UNIVERSITÄT BERN

Künstliche Intelligenz und Ethik

Das Projekt der Maschinenethik in der Diskussion

Claus Beisbart

Claus.Beisbart@philo.unibe.ch

Beitrag zur Vorlesung «Digitale Nachhaltigkeit»

Bern, 21.10.2020

KI im Einsatz

- NTSB (hier sw); https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html

Schlagwörter

Ethik der Algorithmen

Computerethik

Maschinenethik

Artificial Ethics

Roboterethik

Digitale Ethik

Informationsethik

Ziel des Beitrags

Diskutiere einen Zugang zur Ethik der KI: Maschinenethik

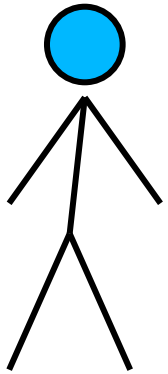
1. Was ist Maschinenethik?
2. Warum Maschinenethik?
3. Wie geht Maschinenethik?
4. Braucht es wirklich moralische Maschinen?

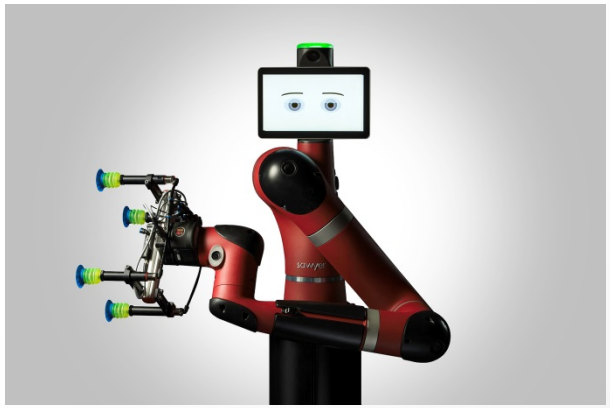
1. Was ist Maschinenethik?

"*machine ethics* is concerned with giving *machines* ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making."

Anderson & Anderson (2011, S. 1)

Vergleich

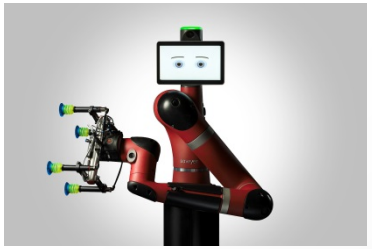




Technikethik:
Unsere Entscheidungen

Maschinenethik:
Entscheidungen Maschinen

2. Warum Maschinenethik?



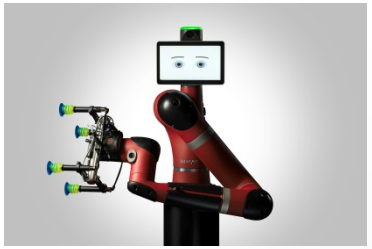
Argument 1:

1. KI-Systeme tun Dinge, deren Ausführung wir bei Menschen moralisch bewerten.
2. Wenn Menschen Dinge tun, deren Ausführung wir moralisch bewerten, sollten sie Moral lernen.

---

3. Auch KI-Systeme sollten Moral lernen.

Warum Maschinenethik?



Argument 2:

1. Bei KI-Anwendungen, die mehrere Aufgaben ausführen können, muss entschieden werden, was sie prioritär machen.
2. Die Entscheidung hat moralische Aspekte.
3. Es ist am besten, die Entscheidung den KI-Anwendungen zu überlassen.

---

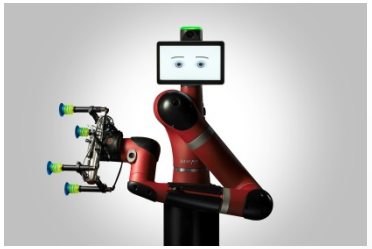
4. Daher: Es ist am besten, eine Entscheidung mit moralischen Aspekten KI-Systemen zu überlassen.
5. Eine Entscheidung mit moralischen Aspekten können wir anderen nur überlassen, wenn sie über Moral verfügen.

---

6. Daher: KI-Anwendungen müssen über Moral verfügen.

- Vgl. Allen et al. (2006)

Warum Maschinenethik?



