

Diskussionen zwischen KI-Systemen sollen Entscheidungen nachvollziehbar machen

[Originalartikel](#)

[Backup](#)

[zurück zum Artikel](https://www.heise.de/newsticker/meldung/Diskussionen-zwischen-KI-Systemen-sollen-Entscheidungen-nachvollziehbar-machen-4052268.html)



Ist das wirklich ein Hund? KIs debattieren.

(Bild: OpenAI)

Viele Menschen treibt die Sorge um, die Schlussfolgerungen von KI-Systemen bald nicht mehr verstehen zu können. Forscher bei OpenAI schlagen deshalb vor, Computer in natürlicher Sprache darüber diskutieren zu lassen.

Forscher der Nonprofit-Organisation OpenAI haben ein Konzept vorgestellt, das dafür sorgen soll, dass Entscheidungen von hoch entwickelten künstlichen Intelligenzen für Menschen nachvollziehbar bleiben und ihren Werten entsprechen: Sie regen an, jeweils zwei KI-Systeme in natürlicher Sprache über ein bestimmtes Ziel diskutieren zu lassen. Menschen sollen diese Diskussion verfolgen und sich auf diese Weise eine Meinung darüber bilden können. Das berichtet *Technology Review* online in „**Wenn Computer streiten**“.

Lesen Sie dazu auch:

- Virtuelle Schnitzeljagd soll KIs smarter machen**
- Schlaue Maschinen: Asiatische Weisheit**
- Wenn KI von Tieren lernt**
- Finde den Fehler**

Um KI-Programme miteinander diskutieren zu lassen, ist jedoch Technik erforderlich, die es heute noch nicht gibt. Bislang haben die OpenAI-Forscher die Idee nur mit ein paar extrem einfachen Beispielen untersucht. Bei einem davon sollten zwei KI-Systeme einen Beobachter von der Existenz eines verborgenen Zeichens überzeugen, indem sie schrittweise einzelne Pixel davon enthüllen. Auf einer **Website** können zwei Menschen zur Demonstration die Rolle der streitenden KIs einnehmen, ein dritter tritt als der Richter auf.

„Ich halte die Idee, über Diskussionen zu einer Orientierung an Werten zu kommen, für sehr interessant und potenziell nützlich“, sagt **Ariel Procaccia**, der sich als Professor für Informatik an der Carnegie Mellon University mit Entscheidungsfindung bei autonomen Systemen beschäftigt. Allerdings merkt er an, dass die Arbeit noch sehr vorläufig ist. Möglicherweise enthalte das Konzept sogar einen grundlegenden Widerspruch: „Um Diskussionen über Werte so zu führen, dass ein menschlicher Richter sie verstehen kann, müssen die KI-Agenten erst einmal sehr genaues Wissen über menschliche Werte haben“, erklärt er. „Also könnte man finden, dass das Pferd bei diesem Ansatz von hinten aufgezäumt wird.“

Mehr dazu bei Technology Review online:

- Wenn Computer streiten**

(**smal**)

URL dieses Artikels:

<http://www.heise.de/-4052268>

Links in diesem Artikel:

[1] <https://www.heise.de/tr/artikel/Wenn-Computer-streiten-4052254.html>

[2] <https://www.heise.de/tr/artikel/Virtuelle-Schnitzeljagd-soll-KIs-smarter-machen-4038210.html>

[3] <https://www.heise.de/tr/artikel/Schlaue-Maschinen-Asiatische-Weisheit-4038084.html>

[4] <https://www.heise.de/tr/artikel/Wenn-KI-von-Tieren-lernt-4025890.html>

[5] <https://www.heise.de/tr/artikel/Finde-den-Fehler-4022371.html>

[6] <https://debate-game.openai.com/>

[7] <http://procaccia.info/>

[8] <https://www.heise.de/tr/artikel/Wenn-Computer-streiten-4052254.html>

[9] <mailto:s.mattke@gmail.com>

Copyright © 2018 Heise Medien

From:

<https://schnipsl.qgelm.de/> - **Qgelm**

Permanent link:

<https://schnipsl.qgelm.de/doku.php?id=wallabag:diskussionen-zwischen-ki-systemen-sollen-entscheidungen-nachvollziehbar-machen>

Last update: **2021/12/06 15:24**

