ALGORIHTMIC BIAS IN SOUTH AFRICA

How incomplete and unrepresentative data caused by a history of Apartheid can be addressed to prevent systemic racial discrimination

by

Alecia Vermeulen

Student number: 160152

Major: Interactive Development

Research report submitted in partial fulfillment of the requirements for the Bachelor of Arts (Honours) degree in Visual Communication

The Open Window

November 2019

Supervisor: Carly Whitaker

Co-supervisor: Lala Crafford

ABSTRACT

In this paper I explore the concept of algorithmic bias, specifically a training data bias, which presents a racial bias. I investigate how the aforementioned algorithmic bias can manifest in South Africa due to a legacy of Apartheid that has rendered the counties data to be incomplete and unrepresentative in relation to non-white groups. I frame how the aforementioned algorithmic bias can perpetuate systemic racial discrimination in South Africa, before investigating the mitigation methods transparency and algorithmic affirmative action as a means to address and prevent the continuation of the aforementioned algorithmic bias.

This research looks into three case studies; the first case study is a game called *Survival of the Best Fit* that educates players about racial discrimination in a hiring process that is caused as a result of bias training data. The second case study that this research examines is a game called *Monster Match* that simulates how dating apps can present a racial bias based on learning from the user's input data. Lastly a driver drowsiness detection system that presents a racial bias when placed in a South African context is reviewed. All three case studies present the consequences of using unrepresentative training data to further train the algorithmic model, either from the initial training data or from contributing to the training data. The value of transparency and algorithmic affirmative action is then analysed in relation to each case study to determine if a training data bias, presenting a racial bias can be prevented.

With this information, a practical response to these findings, in the form of an educational mobile game about the aforementioned algorithmic bias, is conceptualised in order to highlight and explore how South African citizens can prevent a training data bias, which presents a racial bias.

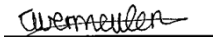
The research suggests that the value of transparency and algorithmic affirmative action is dependent on the context of its use. A training data bias, which presents a racial bias can be prevented through a blended approach. The research concludes by acknowledging that human intervention is key to address the concern of algorithmic bias in South Africa.

Keywords

AI
Algorithmic model
Algorithmic bias
Training data bias
Racial bias
South Africa
Apartheid

DECLARATION OF AUTHORSHIP

With this declaration I wish to state that the research report submitted for the degree Bachelor of Arts (Honours) in Visual Communication at The Open Window School of Visual Communication is my own work. I further declare that a comprehensive list of references in this research report contains all sources cited or quoted.



Alecia Vermeulen

ACKNOWLEDGEMENTS

Writing this research paper was like growing a fragile and endangered plant without the necessary green thumb. With the help and guidance of my study leaders, Carly Whitaker and Lala Crafford a little seed was nurtured to grow and flourish. Thank you for the endless feedback and support.

I would also like to acknowledge my parents. Thank you for allowing me the opportunity to do my honours. I love you both form the bottom of my heart.

TABLE OF CONTENTS

LIST OF FIGURES

Page

CHAPTER ONE: AN INTRODUCTION TO ALGORITHMIC BIAS

1.1.1 Introduction to study

Our everyday lives are ruled by the use of algorithmic models. We go on social media platforms such as Facebook where an algorithmic model filters what post is seen first. We do online shopping and an algorithmic model predicts what other products we are likely to want to buy too. We apply for credit where an algorithmic model evaluates whether we are credit-worthy. Kemper a PhD candidate at the University of Amsterdam and Kolkman an academic at the Jheronimus Academy of Data Science states that we live an 'algorithmic life' (2018:2). Algorithmic models can consist of a singular algorithm or a collection of different algorithms grouped together to be used for a variety of tasks (Kemper & Kolkman 2018:2).

There are many different definitions of an algorithmic model. In this research the following definition will be used: an algorithmic model represents an operation where an input is processed according to a set of rules in order to return an output (Kemper & Kolkman 2018:2). These operations are used to retrieve information, filter information, provide recommendations and execute image recognition amongst other tasks (Kemper & Kolkman 2018:2). We find these algorithmic models embedded in artificial intelligence (AI) systems such as amazon's Alexa, smart watches, and Siri an automated feature on the iPhone. These AI systems consist of many different algorithmic models which generate and use our data to predict habits and patterns, even suggesting new behaviours. According to Kenney (2018:4) a distinguished professor at the University of California some examples of where these algorithmic models are being used currently include software programs that predict where future crimes might occur, which job candidate is the most suitable to employ and distinguishing between individuals who are credit worthy and those who are not. The application of these algorithmic models is diverse and expanding continuously.

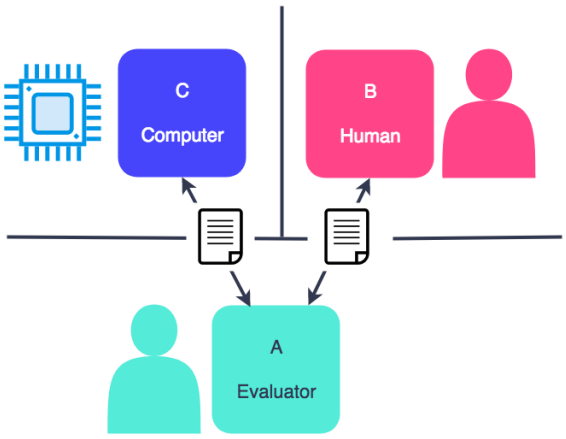


Figure 1: Vittorio Banfi, *What is the Turing test.* Digital illustration. (Botsociety Blog 2018).

I refer to the work of Ashley Nicole Shadowen (2017:4) called *Ethics and Bias in Machine Learning: A Technical Study of What Makes Us "Good"* who has a Master of Science degree to define artificial intelligence (AI) as technology that appears to succeed at making human-like decisions. It is when a piece of software or a device appears to respond or behave like a human, such as Siri's responses. The Turing test, which was developed in 1950 by Alan Turing a computer scientist who helped break Enigma (Hom, 2019) can be used in order to test whether artificial intelligence has been achieved as diagrammatically explained in Figure 1. The test consists of an evaluator (A) interacting with a messaging system with two other agents, where one agent is a human (B) and the other a machine (C). The human tester asks questions receiving responses from both the human and computer agents. If the evaluator fails to identify which one is the human response AI has been achieved.

A component of AI is machine learning which is when a machine, computer or software has the capabilities of AI technology to learn without having to be programmed (Shadowen 2017:4). Machine learning takes input or training data and 'learns' how to identify patterns in order to return an output that has not been hardcoded beforehand through different algorithmic models. This initial data is crucial for the way the machine learns with the algorithmic model and will determine how the algorithmic model evolves and the AI then responds. Discriminatory results occasionally arises as a result of machine learning. Kemper and Kolkman (2018:6) state that the discrimination cannot be seen in the code, but is a pattern that is identified in the training/input data and reproduced by the algorithmic model. Therefore although algorithmic models are praised for their ability to provide reliable

and objective results, it is possible for an algorithmic model to be biased (Kemper & Kolkman 2018:2). This can occur for a variety of reasons, one of which is the initial data that the algorithmic model develops off of. This means that the output of algorithmic models should not and often cannot be regarded as completely accurate or neutral, because it can often reflect a bias that exists in society or a specific context.

According to Cherry a psychosocial rehabilitation specialist (2019) numerous psychologists believe that humans are inherently biased in order to allow our species to make decisions as quickly as possible. The inherent bias, a cognitive bias, is therefore necessary in order for the human species to survive and evolve. However, although cognitive bias is helpful in an attempt to make quick decisions, it can unfortunately also lead to negative acts such as discrimination. For instance if you believe that a certain area is dangerous it might lead to a cognitive bias based on assumptions where you assume that any individual from this area is dangerous, however the individual might very well be harmless. Such acts of inherent human bias becomes of great concern when considering the future of emerging AI and machine learning technologies. For instance, a hiring software that might screen applicants for a specific job description, might not hire individuals from a certain area because it has learned from past decisions (data) not to. When an algorithmic model is inaccurate and reflects a bias, this is referred to as algorithmic bias (Kemper & Kolkman 2018:2).

The algorithmic bias classification model outlined by Danks and London in their article *Algorithmic Bias in Autonomous Systems* (2017) can be used in order to explore the different types of algorithmic bias that can arise in an algorithmic model. Danks and London (2017: 4692-4694) identify five different types of algorithmic bias; algorithmic focus bias, algorithmic processing bias, transfer context bias, interpretation bias and training data bias. This study will focus on a training data bias. A training data bias occurs due to inconsistency in the training or input data used to train an algorithmic model (Danks & London 2017: 4692-4693). In other words the data that the algorithmic models trains on might be unrepresentative of the true value or otherwise incomplete. The proposed algorithmic bias, training data bias can present subsequent biases such as gender, race, age, religion, skill or location.

This research however will focus on any racial bias that is presented consequently. Finally the proposed algorithmic bias will be evaluated in relation to emerging AI technology in South Africa. The reasoning behind the formerly mentioned decisions are to ensure that the research leads me to determine how systemic racial discrimination can be caused by algorithmic models that exhibit a training data bias and furthermore determine how it can be prevented. In order to explore how algorithmic bias might manifest through AI technology in South Africa and further perpetuate racial discrimination it is important to first analyse South Africa's history of bias.

Anupam Chander, a professor of Law at the University of California, states in his paper *The Racist Algorithm* (2017:1045) that when we review "emerging technologies, we must be careful not to romanticize a pretechnological past". I believe this statement is particularly relevant to South Africa because our pretechnological past is influenced by the ideologies and legacy of Apartheid. Apartheid was a system in South Africa that enforced racial segregation (Smith 2019). This segregation occurred in many aspects of daily life such as education, the different neighbourhoods where people lived, access to beaches, public facilities, medical care and more. Apartheid not only enforced segregation, but discriminated against the largest part of South Africa's population, which were classified as non-white and second class citizens (Smith 2019). These people were not included and due to Apartheid and it's discriminatory practices, a lasting bias affect was left on South Africa's data and information about the history of our population. According to Christopher a professor at the Nelson Mandela Metropolitan University (2011:16-17) in the past, the statistical system of South Africa recorded little to no data on individuals who were classified as non-white. As a result large amounts of incomplete and unrepresentative data currently exists in South Africa. The historical data reflects a human bias and an historical injustice which developed over the years. This bias can therefor affect the type of data that is used in different AI and machine learning systems for South Africa. For example, historical data on employment may show non-white individuals getting promoted less than white individuals in South Africa. If an AI system trained on such data identifies such a pattern through machine learning, the system will conclude that non-white individuals are worse hires, and perpetuate racial discrimination in the future.

AI technology driven by algorithmic models should therefore be carefully evaluated before being embedded in our society so as to prevent harmful biases from being automated and perpetuated through machine learning. This paper will investigate different methods and solutions for mitigating this bias in algorithmic models and propose a solution to prevent this source of algorithmic bias from perpetuating racial discrimination in South Africa.

1.1.2 Research Question

The research question I propose to investigate is:

How can algorithmic models that presents a racial bias as a result of incomplete and unrepresentative data due to a history of Apartheid, be circumvented to prevent systemic racial discrimination in South Africa?

1.1.3 Rationale

According to Chander (2017:1027) “In a society where discrimination affects opportunities in innumerable ways, we must worry about the migration of discrimination to decision-making by algorithms”**.** I agree with Chander that we should be concerned about algorithmic models that discriminate because it can have a negative impact on society to disadvantage and exclude certain individuals in the use of technology. I am specifically concerned about how algorithmic models that discriminate can negatively impact and violate the human rights of South Africans. Therefore I have chosen the proposed research question because I believe that the research involved will help to uncover how systemic racial discrimination, that is perpetuated through AI technology in South Africa, can be prevented.