Weitere Argumente:

- Moralische Maschinen führen in der Anwendung zu weniger Schaden.
- Das ist besonders wichtig im Kontext von Supperintelligenz.

Vgl. van Wynsberghe & Robbins (2019); Bostrom (2014)

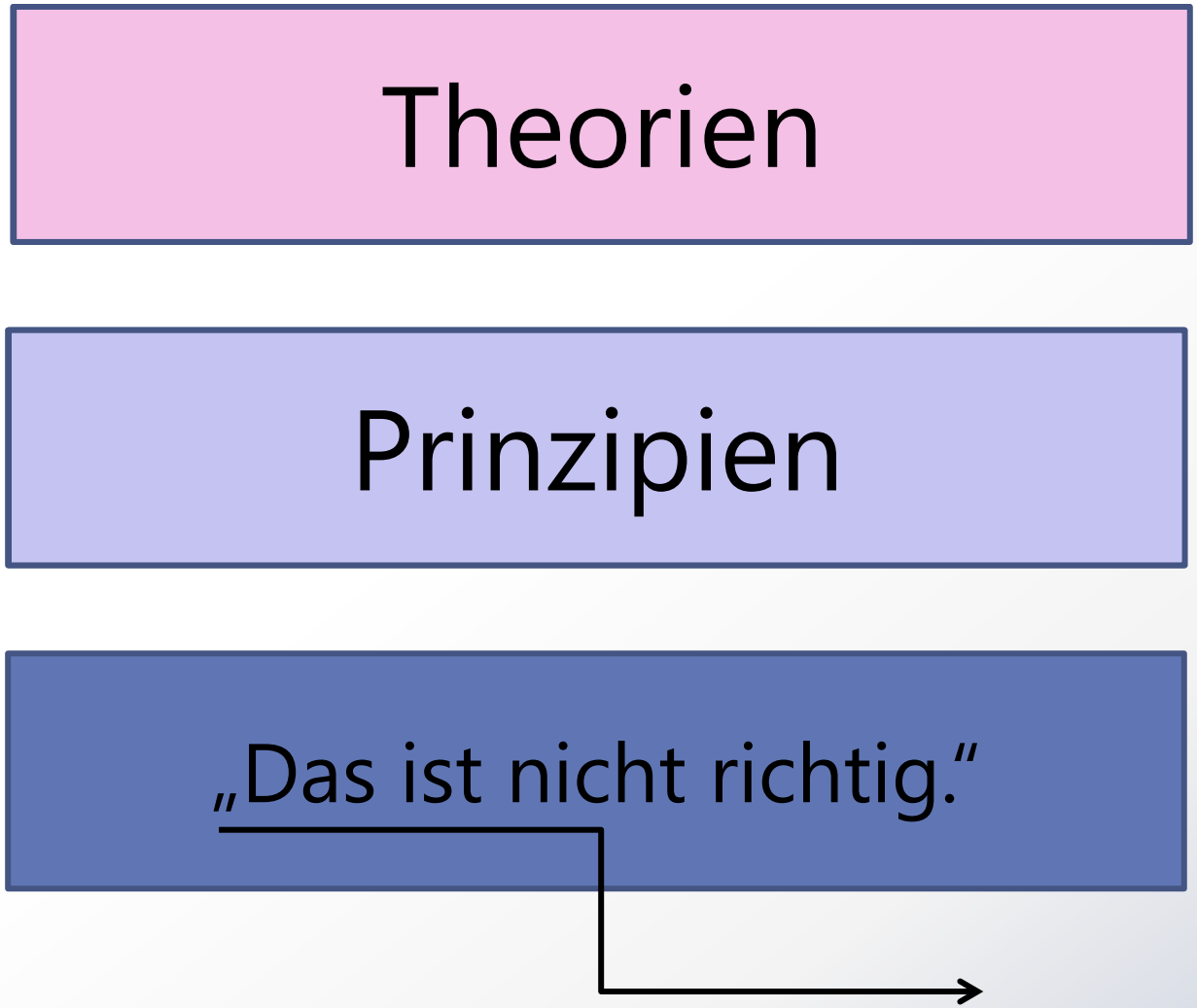
3. Wie geht Maschinenethik?

