Research Master Thesis

Mitigating Gender Bias with Deep Reinforcement Learning

Mojca Kloos

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Linguistics and Text Mining Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Prof. dr. Antske Fokkens, dr. Pia Sommerauer
2^{nd} reader: Dr. Lucia Donatelli

Submitted: June 27, 2023

Abstract

In reinforcement learning (RL), the developer of the RL algorithm can reward certain behaviour, which encourages it to exhibit this behaviour in order to maximize the reward at the end of a task. This thesis presents a study on the use of reinforcement learning as a debiasing method for reducing gender bias in occupation prediction. In occupation prediction, it has been found that classifiers often compound gender distributions of occupations that are traditionally done by men or women. This behaviour may be harmful when these systems are applied in e.g. automated hiring practices. Thus, it is key to find methods that debias these types of occupation prediction systems. This thesis compares two different classifiers trained with reinforcement learning, each with a different reward system, to a non-RL classifier. The results indicate that RL can effectively reduce gender bias in occupation prediction, without harming the performance of the classifier too much. Comparative analysis with other debiasing approaches reveals that the RL-based method achieves comparable or superior results in bias reduction and accuracy. Moreover, the thesis demonstrates that combining RL with input-based debiasing techniques enhances bias reduction. The findings highlight the potential of RL as a suitable approach for addressing gender bias in occupation prediction tasks, but further research is necessary to optimize the RL system and explore its applicability in different contexts.

Declaration of Authorship

I, Mojca Catharina Kloos, declare that this thesis, titled *Mitigating Gender Bias with Deep Reinforcement Learning* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: June 27, 2023

Signed: 

Acknowledgments

I would first like to thank my supervisors Antske Fokkens and Pia Sommerauer: You most often had more trust in me than I did in myself. I would not have been able to tackle a topic that was so out of my comfort zone if it wasn't for your enthusiasm. Your guidance and feedback kept me motivated to execute this project, and I truly feel like I have learnt a lot from you, not only during the thesis but also during the courses you taught.

Secondly, I want to thank Shihan Wang for her guidance on reinforcement learning, a field that I was completely unfamiliar with at the start of this thesis. She took the time to meet with me, and made me feel more comfortable with the topic. I would also like to thank Pantea Haghighatkhah for her help with the data set.

Of course, I cannot forget my classmates, who have made this master's program, and this thesis, a lot more enjoyable. Aga, I want to thank you for listening to me every week and making our joined thesis meetings more fun. And to my family and friends, thank you for listening to my attempts to explain my thesis, and being truly interested in it, or at least pretending to be; for being a distraction when I needed to clear my mind from writing this thesis.

And lastly: Marek, thank you for your endless support and patience, and your Seaborn skills - You have helped me finish this project in many ways.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Machine learning has been used more and more for a variety of tasks. With this development come several pitfalls: Machine learning systems "learn" to tackle tasks by training on great amounts of data, and viewing the statistical patterns that allow them to make accurate predictions. Since the training data is made by people, it will inevitably contain human flaws. One such flaw is bias, where in the real world, there is almost always some prejudice towards one group at the cost of another (Lloyd, 2018). Examples of biases that can be picked up by automatic systems are age bias, gender bias and racial bias (Rupp et al., 2006; Isaac et al., 2009; Nelson, 2002). Since we encounter machine learning systems in many aspects of daily life, the biases that they contain could have serious implications for those whom these systems are prejudiced towards.

Mitigating bias in machine learning is a task that often requires a great deal of manual labor. Reinforcement learning (RL) is a machine learning paradigm that may lend itself well to the task of bias mitigation, since RL algorithms are able to "teach" themselves desired behaviour. In the making of a reinforcement learning algorithm, the developer can specify what behaviour is rewarded, and the algorithm will strive to obtain the highest reward and thus, ideally, exhibit the desired behaviour (Sutton and Barto, 2018). Although it is expected that RL can be a suitable method for developing a classifier with decreased bias, employing reinforcement learning for the mitigation of bias in NLP tasks is understudied: At the time of writing this thesis, only one study was found that applied reinforcement learning to an NLP debiasing task, which was in the case of hate speech detection (Cheng et al., 2021). In this study, not one specific type of bias was investigated, but rather the fact that social media sessions were more likely to be classified as hate speech if they contained demographic identity terms such as *gay*, *fat* or *black*. This study demonstrated promising results, showing that employing RL was successful in debiasing hate speech detection, without being detrimental to the performance of the classifier. It also leaves room for the question as to whether reinforcement learning is a suitable approach for mitigating specific types of bias, such as gender or racial bias, and whether this approach works in a different task.

It is this question that this thesis will address: Whether leveraging reinforcement learning is a successful strategy in the mitigation of a specific type of bias, namely gender bias, in a task that reinforcement learning as a debiasing strategy has not been applied to before, namely occupation prediction. To investigate gender bias in occupation prediction, the BiasBios data set is used, as developed by De-Arteaga et al. (2019). This data set consists of short biographies of people, with a corresponding

occupation label, such as *teacher* or *rapper*, and a gender label (male/female). It is often used to study gender bias, as the task of occupation prediction is known to be biased due to stereotypical gender distributions being present in the data set. For example, in the occupation *nurse*, only 9.4% of the individuals who have this occupation in the data set is male. Therefore, classifiers are at risk of overgeneralizing and thus being biased towards females when predicting this occupation. When these types of systems are used in a real-world application, they may be harmful, as the systems may compound existing gender imbalances. Thus, research needs to be conducted into how the gender bias can be mitigated, which is the target of this thesis. In order to investigate this, the main research question of this thesis is the following:

Can reinforcement learning be employed for the mitigation of gender bias in the task of occupation prediction?

In order to answer this main question, a second research question is posed:

What is the influence of the reward function on the use of reinforcement learning for the mitigation of gender bias in occupation prediction?

The focus of the thesis will be on formalizing the task of occupation prediction in such a manner that reinforcement learning can be applied when training the classifier. I hypothesize that reinforcement learning can lead to a classifier that contains less bias than its non-RL counterpart. The reward function will be most important in developing an RL algorithm for occupation prediction, as this determines the behaviour of the algorithm. I expect that bias will not be eliminated completely, but rather that RL provides a method to develop a machine learning system that is less biased, compared to a vanilla classifier.

The experiments are set up as follows: The impact of reinforcement learning as a debiasing strategy is investigated by implementing a vanilla neural classification system, which is compared to two RL systems. One of those systems includes a simple reward function, that only rewards the RL algorithm when it correctly classifies an instance. The second RL system implements a more elaborate reward function, which pays attention to the minority gender when classifying a biography.

It is also investigated how to evaluate the impact of the RL algorithm on the task of occupation prediction, and see whether RL is a suitable approach for mitigating gender bias in the task of occupation prediction. The only study that investigated using reinforcement learning for an NLP debiasing task (Cheng et al., 2021) did not regard one specific type of bias, and it only focused on the task of hate speech detection. Moreover, Cheng et al. (2021) did not compare the results of RL as a debiasing approach to other debiasing methods, in contrast to the results found in this thesis.

1.1 Main Contributions of the Thesis

Previous work has investigated the influence of gender bias in various NLP tasks thoroughly, with a variety of methods and tasks. Moreover, different approaches for mitigating bias in machine learning have been developed. In the task of occupation prediction, and the mitigation of gender bias, several studies have investigated methods for reducing bias, to varying degrees of success. In contrast, the use of reinforcement learning as a means of mitigating bias in general is understudied in the field of NLP, especially in

the case of occupation prediction. As it is expected that reinforcement learning can be a successful method for the mitigation of gender bias, this thesis investigates whether it is possible to mitigate gender bias in NLP, regarding the use case of the occupation prediction task, using reinforcement learning. The main contributions of this thesis include the formalization of an NLP task in such a manner that reinforcement learning can be applied. Moreover, it investigates the influence of the reward function. Lastly, it demonstrates that reinforcement learning is an effective method for the mitigation of gender bias in occupation prediction.