As a South African developer I feel it is important to understand how algorithmic models may contain biases that can negatively impact my country. This knowledge will enable me to recognise when this occurs in my field and ensure that I do not contribute to this problem through my own work. Algorithmic models that present a training data bias and consequently a racial bias, will result in marginalising specific

racial groups by excluding them in the use of certain AI technologies such as voice recognition, facial recognition, and more. The exclusion of certain racial groups form emerging AI technologies could consequently negatively influence South Africa's economy. Additionally, understanding how algorithmic models may contain biases is the first step to uncovering how it can be prevented. An in depth analysis and understanding of mitigation methods and solutions for algorithmic bias can help to encourage public trust which is important to ensure that South Africa can stand to benefit from the social and economic potential of AI technology.

1.1.4 Aims and objectives

The main aim is to critically evaluate how problematic algorithmic bias such as a training data bias, which presents a racial bias on manifestation in a South African context, can be circumvented in order to prevent systemic racial discrimination in South Africa.

In order to realise this aim, the following objectives are necessary:

- Gather a basic understanding of how algorithmic models can contain biases inherent in biased data, specifically racial biases.
- Review examples of how a training data bias can present a racial bias in algorithmic models in a South African context according to available literature.
- Evaluate how a training data bias algorithmic model presenting a racial bias, enforces systemic racial discrimination in South Africa.
- Investigate how biases in algorithmic models can be circumvented by critically evaluating the mitigation methods transparency and algorithmic affirmative action.
- Apply findings from analysing mitigation methods to three case studies where a training data bias, presenting a racial bias is evident in order to further evaluate the value of each mitigation method.

1.1.5 Methodologies and chapter outline

This research will be conducted from a qualitative approach as I will research different methodologies and solutions as part of uncovering how a training data bias presenting a racial bias in algorithmic models can be prevented.

The preceding elements of the first Chapter serves to outline an introduction to the key definitions and concepts involved in the proposed research question. It states why the study is important to undertake and evaluates the delimitations of the study. The rest of the Chapter will serve to outline in more depth how algorithmic models can contain biases. Lastly the chapter will contextualise the dangers of algorithmic bias in a South African context.

Chapter Two will explore different mitigation methods to prevent algorithmic bias. A more in depth review of transparency and algorithmic affirmative action will be explored. The benefits and disadvantages of both mitigation methods will be discussed. Lastly the mitigation methods will be reviewed in a South African context.

In Chapter Three I will investigate three case studies that reveal a training data bias, presenting a racial bias. Two of the three case studies will present a training data bias, presenting a racial bias in a game context. The last case study focuses on a training data bias, presenting a racial bias in South Africa specifically. The Chapter will explore key concepts of algorithmic bias as outlined in Chapter One in relation to each case study. As part of developing an answer to the research question, the benefits and disadvantages of the mitigation methods transparency and algorithmic affirmative action as outlined in Chapter Two will be applied to each case study.

Chapter Four will serve to describe and explain the practical component of the research. It will examine how it contributes to the research component of the study by detailing how the findings of the proposed research is practically explored and implemented.

Chapter Five will summarise the process and methodology of the study. In addition important points from the research will be restated. The conclusion and outcome of

the research will be presented. Additionally this Chapter will question and speculate what future research on the matter might reveal.

1.1.6 Delimitations

In this research paper, the following delimitations are set in order to ensure that the outcome of the study provides efficient information to address the research question.

- This research paper acknowledges that there are numerous different types of algorithmic bias, but will only focus on a training data bias.
- This research paper acknowledges that a training data bias can present numerous subsequent biases, but will only focus on a racial bias.
- This paper recognizes that there are various mitigation methods that can be used to prevent a training data bias presenting a racial bias from enforcing systemic racial discrimination in South Africa, but only transparency and algorithmic affirmative action will be explored in depth.

1.2 Algorithmic bias

This research will use the definition of 'bias' as the authors Silberg & Manyika (2019:2) in their article *Notes from the AI frontier: Tackling bias in AI (and humans)*, who define bias as "any form of preference, fair or unfair", but go on to further define the term bias by referring to work of Friedman & Nissenbaum (1996:332) who defines it as "systemic discrimination against certain individuals or groups of individuals based on the inappropriate use of certain traits or characteristics". Following this understanding the term algorithmic bias can be defined as an algorithmic model that returns a result which contains biases (Kemper & Kolkman 2018:2). As previously mentioned the algorithmic bias classification model by Danks and London (2017: 4692-4694) state that the five different types of algorithmic bias are; algorithmic focus bias, algorithmic processing bias, transfer context bias, interpretation bias and training data bias.

The first type of algorithmic bias, algorithmic focus bias occurs when protected traits (such as race, gender, religion, and disability) are included or excluded in the training

or input data, which is then used to make predictions and judgements in decision-making procedures (Danks & London 2017:4693). According to Kenney (2018:16) an example of where including a protected trait such as gender is beneficial can be seen when referring to a health diagnostic algorithm. Individuals can have a higher risk for certain health conditions based on their gender which therefore means that including this trait would be more beneficial and would actually prevent a bias outcome. Moreover Kenney (2018:16) elaborates that an example of where including a protected trait such a race or gender is disadvantageous can be seen when referring to a sentencing algorithm as it can lead to discriminatory results seeing as these traits do not determine how likely a criminal is to reoffend. Developers should therefore consider and be aware of the harm that can be caused both by including or excluding protected traits by considering the context that the system will be used in. Moreover developers should also be aware that certain information can be used as a substitute for protected traits such as location or address. Begbie (2019) a Data Scientist at the company Praekelt.org, describes how the location or address of an individual can be used as a proxy for race in South Africa by referring to Parry and Eeden (2015) who state that SA “[remains] highly segregated” geographically as a result of Apartheid in the major cities. This means that even if protected traits such as race are excluded in the information given to an algorithmic model the location or address of an individual might be used as a proxy variable for race.

A second source of algorithmic bias, algorithmic processing bias arises when the bias is embedded in the code itself through the use of variables (Kenney 2018:18). A booking website which disables users from selecting a ranking below a 2 (out of 10) for a hotel or accommodation is an instance of algorithmic processing bias (Kenney 2018:19). This limitation means that the ranking for the hotel or accommodation will reflect a higher standard than the truth which could be a 1 or 0.

According to Danks & London (2017:4694) transfer context bias occurs when an algorithmic model is used outside of the context that it was created for. For instance if a self-driving car created in America is deployed in a different country such as South Africa, where cars drive on the opposite side of the road, the algorithmic model would be bias in the sense that it gives an unfair/inaccurate prediction

because the system was never trained to be deployed outside of the country that is was created for.

The algorithmic bias, interpretation bias can be characterised as a user error case where the user misinterprets the output of the algorithm (Danks & London 2017:4694). For instance, one judge might look at a prediction result of a 6/10 for the likelihood of a criminal to reoffend and perceive it to be high-risk and give a five year sentence while another judge might perceive the result as a low-risk giving a three year sentence.

Lastly according to Danks and London (2017: 4692-4693) a training data bias occurs due to inconsistency in the training or input data. In other words the data might be unrepresentative of the true value or otherwise incomplete. In this research paper I will be focussing specifically on algorithmic models that reveal a training data bias within a South African context.

1.2.1 Training data bias

Danks and London (2017:4692) state “one route to algorithmic bias is through deviations in the training or input data provided to the algorithm”. In other words it’s a type of algorithmic bias that exists when the input/training data that is used to train an algorithmic model is not diverse enough. To further clarify deviations in the training data usually exists due to incomplete or unrepresentative datasets that are used to train an algorithmic model. This sort of algorithmic bias is referred to as a training data bias. According to Danks and London (2017:4692) a training data bias is not easy to detect because the data used to train the algorithmic model is usually not publicly accessible.

I will focus on investigating this type of bias in algorithmic models in relation to incomplete and unrepresentative data in South Africa due to Apartheid. This view is shared by Christopher (2011:16-17) who identifies that the censuses conducted during Apartheid, when South Africa was called the Union of South Africa from 1910 described the population “essentially that of a White nation to the exclusion of other people, clearly undermining the link noted elsewhere between census taking and

nation building." The incomplete and unrepresentative data concerning non-white groups can therefore result in a training data bias when training algorithmic models. The training data bias can consequently present secondary biases. As previously stated this research paper will focus on a training data bias presenting a secondary racial bias. Such an instance in a South African context is to be discussed in the next section.

1.2.2 Racial bias in a training data bias.

An algorithmic model that exhibits a racial bias is an algorithmic model that returns a result which is discriminatory based on the attribute of race (Kemper & Kolkman 2018:3).

The predictive algorithm called PredPol created by PredPol a predictive policing company is an example of a training data bias, which presents a racial bias in the United States (US). The predictive algorithm is trained on historical crime data in order to predict and prevent future crime in the US (Kenney 2018:12). According to Kenny (2018:12) this algorithmic model is an example of a training data bias, which presents a racial bias due to the fact that it is based on historical data that reflects the real world racial bias of the justice system in the US. This means that the output will reflect the racial bias of the data that it has been trained with.

According to Begbie (2019) a training data bias presenting a racial bias can manifest in a South African context due to our history of apartheid that has limited our diversity of data. An example of this can be seen by referring to our data resources of the eleven official languages in South Africa. English is the only language that has sufficient digitally recorded data in comparison to the other official languages of South Africa (Begbie 2019). It is worth noting at this point that during Apartheid many different groups of people were not formally taught in their mother tongue, but were forced to learn Afrikaans which became the states way of controlling and monitoring educations personal growth (Heffernan 2016). As a result many African descendant languages have been passed down from generation to generation, but has not been formally recorded and taught amongst the population of South Africa. This means that AI technology such as voice recognition software will fail to include a diverse

representation of African descendent languages of South Africa. Even though languages are not race specific, some languages are primarily spoken by certain racial groups and reflect certain cultures and context. An article titled *Languages spoken in South Africa per race group according to the latest General Household Survey (GHS)* (2019) reveals that the most spoken language inside and outside the house in South Africa is isiZulu at 25.3%. The majority of Black Africans contribute to the aforementioned statistic as can be seen by referring to Figure 2 below. Therefore if data sets of specific African descendent South African languages are incomplete or unrepresentative a whole racial group can be excluded from using voice recognition software in their language of choice. Prof Tshilidzi Marwala, a Professor at the University of Johannesburg, experienced such a racial bias when using the voice recognition software Google Assistant. The software was unable to pronounce his name due to unrepresentative data on African descendent languages and accents (Marwala 2018). Another example of a training data bias with a racial bias in a South African context can be seen by referring to Prof Marwala's experience using the application Airbnb. The application required him to take a selfie, but then failed to detect his facial features several times, reinforcing that African faces are less represented in the input or training data when compared to European faces (Marwala 2018).

