On the (Non-) Reliance on Algorithms - A decision-theoric account

Bernard Sinclair Desgagné
CNRS - CREDEG

1 Introduction

The relationship between people and machines, more precisely algorithms in this case, has always been complicated. The tendency of humans to shy way from using algorithms even when algorithms observably outperforms their humans counterpart has been referred to algorithm aversion. On the other hands, the tendency of humans to always want to use algorithms even when it is not really necessary has been referred to algorithm appreciation. The two concepts depends on different factors :

- The domain of interest : moral agents, subjective tasks, leadership, reputation
- The prior exposure : unfamiliarity with machines, having experienced algorithms failure
- The typical human traits : need of control, lack of confidence, emotions

The new thing about this issue is that people display aversion to machines and algorithms not only when they are substitutes for human workers and put certain jobs at stake but also when those same machines are complements to human skills and helps to improve their work or life

2 Methods and results

The paper's goal is to build a formal model of algorithm aversion/appreciation while conceding as possible irrationality.

2.1 Advisory Algorithms

Advisory algorithms are algorithms that just give you an advice that you are free to follow or not. For exemple, Waze (GPS) give you multiple choice of ways with the estimated time that you can chose. The most important thing in advisory algorithms is the value of information. Algorithms like Waze will rarely overestimate the time it will take to go somewhere unlike humans but it might underestimate more often, a thing that human rarely do, so humans and algorithms do not make the same mistakes. Depending of the weight you put on these errors, you may chose or not to work with those algorithms. It could be even worse if you have more than 2 categories of errors. The more ambiguity-averse people are, the less they will value information coming from the advisors algorithms.

2.2 Performative Algorithms

Performative algorithms can execute complex tasks independently from humans. the most important thing in these algorithms is the value of control. It is the the value of being able to chose what an uncertain variable will turn out to be and act in consequence. It is understandable that human may be led to distrust this kind of algorithm which take decisions for them. But ambiguity aversion might be something to overcome in the near-future. Many solutions has been proposed, like a greater exposure to algorithms so the ambiguity will sooner or later disappear but only if users are able to see some concrete results. However, this kind of solution is to be taken with a pinch of salt.

3 Conclusion

The general topic of technology acceptance and avoidance is a long-standing one in Information-Technology, various theories have accordingly been developed to explain the data and address the challenges. This talk seeked to contribute a complementary approach to the subject, based on rational decision-making and focused on the specific case of Artificial Intelligence. There is still many un-explored empirically relevant factors influencing human-AI relationship : demographic (age, gender), societal or cultural that we will need to analyze in the future. At the moment, time seems indeed ripe for additional theoretical work on this subject.

Causal based fairness : a methodology to use counterfactual fairness

Christele Tarnec
Orange

1 Introduction

This talk present a study made by Christele Tamec and Frederic Guyard on the counterfactual fairness. As a company, Orange is engaged in developing fair Artificial intelligence based services, the overall goal being to give everyone the key to a responsible digital world. The responsible AI is an AI that is trustworthy (fair, explainable...) and scalable, this talk is more focused on the fair aspect of AI. Fairness in AI is the absence of any prejudice or favoritism towards individual or a group based on their intrinsic or acquired traits in the context of decision making. The usual way to address this topic is not based on causal reasoning. It is of high interest to study this topic in order to develop models that will be fairer, more explainable and also to develop model with less data.

2 Counterfactual fairness

AI prediction is based on correlation, but correlation is not causality. For example, we can find correlation between a couple eating margarine and the probability that they divorce. There is correlation but not causality, it can lead to miss uses in certain models.

So when we develop a model, how can we be sure that it is a fair model ? There is two type of fairness criteria : the statistical criteria (individual, group...) and the causal based criteria. People mainly train their model on statistical criteria, but the fairness explainability is not guaranteed and a lot of data is needed. As there is more than 20 fairness metrics, we have to make a choice on which metrics is the most relevant for our specific use case, it depends on the context. With statistical criteria you are never sure to have a fair model but you have more chances with the structural causal model.

A model is counterfactual fair if the sensitive attributes are not the cause in any way of the output of the model. Example, the fact that that a black woman has been rejected from a job should not come from the fact that she is black or that she is a woman. To be fair, the counterfactual model should predict the same output for its factual and counterfactual instance. A good way to build a counterfactual model is to eliminate the sensitive attributes and all the variables that are depending on it but it may lead to having too less data. To obtain counterfactual model, you can also use some tools like PC, GES, Lingam on your dataset but with this kind of tool you may obtain different graphs from the same dataset, with different number of feature of variable. Again, you have to study numerous case of application in order to chose the best way to fit your model and make it fair.

Measuring fairness is very difficult, it is the measurement problem. To overcome this issue, we can use the Counterfactual accuracy :

$$\frac{i}{r}$$

With i being the number of instances that were predicted the same for the factual and counterfactual value and r being the number of real counterfactual values

3 Conclusion

Most of the time, dataset are biased because society is biased, it is why we have to work hard in order to create fair and explainable models. However, it is very difficult thing to do as as fairness notion are highly dependant from the context and causal discovery step, also, the fairness is hard to measure.

The general experimental contribution of this paper has been to generate a large variety of situations where everything was under control and fairness was tasted and evaluated. This study also helped the improvement of counterfactual fairness understanding, but a lot of work is yet to be done before making it actionable is AI-Based services.