University of Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

BACHELOR'S THESIS

Study programme/specialisation: Bachelor i data	Spring/ Autumn semester, 20.21. Open / Confidential
Author: Eirik Ødegård	
Programme coordinator: Antorweep Chakravorty Supervisor(s): Antorweep Chakravorty	
Title of bachelor's thesis: Fairness and Ethics in AI	
Credits: 20	
Keywords: Aritificial Intelligence, Fairness, Ethics	Number of pages: 17 + supplemental material/other: …… Stavanger, 15.05.2021 date/year

Abstract

As the complexity and capabilities of AI technologies continue to increase, they will continue to pose a risk for their users. In this thesis, different techniques have been reviewed to see how the current research proposes to introduce concepts such as fairness and ethics in AI. These techniques introduces fairness through interpretability of a complex model and an audit tool that allows for verifying bias and fairness metrics.

Table of contents

Nomenclature

- AI - Artificial Intelligence
- DSRPAI - Dartmouth workshop
- LIME - Local interpretable model-agnostic explana-tions
- XAI - Explainable Artificial Intelligence

Chapter 1

Introduction

Contents

1.1 Motivation

The motivation to write this paper revolves around AI being an integral part of our everyday life. Even though it is found in our phones, laptops, cars, and even appliances around the house, most of us have little knowledge about how they actually work. Due to the complexity of these AIs and their increased capability of what they can achieve, it would be interesting to further understand how they approach concepts such as fairness and operating in an ethical manner, given the optimal goal of an AI would be to mimic the human sentiment.

1.2 Goal

The goal of this thesis will be to take a deep dive into the state-of-the-art techniques that are being proposed to mitigate the risks of AI. This will be achieved, by gaining a more fundamental understanding of AI and what type of risks do they pose today through current research. I will focus on looking at research that looks to implement fairness and how to develop ethical AI. This will then be used to create this thesis serving as a survey report for a few important techniques that are being proposed.

Chapter 2

Theory and Background

Contents

2.1 Introduction

This chapter will provide the theory and background related to this thesis. It presents the history of AI, along with possible risks, and how one does define bringing fairness to AI.

2.2 The history of artificial intelligence

Artificial intelligence (AI) has become an integral part of society and is directing various aspects of our lives from which movie we watch, where to travel, and even how we should vote.

So what actually is AI? A definition is " the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent being" [1] and AI has already been around for some time. AI dates back to the 1950s where the general public familiarized themselves with the topic through science fiction movies and was established as a field of study at the Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI) hosted by John McCarthy and Marvin Minsky in 1956 [2]. The DSRPAI gathered researchers from different disciplines to have an open discussion on the topic, although they failed to decide on a standard method for the field, they deemed it achievable and has been one of the driving forces for the next twenty years of research in AI [2].

Since then, AI has come a far way including new sub-topics such as computer vision, deep learning, natural language processing, machine learning, and more. These technologies are being incorporated into our everyday life through social media, navigational apps, media streaming services, and more. The AI systems gather data, that will be used to make an optimized estimate on how to provide us with the right choice that will influence us in a way.

The questions that arise among these emerging technologies are how does one assure that these technologies are being operated ethically. An AI is devised to attain a goal, whether that would be to correctly identify an animal in a picture or to decide if a bank loan will be approved for a person. In the scenario of an AI identifying an animal, there are no affected bystanders, but letting an AI decide whether a person will receive approval on their loan has the potential to cause harm. By letting an AI make the decision, it opens up to the possibilities of discrimination by the AI having bias. When a human would make this decision the accountability of the decision would be held to that person. Who will hold the AI accountable for that same decision?

Chapter 3

Survey

Contents

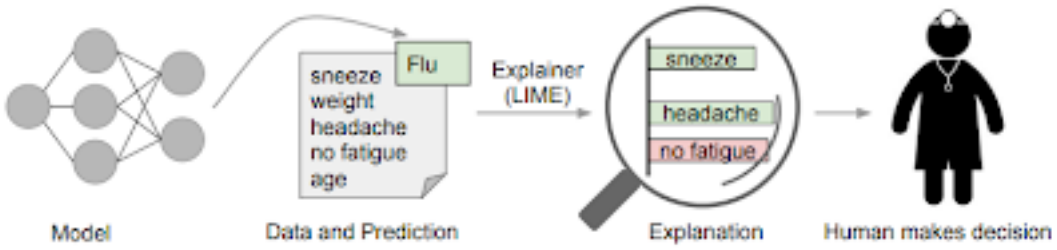
3.1 Introduction

In this chapter, I will present and review some techniques that are being proposed to mitigate risks with AI. There are a fair number of techniques that have been suggested over the last years as the research has increased exponentially after the industry has realized the need for it.