Table 3.1: Percentage of languages spoken by household members inside and outside household by population group, 2018

	Black African		Coloured		Indian/Asian		White		South Africa	
	Inside	Outside	Inside	Outside	Inside	Outside	Inside	Outside	Inside	Outside
Afrikaans	0,9	1,0	77,4	68,8	1,3	1,5	61,2	37,2	**12,2**	**9,7**
English	1,6	8,6	20,1	28,3	92,1	95,8	36,3	61,0	**8,1**	**16,6**
IsiNdebele	1,9	1,6	0,0	0,0	0,3	0,2	0,3	0,1	**1,6**	**1,3**
IsiXhosa	18,2	15,6	1,1	1,3	0,4	0,0	0,1	0,1	**14,8**	**12,8**
IsiZulu	31,1	30,8	0,3	0,3	0,9	1,0	0,5	0,5	**25,3**	**25,1**
Khoi, Nama and San languages	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	**0,1**	**0,1**
Sepedi	12,4	12,0	0,3	0,2	0,5	0,2	0,1	0,3	**10,1**	**9,7**
Sesotho	9,7	9,6	0,1	0,2	0,1	0,3	0,0	0,1	**7,9**	**7,8**
Setswana	11,1	11,5	0,7	0,8	0,2	0,2	0,4	0,4	**9,1**	**9,4**
Sign Language	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,0**	**0,0**
SiSwati	3,5	3,2	0,0	0,0	0,0	0,0	0,0	0,0	**2,8**	**2,6**
Tshivenda	3,1	2,7	0,0	0,0	0,2	0,0	0,0	0,0	**2,5**	**2,2**
Xitsonga	4,4	2,9	0,0	0,1	0,1	0,1	0,0	0,0	**3,6**	**2,4**
Other	2,1	0,5	0,1	0,0	4,0	0,7	1,1	0,5	**1,9**	**0,5**
Total Percentage	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
Total (Thousands)	46 307	46 135	4 961	4 930	1 430	1 426	4 442	4 420	57 143	56 914

Figure 2: South African Market Insights, *Languages spoken in SA.* Digital illustration. (South African Market Insights 2019).

1.3 Chapter conclusion

In conclusion, this Chapter serves to introduce key concepts and definitions involved in the proposed research question such as algorithmic models, AI and machine learning. It outlines how humans are inherently biased and details how a system of Apartheid was established in South Africa due to human bias. It further introduces the argument that historic events such as Apartheid can have a negative impact on emerging AI technologies in South Africa as a result of incomplete and unrepresentative data which can lead to these technologies presenting algorithmic bias. The Chapter explores the different types of algorithmic biases, specifically a training data bias presenting a racial bias. Lastly the Chapter contextualises how this type of algorithmic bias can present itself in a South African context. This information will be used in Chapter Two to analyse mitigation methods as a way to determine if the aforementioned bias can be prevented in a South African context.

CHAPTER TWO: MITIGATING METHODS

This Chapter will discuss various mitigation methods and solutions that can be used as a means to stop AI systems that present algorithmic bias from perpetuating discrimination. The method 'transparency' and 'algorithmic affirmative action' will then be looked at in relation to a training data bias presenting a racial bias. This will then further be applied to a South African context.

2.1 General overview of mitigation methods

It seems the most obvious solution to a training data bias, which entails that the bias is inherent in the data being used to train an algorithmic model, would be to simply use unbiased data. In other words, the solution would be to use representative and complete datasets to train the algorithmic model. However, the training data used is usually data accumulated historically which in itself could be subjected to historical injustices. Take for instance South Africa's history of Apartheid where historical injustices mean that gathering unbiased data is nearly impossible as identified in the previous Chapter. Therefore, it seems the most obvious solution might not be the most achievable. Perhaps the solution is then to ensure that our future data is not biased? Could the real solution to a training data bias then be to eradicate inherent cognitive bias?

Yaël Eisenstat, a former CIA officer who worked at Facebook as head of the integrity operations unit, disagrees with the aforementioned solution and states in her article *The Real Reason Tech Struggles With Algorithmic Bias* (2019) that "Humans cannot wholly avoid bias, as countless studies and publications have shown. Insisting otherwise is an intellectually dishonest and lazy response to a very real problem". Eisenstat (2019) suggests that tech companies should educate their employees on what a cognitive bias is so that they will be able to detect it in the data being used to train algorithmic models. Perhaps the real solution is then not to focus on gathering pure unbiased data, but to ensure that humans (developers, anthropologists or ethnographers) are continually involved in reviewing AI systems to detect the inherent cognitive biases of humans and proceed to use technical methods to remove them. Some of these technical methods include tool kits such as the AI

Fairness 360 created by IBM. This tool kit is a python open-source tool kit meant to help developers detect, understand and mitigate unwanted algorithmic biases by referring to a fairness metric system for datasets and models and using other algorithmic models to mitigate any exiting algorithmic biases (Silberg & Manyika 2019).

There are many different approaches to mitigate algorithmic bias as discussed in the section above, this list is however inexhaustive. This research paper will focus on the following two mitigation methods in detail: transparency and algorithmic affirmative action. By doing so I hope to determine with certainty whether applying these mitigation methods can help to answer my proposed research question: How can algorithmic models that presents a racial bias as a result of incomplete and unrepresentative data due to a history of Apartheid, be circumvented to prevent systemic racial discrimination in South Africa?

2.1.1 Algorithmic transparency

2.1.1.1 What is transparency?

As algorithmic models become more embedded in society through AI technologies and return results that reveal how biases can be reproduced through these systems a call for accountability becomes evident. However, questions arise such as who should be held accountable when systems discriminate? How do we hold them accountable? One highly debated approach to address these questions is the mitigation method transparency. According to Kemper and Kolkman (2018:3) in *Transparent to whom? No algorithmic accountability without a critical audience* who provide a critical perspective on transparency as a mitigation method, transparency refers to the ‘understandability of a specific model’. This means that transparency helps the public to understand algorithmic models by enabling them to see and have access to all of the stages of an algorithmic model such as input, instructions and output. Transparency can be implemented by publishing datasets and data sources, ensuring that the public has access to review systems that employ algorithmic models for decision making procedures (Kemper & Kolkman 2018:3).

2.1.1.2 Advantages and disadvantages of transparency

I believe that by applying transparency to algorithmic models, an increase in trust surrounding AI systems can be achieved because users will trust emerging technology more if they have access to understand the inner workings of how the technology works. At this point it is important to note that having access will not necessarily translate to comprehension regarding the available material. This view is shared by Kemper and Kolkman (2018:6) who state:

> It is possible for organizations to share all available documentation, procedures, and code, yet this will not constitute transparency if that information is not understood by the relevant audience.

Therefore, it seems that the value of transparency is depended on us as a society to be involved, reflect and contribute to the data and the algorithmic models used. Transparency alone is therefore not a single solution, but part of the solution.

Ananny and Crawford (2016:9) state in their paper *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability* that for transparency to have value means "making sure that a system is not only visible but also debated and changeable by observers who are able to consider how they know what they know about it". Following this understanding in order to learn about complex systems society should look beyond just having access to see the information, but interact in such a way with the system that an understanding of it in relation to its environment can be achieved. This could mean presenting users with a system where they have the ability to influence and change results given by an algorithmic model so as to better understand how our choices and data as people influences technology.

An advantage of transparency as a mitigation method is that it can be used to hold developers who create algorithmic models responsible and accountable for the algorithmic models that they create (Kemper & Kolkman 2018:3). However, Ananny and Crawford state (2016:6):

> If transparency has no meaningful effects, then the idea of transparency can lose its purpose. If corrupt practices continue after they have been made transparent, 'public knowledge arising from greater transparency may lead to more cynicism, indeed perhaps to wider corruption'. Visibility carries risks for the goal of accountability if there is no system ready and 'capable of processing, digesting and using the information' to create change.

Therefore, it seems the advantage of transparency to bring about gains in accountability is dependent on if there are procedures in place to ensure that it is not only seen, but changed.

Algorithmic models have become more complex due to the use of big data (Kemper & Kolkman 2018:3). Frequently the bias exists in the training data. Ananny and Crawford (2016:9) describe how in 2015 Google's Photo app misidentified black people as "gorillas" as a result of having unrepresentative data concerning black people's faces, a blatant racial bias as a result of a training data bias. However, other instances have occurred where the training data could not be identified as the cause of the bias such as Google's system which misidentified people of all races as dogs (Ananny & Crawford 2016:9). The engineers behind the Google system could regardless of having access to the entire system not identify why the system returned such a result (Ananny & Crawford 2016:9). Following this understanding Kemper and Kolkman (2018:10) reveal that some algorithmic models cannot be understood because of complexities that arise from applying concepts such as machine learning to these models. The aforementioned brings about concern that transparency then as a mitigation method might be limited if seeing the model will not bring any clarity as to why it functions as it does. Should developers then be held accountable for deploying systems which can't be explained or predicted? Or should we as a society hold ourselves accountable for integrating these systems which we cannot predict into our lives? Perhaps a bit of both is valid.

Nathan Begbie's article *Problematic algorithmic bias; on manifestation in a South African context and methods for mitigation* (2019), which contextualises the value of the mitigating method transparency in South Africa states that transparency will allow the public to review algorithmic models which in turn can help to expose any biases that might have been missed by the developers behind these algorithmic models. However, algorithmic models are being created at an unprecedented speed which

means that quality checking and control has become limited (Kemper & Kolkman 2018:3). Therefore, although transparency will allow the public to review algorithmic models, it is likely that it will be impossible to review these models at the same speed at which they are being released and deployed within society. This is a clear challenge that arises when reviewing the mitigation method of transparency. The alternative is to accept that humans cannot control or censor these systems which seems to be far worse as it shapes an exclusionary future of computing.

According to Annany and Crawford (2016:6) “full transparency can do great harm. If implemented without a notion of why some part of a system should be revealed”. Following this understanding companies cannot always disclose their entire code to the public because they have to keep certain parts protected to prevent their intellectual property (IP) from being misused by individuals or competitors (Kemper & Kolkman 2018:6). However, Begbie (2019) does state that companies can limit access to their Application Programming Interface (API) through an auditing agreement in order to protect their intellectual property. Lastly, according to Begbie (2019), exposing datasets that are used in algorithmic models to the public can violate the trust of users who have been told that their data will not be shared. A user’s personal data needs to be treated with sensitivity. Although this might prove to be a challenging aspect for companies to overcome, there are ways to anonymize data in order to ensure the ethical sharing of user data (Begbie 2019). Do the disadvantages of the mitigation method transparency outweigh the advantages? I think the value of this mitigation method depends on the context of its use.

Although there are clear challenges and limitations when reviewing transparency as a method to prevent algorithmic bias, the core principle of this method is to create awareness of the training data and to allow the public to have access to question and review systems that employ algorithmic models for decision making procedures.

2.1.1.3 Transparency in SA

In South Africa there are unfortunately no clear regulatory guidelines which discourages companies from sharing and making datasets publicly available (Begbie 2019). This can prove challenging when trying to apply transparency to prevent

algorithmic bias in training data in South Africa. Accordingly, this challenge can be solved and eliminated if regulatory guidelines are addressed in the future which then would mean that the value of transparency can be utilised to address public concern about algorithmic bias in South Africa.

2.1.2 Algorithmic affirmative action

2.1.2.1 What is affirmative action?

If the mitigation method transparency discussed above is limited to prevent algorithmic bias in South Africa what might an alternative mitigation method be? One such mitigation method called algorithmic affirmative action has been suggested by Chander as a better alternative to address algorithmic bias. Chander (2017:1041) defines algorithmic affirmative action "as a set of proactive practices that recognises deficiencies in the equality of opportunity and act in a multiplicity of ways to seek to correct for those deficiencies". Moreover, the goal of algorithmic affirmative action is not to focus on solving the problem of discrimination in our society, but to actively attempt to rectify the problem by designing "algorithms for a world permeated with the legacy of discriminations past and the reality of discriminations present" (Chander 2017:1025). This mitigating method seeks to effectively recognize in society where women and minority groups have suffered as a result of discrimination and through awareness of these injustices prevent biases from forming during the design phase of algorithmic driven AI systems.