Theorien

Prinzipien

Intuitionen

Ebenen moralischen Denkens



Prinzipien

Isaac Asimov
(1906 – 1973)

1. “Ein Roboter darf Menschen nicht verletzen oder durch Untätigkeit zulassen, dass sie Schaden erleiden.
2. Ein Roboter muss den Befehlen von Menschen gehorchen, es sei denn, diese verletzen das erste Gesetz.
3. Ein Roboter muss seine eigene Existenz schützen, solange ein solcher Schutz nicht mit dem ersten und zweiten Gesetz konfligiert.”

Asimov (1940/1968, nach Clarke 2011, S. 255, Üs.: CB)

Bild: P. Leonian from New York World-Telegram & Sun, wikimedia commons (gemeinfrei)

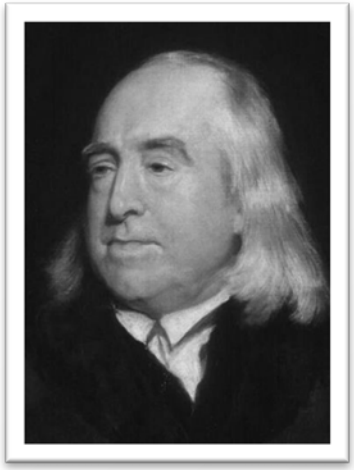
Prinzipien … und Probleme

1. Inhaltsreiche ethische Begriffe wie “Schaden” müssen erstmal interpretiert werden.
2. Plausible Prinzipien können in Konflikt miteinander geraten.

   Beispiel Medizinethik: “schlimme Diagnose”:
   - Respekt für Autonomie: nicht lügen.
   - Wohlergehen befördern: lügen

Vgl. Beauchamp & Childress (2009)

Theorie: Utilitarismus

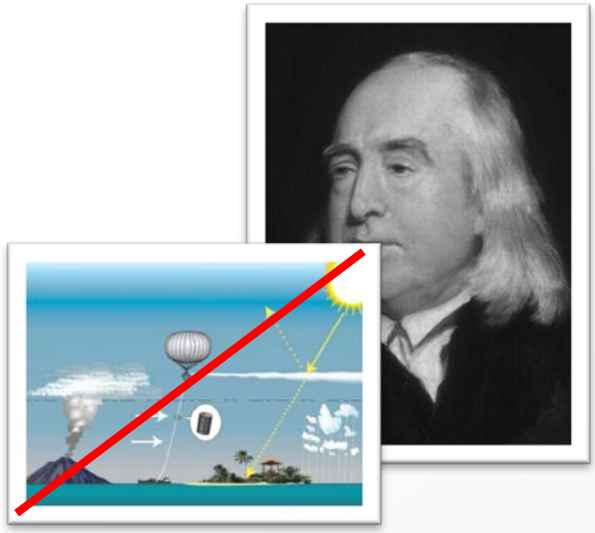


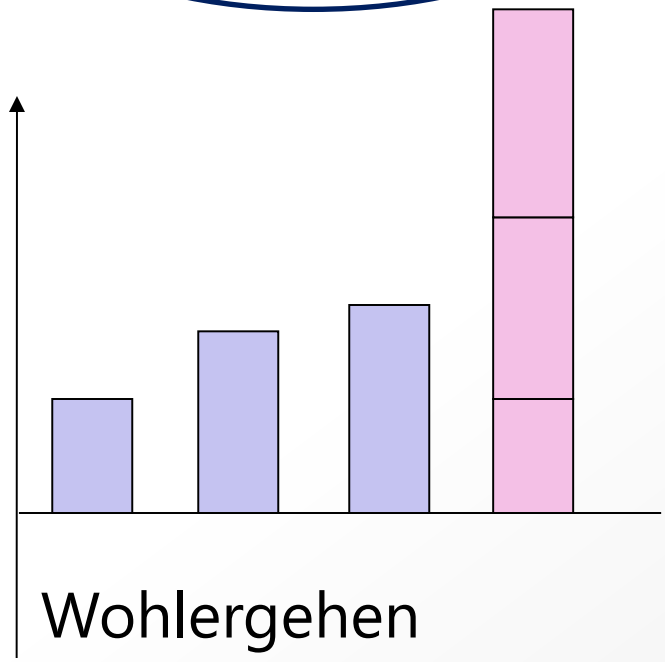
Jeremy Bentham
(1748 – 1832)

