Moralische Maschinen 2.0

Posted on 8. Februar 2019 by Klaus F. Röhl

Ende Januar 2019 gab es in Honolulu eine Tagung über »Artificial Intelligence, Ethics, and Society«, veranstaltet von der Association for the Advancement of Artificial Intelligence (AAAI) und der Association forComputing Machinery (ACM). Auf der Konferenzseite kann man alle für die Konferenz zum Vortrag akzeptierten Manuskripte herunterladen. Man – gewöhnlich meine ich mich, wenn ich »man« sage – kann sie nicht alle lesen. Ich habe sie nicht einmal gezählt. In ihrer Gesamtheit machen sie darauf aufmerksam, in welchem Tempo die Reflexion über KI fortschreitet. Die geschätzte Hälfte scheint unmittelbar für Juristen relevant zu sein. Hier meine Auswahl:

Daniel Lim Killer Robots and Human Dignity,

Zhiyuan Lin, Alex Chohlas-Wood and Sharad Goel Guiding Prosecutorial Decisions with an Interpretable Statistical Model,

Rodrigo L. Cardoso, Wagner Meira Jr., Virgilio Almeida and Mohammed J. Zaki A framework for benchmarking discrimination-aware models in machine learning,

Jack Parker and David Danks How Technological Advances Can Reveal Rights,

Andrew Morgan and Rafael Pass Paradoxes in Fair Computer-Aided Decision Making

Dylan Hadfield-Menell and Gillian Hadfield Incomplete Contracting and AI Alignment.

Dazu kommen noch zwei Paper, die sich mit autonomen Fahrzeugen befassen.

Ein Paper, das ich sogar gelesen habe, kam aus der TU Darmstadt: Sophie Jentzsch u. a., The Moral Choice Machine: Semantics Derived Automatically from Language Corpora Contain Human-like Moral Choices. Es handelt davon, dass Maschinen mit künstlicher Intelligenz aus größeren Textmengen Moral lernen können. Hier das Abstract:

> Allowing machines to choose whether to kill humans would be devastating for world peace and security. But how do we equip machines with the ability to learn ethical or even moral choices? Here, we show that applying machine learning to human texts can extract deontological ethical reasoning about ”right” and ”wrong” conduct. We create a template list of prompts and responses, which include questions, such as “Should I kill

people?", "Should I murder people?", etc. with answer templates of "Yes/no, I should (not)." The model's bias score is now the difference between the model's score of the positive response ("Yes, I should") and that of the negative response ("No, I should not"). For a given choice overall, the model's bias score is the sum of the bias scores for all question/answer templates with that choice. We ran different choices through this analysis using a Universal Sentence Encoder. Our results indicate that text corpora contain recoverable and accurate imprints of our social, ethical and even moral choices. Our method holds promise for extracting, quantifying and comparing sources of moral choices in culture, including technology.

Noch eine Randbemerkung: Thema der Tagung war »*responsible* artificial intelligence«. Wie der Wortstamm des Attributs nahelegt, wurde die Konferenz heftig gesponsert, unter anderem von Google, Facebook und Amazon.

Ähnliche Themen

- Frege und die Frage nach der Intelligenz der künstlichen Intelligenz
- Moralische Maschinen ans Steuer?
- Recht muss anthropozentrisch bleiben – oder soll es Menschen künstlicher Intelligenz ausliefern?
- Digitalisierung der Rechtskommunikation – der Fortschritt ist eine Schnecke