1.2 Thesis structure

This thesis is structured as follows: Firstly, Chapter 2 provides an overview of the key research in the field of bias in machine learning, and more specifically in NLP. In addition, this chapter will demonstrate how reinforcement learning works and explain why this may be a suitable approach for mitigating gender bias. Subsequently, a synopsis of studies that employ RL for the mitigation of bias is given, along with the description of the task of occupation prediction. This chapter is followed by Chapter 3, which provides an overview of the task that this thesis focuses on, the data set, and the corresponding evaluation metrics that are used to measure how well reinforcement learning works in mitigating gender bias. Chapter 4 outlines the methodological elements of this thesis, including the formalization of occupation prediction for RL, and the setup of the experiments. The results of the experiments are shown in Chapter 5, along with the results of an additional experiment which combines RL with an input-based debiasing strategy. In this chapter, the findings of this thesis are also compared to related work; the results will be discussed further in Chapter 6, as well as a conclusion to this thesis, and recommendations for future work.

Chapter 2

Background

This chapter discusses the theoretical framework in which this thesis is written. The main research question of the thesis concerns the mitigation of gender bias in occupation prediction. Bias is defined in this thesis as undue prejudice which inflicts an unjust outcome on a group or population. In machine learning, this negative outcome can be caused by skewed statistics that favour one group over another (Lloyd, 2018). This chapter provides an overview of studies on the impact of (gender) bias in society, bias in machine learning and NLP, as well as efforts that have been made to mitigate the harmful effects of bias in NLP tasks. Additionally, the intuition and mathematical formalization behind the method of reinforcement learning is explained. Subsequently, studies that have leveraged RL for the mitigation of bias are discussed. At the end of the chapter, a brief synopsis is given of the information provided in the following sections.

In recent years, an array of studies has investigated the influence of social bias and stereotyping on artificial intelligence. Bias may occur in different forms in society, such as negative treatment of individuals due to their age. An example is the bias towards older people that is often present in the workplace, due to the widespread belief that work performance decreases with age (Rupp et al., 2006). Bias is not only found in the workplace, but also in the educational system. Tenenbaum and Ruck (2007) found that in the classroom, racial bias is present: Teachers had highest expectations for Asian American students, and held higher expectations for white pupils than their peers with Latin or black backgrounds. Another example of racial bias can be observed in the clinical practice, where black individuals and other minority groups receive care of poorer quality than white individuals (Nelson, 2002). Here, bias based on gender is also found: A variety of conditions, such as Parkinson's disease, are investigated and treated to a larger extent in men than in women (Hamberg, 2008). Moreover, Isaac et al. (2009) found that society holds negative views towards women who applied for jobs that are traditionally done by men. It is evident that social bias is present in daily life and poses issues for the individuals or groups that are viewed or treated through a biased lens. As artificial intelligence becomes more integrated in people's lives, the risk of bias by machines grows. Automated systems that make decisions are trained on large data sets, which often contain biases. Although there have been attempts to remove bias from data, AI systems have the potential to negatively impact groups that are already marginalized (Lloyd, 2018).

The focus of this thesis will be on gender bias in machine learning. This chapter will provide a definition of gender bias in the first section. Subsequently, an overview

of gender bias in machine learning, and more specifically in NLP, will be presented. In addition, a synopsis of studies that have attempted to mitigate bias in machine learning is given. Since reinforcement learning will be employed in this thesis as a mitigation method for gender bias, this type of learning will be discussed in more detail in this chapter, as well as previous studies that have used reinforcement learning for the mitigation of bias.

2.1 Gender Bias

Gender bias can be defined as the prejudice or preference of one gender over another (Moss-Racusin et al., 2012). In general, bias may be categorized in different ways, but the categorization as proposed by Crawford (2017) is often employed in the field of NLP. Crawford (2017) divides bias in terms of allocation bias and representation bias. Allocation bias is stated to be an economic issue where a system unfairly divides resources over certain groups at the cost of others. Representation bias occurs when a system takes away from the social identity and representation of certain groups. When applying this distinction to natural language processing, allocation bias can be seen when models perform better on data associated with a majority gender; representation bias may be observed when connections between gender and a certain concept are captured in model parameters or word embeddings. Crawford (2017) states that bias in NLP comes in four different types: (1) Denigration, i.e. using historically or culturally derogatory terms; (2) recognition bias being the inability of models to have high accuracy in a recognition task; (3) stereotyping, meaning the reinforcement of already existing societal stereotypes, and (4) under-representation, i.e. the disproportionately low representation of a particular group. Sun et al. (2019) argue that both allocation and representation biases are often present in NLP applications because of biased statistical patterns that arise in training data.

Gender Dichotomy. An important issue to note is that often in studies investigating gender bias in machine learning, gender is viewed in binary terms: Unfair treatment of women versus men is usually the focus of these studies. However, gender is a concept that spans more than the simple dichotomy of men and women. According to Matsuno and Budge (2017), *non-binary* is an umbrella term that may include individuals whose gender identity falls outside or in between female and male identities; those who experience being a man or woman at different times and people who do not experience having a gender identity. The binary division of gender that is often employed in NLP studies should thus be regarded critically. There has been some work done outside of the gender dichotomy in NLP, such as Cao and Daumé III (2021), where it was identified that systematic biases in coreference resolution models can harm binary, non-binary, cis and trans stakeholders. These authors also argue that developing systems that are ignorant of the complexity of gender is dangerous to various groups, including transgender people.

With regard to real-world data, there are certain challenges that arise with regard to how gender is represented in language: Dev et al. (2021) states that neopronouns such as singular *they*, or *xe/xem* do not have sufficient coverage to be able to train a model. In the case of *they*, it is also difficult to distinguish between the singular and plural form of the word. Moreover, as gender is a fluid concept, representing it in discrete categories may be detrimental to individuals who identify as non-binary. Although I am

aware of the challenges of regarding gender in binary terms, the purpose of this thesis is to develop a method that mitigates gender bias using reinforcement learning. If it is found that this approach works with the dichotomous man/woman gender distinction, follow-up research can be done to apply the method to the wider spectrum of gender. Therefore, this thesis will analyse machine learning bias towards women, compared to men.

2.2 Gender Bias in NLP

Bias that humans have towards certain groups of people may be copied or even amplified by machine learning algorithms. Since they are trained on data that contain certain biases, the algorithms can replicate this pattern and propagate it. In the field of natural language processing, gender bias has been found in many subtasks. For example, Prates et al. (2020) found that Google Translate displayed a strong favor of male pronouns when translating them in relation to jobs typically associated with men. In addition, it was found in this study that the algorithm propagated this bias more than is seen in real-world data, suggesting that the model amplified the bias learnt from the data. Another example of gender bias in automatic translation can be seen in Stanovsky et al. (2019), who found popular machine translation algorithms are more prone to biased translation errors, often connecting male pronouns to stereotypically male roles. Bias towards males is also found in word embedding models, as demonstrated in Bolukbasi et al. (2016) and Caliskan et al. (2017). Word embeddings are a means of representing words as vector spaces, so they can be processed by a machine. Bolukbasi et al. (2016) observed that embedding models trained on Google News data implicitly amplify biased information captured in text. For example, if the system is prompted the statement "man is to computer programmer as woman is to x", it will answer with x=*homemaker*. Additionally, Caliskan et al. (2017) found that gender bias was present in widely used word embedding models such as GloVe and Word2Vec. Nissim et al. (2020) demonstrate that these types of analogies are not completely accurate diagnostics for bias in word embedding models, though they show that there is still bias in word embeddings.

Gender bias can also be seen in other NLP tasks, for example in caption generation. Hendricks et al. (2018) demonstrated that image captioning models exaggerate biases that are present in training data, by e.g. incorrectly predicting that an agent is male because there is a computer close to the agent. Moreover, speech recognition models were shown to have lower accuracy for female voices. Tatman (2017) found that the automatically generated captions on YouTube exhibit significant differences in accuracy between male and female voices, a consequence of an imbalanced training data set favoring male voices. Another task where gender bias is shown to be present is sentiment mining, where sentences containing female noun phrases are often ranked to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018).