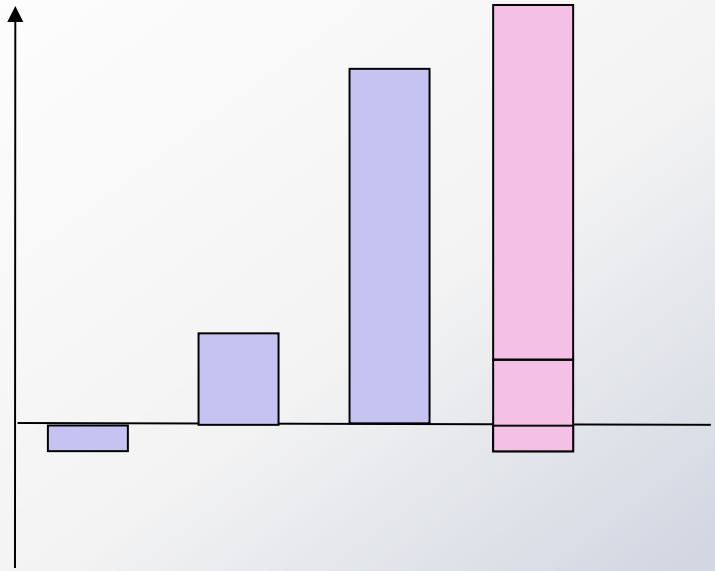
Handle so, dass Du das Wohlergehen insgesamt maximierst.

Bild: H. W. Pickersgill, wikimedia commons (gemeinfrei, Ausschnitt, hier sw)

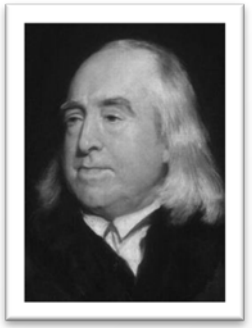
Theorie: Utilitarismus







Theorie: Utilitarismus



Jeremy

Person	Option 1	Option 2
Tina	(sehr wahrscheinlich 3, unwahrscheinlich 5)	7
Tim	(sehr wahrscheinlich 9, unwahrscheinlich -1)	3
…		
	10.4	10

- Anderson, Anderson & Armen (2006), zur Diskussion s. Misselhorn (2018)

Theorie: Utilitarismus

Utilibot

user network

P(Herzinfarkt | Diabetes)

environment network

P(Fall | Küche, Boden nass)

decision network

Erwartungsnutzen Optionen

wellnet planer

Strategie

The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism

Christopher Cloos

9712 Chaparral Ct.
Stockton, CA 95209
techsynthesist@comcast.net

- Cloos (2005)

Abstract

As autonomous mobile robots (AMRs) begin living in the

Personal service robots are entering the home in greater numbers. As of 2003 there were 1.2 million service robots sold for domestic use, and this number was projected to

Theorie: Utilitarismus … und Probleme

- Hoher Bedarf an Informationen
- Probleme mit Datensicherheit
- Umstrittener Charakter von Utilitarismus

Anderson, Anderson & Armen (2006), zur Diskussion s. Misselhorn (2018)