Affirmative action in data training encourages the inclusion of protected traits such as race and gender to be included in the structure of algorithmic models so that women and minority groups can be fairly represented (Chander 2017:1041). According to the paper *Is Algorithmic Affirmative Action Legal?* By Jason R. Bent (2019:4) "Machine learning scholars increasingly agree that the best way to get fair algorithmic results is not by hiding the protected trait, but instead by using the protected trait to set a fairness constraint within the algorithm design". This means that the data intended to be used as the training dataset will be reviewed to see if historically data on some groups are unrepresented or incomplete and then altered so that the output result will not discriminate based on the bias in the data. For example, the training dataset for a hiring algorithmic model can be reviewed to

reveal that more men have been hired in the past by a specific company than women. This could be the case as a result of human bias against women in the past. Software engineers would then encourage the inclusion of gender information on CV applications and would further set a variable in the code to ensure that equal men and women are hired by the algorithmic model to ensure diversity in the company.

2.1.2.2 Advantages and disadvantages of affirmative action

It is important to note that Chander states (2017:1042) that corporations are hesitant to apply algorithmic affirmative action because they fear that they might later be accused of reverse-discrimination. The possibility of reverse-discrimination has prompted some to take an anti-affirmative action viewpoint. Chief Justice John Roberts reflects such a viewpoint in his statement "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race" (Chander 2017:1041). He believes that affirmative action is not the solution, but in fact a hard-coded form of racism. Others disagree such as Justice Sonia Sotomayor who believes that "The way to stop discriminating on the basis of race is to speak openly and candidly on the subject of race, and to apply the Constitution with eyes open to the unfortunate effects of centuries of racial discrimination" (Chander 2017:1041). The idea of algorithmic affirmative action is therefore not to harm, but to address the harm that has been caused by past injustices and to provide support for groups that have previously been discriminated against. However, even though the idea of algorithmic affirmative action is to promote equality, implementing this method can prove to be complicated due to uncertainty regarding what constitutes fairness.

Silberg and Manyika (2019:3) describes an image search for a CEO scenario in order to articulate that 'fair' is a complex result to achieve. The scenario raises questions such as: Should the algorithmic model return a result representing the true statistical value of how many women are CEO's in our society in comparison to men? Or should it return a result reflecting an ideal future where 50% CEO's are women and 50% CEO's are men? I believe these questions amount to a bigger question with regards to algorithmic affirmative action, which is whether we should intentionally modify algorithmic models if we have not yet agreed upon a clear definition of fairness. Some metrics for fairness exists such as group fairness and

individual fairness amongst countless others (Silberg & Manyika 2019:4). However, algorithmic models can not conform to all metrics of fairness. According to Silberg & Manyika (2019:4) “Other research has shown that ensuring an AI system satisfies measures of group fairness could create trade-offs with measures of individual fairness or could reduce the utility of the model”. For example, imagine a credit scoring system where algorithmic affirmative action has been applied. A group fairness metric would mean that equal loans should be granted across different race groups. On further investigation, it seems that the biggest disadvantage of a group fairness measure might be that it can potentially harm the people that the measure seeks to protect. Group fairness could ultimately lead to hurting the credit scores of individuals of a particular race if they have been granted a loan and then are unable to repay as they might be economically unstable due to past injustices (Silberg & Manyika 2019:4). Alternatively, to a group fairness measure is an individual fairness measure. Bent (2019:18) states:

> The critical difference is that instead of focussing on comparison at the group level, individual fairness approaches look to measure disparities in treatment at the individual level for individuals who are similar in their non-sensitive features

This means that individuals of different races who share similar information excluding race should equally be granted loans. However historical injustices might mean that one group is financially disadvantages compared to another. Therefore, it has been suggested to create different decision thresholds for different groups of individuals (Silberg & Manyika 2019:4), but some disagree and believe that a single threshold is fairer. Silberg & Manyika (2019:4) conclude “As a result of these complexities, crafting a single, universal definition of fairness or a metric to measure it will probably never be possible”. Is it then futile to turn to algorithmic affirmative action as a solution to prevent algorithmic bias outcomes from perpetuating discrimination? Regardless of conflicting views and clear obstacles, it seems that algorithmic affirmative action will continue to be debated with the hope that it might be the answer to prevent algorithmic bias in the future. More research on the use of algorithmic affirmative action to prevent algorithmic bias might help to resolve existing problems that the method presents. Some research has begun to emerge such as explored in the paper *Is Algorithmic Affirmative Action Legal?* by Jason R.

Bent (2019:6), a Professor of Law at Stetson University College of Law who investigated whether algorithmic affirmative action is lawful under United States antidiscrimination law. His findings reflect that there is considerate leeway with regards to both statutory and constitutional antidiscrimination law that allows race-aware affirmative action as a solution to combat algorithmic bias (Bent 2019:6).

2.1.2.3 Affirmative action in SA

The paper *Affirmative action in South Africa: an empirical assessment of the impact on labour market outcomes* (2010) by Rulof Burger and Rachel Jafta who are both professors in the Department of Economics at the University of Stellenbosch will be used to gain further insight into how algorihtmic affirmative action as a mitigation method for algorithmic bias might prove to be useful in a South African context.

Affirmative action was first introduced in South Africa in 1998 as a method that attempts to achieve employment equality (Burger & Jafta 2010:3). It was initialised as a way to address the imbalances caused by Apartheid (Burger & Jafta 2010:4). It was established with the Employment Equity Act and the Broad Based Black Economic Empowerment Act (Burger & Jafta 2010:4). Both of these acts strive to achieve equality in the workplace by providing opportunities to previously disadvantaged groups such as black people, women and individuals with disabilities (Burger & Jafta 2010:5). Burger and Jafta (2010:23) found in their research that only a slight improvement in reducing inequality in employment has been achieved with affirmative action. Moreover, they suggest that affirmative action might have helped some previously disadvantages individuals, but that these individuals were those who had the most skills and education prior to the establishment of a post-apartheid South Africa (Burger and Jafta 2010:23). As a result, Burger and Jafta (2010:23) conclude that affirmative action has been ineffective to achieve employment equality as the average previously disadvantages individual has still not been helped through its implementation. It seems questionable then whether affirmative action would be helpful in an attempt to prevent algorithmic bias if the non-technological application of this method in South Africa has failed to achieve equality.

An article titled *New court case can end affirmative action based on race in South Africa* (2019) states that affirmative action based on race might be brought to an end in South Africa. The article (2019) proclaims that “according to the SAHRC [South African Human Rights Commission], certain parts of the Employment Equity Act (EEA) do not comply with the Constitution and international conventions”. It is therefore uncertain what the future state of affirmative action in South Africa will be.

Perhaps changes will eliminate some limitations and challenges previously mentioned in this Chapter. Furthermore, a deeper investigation concerning the application of algorithmic affirmative action to algorithmic bias models in a South African context are necessary in order to fully understand both the implications and possibilities of algorithmic affirmative action as a method to answer the proposed research question.

2.2 Chapter conclusion

The Chapter discussed different methods and solutions that can be used to address an algorithmic bias, specifically a training bias presenting a racial bias. The mitigation methods transparency and algorithmic affirmative action were explored in detail. The Chapter concluded that transparency proves to be a method that could potentially lead to a solution if limitations such as accessibility, comprehensibility, and interpretability are adequately addressed. Finally, algorithmic affirmative action seeks to actively prevent a discriminatory outcome by modifying neutral algorithmic models so that it cannot reflect a training data bias, however concerns about what constitutes as fair have to be addressed in the future.

Both of these mitigation methods could potentially help to alleviate algorithmic bias if implemented in a South African context, however the limitations of each method render them both ineffective to completely eradicate algorithmic bias. Further investigation is needed to determine with certainty the value of the aforementioned analysed methods.

The mitigation methods discussed will be used in Chapter Three to analyse three case studies that show a training data bias, presenting a racial bias.

CHAPTER THREE: CASE STUDIES

This Chapter will critically investigate three case studies; the first case study is an educational game about racial discrimination in a hiring process that is caused as a result of bias training data. The second case study is a game that simulates how a dating app can develop a racial bias based on learning from the input data of the user. Lastly a driver drowsiness detection system that presents a racial bias when placed in a South African context will be reviewed. All three case studies present the consequences of using unrepresentative training data to further train the algorithmic model, either from the initial training data or from contributing to the training data. I will analyse the value of transparency and algorithmic affirmative action as outlined in Chapter Two in relation to each case study in aid of answering the research question: How can algorithmic models that presents a racial bias as a result of incomplete and unrepresentative data due to a history of Apartheid, be circumvented to prevent systemic racial discrimination in South Africa?

3.1 Algorithmic bias game: Survival of the Best Fit

3.1.1 Game description

Survival of the Best Fit (SOTBF) is an educational game which uses simulated artificial intelligence in a hiring process to educate users about algorithmic bias. The game was created by Gabor, Jihyun, Miha, and Alia a team of software engineers, designers and technologists (survivalofthebestfit.com 2019).

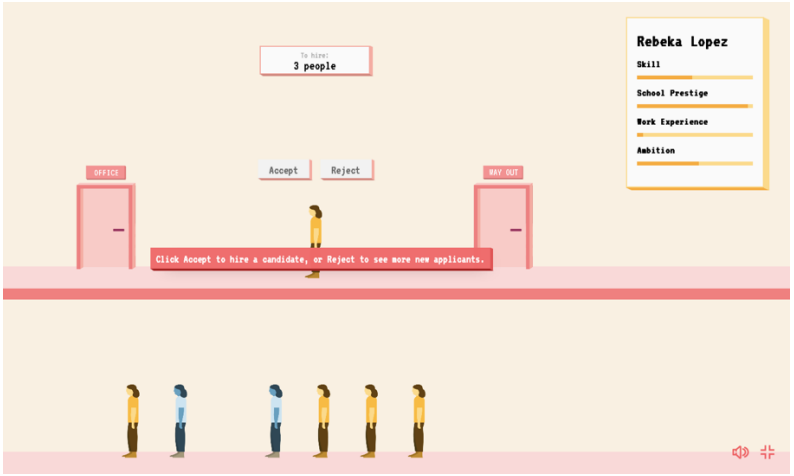


Figure 3: 'Hire candidates'. Screenshot from the game Survival of the Best Fit (2019).

The player takes on the role of a CEO of a start-up tech company who has to select candidates to hire as shown in Figure 3. As the CEO you have to ‘accept’ or ‘reject’ applicants. These decisions can be based on either reviewing the applicants CV (top right of the screen) where different skills of each applicant are measured and presented to you as the CEO, for evaluation. A more obvious differentiating factor of the applicants is their visual colour, they are either blue or yellow.