Lastly, large language models are often a source of gender bias. Lu et al. (2020) discovered that recurrent neural network (RNN) based language models trained on benchmark data sets display significant gender bias in how the models view occupations. In a coreference resolution task, it was found that female pronouns were often assigned to stereotypically female occupations. Similar forms of bias can also be observed in transformer models, such as BERT (Bhardwaj et al., 2021).

The studies discussed above demonstrate that social bias in data sets can be picked

up by machine learning algorithms, which in turn may propagate and sometimes compound these gender imbalances. This behaviour may be harmful when these automated systems are applied in real-world situations, such as job recommending systems or automated hiring practices, since these systems may amplify existing stereotypes and biases. Many studies have investigated methods to mitigate bias; the following section will address several that made an effort to reduce gender bias in a variety of NLP tasks.

2.3 Mitigating Gender Bias in NLP

A vast body of literature has been written discussing bias in language models, although Blodgett et al. (2020), among others, argue that the concept is a difficult one and often not well-defined in studies. In previous work, biases are quantified through metrics, which function as proxies for measuring bias. It is the question whether these metrics actually evaluate the social notion of bias. As became evident from the previous section, gender bias may surface in varying forms in each NLP task. Therefore, every effort to mitigate bias ought to include thorough reflection on what type of bias may surface in what task, what assumptions underlie the method chosen for the mitigation, and whether the measure reflects the social concept of bias. This section will analyze previous work done to resolve gender bias in different NLP tasks, from developing measures for quantifying bias to the augmentation of training data.

2.3.1 Observing Bias

Word embeddings are a means of representing words as vector spaces. As word embeddings are often a fundamental element of NLP experiments, bias that is present in this representation may be propagated downstream in further tasks. Therefore, being able to measure bias in word embeddings is vital. Caliskan et al. (2017) developed a test for determining to what degree bias is present in this type of representation. The Word Embedding Association Test (WEAT) was based on a psychological measure (the Implicit Association Test - IAT), a cognitive reaction task which is supposed to measure associations between mental representations of concepts in memory (Greenwald et al., 1998). Caliskan et al. (2017) adopt the concept of the IAT and apply it to word embeddings to create WEAT. In this metric, the proximity between word embedding vectors is used as a proxy for measuring social bias. May et al. (2019) use the WEAT metric to develop a similar metric that is capable of measuring bias in sentence encoding models such as ELMo, namely the Sentence Encoder Association Test (SEAT).

Gender bias in word embeddings can also be analyzed in terms of the difference in performance across genders. Predictions by a non-biased model should not be influenced by the gender of the entities in the input. Gender-swapped sentences, i.e. sentences in which the gender of the gendered noun is changed to the opposite, should lead to equal performance by a model (Zhao et al., 2018a; Lu et al., 2020). If there is a difference in performance, the difference in performance scores can be viewed as a measure for the amount of gender bias in a system. An example of this method is Dixon et al. (2018), who developed two metrics to measure performance difference: False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED). These have been used mostly in the task of abusive language detection (Cheng et al., 2021; Park et al., 2018), and are defined as the difference in false positive and false negative rates of predictions between the original and gender-swapped inputs. Addi-

tionally, De-Arteaga et al. (2019) measure gender bias through the true positive rate (TPR) gender gap in the task of occupation prediction, i.e. the difference in the true positive rate between genders. A model with lower bias is expected to have lower TPR gender gaps. This measure will be used to evaluate the effectiveness of the bias mitigating method developed in this thesis, and will be discussed in more detail in Chapter 3.

In addition to the measures that quantify bias within embeddings, and those that regard the performance of the classifier, test sets can also provide a means of quantifying bias. Gender bias evaluation test sets (GBETs) are designed to check whether language models make mistakes as a result of gender bias. Most of the publicly available GBETs are focused on the task of coreference resolution, such as the Winogender Schemas data set and the WinoBias data set, which probe the concept of occupation prediction (Rudinger et al., 2018; Zhao et al., 2018a). For sentiment analysis, the Equity Evaluation Corpus was designed as a gender bias test set (Kiritchenko and Mohammad, 2018).

2.3.2 Debiasing Methods

As bias may occur in different stages in a pipeline, from the input representation or the classifier itself, different debiasing methods also address varying parts of a pipeline. The following section discusses the different types of debiasing strategies that have been previously studied.

Regarding gender bias, Zhao et al. (2018b) address the issue of word embeddings learning social stereotypes from human data by developing gender-neutral word embeddings. They do so by preserving gender information in some dimensions of the embeddings, but training other dimensions to be gender-free. The authors found that this approach was successful in isolating gender information without damaging the functionality of the embedding model. Another method to debias word embeddings is removing their gender subspace: Bolukbasi et al. (2016) proposed to change the embedding space by making gender-neutral words orthogonal to the gender direction. Little research has been done that investigates whether these debiasing methods can be extended to other languages than English, since English does not have grammatical gender which could make the task more complicated.

In addition to bias that may be found in input representation, algorithms often copy bias they find in the data they train on. For example, Zhao et al. (2018b) demonstrate that a data set may have a disproportionate number of references to one particular gender. These authors introduce an approach to mitigate this: they created an augmented data set that is biased towards the opposite gender to the original data set. The data augmentation works by gender-swapping every sentence in the original data set. Subsequently, name anonymization is applied to both versions of the sentence. This procedure removes the associations between gender and named entities. The model is then trained on the union of the two sets, which together form a more balanced data set. This method has been found to be successful in the mitigation of gender bias in a variety of tasks. However, the augmentation of data may be time-consuming and expensive, if there is high variability in the data, or if the data set is large. Moreover, combining the gender-swapped and the original data set doubles the amount of training data, which may increase the time that it takes to train a model. Madaan et al. (2018) also show that gender-swapping sentences can lead to nonsense inputs, such as *he gave birth*, as a counterpart to *she gave birth*.

In tasks such as machine translation, models disproportionally often predict the source of a data point to be male. This is caused by training sets being dominated by male-sourced inputs, so a model learns statistical patterns that favour males (Vanmassenhove et al., 2019). Gender tagging can provide a solution for this issue, as it consists of adding a tag that indicates the gender of the source of the data point. The goal is to preserve the gender information of a data point so a model can make more accurate predictions (Vanmassenhove et al., 2019). This method was found to be effective in mitigating gender bias, but it is a costly operation to manually add gender tags to all ambiguous data points.

Another method that was created for the mitigation of gender bias is that of bias fine-tuning. This approach incorporates transfer learning, i.e. the knowledge that a model gained from training on different data, from an unbiased data set . This ensures that the model has a minimal amount of bias before fine-tuning it on a more biased data set (Park et al., 2018). Bias fine-tuning has not been extensively studied, as unbiased data sets are rare. Therefore, more investigation into the effectiveness of this biased mitigation method ought to be done. On the task of occupation prediction, which is the use case of this thesis, De-Arteaga et al. (2019) employ a method for reducing gender bias in occupation prediction by removing gender indicators from the input of the classifier. They found that this approach was successful in mitigating gender bias in this task, although they note that some amount of bias still remains in the systems. Using these systems in a pipeline that is used for a real-world application may still be harmful for individuals, even when there is less bias compared to the non-debiased classifier. On the same task of occupation prediction, Ravfogel et al. (2020) develop a different debiasing approach, which is a post-hoc method, meaning that the debiasing method is applied after the models are trained: this approach, called iterative null-space projection (INLP) removes specific information through iterative null-space projections, meaning that the embedding space is altered. The authors found that INLP was able to reduce gender bias in occupation prediction. When this method goes through multiple iterations, however, the risk of other information besides the target being affected increases. Haghighatkhah et al. (2022) address this issue by creating a similar method, called Mean Projection (MP). MP, according to Haghighatkhah et al. (2022), targets only the target attribute, and does not alter the remainder of the embedding space. The results of this study show that MP can mitigate gender bias to a similar extent than INLP, and that the classifiers where MP was applied maintain marginally higher accuracy.