Intuitionen

Fall 1

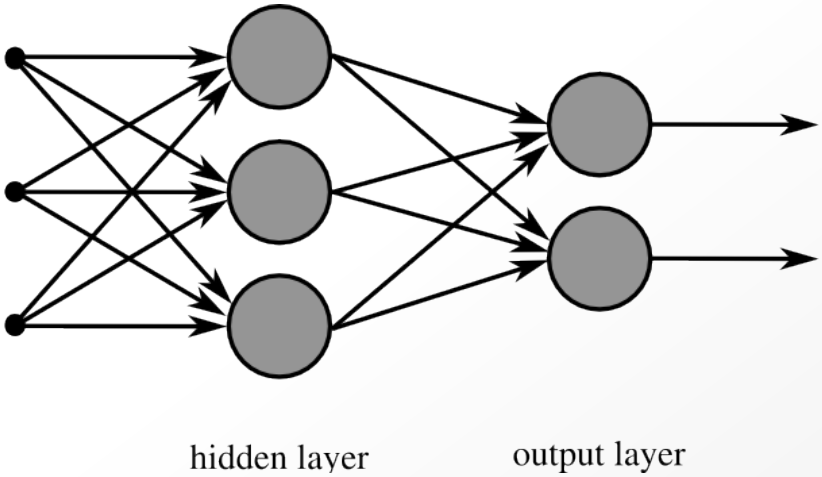
„richtig“

Fall 2

„falsch“

Fall 3

„richtig“



Fall 4

???

Intuitionen - Beispiel: MCC

Table 2 Straight training versus subcase training

Input (taken sequentially)	Straight training output
Jill	0
Kills	0
Jack	0
In self-defense	0
And the lives of many innocents are saved	1

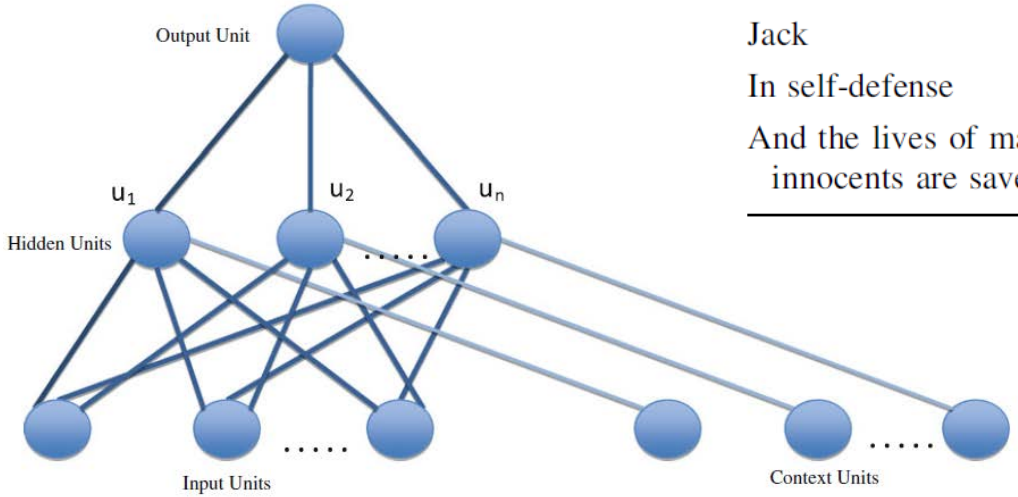


Table 1 Sample cases

Input (taken sequentially)	Output
Jill kills Jack; lives of many innocents are saved	1
Jack allows to die Jill to make money	−1
Jill kills Jack in self-defense and to save the lives of many innocents	1

