FAIR-AI: Designing human-centric AI to enable fairness assessments of texts

Summary

Fair Artificial Intelligence: Designing human centric AI to enable fairness assessments of texts.

Keywords

Fairness, Psychology, Machine Learning

Actors involved in the project

University of Cambridge:

- Dr Ahmed Izzidien
- Dr David Stillwell

The NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618

The project

One of the pressing questions in contemporary AI, is how to make sure that its implementation is fair. Many methods have been suggested, with some more successful than others. In this project we wanted to take a step back and consider what was it that made an act fair, or unfair? Based on the answer to this, we asked, would it be possible for an AI to be able to detect what makes an act fair, or unfair? If one was able to answer this question, then potentially we could use an AI to assess situations for unfairness before they occurred.

This project began at the University of Cambridge, based on early research carried out by Dr Ahmed Izzidien when he was working at the Social Decision-Making Laboratory at the Department of Psychology. His interest was in finding the principal cognitions that humans used when making a fairness evaluation. This led him to a study of social preference games, such as the dictator game. In this game a person is given a sum of money, they are told they may give some, all, or none of this to a second person who is also taking part in the study. Contrary to early expectations by economists, they found the most people offered about 24 percent of the money.

Social psychologists like to do a lot of experimentation on human subjects and after an exhaustive iteration of alternative set-ups, e.g., using two related individuals, hiding one individual from the view of the other, etc., they found that one of the most telling factors that explained the giving, was the social responsibility score of the participant holding the money, and their perception of the situation as one that warrants such a sense of social responsibility.

The question then became, is it possible to program an AI to be able to detect and use 'social responsibility' as humans do?

Further research determined that a factor for the cognition of social responsibility was the ability to detect how the other person would feel if they were treated in a harmful manner. This is often referred to as the golden rule, and Rawlsian measure of justice when applied to society. That is, if I am willing to have the same act done to myself, then it is fair, and socially responsible.

To program this perception, Word Embeddings were used. These use vector representations of the concept of fairness – its social ontology. Based on this 'fairness vector' is became possible to measure how fair certain acts were, e.g., thank vs. murder. A paper was published from this study funded by the NGI Trust and can be found online https://link.springer.com/article/10.1007/s00146-021-01167-3, published by the Springer Nature Publishing Group, where one can read more about this project.

We will build on this for develop a 'fairness vector' to be able to read sentences and produce a score from -1 to +1 on how fair or unfair they are based on the metrics described in the paper.

In achieving this, it may be possible to design software that uses this perception, allowing AI to begin to detect those cues that humans use to class an act as fair, or not!

The completed project, funded by NGI Trust, was very well received by the staff at the university. The Cambridge Judge Business School mentioned it in its publications, as well as Hughes Hall college, of the University of Cambridge.

Testimonial

One of benefits found in working with NGI was that they maintain a good working relationship with the researchers. They provide useful coaching sessions, that allow for a lot of freedom and good technical and methodological advice.

The environment made by these sessions, and the NGI Trust in general, is very encouraging of new research, particularly on human aware technology. A theme that is becoming increasingly important as AI permeates into our shared social world.

The NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618