A possible approach for debiasing machine learning algorithms that has not been studied widely is reinforcement learning: A machine learning paradigm that is based on the algorithm striving towards obtaining a high reward for good behaviour, rather than finding hidden statistical patterns (Sutton and Barto, 2018). This method may be appropriate for the mitigation of bias, as the maker of the algorithm can define a reward for behaviour that pays attention to minority classes in a classification task. Regardless of the potential that reinforcement learning has as a debiasing method, it has not been studied as such. By rewarding behaviour that pays attention to the underrepresented gender, the algorithm may find a way to act in such a manner that it "teaches" itself to make debiased predictions. Thus, there is no need for manual interference in data or algorithms, as one can specify in an RL algorithm what behaviour is desired. Few studies have been executed that investigate leveraging reinforcement learning for gender bias in machine learning. In NLP in particular, to my knowledge at the moment of

writing this thesis, only one study was done that uses reinforcement learning for the mitigation of bias, namely Cheng et al. (2021), which concerned general bias in hate speech detection. Thus, there is a gap in the literature pertaining to the mitigation of specific types of bias in different NLP tasks.

In the following section, an explanation of reinforcement learning will be given. Firstly, the intuition behind RL is explained with an example. The intuition is then elaborated on in the mathematical formalization of reinforcement learning. Subsequently, an overview will be given of studies that have employed RL for the mitigation of bias.

2.4 Reinforcement Learning: Intuition

In machine learning, there used to be two different paradigms. Firstly, there is *unsupervised learning*, which entails a model finding patterns from unlabeled data. Secondly, there is *supervised learning*, where a machine learns from data that is labeled by some external expert supervisor. In addition to these two types of learning, there is a third type of machine learning, namely *reinforcement learning* (Sutton and Barto, 2018). In reinforcement learning, an agent (i.e. some algorithm or classifier) interacts with its environment, and tries to maximise a reward signal rather than trying to find a hidden pattern. The intuition behind reinforcement learning is firstly discussed according to the example of a game. Subsequently, a formal definition of the elements of an RL algorithm is given.

In Figure 2.1, the *lizard game*[1] is shown. In this game, the lizard is the decision maker (who in reinforcement learning is called the *agent*), whose goal it is to eat as many crickets as possible in the least amount of time. If the lizard comes across the bird, the lizard itself will be eaten. The lizard has four possible decisions to make in terms of movement. These decisions are called *actions* in RL and, in this game, consist of moving left, right, up or down. The lizard is given a *reward* of one point if it lands on a tile with one cricket; an empty tile results in minus one point. If the lizard comes across the tile with five crickets, it receives a reward of plus ten points, and the *episode* of the game ends. If the lizard lands on the tile with the bird, this results in minus ten points and this also ends the episode.

When thinking of the lizard game as a reinforcement learning problem, the playing board is called the *environment*. In the lizard game, the lizard (i.e. the agent) interacts with the environment in order to maximize the reward before the end of the episode. The lizard starts by choosing which tile to go to from the current tile. The tiles represent the current *state* of the agent (i.e. its current perception of the environment), and by choosing an *action* the agent moves from one state to the next, i.e. from one tile to another. When the agent has moved to a new state, it receives a reward based on the action that it took. In the example, the first two actions that the lizard can take are either up or right (it cannot move away from the board), which will both result in minus one point since the lizard will land on an empty tile. From that tile, i.e. the next state, the lizard will choose the following tile to go to. At first, choosing actions will be done on a trial and error basis, as the agent has no knowledge on the environment. As the agent goes through more episodes, it will choose its actions based on its knowledge of actions that resulted in the highest reward in the previous episodes.

[1] example inspired by `https://deeplizard.com/learn/video/qhRNvCVVJaA`

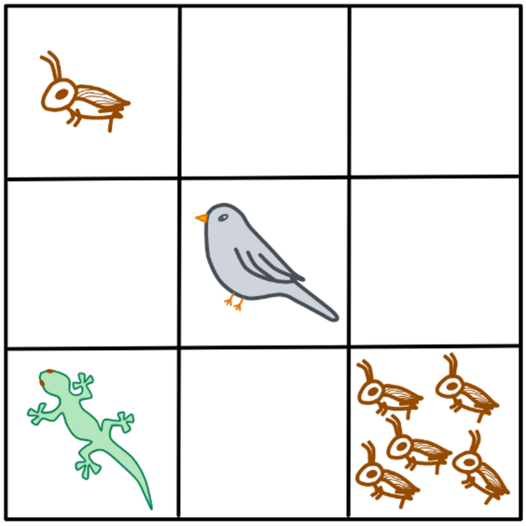

Figure 2.1: Example of Reinforcement Learning: The Lizard Game

In reinforcement learning, the *policy* determines the way the agent behaves at a given time. It entails a mapping from a perceived state of the environment to the action that is taken when the agent is in that particular state. The policy can be stochastic, where probabilities for every action are calculated. In addition to a policy, a reinforcement learning model needs a reward signal, which defines the goal of the problem. With every step, the environment sends a signal to the learning agent called the *reward*, which is a single number (Sutton and Barto, 2018). The goal of the agent is to maximize the reward over a longer period of time. The reward signal is the basis for changing the policy: If an action is followed by a low reward, then the policy may be changed. When applying the notion of the policy to the lizard game, it defines the way in which the lizard decides on the direction to go in from its current tile, in order to get as many crickets as possible. The reward the lizard receives depends on whether there is a cricket, a bird or nothing on the tile, and this will determine the movement of the lizard.

Where the reward signal determines what actions are good in the short term, the *value function* defines what is favorable in the long term. The value of a specific state is the total reward that an agent can achieve over a long period of time, starting from that state. Values indicate the appeal of a state, the probability of the following states and the rewards that correspond to them. In the lizard game, the tile with the five crickets on it has a high value; the tile with the bird a low value. The last element of reinforcement learning, which is optional, is the *model* of the environment. The model allows inferences to be made about how the environment may behave, and is used for planning, meaning that the model decides on a course of action before future situations are experienced. Model-based reinforcement learning (also called deep reinforcement learning or DRL) is more complicated than model free methods, which are essentially based on trial and error.

2.5 Reinforcement Learning: Formalization

Decision making by an agent can be mathematically formalized through Markov Decision Processes (MDPs), which are an idealized way to represent sequential decision making. Here, actions influence immediate rewards as well as following states, and thus subsequent rewards. MDPs make use of delayed reward, i.e. the long-term reward, and need to find the balance between the delayed reward and the immediate reward. The following section will explain the mathematical elements of MDPs.[2]

As we saw in the lizard game example, reinforcement learning consists of an agent interacting with its environment. This interaction occurs in steps over time. At each step, the agent receives some representation of the *state* of the environment, and decides on an *action,* given the state. The environment then moves to the subsequent state, and the agent is given a *reward* based on the previously selected action. Mathematically, MDPs can be notated as follows: In an MDP, there is a set of states $\boldsymbol{S}$, a set of actions $\boldsymbol{A}$, and a set of rewards $\boldsymbol{R}$. At each time step, the agent receives a representation of the state $S_t \in \mathbf{S}$. Based on this, the agent then selects an action $A_t \in \mathbf{A}$. This yields the state-action pair (S_t, At). Then, the time is incremented to the next step *t+1*, and the environment transitions to a new state $S_{t+1} \in \mathbf{S}$. The process can be seen as a function that maps state-action pairs to rewards, meaning that at each time t, we have $f(S_t, A_t) = R_{t+1}$. In Figure 2.2, this process is illustrated in a diagram. The trajectory of the agent selecting an action from the state is in, transitioning to the next state, and obtaining a reward can be formalized as follows: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, ...$

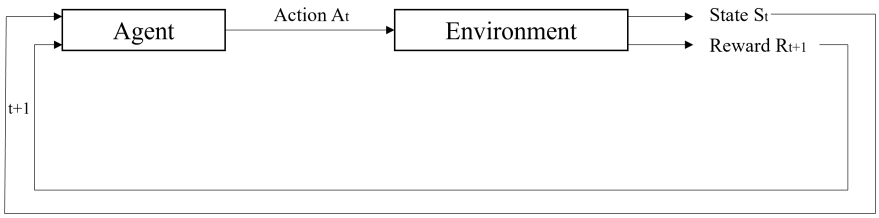

Figure 2.2: Illustration of the workings of MDP-based reinforcement learning systems