3.2 LIME - Local interpretable model-agnostic explanations

In this chapter, LIME will be reviewed, also known as the local interpretable model-agnostic explanations[4]. It is a post-hoc explanation technique that is a way of enhancing the AI's interpretability by giving a type of descriptive output that simplifies and further explains how the model reached its outcome [3]. This descriptive output may vary depending on the input, but will often be a visual or text explanation with a simplified pointer as to how it reached its conclusion.

To further explain how the LIME framework functions, see figure 3.1. As the name suggests, the technique is model-agnostic, which means that it is a post-hoc explainability technique that is usable for any sort of classifier and regression problems [4].

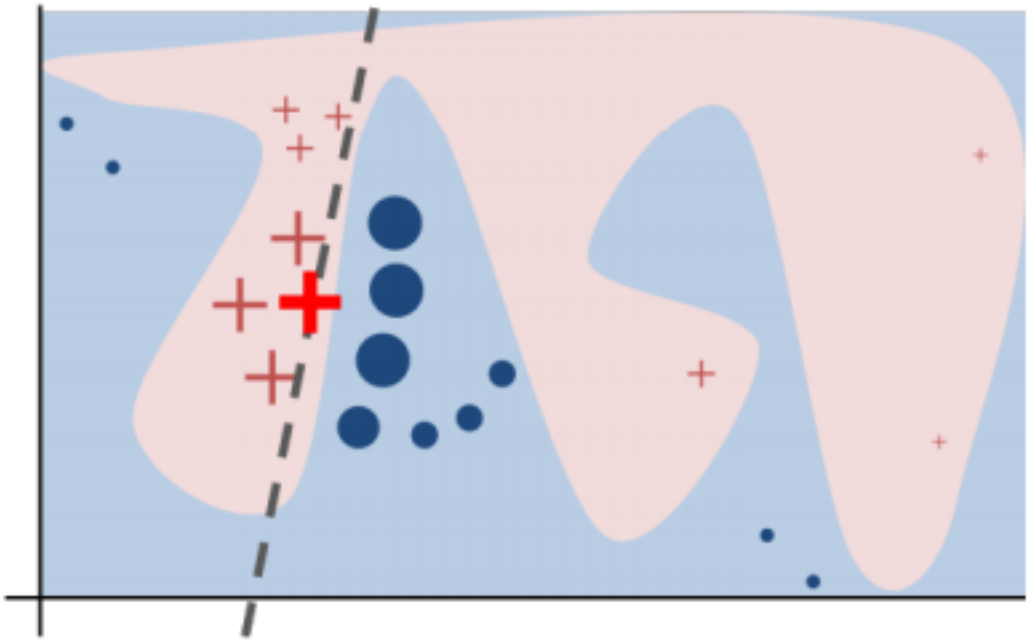


Figur 3.1: Text description provided by the LIME framework to introduce interpretability of the predictor.[4]

Figure 3.1 shows how a model predicts an outcome based on a set of symptoms. The explainer LIME [4] is then used to provide a text explanation as to how it reached its outcome, where it shows that the sneezing and headache were factors that contributed to the model predicting flu as the outcome, while a lack of fatigue is a factor that could contribute to it not [4]. This allows for a simpler way of enabling trust and interpretability between the user and the model by showing how it reached its outcome in a simplified manner and allowing the user to take a qualified decision.

To further explain how this works, we will have a look at figure 3.2 below. The graph is showing an

unknown decision function of a black-box model[4]. The model will make a prediction that will be placed within this decision function that will predict an outcome, as in an object will be predicted to have a certain trait or not, by whether the prediction will be placed in the blue area or the pink area. However, the two areas are not linear and given it is a black-box model, simply looking into the model to see what input is contributing to the prediction is not an option [5]. To further investigate how the bright red cross lands in the pink area, LIME will perturb the information around its neighborhood, which means it will input similarly adjacent input and see how the black box model will change its predictions [5]. These new data points will then be weighed by the proximity to the original prediction bright red cross and create an interpretable model, as shown by the red crosses and blue dots below, that shows a linear model that simply explains by having a prediction landing on the left side will contribute to a prediction equal to cross while landing on the right side will result in a prediction equal to dot [5].