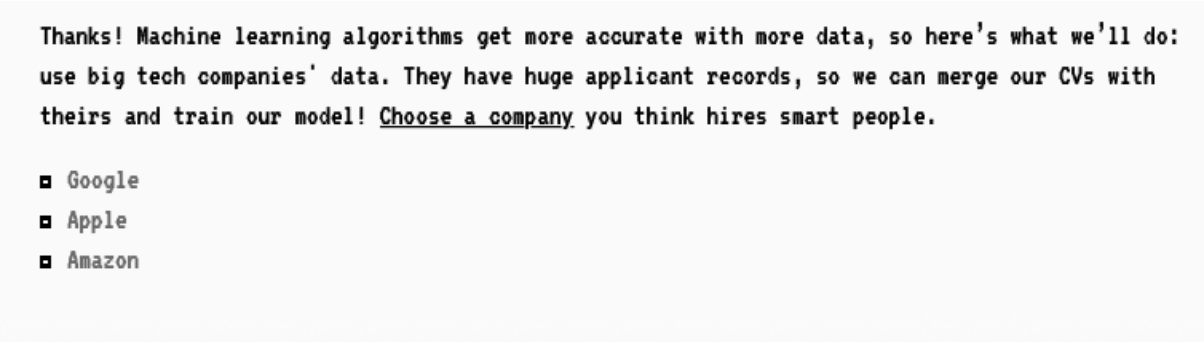


Figure 4: ‘Choose company’. Screenshot from the game Survival of the Best Fit (2019).

As the company starts to grow fast the game prompts the player as CEO to make the decisions faster regarding the applicants. The pace increases again and the game prompts you, via an investor, to hand over the data containing your past hiring decisions to the software engineers so that they can train an algorithmic model to take over the decision making process of hiring new workers. The game further prompts the player to select an existing company’s data (see Figure 4) to be used as further training data to train the algorithmic model. At first the algorithmic model performs well and reduces overall costs for the company, however eventually concerns arise when a past blue applicant named Elvan Yang enquires about why he was rejected to be hired. Upon further investigation is becomes apparent that the data used to learn the algorithmic model is unrepresentative of the blue applicants. The AI starts discriminating against blue people in the game because the training data reflected that these individuals were either not chosen in the past by the player as CEO or the selected company’s data revealed a bias where blue applicants have been hired less in the past and so continues this pattern.

As outlined in Chapter One a training data bias can exist because of unrepresentative data concerning one group. In this case, the blue applicants were unrepresented. Furthermore the training data bias presents a racial bias as a result of a group being excluded on the basis of race which in the game are any applicants that are blue. The game ends with the company being sued for hiring discrimination. SOTBF serves to illustrate how AI can exhibit inherit human biases and perpetuate inequality.

3.1.2 Applying transparency to case study

Creators of the game *SOTBF* state (survivalofthebestfit.com 2019):

> With SOTBF, we want to reach an audience that may not be the makers of the very technology that impact them everyday. We want to help them better understand how AI works and how it may affect them, so that they can better demand transparency and accountability in systems that make more and more decisions for us.

SOTBF illustrates the mitigation method transparency because they explain how the bias in their hiring AI works and present documentation and resources that they recommend to interested users and the players. The creators give users the chance to engage with algorithmic bias through a game scenario which improves understandability regarding how AI systems can present a racial bias. The use of a game as a context for education allows the player to experience a bias through the eyes of a CEO and consequently how automating the hiring process can result in a discriminatory outcome.

Unfortunately the game is hard-coded to always return a bias towards blue people, the AI isn't a real algorithmic model. The game creates the illusion that the player's choices are taken into account whereas regardless of personal choice a bias against blue applicants are always the end result. This result is brought by the additional data from an external selected company, such as Google, Apple or Amazon as shown in Figure 4. As a result the game is not as meaningful and impactful as it potentially could be. The project is open-sourced, but there is no corresponding bias machine learning algorithm that users can review in order to learn about algorithmic

bias. The additional data from the selected external company as shown in Figure 4 only creates the illusion that data is being gathered whereas the game does not make use of real data from these companies to train an algorithmic model in the game itself. Therefore users are unable to review the training data used that resulted in the game presenting a racial bias in the end. This illusion has led to many critical reviews of players who are disappointed that the game does not allow for real human bias to be reflected whilst playing. A comment reflecting such a view is made by a former player Joaquin who states “This isn’t a game, because games have different outcomes based on the player’s input. This is propaganda at best.”(Paul van der Laken 2019).

3.1.3 Applying algorithmic affirmative action to case study

Applying the mitigation method of algorithmic affirmative action to the game *SOTBF* would mean that the inclusion of race information on CV’s would be encouraged by the company in order to ensure that a diverse set of individuals would be able to be selected by the AI system. In other words the AI would be well aware of the race of each applicant and would make decisions so that a fair representation across different races are selected to be employed, such as 50% blue and 50% yellow or if blue has been unfairly discriminated against 30% yellow and 70% blue. By doing so an unfair racial bias could potentially be prevented. However as stated in Chapter Two defining what constitutes fairness is complex. It is also important to state that applying algorithmic affirmative action means that a company might not be receiving the best qualified candidates, but rather a diverse set of candidates. This could discourage players of the game from choosing to apply algorithmic affirmative action seeing as it might not be profitable.

3.2 Algorithmic bias game: Monster Match

3.2.1 Game description

Monster Match is a game that simulates how a dating app algorithmic model works and exposes how it can contain or perpetuate hidden biases. The game was created by Benjamin Berman & Miguel Perez and uses an algorithmic model called collaborative filtering (MonsterMatch 2019).

A collaborative filtering algorithmic model is used by many dating apps such as OkCupid and Tinder, it is also used by Facebook, Twitter, Google, Amazon and Netflix to provide recommendations to users (MonsterMatch 2019). Collaborative filtering is an algorithmic model that makes predictions by gathering data from users to predict what content those users and other users might also enjoy and like (MonsterMatch 2019). It can be very beneficial for recommending content such as songs, movies, news articles, products and more to users that has been indicated through past behaviour to most likely also be liked.

According to the game (MonsterMatch 2019):

> Users of dating apps make “yes” or “no” decisions on other users, one-by-one, and that data is counted to figure out whose preferences that user most resembles. Then, data from the older user is borrowed to make recommendations for the newer user: that’s how the next profile to show is determined.

If one user swipes left (rejecting) on a profile because they do not like a profile, the algorithmic model assumes the next user will also swipe left or reject that profile and therefore refrains from presenting that profile to the next user. As a result some profiles will be shown less than others based on people’s own personal biases or preferences. This could be anything from the type of photo taken to more specific characteristics such as race, religion or gender.

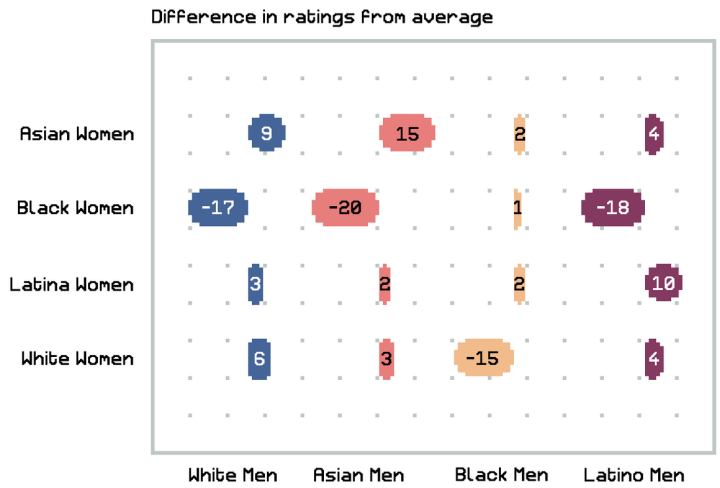


Figure 5: MonsterMatch, *Differences in ratings*, 2019. Digital illustration. (MonsterMatch 2019).

According to the dating app OkCupid (Medium 2019), black female profiles are consistently rated lower by all male profiles when looking for a suitable match on dating apps as illustrated in Figure 5. The collaborative filtering algorithmic model of OkCupid then takes these decisions (past data) as the training data to learn which profiles to recommend going forward, in comparison to others. The more liked a profile, the more likely it is to be shown again. A training data bias becomes evident as certain profiles are continuously less represented in the input data from which the algorithmic model learns from. Moreover the training data bias of unrepresentative data of black female profiles can perpetuate a racial bias because these women will be excluded from making matches with other dating profiles. Their profiles will not be shown as a result of an alleged identified dislike pattern for their race.

Figure 6: 'Create profile'. Screenshot from the game Monster Match (2019).

Figure 7: 'Swipe'. Screenshot from the game Monster Match (2019).

Monster Match uses a similar scenario to educate the player about algorithmic bias. The game starts by prompting the user to create a dating profile as can be seen in Figure 6. The user can then start to match with creature or humanoid monster profiles by either swiping left to reject a profile or right to match with a profile as can be seen in Figure 7. Figure 8 below, Collaborative filtering result (2019), conveys how the players previous swipes will further be used by the collaborative filtering algorithmic model to recommend either creature or humanoid monster profiles going forward.

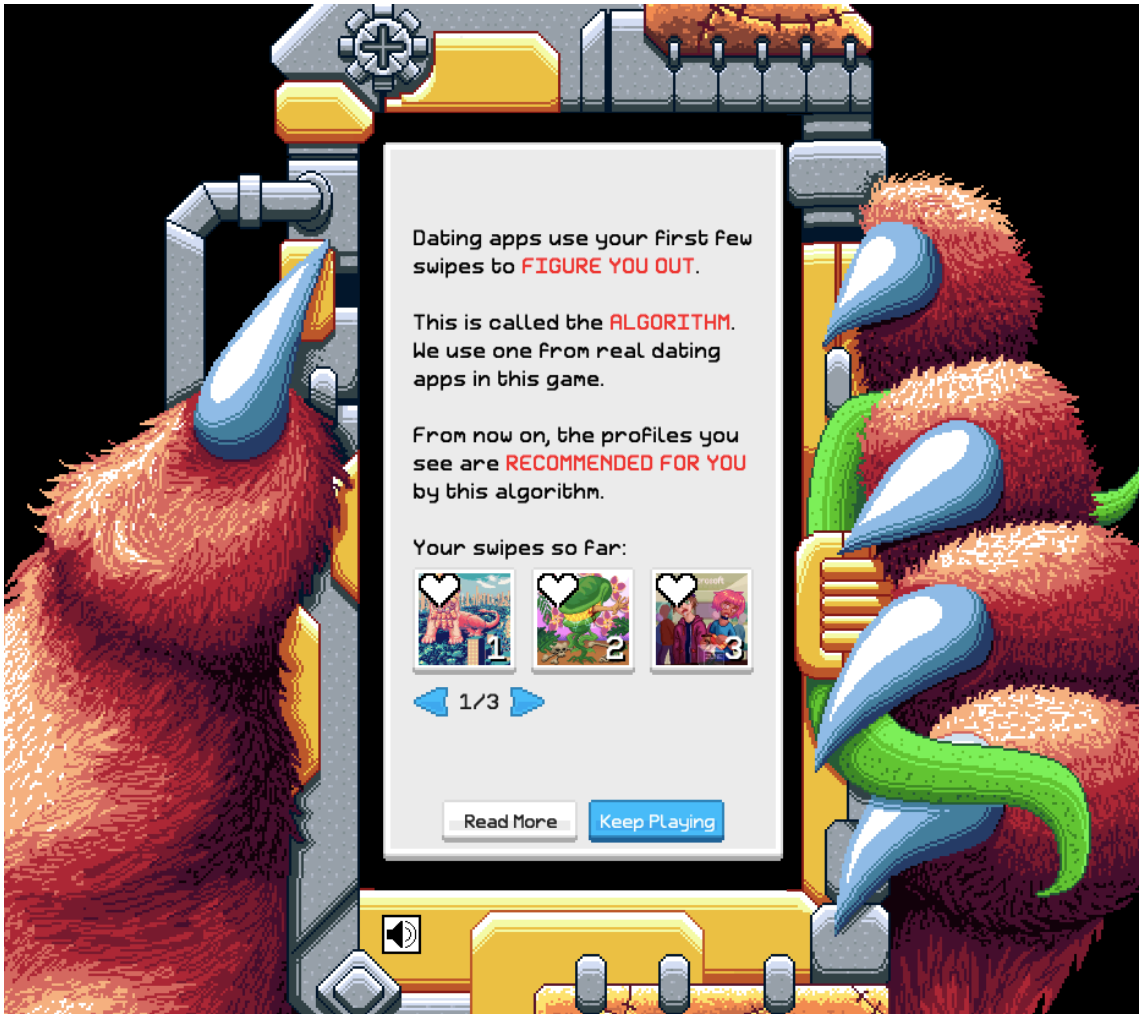


Figure 8: 'Collaborative filtering result'. Screenshot from the game Monster Match (2019).