The objective of the agent in an RL task is to maximize the long-term cumulative reward. The sequence of rewards after every time step t is denoted as $R_{t+1}, R_{t+2}, R_{t+3}, ...$ In the case of episodic tasks, which are tasks where the agent-environment interaction naturally breaks into sub-sequences called *episodes* (as was the case in the lizard game), the *expected return* (G_t), is the sum of all rewards:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + ... + R_T, \tag{2.1}$$

where T is the final, terminal state (e.g. landing on the tile with the bird in the lizard game). In continuing tasks, i.e. tasks that cannot be broken up into episodes and have no final state, Formula 2.1 is an issue since the return could be infinite. For this type of task, we need the additional concept of *discounting.* In this approach, the agent attempts to choose actions in such a way that it maximizes the sum of discounted rewards it receives over the future (Sutton and Barto, 2018). This means that the agent selects A_t in order to maximize the expected *discounted return*:

[2]explanation inspired by Sutton and Barto (2018)

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{2.2}$$

where γ is a parameter between 0 and 1 called the *discount rate*. If γ is closer to 1, the agent cares more about the immediate reward over future rewards, which are more heavily discounted.

Policies and Value Functions. Most reinforcement learning algorithms involve functions that estimate how good it is for an agent to be in a state or to perform an action in a given state. The notion of *good* can be defined in terms of the expected return. These so-called *value functions* are defined regarding *policies*, which are ways for the agent to act. A policy is a function that maps states to the probabilities of selecting each possible action. If an agent follows policy π at time t, then $\pi(a|s)$ is the probability that $A_t = a$ if $S_t = s$. This means that for each state $s \in \mathbf{S}$, π is the probability distribution over $a \in \mathbf{A}$. In the lizard game example, the agent will choose an action based on the probabilities of each different move, as yielded by the value function, from its current tile. The *value function* of a state s under policy π, denoted $v_\pi(s)$, is given in terms of the expected return when starting in s and following π thereafter. In an MDP, v_π is defined as

$$v_\pi(s) = E_\pi[G_t|S_t] = E[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s], \tag{2.3}$$

where $E_\pi[.]$ denotes the expected value of a random variable, given policy π. This function is called the *state-value function* for policy π. Thinking back of the lizard game, the state-value function shows the lizard a high value for a tile with a cricket, and a low value for the tile with the bird. The value of taking action a in state s under policy π, written as $q_\pi(s, a)$, can be defined as the expected return starting from s, taking action a, and thereafter following policy π. This is called the *action-value function* for policy π:

$$q_\pi(s, a) = E_\pi[G_t|S_t = s, A_t = a] = E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = S, A_t = a] \tag{2.4}$$

This function is referred to as the *Q-function*, which represents the quality of an action. An example of the use of the action-value function in the lizard game is that this function will yield a higher value for moving from an empty tile to one with the crickets, than moving from an empty tile to the bird or to another empty tile. The value functions thus work together under some policy to find the optimal actions for the agent to take, given its current state.

Reinforcement learning algorithms seek to find an optimal policy, i.e. some policy that is better than all others. "Better" is defined in terms of return: a policy π is considered better than another policy π' if the expected return of π is greater than the expected return of π' for all states. The optimal policy has an optimal state-value function, which is denoted as:

$$v_*(s) = \max_\pi v_\pi(s) \tag{2.5}$$

v_* thus gives us the largest expected return possible by any policy π for each state. The optimal policy also has a corresponding optimal action-value function, defined as:

$$q_*(s, a) = \max_\pi v_\pi(s) \tag{2.6}$$

q_* gives us the largest return possible by any policy π for each state-action pair. An essential property of q_* is that is must satisfy the *Bellman optimality equation*. This equation states that for any state-action pair (s, a), at time t, the expected return from the starting state s_t, selecting action a and subsequently following the optimal policy (= the Q-value of this pair) is going to be the expected reward from selecting action a in state s (which equals to R_{t+1}), plus the maximum expected discounted return from any next state-action pair (s', a').

$$q_*(s, a) = E[R_{t+1} + \gamma \max_\pi q_*(s', a')] \tag{2.7}$$

Thus, the Bellman optimality equation expresses the idea that the value of a state under an optimal policy needs to be equal to the expected return for the best action chosen from that state. This function allows optimal actions to be chosen without knowing anything about the subsequent states and their values.

Q-learning. Returning to the example of the lizard game, the agent's goal is to find the policy that leads to it finding the most crickets in the least amount of actions. When starting the game, the lizard does not know anything about the environment and thus knows nothing about what action will lead to what reward. The question is how the lizard is able to find the optimal policy. One such way to do so, without the use of a model, is *Q-learning*. This algorithm iteratively updates the Q-values for each state-action pair, making use of the Bellman equation, until the Q-function converges to the optimal Q-function q_*. This is called *value iteration*. The agent tries to find a balance between *exploration*, the act of exploring the environment to find information, and *exploitation*, the act of exploiting known information to maximize the return. In order to do so, an *epsilon greedy strategy* is used. Epsilon is a parameter between 0 and 1 that represents the probability rate that the agent will explore rather than exploit the environment. In the lizard game, this means that if epsilon is set to 1, the lizard will chose to explore the environment rather than exploit it. This is the case in the first episode of the game, because it has no knowledge of the environment. Once the agent knows more about the environment, ϵ will decay by some factor, that the maker of the algorithm sets. The agent thus becomes more and more greedy concerning the exploitation of the environment, once it has the opportunity to learn more of it. This means that when the lizard is exploring, it will randomly select an action and see what the resulting reward is. Once it gains experience in the environment, it will exploit more, meaning that it chooses actions that will most likely result in higher rewards, based on the knowledge it gained while exploring. The lizard will eventually know that it can gain a higher reward by moving to the tile with the five crickets, and will avoid moving to the tile with the bird. After the agent chooses an action based on ϵ, the Q-value is updated, and the loss between the Q-value and the optimal Q-value is iteratively compared for each state-action pair. The lizard thus chooses an action, observes the following state, the reward it received for choosing the action, and updates the Q-value. Eventually, the Q-value function converges with the right-hand

side of the Bellman equation, and an optimal policy is found. Q-learning is an instance of model-free reinforcement learning.

Policy Approximation. There are issues with representing a reinforcement learning in the manner described above. Firstly, it assumes that there is complete knowledge of the environment, which would ensure that there is a perfect model of the environment, which is often not the case. Even if the agent had access to a perfect model, it would usually be unable to use it to its full extent due to computational and memory limitations. Therefore, the descriptions above provide an idealized way to understand the elements of reinforcement learning. In practice, agents can often only approximate an optimal solution, given limited computation resources.

A more recent development in the field of Q-learning is the deep Q-network (DQN, Mnih et al. (2015)). DQN consists of a neural network that is trained with Q-learning. It was originally designed to play a set of Atari games (i.e. simple games on game consoles), but this method can also be applied to various other tasks. DQN is an example of deep reinforcement learning (DRL), which allows traditional reinforcement learning methods to use the generalization power of neural networks. Deep neural networks allow reinforcement learning algorithms to approximate complex value functions, state-action functions and policy functions. DRL has been a source for a variety of studies, applying it to different tasks, including NLP problems (Uc-Cetina et al., 2022). An overview of these is presented in Section 2.7.

2.6 Types of Reinforcement Learning

Within reinforcement learning, there are three different major groups, namely dynamic programming, Monte Carlo methods, and temporal difference methods. When using dynamic progamming, a large problem is broken up into incremental steps so solutions to sub-problems can be found. The agent must learn from the environment by interacting with it and gaining experience. By doing so, the agent can evaluate its current strategy and can iteratively obtain the optimal value function. A model of the environment is used to do the policy evaluation, meaning that the current value function is updated based on the current policy. When the agent is in its current state, only the optimal path of coming to that state is retrieved. Dynamic programming methods make use of bootstrapping, meaning that they update estimates based on other estimates, which requires a full model of the environment.

