

»Die größte Herausforderung besteht nicht darin, künstlicher Intelligenz menschliche Werte einzuprogrammieren, sondern festzulegen, welche Werte das sein sollen.«

Iason Gabriel: Artificial Intelligence, Values, and Alignment¹

I Werte

Ausrichtung auf menschliche Werte (value alignment)

„Ein Roboter darf die Menschheit nicht verletzen oder durch Passivität zulassen, dass die Menschheit zu Schaden kommt.“

So formulierte Isaac Asimov das erste seiner drei berühmten Robotergesetze.

Es liegt nahe, dass auch künstliche Intelligenz - egal ob als Chatbot, Industrietool oder sonstiger Algorithmus - gewisse grundlegende Regeln befolgen sollte, beispielsweise Menschen nicht zu schaden.

Mit Blick auf eine KI, die der menschlichen Intelligenz vollumfänglich überlegen ist („AGI“), oder gar Bewusstsein entwickelt, gilt dies umso mehr.



„Make Paperclips!“

Warum KI nicht tun sollte, was wir ihr auftragen.

I Werte

Ausrichtung auf menschliche Werte (value alignment)

Dieses Ziel wirft zwei wesentliche Fragen auf, von denen die zweite im Folgenden näher beleuchtet werden soll:

- Wie lassen sich Werte in Algorithmen umsetzen?
- Welche Werte sollen das sein?

I Werte

Utilitaristische Ethik

Die Mehrzahl gängiger Machine-Learning-Algorithmen arbeitet mit Nutzen- bzw. Verlustfunktionen. Das sind konkret mathematische Funktionen, die beispielsweise eine Entfernung (Distanz) zwischen zwei Punkten oder Vektoren beschreiben. Minimiert man die Funktion, hat man den Ort größter Nähe gefunden, was oft etwa mit „genaueste Vorhersage“ oder „bester Match“ übereinstimmt.

Für diesen Mechanismus gibt es eine relativ passgenaue ethische Entsprechung: Die Theorie des Utilitarismus, dessen Ansatz es ist, Glück zu maximieren bzw. Leid zu minimieren. Utilitaristische Ethik ist durchaus nützlich und wirkmächtig, bringt aber viele offene Fragen mit sich, zuvorderst: Was ist Glück? Und: In welchem Modus genau wird es für welche Personengruppe maximiert?

I Werte

Ethik durch Vorbilder

Ein weiterer Ansatz, der durch die Frage der Umsetzung motiviert ist, beinhaltet den Einsatz von *unsupervised learning*: Man überlässt dem Computer die Aufgabe, aus den zur Verfügung stehenden Daten selbst moralische Prinzipien bzw. moralisches Verhalten zu destillieren.

Obschon die technischen Möglichkeiten dazu mehr oder weniger vorhanden sind, werden schnell zwei Probleme offensichtlich:

- Der Korpus „Alle Daten“ enthält schlicht zuviele Beispiele moralisch fragwürdigen Verhaltens.
- Begrenzt man den Korpus auf eine Auswahl, ergibt sich die Frage, wer diese Auswahl auf Basis welcher Kriterien vornimmt.

I Werte

Globale moralische Konvergenz

Auch wenn es - Stichwort Zensur - durchaus nationale Unterschiede in der Anwendung bzw. Nutzbarkeit von Technologie gibt, so scheint es doch unumgänglich, Werte in der KI als globales Problem zu begreifen.

Da sowohl Individuen wie auch Gruppen über unterschiedliche ethische Systeme und Schwerpunktsetzungen verfügen, könnte ein Ansatz darin bestehen, den größten gemeinsamen Nenner ethischer Prinzipien zu bestimmen. So wurden von der OECD die folgenden vier ethischen KI-Prinzipien als quasi universell identifiziert:



I Werte

Hypothetische Übereinstimmung

Eines der wichtigsten philosophischen Gedankenexperimente zur Beantwortung ethischer Fragen ist der „Schleier des Nichtwissens“ (veil of ignorance), ersonnen von John Rawls²: Die Teilnehmer haben keinerlei Wissens über ihren Hintergrund, kennen weder ihre Nationalität, ihre Geschlecht, ihre Religion, oder ihren sozialen Status. Nun legen sie gemeinsame die ethischen Prinzipien fest, die für alle gelten sollen - und mittels der Rahmenbedingungen sind diese gerecht für alle.

Neben den bereits diskutierten Prinzipien wäre eine Ergänzung vielleicht die Festlegung, dass der Nutzen von künstlicher Intelligenz weitestgehend gleich verteilt sein sollte unter allen Menschen (oder zumindest allen Betroffenen).

I Werte

Demokratische Legitimation (Werteaggregation)

Zu guter Letzt wäre, basierend auf dem Ansatz der Sozialwahltheorie, auch denkbar, mittels demokratischer Wahl auf Basis individueller Präferenzen über ethische Werte künstlicher Intelligenz abstimmen zu lassen.

Gegen diesen Ansatz sprechen aber verschiedene Argumente: Individuelle Präferenzen sind unter anderem oftmals nicht rational, nicht ethisch einwandfrei, und nicht ohne Paradoxien aggregierbar. Darüber hinaus wäre eine solche Wahl praktisch kaum durchzuführen.

| Werte

Referenzen

1

Iason Gabriel
Artificial Intelligence, Values, and Alignment (2020)
<https://link.springer.com/article/10.1007/s11023-020-09539-2>

2

John Rawls
A Theory of Justice

3

Anna Jobin, Marcello lenca, Effy Vayena:
The Global Landscape of AI Ethical Guidelines
<https://www.nature.com/articles/s42256-019-0088-2>

Image Prompts

Dall-E Prompts, ohne Tuning, erster Versuch, zu:
„make paperclips, dystopian“