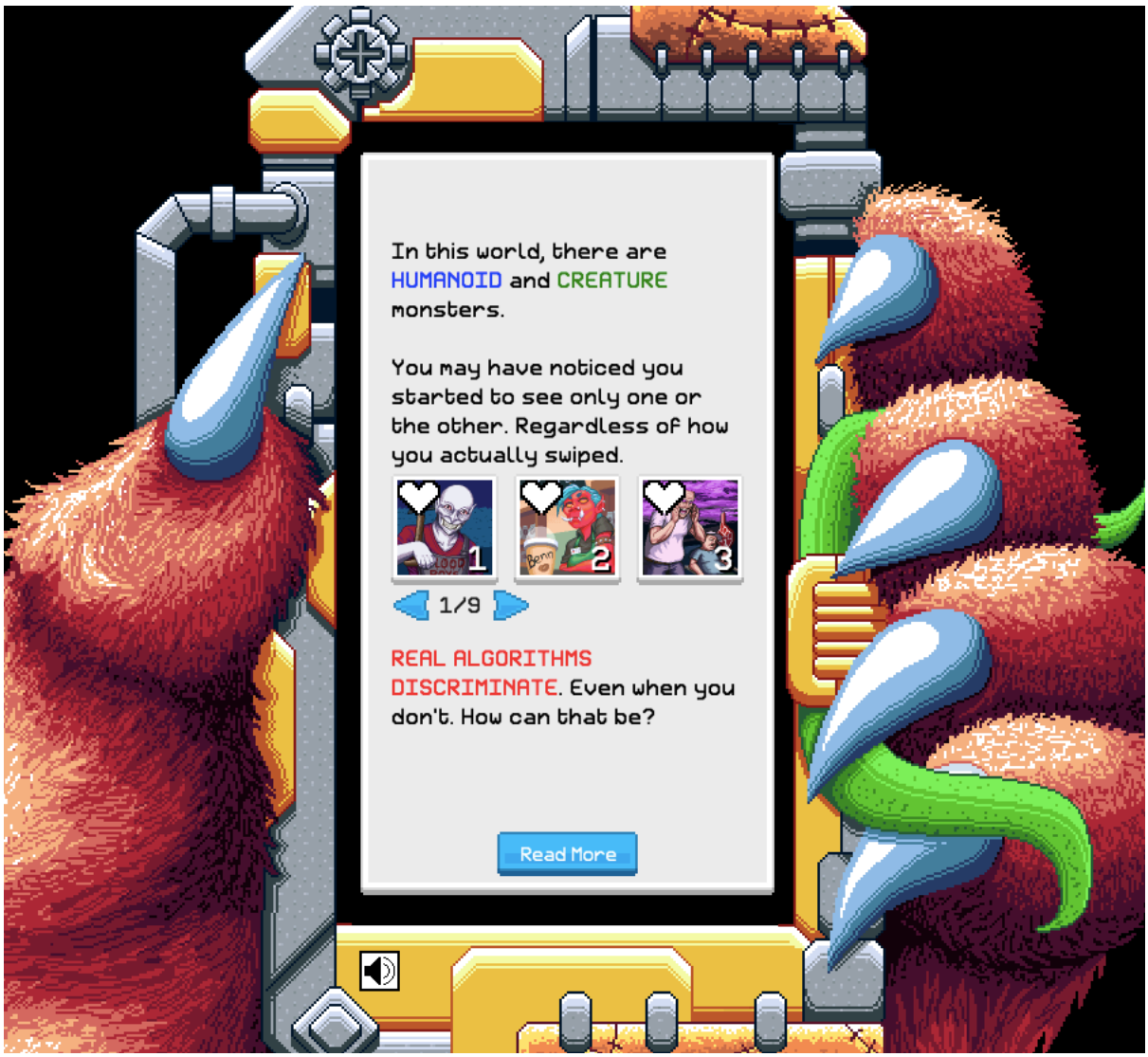


Figure 9: 'Algorithm discriminates'. Screenshot from the game Monster Match (2019).

An additional factor that influences and informs the algorithmic model in the game is whether the player receives any phone numbers. If they have, then the dating app assumes that the player's profile is liked and starts to further recommend the profile to other users more. It is important to note at this stage that this factor further elevates already likes profiles because a user has to match before starting a conversation and consequently receive any phone numbers. As a result the aforementioned factor ends up further perpetuating the gap between generally liked profiles and disliked profiles. Finally the game reveals as shown in Figure 9 that the dating app simulation ends up discriminating either against humanoid or creature monsters.

3.2.2 Applying transparency to the case study

The game *Monster Match*, similarly to the *SOTBF* game illustrates transparency by providing users with resources and documentation to learn more about algorithmic bias. The game also allows users to engage with algorithmic bias through a game scenario which improves understandability regarding how AI can present a racial bias outcome as the player can actually experience the consequences of their choices . A real collaborative filtering algorithmic model is used to present the bias to the user. The project *Monster Match* is open source and can therefore be reviewed by interested users who would like to view the code and algorithmic model in question. This transparency means that users can completely review how the dating app *Monster Match* uses a collaborative filtering algorithmic model to present a training data bias and consequently a racial bias. Lastly users can use the code to create their own bias game or application, however the creative assets are licenced and may not be used for any commercial purposes. As stated in Chapter Two by Ananny and Crawford (2016:9) “making sure that a system is not only visible but also debated and changeable by observers who are able to consider how they know what they know about is” increases the value of the mitigation method transparency.

3.2.3 Applying algorithmic affirmative action to the case study

Applying algorithmic affirmative action to the dating app simulation game *Monster Match* could mean creating different dating apps for different users. Identifying that one group has in the past been discriminated against on dating apps and prevent the continuation of this discrimination by creating a separate dating app to cater for the disadvantageous group. In real life there exist such niche apps for instance JDate for Jewish users and Grinder for LGBT users (MonsterMatch 2019). The downside of separating people according to religion, sextual orientation and race on dating apps is that users will have limited choices when it comes to diversity. Algorithmic affirmative action then might prove to eliminate racial bias results in dating apps, but as a result will also limit available options to users who are looking for love. Finally applying algorithmic affirmative action to dating apps might also encourage racial segregation which as a result reinforces racial prejudice. An alternative method of applying algorithmic affirmative action to evoke a more positive result might be too

hard-code the collaborative filtering algorithmic model to present a diverse set of profiles to users, such as 50% creature monsters and 50% humanoid monsters. Unfortunately this might mean that users are less likely to be presented with profiles that match their desired preferences.

3.3 Driver drowsiness detection algorithmic model in SA

3.3.1 Description

The article *Detecting inter-sectional accuracy differences in driver drowsiness detection algorithms* written by Mkhuseli Ngxande, Jules-Raymond Tapamo and Michael Burke (2019) describes how a driver drowsiness detection algorithmic model can present a racial bias in a South African context when publicly available datasets that do not represent people of all races and ethnicities are used. The three publicly available datasets that were used were; ULg Multimodality Drowsiness Dataset (DROZY), the National Tsinghua University Drowsy Driver Detection database (NTHU-DDD), and the Closed Eyes in the Wild dataset (CEW) (Ngxanda, Tapamo & Burke 2019:2). The authors used these datasets to highlight the consequences of using unrepresentative training data to train vision-based driver drowsiness detection algorithmic models. They then compare these results with a more diverse dataset representative of South Africa's population in order to illustrate how publicly available datasets can exhibit a racial bias in a South African context (Ngxanda et al. 2019:1). As a solution to the aforementioned problem Ngxande et al. (2019:1) state:

> We propose a novel visualisation technique that can assist in identifying groups of people where there might be the potential of discrimination, using Principal Component analysis (PCA) to produce a grid of faces sorted by similarity, and combining these with a model accuracy overlay.

Figure 10: Ngxande et al. *Visualisation technique*. 2019. Digital illustration. (Detecting inter-sectional accuracy differences in driver drowsiness detection algorithms 2019).

Figure 10 illustrates the novel visualisation technique. Faces of a similar skin tone are grouped together in order to establish if there is a fair representation of different skin types. Dark-skinned individuals are located at the top whilst lighter-skinned individuals are located at the bottom of the image. Each image is evaluated with the algorithmic model in order to determine the overall accuracy of the dataset and if it can detect each face and subsequently differentiate between each skin type and drowsiness probability.

According to Ngxande et al. (2019:1):

> Statistics South Africa report that about 76% of citizens in the country use public transportation to get to their destinations, with private minibus taxis a primary mode of transport (51.0%), followed by busses at 18.1% and trains at 7.6%

Systems using driver drowsiness detection algorithmic models are being reviewed with the hope that implementing these systems could reduce the number of road accidents in South Africa. Ngxande et al. (2019:1) further states that the majority of South African citizens identify as black. Therefore with regards to the visualisation

technique discussed above a fair representation of dark-skin individuals vs light-skin individuals have to be represented in the dataset otherwise the majority of South African citizens will not benefit from driver drowsiness detection systems.

The work of to Ngxande et al. therefore contributes immensely to prevent systems that use driver drowsiness detection algorithmic models from exhibiting discriminatory results in South Africa.

3.3.2 Applying transparency to case study

The work of Ngxande et al. (2019: 4) illustrates transparency by training their classification models on publicly available drowsiness detection datasets. As a result of using publicly available datasets their project can be re-produced and tested by any individual or organisation that would like to compare their own findings.

3.3.3 Applying algorithmic affirmative action to case study

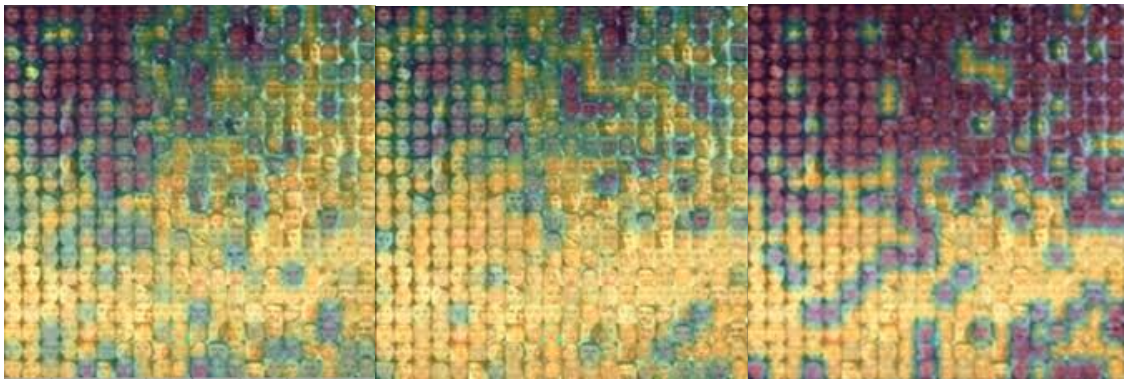


Figure 11: Ngxande et al. *Visualisation technique on CEW model.* 2019. Digital illustration. (Detecting inter-sectional accuracy differences in driver drowsiness detection algorithms 2019).