In a Markov Decision Process, an entire problem can be broken down into smaller problems with five tuples: S (set of states), A (set of actions), P (probability of moving from one state to another), R (reward given after transition from one state to another) and γ (discount factor used to generate a discounted reward). When all five are known, it is easy to calculate an optimal strategy to obtain the maximum reward. In most problems, however, usually not all information is present at the same time. For example, transition probabilities P are difficult to know and thus, the Bellman equation cannot be used to solve V and Q values. Monte Carlo methods learn from sample returns, rather than immediate rewards. The agent runs trials, where it collects samples, receives rewards and by doing so, evaluates the value function. The final estimated value function will be close to the real value function. In this type of reinforcement learning, each try is called an episode. Therefore, Monte Carlo methods only learn from complete, terminated episodes and can thus only be used in episodic tasks.

Temporal difference methods counteract the issue that Monte Carlo methods can only be applied to episodic MDPs. It combines the bootstrapping from dynamic programming and the sampling from Monte Carlo methods. Temporal difference methods do not need to wait until the end of an episode to update the value function, it simply needs to wait until the next step. They adjust predictions to match more accurate predictions about the future, before the final outcome is known. Q-learning is an example of a temporal difference method, which is the chosen method in this thesis.

2.7 Reinforcement Learning for Bias Mitigation

Reinforcement learning has also been employed in a variety of studies that attempt to mitigate bias. In the field of facial recognition, Wang and Deng (2020) utilised a reinforcement learning based neural network to resolve the bias towards non-Caucasian faces. This study demonstrated that this approach was successful in mitigating racial bias, and learning a more balanced performance. In addition, Lin et al. (2020) used deep reinforcement learning to classify imbalanced data sets. This approach was developed for the task of image classification, but the authors also applied their approach to a text-based data set, which consisted of movie reviews for sentiment classification. The reward in the deep reinforcement learning algorithm was formulated as such that a prediction by the agent that belonged to the minority class received a higher reward than if it belonged to the majority class. It was found by Lin et al. (2020) that their approach outperformed other imbalanced classification algorithms.

As mentioned earlier, only one study was found that used reinforcement learning to mitigate social bias in an NLP task specifically, namely Cheng et al. (2021). The goal of this study was to investigate whether reinforcement learning would be applicable to the task of mitigating cyberbullying. The authors state that social media posts containing certain social biases, such as the occurrence of the words "black" and "gay", are more likely to be classified as instances of cyberbullying. In this work, a standard classifier is used as an RL agent, and a sequence of comments observed at time $\{1, 2, ..., t\}$ as state s_t. The RL agent then selects an action $a_t \in \{not - bullying, bullying\}$ according to a policy (Cheng et al., 2021). The reward function in this work is based on how successful the agent is in predicting the label for a given social media session, and how much bias the classifier has at that point in time. The reward function is not formalized in this paper, and the code of the implementation is also not made public, so it is difficult to know the exact reward function. The effectiveness of RL as a debiasing strategy is evaluated using FNED and FPED (for details on these measures, see Section 2.3.1. Cheng et al. (2021) conclude that reinforcement learning was successful in mitigating bias, without compromising on the accuracy of the classification of comments containing cyberbullying.

2.8 Summary

This chapter has explained that social bias is present in our society. Machine learning algorithms that train on the data that contain bias may propagate and sometimes amplify this bias. Gender bias was seen to be present in a variety of NLP applications, from caption generation to coreference resolution. As gender bias in automatic recognition and classification is often harmful for the groups or individuals that the systems are biased towards, it is key to find methods that mitigate not only gender bias, but

other types of bias as well. Previous work has investigated ways to measure bias, as well as methods to avoid or mitigate gender bias. Several metrics, such as FPED/FNED, WEAT and the TPR gender gaps, have been shown to capture the amount of bias that a system has. Moreover, different approaches for mitigating gender bias have been proposed. These methods mostly focus on the removing bias from word embeddings, altering the training data or adjusting the algorithms. Most methods were found to be successful in mitigating bias to some extent. Nevertheless, they often require a great deal of manual labour, or are costly and time-consuming to implement. Although reinforcement learning has the potential to reduce bias in machine learning algorithms, there are few studies that investigate the possibilities of doing so. Reinforcement learning is defined in this thesis as a type of machine learning, where an agent interacts with its environment, which it looks at in separate states. Oftentimes, the decision making by the agent is formalized as a Markov Decision Process: an idealized way to represent sequential decision making. The agent chooses actions based on those states, and receives a reward based on how desirable that action was. The agent develops a policy that determines how it behaves, in order to maximize the cumulative reward over time. Reinforcement learning has been applied used to mitigate bias in a handful of studies. However, only one study specifically employed this method for the mitigation of bias in NLP, which demonstrated promising results but did not investigate one specific type of bias. Thus, the question arises whether RL can be a suitable approach for the mitigation of gender bias, which is the focus of this thesis.

Chapter 3

Task, Data and Evaluation

In this thesis, reinforcement learning is tested as a viable solution for bias mitigation. This section firstly discusses the task of occupation prediction, which is the use case that RL as a mitigation method is applied to. Secondly, the dataset that is employed in this thesis is presented, along with its corresponding measure for quantifying gender bias.

3.1 Task

The use case of this thesis consists of the task of occupation prediction. Machine learning is increasingly used for automated recruiting and hiring. The information that people show of themselves online is used as input for decision-making systems that recruit candidates for job openings. These system assess people's current jobs, skills and interests. However, even this first step in these automated processes, occupation prediction, can be a source for a variety of issues. Occupation prediction is susceptible to different types of bias, as the algorithms are trained on data that contain biases towards people's gender, but also e.g. their ethnicity (De-Arteaga et al., 2019). In the case of gender bias, the training data for a classifier often contain gender imbalances, as there are occupations that are typically or traditionally done by a specific gender. An algorithm may reproduce, but sometimes also compound these imbalances, which has real-world implications for the candidates that are (not) being recruited by automated systems. Especially when an occupation classifier is used in a pipeline, the bias that is propagated in the first step may be compounded in later applications, and existing gender imbalances may be amplified by these automated systems (De-Arteaga et al., 2019).

3.2 Data

For the task of mitigating gender bias in occupation prediction, the BiasBios dataset is used (De-Arteaga et al., 2019). The dataset consists of online biographies, written in English, stemming from the Common Crawl corpus. De-Arteaga et al. (2019) selected biographies that began with a name-like pattern, i.e. a sequence of two capitalized words, followed by the string "is a(n) (xxx) *title*. The title is an occupation from the BLS Standard Occupation Classification system[1]. The 28 most common occupations

[1] `https://www.bls.gov/soc/`

Profession	Nr Biographies	Minority Gender (%)
professor	118.110	f (45.9)
physician	38.565	f (43.2)
attorney	32.607	f (39.6)
photographer	24.324	f (37.4)
journalist	19.950	m (49.0)
nurse	18.971	m (9.36)
psychologist	18.295	m (36.3)
teacher	16.196	m (38.1)
dentist	14.479	f (36.0)
surgeon	13.273	f (13.6)
architect	10.113	f (26.2)
painter	7.736	f (46.2)
model	7.502	m (15.8)
poet	7.011	f (49.8)
filmmaker	7.009	f (33.7)
software engineer	6.906	f (17.8)
accountant	5.652	f (38.0)
composer	5.600	f (17.7)
dietitian	3.978	m (7.09)
comedian	2.799	f (22.8)
chiropractor	2.598	f (25.4)
pastor	2.532	f (24.8)
paralegal	1.767	m (15.7)
yoga teacher	1.633	m (15.7)
dj	1.485	f (15.4)
interior designer	1.463	m (19.1)
personal trainer	1.432	f (45.2)
rapper	1.407	f (11.4)

Table 3.1: Distribution of professions and gender in the BiasBios data set

were identified from this system, and only those biographies corresponding to these occupations, written between 2014 and 2018, were crawled. As a final step, de-duplication was performed by treating biographies as duplicates when they had the same first name, last name and occupation.