Figur 3.2: Decision function with the LIME framework[4]

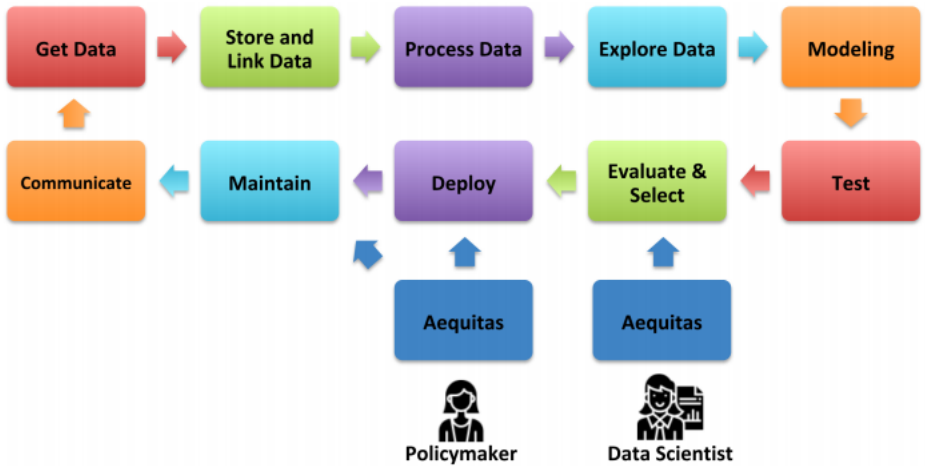
Figure 3.2 displays how LIME achieves its goal of identifying "an interpretable model over the interpretable representation that is locally faithful to the classifier"[4]. Although it does not have an understanding of the inner workings of this model, it will create a simple and interpretable model that will give provide insight to the user into how the model behaves and will allow the user to establish trust as to its performance. The interpretable model created by LIME will remain locally faithful in how it has drawn the dotted line above that holds true for this specific instance

of a prediction, but that will not necessarily remain true for other predictions as this is a simple interpretation of one prediction in a complex model.

LIME has displayed to be an insightful tool that allows users to further gain trust from their models by increasing the interpretability and could be used to establish whether a model actually accomplishes what it is supposed to do. As the tool is model-agnostic it shows high usability for AI applications.