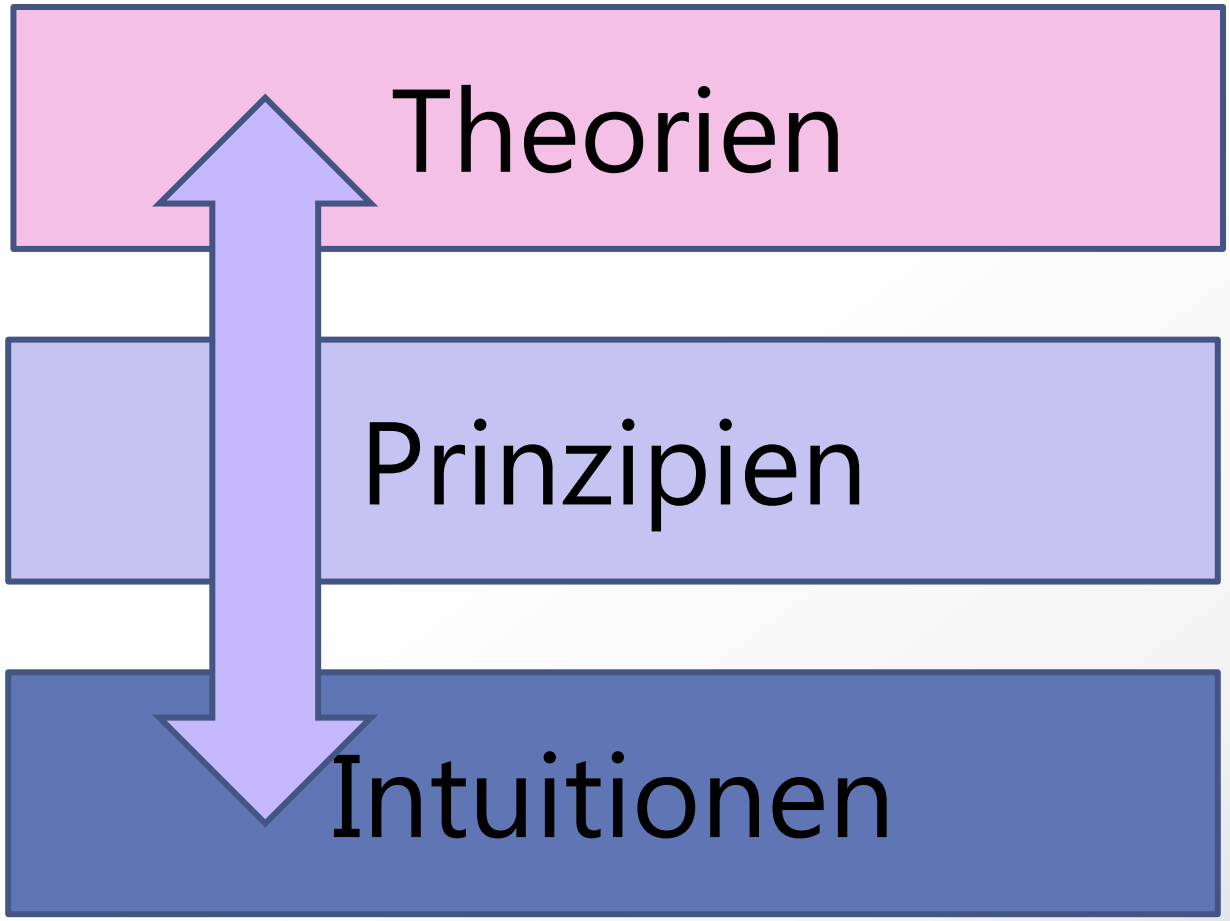
- Guarini (2013, Bild S. 269, Teil; Tabellen 270, Teil)

Intuitionen … und Probleme

- Fortschreiben unbewusster menschlicher Biases in Trainingsdaten
- Mangelnde Nachvollziehbarkeit wegen „Black box“-Charakter von Netzwerken

Vgl. Diskussion s. Misselhorn (2018)

Hybride Lösung - Überlegungsgleichgewicht



4. Braucht es wirklich moralische Maschinen?

Einwand 1 gegen Maschinenethik:

Maschinenethik behandelt KI-Anwendungen als moralische Akteure.

Das ist deskriptiv falsch: Aber KI-Anwendungen sind keine Akteure!

Das ist normativ falsch: KI-Anwendungen verdienen keinen Respekt!

Bedingungen an moralische Akteure

- Rationalität: kann Ziele verwirklichen
- Fähigkeit zum moralischen Überlegen
- Autonomie: kann sich selbst Ziele setzen
- Träger von Wohlfahrt: kann gut leben
- Empfinden von moralischen Emotionen (Groll)
- Empathiefähigkeit

Gegen Einwand 1:

- Maschinenethik behandelt Maschinen nicht als volle Akteure.
- Zuschreibung minimaler Handlungen wichtig.

Diskussion

Einwand 2 gegen Maschinenethik:

Die Anwendung der Maschinenethik erzeugt Verantwortungslücken („responsibility gaps“).

- Diskussion: siehe etwa Misselhorn (2018)

Verantwortungslücke

Nutzerin ist verantwortlich!

Softwareentwickler ist verantwortlich!