Ngxande et al. (2019:3) state that some believe information on the demographics of participants should be included in datasets to be used to train driver drowsiness detection algorithmic models. This however can be prove to be problematic because it enforces the classification of people. As a solution Ngxande et al. (2019:3) propose a visualisation technique as previously discussed that can identify groups on who the algorithmic model fails, but will not pre-classify them beforehand into groups. An example of using the visualisation technique to test the accuracy of a model is

illustrated in Figure 11. According to Ngxande et al. (2019:7) the visualisation technique shows how the CEW model struggles to predict drowsiness for dark skin individuals through the use of highlighting different section with different shades of colour. The yellow shaded parts indicate where the model performs well, the purple shaded parts indicate where the model fails and the green shaded parts indicate where the model preforms well, but with lower probability (Ngxande et al 2019:8). By analysing Figure 11 it is clear that the purple shaded parts are mainly located at the top right of the image where darker-skinned individuals are located. This indicates that additional training data for dark-skinned individuals are necessary to ensure that the model does not return such a high failure rate regarding darker-skinned individuals in the future. The visualisation technique can help software engineers easily identify where a group is being discriminated against because of unrepresentative training data. Software engineers can then proceed to add more training data for any unrepresented group to prevent a discriminatory outcome. As a result the visualisation technique could be a more successful method to apply algorithmic affirmative action without directly classifying individuals according to their race. The technique enables developers to see if there is an equal representation and diversity in the training data during the design phase of the algorithmic model.

3.4 Chapter conclusion

In conclusion the case studies in combination have proven that the mitigation methods transparency and algorithmic affirmative action have considered value to prevent algorithmic bias, specifically a training data bias, presenting a racial bias. However the value of each mitigation method is dependent on the context of its use. These mitigation methods are now to be evaluated in relation to a practical project called *Tobias Park* a mobile educational game about algorithmic bias, specifically a training data bias, which presents a racial bias.

CHAPTER FOUR: TOBIAS PARK AN ALGORITHMIC BIAS GAME

This Chapter will discuss the proposed practical project in relation to the research conducted in the preceding Chapters. The aim of the practical project is to explore how a training data bias that presents a racial bias, as outlined in Chapter One, can use the mitigation methods of transparency and algorithmic affirmative action, as discussed in Chapter Two to resolve the problem of the aforementioned algorithmic bias. The findings from analysing the case studies presented in Chapter Three will be used to inform decisions made in the practical project.

1.1 Project description

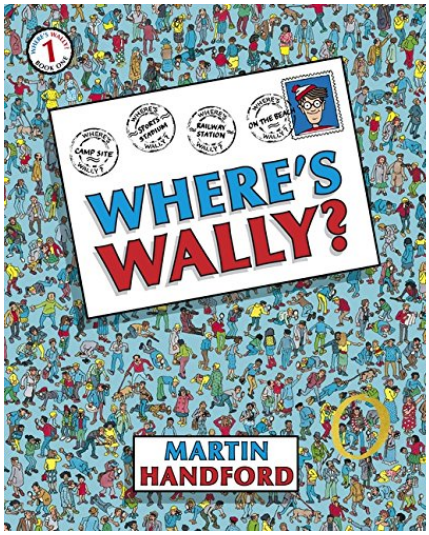


Figure 12: Martin Handford, *Where's Wally?*, 2007, Digital illustration. (Amazon.co.uk).

The proposed practical project is an educational mobile game where users can learn about algorithmic bias through a story-based metaphoric context. The proposed mobile game is inspired by the puzzle game *Where's Wally?* as illustrated in Figure 12. The goal of the *Where's Wally?* puzzle game is for the player to analyse a crowded scene where Wally is hidden and find him. This puzzle game is fun yet can be time consuming. It inspired me to create an AI application that can help to detect a specific object, figure or character using machine learning. I have created my own educational version of the *Where's Wally?* game called *Tobias Park.*

My game, which is an Android application, will ask the player to help a creature in Tobias Park called Penko find his friends Mash and Beau by using the AI technology object detection. The player is prompted to scan various images with Mash and Beau hidden in the different scenes. The game presents an algorithmic bias result when only one type of creature is found by the AI and another is not, due to a training data bias. The player is encouraged to correct the bias by choosing and accumulating the necessary training data of the 'left out' creature, for the algorithmic model. With help from Penko, this additional missing data will ensure that the AI can then detect both of Penko's friends and that no bias is presented. The game also acts as an educational context as the user is gradually informed of the bias and facilitated to correct it.

1.1.1 Technical description

The library that is used to create the machine learning android application is TensorFlow. TensorFlow is an easy to use platform for beginners who want to apply machine learning to a project. The proposed practical project will use TensorFlow's Object Detection API to train a neural network capable of recognizing objects in a frame. This project will specifically be trained to detect the creatures Mash and Beau in an image.

1.1.2 How does the game work?

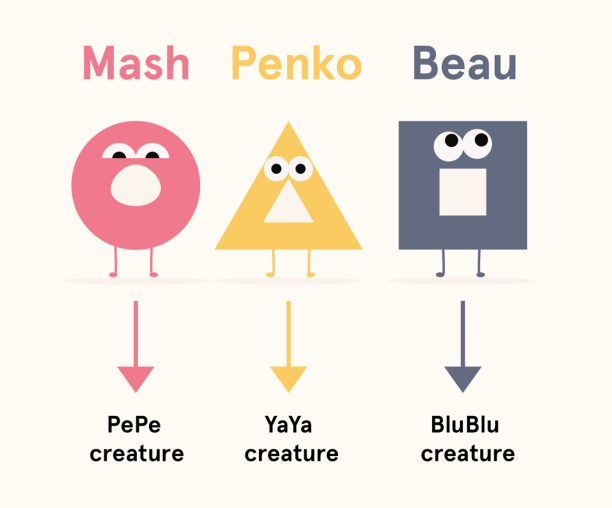


Figure 13: Alecia Vermeulen, *Tobias creatures*, 2019. Artwork in possession of author.

There are three creatures in Tobias Park, each representing a category of creature in Tobias World; Blublu creatures (blue creatures), PePe creatures (pink creatures), and YaYa creatures (yellow creatures). The three main creature characters in the game are Penko, Mash and Beau as illustrated in Figure 13. Penko is a YaYa creature characterised by a yellow glow. Mash is a PePe creature characterised by a pink blush. Beau is a BluBlu creature characterised by a blue hue.

Figure 14: Alecia Vermeulen, *Detect creatures*, 2019. Artwork in possession of author.

The start of the game prompts the player to help Penko find his friends Mash and Beau in Tobias Park. The player uses their phone's camera to access the AI technology object detection and then proceeds to point it to a picture of Tobias Park where Penko's friends are hidden in Tobias park as illustrated in Figure 14. The AI finds Penko's friend Mash, but fails to find Beau because not enough training data

representing BluBlu creatures were selected by the player to access the desired pre-trained model where both creatures are fairly represented. The game then prompts the player to ask Penko for help regarding selecting images of BluBlu creatures in order to be able to then find Beau in Tobias Park. If the player does not select enough images Beau will not be found, but if the player selects enough image resources both friends Beau and Mash will be found. It is important to note however that if the player selects more than 50% images of BluBlu creatures, the image resources for PePe creatures will decrease, resulting in a reverse bias result where PePe creatures like Mash won't be found anymore. After the player has succeeded in helping Penko find both of his friends in Tobias Park the game concludes with a summary of how algorithmic bias was presented throughout the game.

1.2 How is it a training data bias presenting a racial bias?

The goal of the game *Tobias Park* is to explain how a training data bias can present a racial bias by exploring a metaphoric scenario of where one creature is found using the AI technology object detection and another is not. A key differentiator between these creatures is their colour. The case studies discussed in Chapter Three, namely *Survival of the Best Fit* and *Monster Match*, informed this decision as both created a metaphor for race to communicate a real-world context concern. *Survival of the Best Fit* used blue and yellow people to communicate racial discrimination in hiring AI. The game *Monster Match* used humanoid and creature monsters to illustrate racial discrimination in dating apps. As discussed in Chapter Three, both games had a racial discriminatory outcome due to a training data bias. In the game *Tobias Park* a metaphor for race is presented through three different categories of creatures; PePe creatures that are pink, BluBlu creatures that are blue and YaYa creatures that are yellow. A training data bias is presented in the game when Penko a YaYa creature can't find his friend Beau who is a BluBlu creature because of incomplete and unrepresentative training data regarding the BluBlu creature category. By not being able to detect this category of creature a racial bias is evident.

1.3 How is transparency applied?

The game *Tobias Park* will be open source and accessible on github. Any interested user can download the project to review the code and algorithmic model used in the project to learn more about the technology in question. Additionally the training data used to train all three of the custom object detection algorithmic models (biased against BluBlu creatures, biased against PePe creatures and not biased) will be available on github. Users can then clearly see why one model presented a bias and the other did not. As stated in Chapter Two by Ananny and Crawford (2016:9), in order to learn about complex systems society should look beyond just having access to see the information, but interact in such a way with the system that an understanding of it in relation to its environment can be achieved. The game *Tobias Park* presents the complex problem of algorithmic bias in such a way that it can be understood through the means of a metaphor in a gaming context. The player not only experiences a biased result but is then required to fix the bias, encouraging the learning process of the user regarding algorithmic bias, specifically a training data bias, presenting a racial bias.

After analysing the case studies *Survival of the Best Fit* and *Monster Match* in Chapter Three, a key finding/observation was that both games presented the problem of algorithmic bias, but did not present a way for the player to correct the bias. As a result, I decided to present a way for the player to do so. I believe it will add value and highlight not only the problem of algorithmic bias, but a potential solution too, through the means of human intervention.

1.4 How is algorithmic affirmative action applied?

The visualisation technique described in the previous Chapter was used as inspiration to allow the player in the game *Tobias Park* to fix the bias against BluBlu creatures. A slider representing the amount of data of BluBlu creatures in comparison to PePe creatures is shown to the player. This reveals to the player the current state of their algorithmic model that is to be used to help Penko find his friends. The player receives help from Penko who knows what his friends Mash and Beau and the corresponding different categories of creatures look like and proceeds

to fill the data that is incomplete and unrepresentative by selecting and/or removing image resources for the PePe and BluBlu creatures in the game. This visualisation technique allows the player to clearly see if one category of creature is discriminated against and proceed to add images for the unrepresented creature. It is important to note however that the method to fix the bias only simulates the real solution, which is to extend the training data that the object detection algorithmic model is trained on. The player merely unlocks a pre-trained unbiased custom object detection algorithmic model by achieving to represent both categories of creatures, namely PePe and BlueBlue, equally in the game. If the player fails to select enough image resources to represent both PePe and BluBlu creatures, a pre-trained biased custom object detection algorithmic model proceeds to be the one in use when playing the game.