3.3 Aequitas - A Bias and Fairness Audit Toolkit

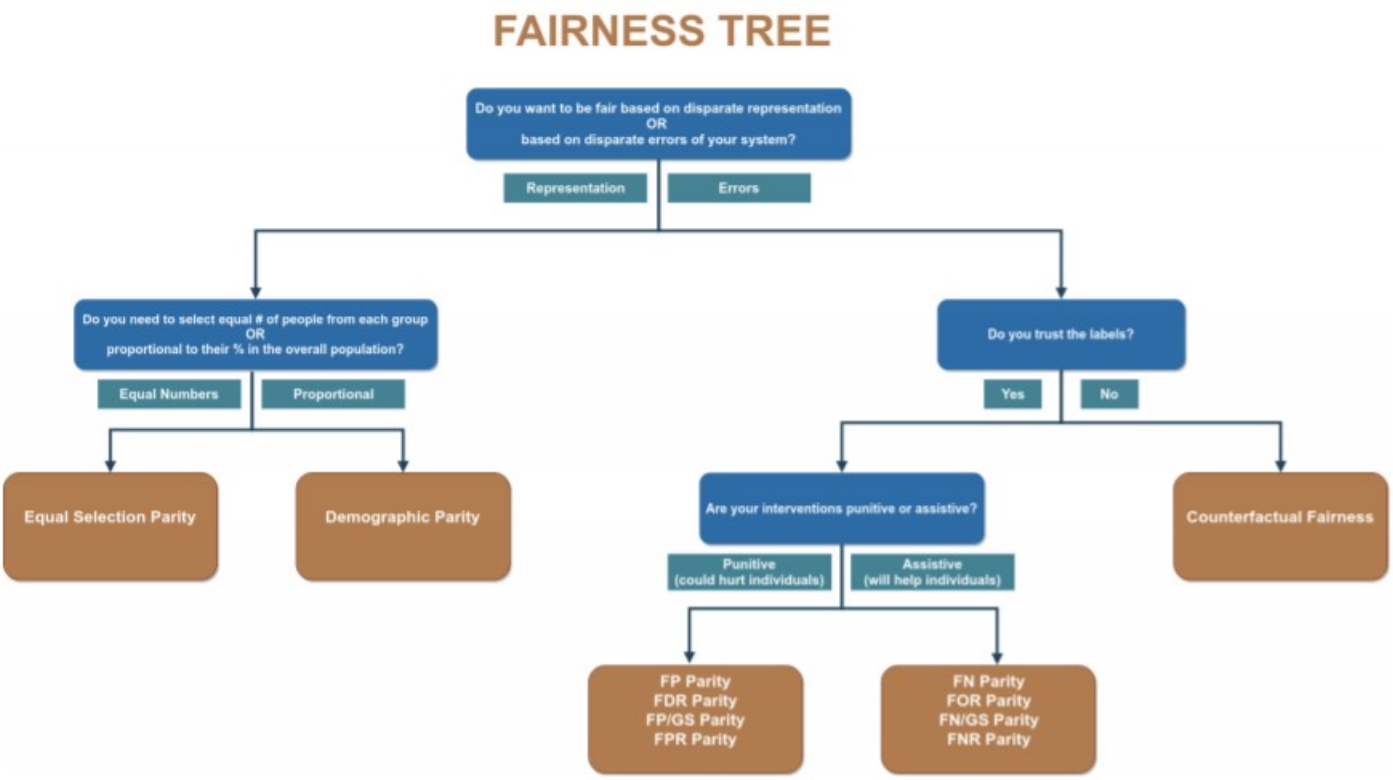
In this chapter, Aequitas [7] will be reviewed, which is a bias and fairness audit toolkit. This toolkit was created for multiple reasons, as to how it could gain a better understanding of bias against groups of people, to explore various bias metrics and fairness definitions in real-world public policy problems, as well as earning the trust of the general public in regards to AI [7]. The tool is created to be used by both computer scientists and policymakers where it would be incorporated into the developing, maintaining, and deploying of AI systems as seen in figure 3.3[7].



Figur 3.3: The pipeline of an ML system, incorporating the Aequitas toolkit.[7]

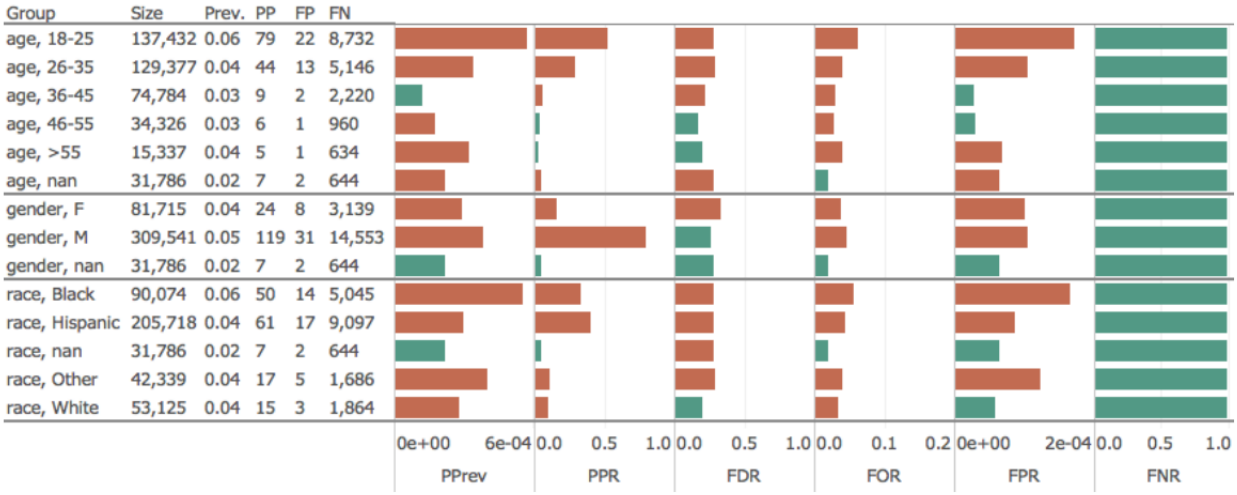
The Aequitas toolkit works as an audit toolkit that will assist in assessing the various bias and fairness metrics within a model. It may be taken into use while reviewing which model one might want to use, to gain a better understanding of the possible bias and fairness metrics associated with a model before continuing to development [7]. It will then fall onto the policymakers to perform an external audit of the model before deployment to verify its use case[8]. Involving several parties in the process will force them to act in consideration of fairness in regards to

