Introduction to Accountability, Evaluation and Obscurity of AI Algorithms Minitrack

Radmila Juric
University of South Eastern Norway
Kongsberg, Norway
Radmila.Juric@usn.no

Robert Steele
Capitol Technology University
Laurel, MD 20708, USA
rjsteele@captechu.edu

This Minitrack has attracted very interesting submissions and we have chosen to accept three papers.

They are very different and illustrate various thinking and ideas which are hidden behind our concerns about the obscurity of AI algorithms, , which shape current AI systems.

The paper "Computation, Rule Following, and Ethics in AIs" comes from the US. It raises widely shared concern about the harmful impact of algorithms on legal, financial, or health initiatives, particularly if they would be able to either marginalize populations or not properly represent it in training datasets. Therefore, both algorithms and data set are directly responsible for these worries, but the authors argue that addressing this problem requires rethinking beyond our practice regarding representation and sampling. They propose a framework centered on rules for examining fundamental issues in ethics, AI alignment, and computational models. The authors' argument, that though all AIs are rule governed, some of them can also be described as rule following. This asks new questions if current AI system's (in practice) are able to follow rules, use language, possess concepts, or reason. As AIs move into areas involving decisions with ethical implications, the authors argue that learning and acting with regard to ethics and morality is simultaneously rule governed and rule following and thus the proposed framework is centered on rules for examining fundamental issues in ethics.

The paper "Managing Temporal Dynamics of Filter Bubbles" comes from Europe and proposes an approach to manage filter bubbles for digital journalism, in order to allow users of the bubble to better co-create value by enhanced interaction possibilities. By conceptualizing and designing the temporal dynamics of filter bubbles, the authors illustrate the way of enabling users to interact with the filter bubble. It is a a continuous value co-creation which aligns foreign image and self-image of the user. The results from this research is valuable for public broadcasters that have a public-service remit to fulfil. However, it can also be of value to private media journalism, as they may increase trust in recommender systems and increase customer loyalty.

The paper Perceptions of Fairness and Trustworthiness Based on Explanations in Human vs. Automated Decision-Making" comes also from Europe. It gives results of an online study with 200 participants, to examine people's perceptions of fairness and trustworthiness towards automated versus human decision making. The results are surprising. People perceive automated decision making as fairer than human. People's AI literacy affects their perceptions, because they do favor automation more strongly if they are AI literate. Low-AI-literacy people exhibit no significant differences in their perceptions. However, the authors warned that this hypothesis will have to be tested further. The dangers of wrongful persuasion and automation biases, i.e., the tendency of people to over-rely on automated decision making must be addressed. Consequently, if too many (compelling) explanations about the inner workings of automated decision making are provided, additional problems may be generated.

URI: https://hdl.handle.net/10125/79463
978-0-9981331-5-7