir die Weichenstellung ändern, um fünf Leben zu retten, aber eines zu opfern?

- Utilitarismus: Ja, wegen Nutzen-Leiden Bilanz.
- (Deontologische) Pflichtenethik: Eher nein, denn Pflicht nicht zu töten meist stärker als Pflicht zu retten.

Trolley Problem

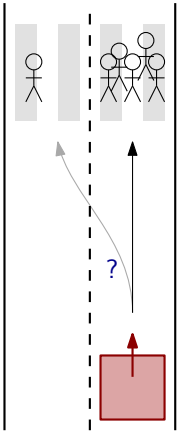


Trolley Problem (1930, 1951, 1967)

Dürfen wir die Weichenstellung ändern, um fünf Leben zu retten, aber eines zu opfern?

- Utilitarismus: Ja, wegen Nutzen-Leiden Bilanz.
- (Deontologische) Pflichtenethik: Eher nein, denn Pflicht nicht zu töten meist stärker als Pflicht zu retten.

Trolley Problem



Trolley Problem (1930, 1951, 1967)

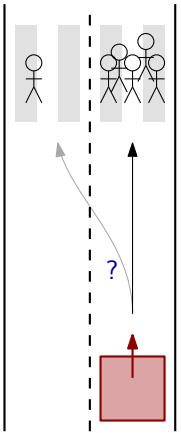
Dürfen wir die Weichenstellung ändern, um fünf Leben zu retten, aber eines zu opfern?

- Utilitarismus: Ja, wegen Nutzen-Leiden Bilanz.
- (Deontologische) Pflichtenethik: Eher nein, denn Pflicht nicht zu töten meist stärker als Pflicht zu retten.

Varianten verändern moralische Bewertung (siehe Moral Maschine):

- Was, wenn 5 Alte gegen 1 Kind? Schwere Krankheit?

Trolley Problem



Trolley Problem (1930, 1951, 1967)

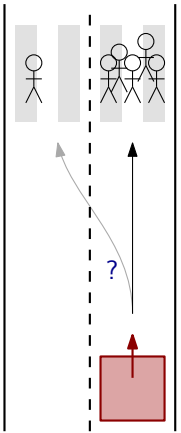
Dürfen wir die Weichenstellung ändern, um fünf Leben zu retten, aber eines zu opfern?

- Utilitarismus: Ja, wegen Nutzen-Leiden Bilanz.
- (Deontologische) Pflichtenethik: Eher nein, denn Pflicht nicht zu töten meist stärker als Pflicht zu retten.

Varianten verändern moralische Bewertung (siehe Moral Maschine):

- Was, wenn 5 Alte gegen 1 Kind? Schwere Krankheit?
- Was, wenn 5 bei Rot über die Ampel gingen oder schlimmer?

Trolley Problem



Trolley Problem (1930, 1951, 1967)

Dürfen wir die Weichenstellung ändern, um fünf Leben zu retten, aber eines zu opfern?

- Utilitarismus: Ja, wegen Nutzen-Leiden Bilanz.
- (Deontologische) Pflichtenethik: Eher nein, denn Pflicht nicht zu töten meist stärker als Pflicht zu retten.

Varianten verändern moralische Bewertung (siehe Moral Maschine):

- Was, wenn 5 Alte gegen 1 Kind? Schwere Krankheit?
- Was, wenn 5 bei Rot über die Ampel gingen oder schlimmer?
- Option, dass die Insaßen des autonomen Autos nicht geopfert werden → Spieltheorie

Danke für die Aufmerksamkeit.

Mehr hier: `https://www.sthu.org/blog/19-ethik-ki/index.html`

References I

[Mis18] Catrin Misselhorn. *Grundfragen der Maschinenethik*. 4th ed. 2018. ISBN: 978-3-15-019583-3.

[RN20] Stewart Russel and Peter Norvig. *Artificial Intelligence. A Modern Approach*. 4th ed. 2020. ISBN: 978-0134610993.