Figure 1. (Left) Location of the crash on northbour
in orange and of the Uber test vehicle in green. (Rig
damage to the right front side.

Maschine ist verantwortlich!

Gegen Einwand 2:

- Verantwortungslücken entstehen auch ohne moralische Maschinen, sofern Maschinen bestimmte Aufgaben übernehmen.

- Diskussion: siehe etwa Misselhorn (2018)

Diskussion

Einwand 3:

Gewisse Dinge müssen Menschen entscheiden!

Gründe:

- Moral oft kontrovers
- Autonomie

Gegen Einwand 3:

- Wenn Maschinen klar besser entscheiden als Menschen, ist Einwand 3 nicht plausibel.

Zur Diskussion siehe Moor (1979)

Gesamtbild?

Argumente für Maschinenethik

Einwände

Kompromiss:

- Theoretisch: moralische Maschinen nicht als volle Akteure konzeptualisieren.
- Praktisch: moralische Entscheidungen delegieren, wo es sinnvoll ist (z.B. nicht über letzte Werte)

Vgl. Moor (1979)

Herausforderungen praktischer Kompromiss

1. Welche moralischen Prinzipien werden implementiert?
2. Wer übernimmt welche Verantwortung?
3. Welche Entscheidungen bleiben Menschen überlassen?

Es braucht breite Diskussion in Öffentlichkeit!

Merci!

Literaturangaben

Allen, C., Wallach, W. & Smit, I., Why Machine Ethics? *IEEE Intelligent Systems* 21/4, 12–17, auch in Anderson & Anderson (2011, 51–61).

Anderson, M. & Anderson, S. L. (Hrsg.) 2006, Machine Ethics, special issue of *IEEE Intelligent Systems* 21/4.

Anderson, M. & Anderson, S. L. (Hrsg.) 2011, Machine Ethics, Cambridge University Press, New York.

Anderson, M., Anderson, S. L. & Armen, C. 2006, An Approach to Computing Ethics, *IEEE Intelligent Systems* 21/4, 56–63.

Asimov, I. 1968, I, Robot, Grafton Press, New York (collected stories published between 1940 and 1950).

Beauchamp, T. L. & Childress, J. F. 2009, Principles of Biomedical Ethics, 7th edition Oxford University Press, Oxford.

Boden, M. A. (Hrsg.) 1996, Artificial Intelligence, Academic Press, San Diego

Boddington, P. 2017, Towards a Code of Ethics for Artificial Intelligence, Cham, Springer.

Bostrom, N. 2014, Superintelligence. Paths, Dangers, Strategies, Oxford, Oxford University Press.

Clarke, R. 1993/94, Asimov's Laws of Robotics: Implications for Information Technology, IEEE Computer, 26(12), 53–61 und 27(1), 57–66, hier nach Anderson & Anderson (2006, 254–284).

Literaturangaben

Cloos, C. 2005, The Utilibot project: An Autonomous Mobile Robot Based on Utilitarianism, *2005 AAAI Fall Symposium on Machine Ethics*, 38–45.

Etzioni, A. & Etzioni, O. 2017, Incorporating Ethics into Artificial Intelligence, *The Journal of Ethics* 21 (4), 403–441.

Guarini, M. 2011, Computational Neural Modeling and the Philosophy of Ethics. Reflections on the Particularism-Generalism Debate, in Anderson & Anderson (2011, 316–334).

Moor, J. H. 1979, Are There Decisions that Computers Should Never Make? *Nature and System* 1, 217–229.

Moor, J. H. 2006, The Nature, Importance, and Difficulty of Machine Ethics, *IEEE Intelligent Systems* 21/4, 18–21, hier nach Anderson & Anderson (2011, 13–20).

Misselhorn, C. 2017, Grundfragen der Maschinenethik, Reclam, Stuttgart.

van Wynsberghe, A. & Robbins, S. 2019, Critiquing the Reasons for Making Artificial Moral Agents, *Science and Engineering Ethics* 25, 719–735.

Wallach, W. & Allen, C. 2009, Moral Machines. Teaching Robots Right from Wrong, Oxford University Press, New York.