The full dataset as used in De-Arteaga et al. (2019) is not available as a whole online, but the authors supply code to crawl it[2]. Since I did not have the computational means to download the entire dataset, I kindly received the dataset as crawled by Pantea Haghighatkhah. The dataset consists of 397,340 biographies in total. The frequency distribution of all 28 occupations in the dataset is shown in Figure 3.2. The occupation *professor* was most common, consisting of 118,110 biographies; *rapper* was the least common occupation, consisting of 1,407 biographies. The total dataset is split into a training set, a development set, and a test set, each consisting of 255,710; 39,369 and 98,344 biographies respectively. The distribution of occupations between the different splits of the dataset is comparable. The longest biography is 194 tokens; the shortest eighteen, and the median length was seventy-two tokens. The biographies

[2] `https://github.com/Microsoft/biosbias`

are typically written in the third-person by the subject themselves, so self-identified (binary) gender could be extracted. Each biography has a corresponding gold label (i.e. one of 28 occupations) and a gender label (i.e. m/f). A classifier can be trained on these biographies to predict individuals' occupations. The data points that the classifiers are trained on in this thesis are stripped of the name-like pattern, and the string "is a(n) (xxx) *title*".

3.3 Evaluation

De-Arteaga et al. (2019) developed a measure specifically for the BiosBias data set, which quantifies the amount of bias in the task of occupation prediction. In this approach, gender bias is quantified by using the test split of the BiosBias dataset to calculate the true positive rate (TPR) gender gap. The TPR score for a gender and occupation is the proportion of people with this gender and occupation that are accurately predicted as having that occupation (see Equation 3.1). The TPR gender gap is the difference in TPRs between binary genders g and $\sim g$ for each occupation y (see Equation 3.2). In order to quantify the overall bias of a classifier, Romanov et al. (2019) take the root mean square of the TPR gaps for all occupations (see Equation 3.3). This measure aggregates the TPR gaps of a particular gender over all occupations, in order to provide a score that reflects the overall bias of a classifier. The intuition behind using the root mean square rather than the average over the biases in all occupations is that larger values have a larger effect, and the purpose of a debiasing method is to mitigate larger biases (Romanov et al., 2019). When talking about the aggregated TPR gender gaps in the following sections, I mean the aggregated scores of the TPR_{female} for every occupation.

$$TPR_{g,y} = P[\hat{Y} = y|G = g, Y = y] \tag{3.1}$$

$$GAP^{TPR}_{g,y} = TPR_{g,y} - TPR_{\sim g,y} \tag{3.2}$$

$$GAP^{RMS}_{g} = \sqrt{\frac{1}{|C|}\sum_{y \in C}(GAP^{TPR}_{y})^2} \tag{3.3}$$

Gender imbalance may also be compounded by a classifier. The percentage of people with gender g in occupation y is defined as $\pi_{g,y} = P[G = g|Y = y]$, where Y represents the target label and G represents the binary gender of a biography's subject. The gender imbalance of an occupation y is defined by De-Arteaga et al. (2019) as $\frac{\pi_{g,y}}{\pi_{\sim g,y}}$. Gender g is underrepresented if $\frac{\pi_{g,y}}{\pi_{\sim g,y}} < 1$, or if $\pi_{g,y} < 0.5$. The gender imbalance is magnified if the underrepresented gender has a lower TPR score than the overrepresented gender. This can be measured by the correlation between existing gender gaps in the data set, and the TPR gender gaps as produced by the classifier. As is visible in Table 3.2, there is no occupation in the BiosBias dataset that has a completely equal gender distribution, although some are close.

In order to investigate whether using RL for occupation prediction is able to provide a debiased method of classification, the TPR gender gaps of the predictions by this method will be compared to the predictions made by a non-RL classifier. A non-biased classifier has TPR gender gaps close to zero. Therefore, it can be concluded that a debiasing method works when the gender gaps move closer to zero, compared to a

biased counterpart. To compare related work, the performance of the debiased classifier is tested according to the correlation between existing gender gaps in the data set and the TPR gender gaps produced by the classifier. Additionally, the aggregated TPR gaps of all occupations is calculated in order to study the performance of the debiasing method of this thesis to comparable methods.

Chapter 4

Method

This section will outline the method that is used to test the research questions of this thesis. Firstly, the hypotheses on the main research question are posed. Subsequently, the formalization of the reinforcement learning algorithm will be presented in Section 4.2. In Section 4.3, the experimental setup of this thesis is explained.

4.1 Hypotheses

The main research question of this thesis is: ***Can reinforcement learning be employed for the mitigation of gender bias in the task of occupation prediction?*** In order to answer this question, a second research question is posed: *What is the influence of the reward function on the use of reinforcement learning for the mitigation of gender bias in occupation prediction?*

To test these questions, two hypotheses are posed below:

1. *A deep reinforcement learning classification algorithm will contain less gender bias, measured in true positive rate (TPR) gender gaps, compared to a classification algorithm that does not use reinforcement learning.*

2. *A reward function based on higher reward for correct classification for minority genders in an occupation will have less gender bias, measured in TPR gender gaps, compared to a reward function based on correct classification only.*

4.2 Mitigating Bias with Reinforcement Learning

Debiasing occupation prediction can be viewed as a sequential decision making process. In this approach, biographies are presented and observed sequentially. At each step, the decision is made to classify the biography into one of 28 occupations. Thus, occupation prediction may be seen as a sequential Markov Decision Process (MDP), as explained in Section 2.5. To summarize, in an MDP, the agent A interacts with the environment over discrete timesteps t. The agent then selects an action a_t as a response to the state s_t. This causes the environment to change into the next state s_{t+1} and the agent receives reward r_{t+1}.

In this thesis, reinforcement learning is employed to mitigate unwanted gender bias in the task of occupation prediction. A standard classifier (i.e. convolutional neural network) is considered the agent, and a biography observed at a timestep as the state

s_t. The agent then chooses an action $a_t \in \{occupations\}$ according to a policy function $\pi(s_t)$. The policy in this RL system is approximated by a neural network, making it a Deep Q-learning algorithm. The reward is then calculated for the state-action pair (s_t, a_t), and the cumulative discounted sum of all rewards G_t is employed to optimize the policy. The environment in the setup of this thesis consists of the BiasBios data set, where a single biography is presented to the agent at each timestep.

The RL systems in this thesis are based on deep Q-learning. In this RL approach, the policy π consists of a function that receives an observation and returns the probabilities of all labels. The goal of the agent in the deep reinforcement learning algorithm is to maximize the reward over time. Since the agent receives a positive reward when is correctly predicts an occupation, it can achieve its goal by maximizing the cumulative reward. The Q-function then calculates the quality of a state-action pair, according to the Bellman equation. The agent is able to maximize the cumulative reward by solving the optimal Q-function. Deep Q-learning fits the optimal Q-function with the neural network. The Adam algorithm is used to optimize the algorithm.

In order to address the hypotheses as stated in Section 4.1, the results from vanilla neural classifier are compared to the two reinforcement learning systems. The first RL system pertains to the efficacy of reinforcement learning as a gender bias mitigation method. This system consists of a classifier comprised of a simple neural network without reinforcement learning, as described in Section 4.3.1. If the results from the RL system result in lower TPR gaps than the vanilla neural classifier, it may be concluded that reinforcement learning is a successful method for bias mitigation in occupation prediction.

The second RL system, discussed in more detail in Section 4.3.2, addresses the effectiveness of the minority gender-sensitive reward function. It is expected that specifying in this RL experiment that the algorithm should pay more attention to the minority gender of a particular occupation would result in lower TPR gender gaps than a reward function that simply rewards correct classification. Thus, if the results of the RL with minority gender-sensitive reward function, described in Section 4.3.3, demonstrate lower TPR gender gaps than those yielded by the RL with reward for correct classification, it may be concluded that a minority gender-sensitive reward function can mitigate the effects of gender bias in occupation prediction. Since neural networks contain random factors that may influence performance, the average result over five runs will be reported in the next chapter.