1.5 Chapter conclusion

In conclusion the creation of the game *Tobias Park* was inspired by the findings in the preceding Chapters. The goal of the game is to explain how a training data bias can present a racial bias as outlined in Chapter One. The game explored how the mitigation methods of transparency and algorithmic affirmative action as discussed in Chapter Two can be applied as a means to highlight and create awareness of the proposed research topic, namely algorithmic bias. Furthermore, the case studies discussed in Chapter Three influenced both the conceptualisation and the final outcome of the game *Tobias Park*. I utilised a metaphor to represent race in the game and presented a means for the player to fix the bias in the game by referring to a method of applying algorithmic affirmative action, namely the visualisation technique that was discussed in Chapter Three. Lastly, the practical project serves to contribute to the research paper as it practically explored both the creation and prevention of algorithmic bias, specifically a training data bias, which presents a racial bias.

CHAPTER FIVE: RESEARCH CONCLUSION

There is no cure or solution to humankinds inherent bias. We form assumptions as a way to survive and evolve as a species. Our human biases have caused harm and often severe damage in the past as they have led to discriminatory practices such as Apartheid in South Africa, where one group of people oppressed another as a result of a racial bias ideology. Concern arises as South Africa enters a future where AI systems which learns from past human decisions and data are becoming a part of our everyday lives. Understanding how algorithmic models can contain biases inherent in biased data, specifically racial biases, is necessary so that we can identify when an AI system perpetuates racial discrimination in South Africa. This paper sought to investigate how algorithmic models presenting a racial bias as a result of incomplete and unrepresentative data, due to a history of Apartheid, could be circumvented to prevent systemic racial discrimination in South Africa. The following is a summary of the findings in the research.

Chapter One introduced the different types of algorithmic bias before detailing how a training data bias can present a racial bias, specifically in South Africa. To restate the findings of the main example discussed in Chapter One our limited resources regarding African descendent languages that forms part of South Africa's eleven official languages create concern that the introduction of AI technology such as voice recognition will mean the exclusion of the majority of South Africa's population and consequently reinforce systemic racial discrimination.

In Chapter Two transparency and algorithmic affirmative action were analysed as a means to uncover whether the aforementioned concern and main topic of the proposed research question which is algorithmic bias, specifically a training data bias, presenting a racial bias can be prevented in a South African context.

The advantages of the mitigation method transparency, include the reviewing of algorithmic models by the public, so that interested individuals can have access to learn more about how a specific algorithmic models functions, increasing awareness regarding algorithmic bias and public involvement in emerging AI technologies in the future is encouraged. Disadvantages of the mitigation method transparency, include

algorithmic models can be too complex to understand as a result of machine learning, so public access does not equal comprehension regarding the available material, time to review algorithmic models are limited, and regulatory guidelines regarding sharing datasets and code are both unclear and in some situations unpractical.

Advantages of the mitigation method, algorithmic affirmative action include prioritising previously disadvantaged groups as a means to rectify past injustices, past data that reflects bias can be used in combination with hard-coding variables in the structure of the algorithmic model, to ensure that the bias data does not perpetuate discrimination when used to train an algorithmic model for decision making procedures. Disadvantages of the mitigation method, algorithmic affirmative action include complexity regarding what constitutes as fair arise, some believe it is a hard-coded form of discrimination, lastly in South Africa the value of affirmative action debatable indicating that algorithmic affirmative action might have the same result.

Chapter Three further evaluates the aforementioned mitigation methods by applying the findings from Chapter Two to three case studies where a training data bias, presenting a racial bias is evident. A key takeaway from the case study *SOTBF* is that transparency is limited if there is no underlying algorithmic model that players and users can access and review for further comprehension. With regards to algorithmic affirmative action the case study proved that racial discrimination can be prevented in the game, but whether this method is profitable for a company remains questionable. The second case study *Monster Match* went beyond giving users and players just access to review a bias algorithmic model, but presented the concern of algorithmic bias in a gaming context for the material to be comprehensible. This case study proved that limitations to the mitigation method transparency such as public access does not equal comprehension can be overcome. With regards to algorithmic affirmative action, this case study proved that the value of this mitigation method is dependent on the context of its use. In this particular instance algorithmic affirmative action did not prove to alleviate racial discrimination, but in fact would only either perpetuate it or provide users on the dating app with profiles that might not reflect their personal preference. The last case study *Detecting inter-sectional accuracy*

differences in driver drowsiness detection algorithms proved that transparency has considered value to allow the public to reproduce findings as outlined in the case study's findings. With regards to algorithmic affirmative action the case study serves to prove that regardless of the clear disadvantages of the mitigation method algorithmic affirmative action in a South African context as discussed in Chapter two, the value of this mitigation method is dependent on the context of its use and application. The case study which presented a training data bias, presenting a racial bias in South Africa proved that the mitigation methods transparency and affirmative action can indeed eliminate driver drowsiness detection AI systems from perpetuating systemic racial discrimination in South Africa.

It can be concluded that a training data bias, presenting a racial bias in South Africa can be prevented by applying the mitigation methods, transparency and algorithmic affirmative action, if the context of use allows for it. The research does however acknowledge that algorithmic model are as complex as human biases and can perpetuate them therefore a single solution is not often the single solution and that a context or case specific solution would often be required, a blended approach.

To further investigate this, and support the proposed answer to the research the practical component was made as a way to investigate if South African citizens can eliminate a training data bias, presenting a racial bias in a game context if the mitigation methods transparency and algorithmic affirmative action is presented in the game as a way for the user to fix the aforementioned bias. User feedback on the final presented practical project could help to uncover further findings regarding the proposed research question.

To conclude algorithmic models have become a part of our everyday reality and will continue to be integrated into our lives in the future. While writing this research paper I constantly thought about the novel Frankenstein by Mary Shelley that conveys that science should be cautious. A key takeaway from the novel is that the creation of Frankenstein's monster was not the problem, but in fact the problem was the way that he reacted to it. We, humans are Frankenstein and the algorithmic model is our monster which can harm and destroy us, its creator, if we do not take responsibility for our creation and the consequences of its choices which we programme. We have

to nurture and work together to ensure that these algorithmic models imbedded in AI systems do not perpetuate racial discrimination.

To quote the concluding words of Chander's (2017:1045) paper *The Racist Algorithm?* "[O]nly humans can perform the critical function of making sure that, as our social relations become ever more automated, domination and discrimination aren't built invisibly into their code". Human involvement is therefore key to ensure that the mistakes of the past will not be automated in our future. Although we cannot escape our past or our future human biases, we can learn how to identify them so that we can ensure that we do not transfer these biases through technology into a future that we all in vison.

LIST OF SOURCES

Amazon.co.uk. 2007. *Where's Wally?.* https://www.waterstones.com/book/wheres-wally/martin-handford/9781406305890 (Accessed 12 October 2019).

Ananny, M & Crawford, K. 2016. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 0:1-17.

Anne Heffernan. 2016.*Strategic lessons South Africa's students can learn from the leaders of 1976.* https://theconversation.com/strategic-lessons-south-africas-students-can-learn-from-the-leaders-of-1976-60976 (Accessed 13 October 2019).

Banfi, V. 2019.*What is a Turing test and how to run one using Slack and Zapier*. https://botsociety.io/blog/2018/03/the-turing-test/ (Accessed 24 July 2019).

Begbie, N. 2019. *Problematic algorithmic bias; on manifestation in a South African context and methods for mitigation*. https://medium.com/mobileforgood/problematic-algorithmic-bias-on-manifestation-in-a-south-african-context-and-methods-for-e9c9b7b19586 (Accessed 26 September 2019).

Bent, J.R. 2019. Is Algorithmic Affirmative Action Legal? *Georgetown Law Journal 108:1-59*

Burger, R & Jafta, R. 2010. *Affirmative action in South Africa: an empirical assessment of the impact on labour market outcomes. 1-26.*

Chander, A. 2017. The Racist Algorithm. *Michigan Law Review* 115(6):1023-1045.

Cherry, K. 2019. *How Cognitive Biases Influence How You Think and Act).* https://www.verywellmind.com/what-is-a-cognitive-bias-2794963 (Accessed 22 October 2019).

Christopher, A.J. 2011. The Union of South Africa censuses 1911-1960: an incomplete record. *Historia.* 56(6). 01-18.

Danks, D & London, A,J. 2017. *Algorithmic Bias in Autonomous Systems.* Charles Sierra(Ed). Melbourne: AAAI press.

Eisenstat, Y. 2019. *The Real Reason Tech Struggles With Algorithmic Bias. https://www.wired.com/story/the-real-reason-tech-struggles-with-algorithmic-bias/* (Accessed 22 September 2019)

Freepik.com. 2019. *Detect creatures.* http://www.freepik.com (Accessed 02 November 2019).

Friedman, B & Nissembaum, H. 1996. Bias in computer systems*. ACM Transactions on Information Systems (TOIS)* 14(3):330-347.

Hom, E. 2019. *Alan Turing Biography: Computer Pioneer, Gay Icon*. https://www.livescience.com/29483-alan-turing.html (Accessed 24 July 2019).

Kemper, J & Kolkman, D. 2018. Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society.*1-16.

Kenney, M. 2018. *Algorithms, Platforms, and Ethnic Bias: An Integrative Essay Forthcoming in Phylon: The Clark Atlanta University Review of Race and Culture.* 2-41.

Marwala, T. 2018. Opinion: Tackling bias in technology requires a new form of activism. https://www.uj.ac.za/newandevents/Pages/Opinion-Tackling-bias-in-technology-requires-a-new-form-of-activism.aspx (Accessed 24 July 2019).

Medium. 2019. *Race and Attraction, 2009-2014.* https://theblog.okcupid.com/race-and-attraction-2009-2014-107dcbb4f060 (Accessed 7 October 2019).

MonsterMatch. 2019. *MonsterMatch.* https://monstermatch.hiddenswitch.com

(Accessed 7 October 2019).

Mybroadband.co.za. 2019. *New court case can end affirmative action based on race in South Africa.* https://mybroadband.co.za/news/business/320085-new-court-case-can-end-affirmative-action-based-on-race-in-south-africa.html (Accessed 14 October 2019).

Ngxande, M & Tapamo, J,R & Burke, M. 2019. Detecting inter-sectional accuracy differences in driver drowsiness detection algorithms.

Parry, K & Eeden, A. 2015 Measuring racial residential segregation at different geographic scales in Cape Town and Johannesburg. *South African Geographical Journal* 97(1):31-49.

Shadowen, A.N. 2017. Ethics and Bias in Machine learning: A Technical Study of What Makes Us "Good". MS thesis. New York: University of New York.

Silberg, J & Manyika, J. 2019. *Notes from the AI frontier: Tackling bias in AI (and humans)*. https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans (Accessed 25 September 2019).

Smith, A. 2019. *Apartheid: The Beginning, What Happened, And When It Ended.* https://buzzsouthafrica.com/apartheid/ (Accessed 24 July 2019).

South African Market Insights. 2019. *Languages spoken in South Africa per race group according to the latest General Household Survey (GHS).* https://www.southafricanmi.com/sa-language-ghs-28may2019.html (Accessed 13 October 2019).

Survivalofthebestfit.com. 2019. *Survival of the Best Fit*. https://www.survivalofthebestfit.com (Accessed 7 October 2019).

van der Laken, P. 2019. *Survival of the Best Fit: A webgame on AI in recruitment.*

https://paulvanderlaken.com/2019/06/30/survival-of-the-best-fit-a-webgame-on-ai-in-recruitment/ (Accessed 7 October 2019).