the different aspects of their model [7]. To look further into how Aequitas we will review a case study that was performed in [7] involving criminal justice, where the goal was to intervene these people that were cycling through the system, not only within the criminal justice system but also emergency medical services, hospital emergency rooms, and other public services. With Aequitas there are guidelines that display how one finds the appropriate metrics based on the scope of the audit [7] as seen in figure 3.4.



Figur 3.4: Fairness tree - assisting data scientists and policymakers in finding the appropriate fairness metrics that are considered relevant for their scope. [7]

In the case study [7], they wanted to match the 150 highest risk individuals with preventative interventions. The data set used contained 1,5 million individuals over the past 10 years and focused on around 400 000 individuals who had repeated offenses with the criminal justice system [7]. In the analysis, they generated over 3000 features to be used within the data set that ranged from demographic to behavioral attributes using record linkage techniques[7].



Figur 3.5: Criminal justice - case study with Aequitas[7]

The Aequitas output is shown in figure 3.5 for this case study. Figure 3.4 Aequitas presenting the results from running the data sets. The Aequitas tool has shown to be a valuable tool that will provide insight into fairness and bias when using a model. By involving both data scientists and policymakers in selecting and evaluating the criteria for both the models and how we collect data, forces both parties to have fairness in mind.

Chapter 4

Discussion

Contents

4.1 Challenges

There have been some challenges during writing this thesis. This would, first of all, include poor planning that allowed for less than optimal time to review and gain the required knowledge of the fundamentals before heading into the state-of-the-art research that often included advanced mathematics, statistics, and programming.

The selection process of relevant articles was harder than I initially thought as the number of available resources on the topic has exponentially increased over the last few years. In addition, finding something that was relevant and understandable could be a challenge at times.

4.2 Conclusion

During this thesis, I have created a brief survey report about two techniques that have been proposed and reviewed to mitigate a sort of risk with AI. They do so by implementing additional steps to the normal development process of AI models, to let the developers and other bystanders have concepts such as interpretability, fairness, and ethics in mind while doing so.

Mitigating risks in AI has proven to be a lot more complex than I initially thought, where the challenges are so diverse and therefore require a multitude of approaches to do so. I have seen that due to the topic being more applicable in our times as people are being more aware of AI, the vast amount of research will continue to further our understanding of the topic and help us become more pro-active while developing tomorrow's AI.

4.3 Further work

As I have mentioned there is a lot of available research on the topic and simply investigating ways to mitigate the risks with AI, I believe it would prove useful to narrow the scope down further. I did find a lot of recent research and the field seems to be quickly growing.

Sources

[1] B.J. Copeland. Artificial intelligence https://www.britannica.com/technology/artificial-intelligence [2021 May]

[2] Rockwell Anyoha. Can Machines Think? https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/ [2021 May]

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik , Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. https://www.sciencedirect.com/science/article/pii/S1566253519308103 [2021 May]

[4] Marco Tulio Ribeiro , Sameer Singh and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. https://arxiv.org/pdf/1602.04938.pdf [2021 May]

[5] Marco Tulio Ribeiro. LIME - Local Interpretable Model-Agnostic Explanations. https://homes.cs.washington.edu/~marcotcr/blog/lime/ [2021 May]

[6] Filippo Remonato. Explainable AI (XAI). https://www.sintef.no/ekspertise/sintef-ikt/explainable-ai-xai/ [2021 May]

[7] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa and Rayid Ghani. Aequitas: A Bias and Fairness Audit Toolkit . Explainable AI (XAI). https://arxiv.org/pdf/1811.05577.pdf [2021 May]