4.2.1 Input Representation

The data that is used for the experiments in this thesis was sourced from De-Arteaga et al. (2019), namely the BiosBias data set, as described in Chapter 3. The input biographies are represented in terms of word embeddings. These embeddings are created during the training process by the embedding layer of the neural network that is used as the classifier in this thesis. The choice for this type of embeddings was based on Lin et al. (2020), whose code was the basis for the RL setup of this thesis. The input text is first tokenized by the pretrained BERT tokenizer. The choice for this tokenizer was based on the idea that it could deal well with unknown words and thus improve the performance of the classifier. The inputs that are used in this thesis are thus not BERT embeddings, the BERT tokenizer is simply used to tokenize the texts. All inputs are padded to match the length of the longest sentence, i.e. 194 tokens. They are then fed into the embedding layer of the classifier which transforms them into embeddings.

4.2.2 Code and Computational Requirements

The debiasing method of thesis is inspired by the approach as proposed by Lin et al. (2020)[1], who developed a deep reinforcement learning algorithm for imbalanced classification. The reinforcement learning algorithms were set up using OpenAI's Gym [2], an API for reinforcement learning environments. Reinforcement learning was chosen as a gender bias mitigation method for this thesis, as it is expected that this is a suitable approach to prevent language models from propagating bias during training. In order to test the hypotheses posed in the previous section, three different types of experimental setup will be built. The code used in this thesis can be found on GitHub[3]. .All experiments were done on a Google Colab GPU. Each reinforcement learning system took around two hours to train, the non-RL system around ten minutes.

4.3 Experimental Setup

The experimental setup of this thesis is built of three components: First, a vanilla neural classification system is built, which serves as a comparison for the RL models to test whether they show the hypothesised result. This system is discussed in Section 4.3.1. A second system, which consists of a reinforcement learning-trained classifier, is made to be a test for the effect of the reward function. This RL system contains a simple reward function and is discussed in Section 4.3.2. Moreover, a RL system with a more elaborate reward function is made, which is addressed in Section 4.3.3. The explanations of the individual systems also include their hyperparameter settings.

4.3.1 Vanilla Neural Classifier: No RL

In order to test whether reinforcement learning is a suitable approach for mitigation of gender bias, a system without reinforcement learning is built. The classifier in this setup consists of a convolutional neural networkl. This system makes use of the same classification model as in the RL setup, but it consists only of the neural network, with the addition of a softmax classification layer. The choice for this model architecture was based on Lin et al. (2020), whom the code for this thesis was inspired by. The benefit of using this architecture is that it is a simple neural network, which increases its transparency as compared to e.g. Transformer models. The Adam optimizer is used in this setup, as well as a sparse categorical cross-entropy loss function, with a batch size of 64. The models used in this experiment are fine-tuned for optimal performance, and it was found that they performed the best in terms of accuracy with a learning rate of 0.0001, and three epochs.

4.3.2 RL with Reward for Correct Classification

As a comparison system to test the effectiveness of the reward function that pays attention to the minority gender in occupation prediction, a RL system is set up. In this setup, in contrast to the setup for the RL with minority gender-sensitive reward

[1] code can be found at `https://github.com/linenus/DRL-For-imbalanced-Classification.git`
[2] `https://gymnasium.farama.org/`
[3] `https://github.com/cltl-students/Kloos_Mojca_RL_gender_bias.git`

function, the reward function is simply based on correct predictions:

$$R(s_t, a_t, l) = \begin{cases} +1, a_t = l_t \\ -1, a_t \neq l_t, \end{cases} \tag{4.1}$$

The RL system with a reward based on correct classification only can be compared to the RL system in which the takes the minority gender into account for its predictions, which is elaborated upon in the next section. By doing so, it can be tested whether the reward function performs as hypothesized.

This system is also fine-tuned to find the parameters that would result in the optimal performance. Firstly, γ is set to 0.5. This parameter indicates the probability of the agent choosing exploration over exploitation in the ϵ-greedy strategy. This value was chosen because this balances the probability of exploration and exploitation. The learning rate is set to 0.0001, as it was found that this resulted in high accuracy without the need for extremely long training times. The performance in terms of accuracy and reward stabilized around one million steps. Therefore, the number of steps that the agent takes in this setup is set on 1,022,840, which means that the agent sees every training instance exactly four times (as the size of the training data set is 255,710 instances).

4.3.3 RL System with Minority Gender-Sensitive Reward

For the third type of model developed in this thesis, a reinforcement learning algorithm will be created, in which a reward function is used that pays attention to the minority gender in the predicted occupation.

The reward for this RL system is based on the gender distribution of the predicted occupation. The intuition behind the reward is as follows: The goal of the task is to mitigate the biased prediction of occupations. A prediction is seen as biased when the agent classifies an occupation solely on the gender of the individual described in the biography, because the overall gender distribution of that occupation is skewed. For example, if a biography reads *She works in a hospital*, and the agent predicts the occupation *nurse*, not the gold label *physician* because there are more female nurses and less female physicians, the prediction is considered biased.

To mitigate biased predictions, the reward will be higher if the agent correctly classifies an occupation for a biography with a minority gender for that specific occupation, than if it correctly does so for a majority gender. Thus, for the prediction *physician*, the agent will receive a higher reward if the biography reads *She works in a hospital* than if it reads *He works in a hospital*, since *female* is the minority gender in the occupation of *physician*. The reverse will also be implemented in the reward function: if the agent classifies a biography falsely while the gender in the biography is the majority gender, the punishment will be higher than if the gender is the minority gender. The reward function is formalized as follows:

$$R(s_t, a_t, l) = \begin{cases} +0.5, a_t = l_t \text{ and } s_t \in D_{majority} \\ +(1-\lambda), a_t = l_t \text{ and } s_t \in D_{minority} \\ -0.5, a_t \neq l_t \text{ and } s_t \in D_{majority} \\ -(1-\lambda), a_t \neq l_t \text{ and } s_t \in D_{minority}, \end{cases} \tag{4.2}$$

where $\lambda \in [0, 1]$ signifies a parameter that indicates the share that the minority group holds in the predicted occupation (see Table 3.2 for the gender distribution per occupation).

The RL algorithm with reward for minority-sensitive classification is fine-tuned for optimal performance. The learning rate is set to 0.0001, and the number of training steps to 1,022,840. The performance regarding reward and accuracy stabilizes around one million steps and the model is not trained any further to avoid overfitting. The Adam optimizer is applied to optimize this algorithm. It is to be noted that in a deep Q-learning algorithm, there are many more parameters that may be tuned for more optimal performance. Due to the limited time span of this thesis, and the limited computational resources I had access to, only the hyperparameters mentioned above were fine-tuned.

Chapter 5

Results

In this thesis, the potential of reinforcement learning as a debiasing method for reducing TPR gender gaps in occupation prediction is investigated. This chapter presents the results of the experiments as introduced in Chapter 4. In the first part of the chapter, Section 5.1 demonstrates the influence of reinforcement learning on the gender bias of occupation prediction algorithms. The second part presents a more in-depth analysis of the results, and investigates what the influence is of changing the reward function between RL with reward for correct classification and the RL with minority gender-sensitive reward. The discussion of the results is done on the average TPR scores and performance over five runs of the same model, with different seeds. As there was little variation between the results, five runs was deemed enough to perform an analysis on. The results of all individual models can be found in Appendix A.

5.1 Results of Reinforcement Learning for Reducing TPR Gender Gaps

Model	Accuracy (in %)	Aggregated TPR Gap
Vanilla neural classification	79.4	0.178
RL with reward for correct classification	75.4	0.115
RL with minority gender-sensitive reward	75.0	0.193

Table 5.1: Performance of models

A classifier can be considered as containing less gender bias when its TPR gaps moves closer to zero as a result of a debiasing method. The bias may also be considered lower when the correlation between existing gender gaps in the data set, and those produced by the classifier is reduced.

As is visible from Table 5.1, the aggregated TPR gender gap score (for details on this measure, see Section 3), is reduced by RL with reward for correct classification, but not by the RL with minority gender-sensitive reward (in this system, it is increased). The goal of the experiments in this thesis was to lower the scores on this measure compared to a non-debiased system. The accuracy of RL with reward for correct classification, i.e. the best-performing debiasing system, drops by four percent compared to the vanilla neural classifier.

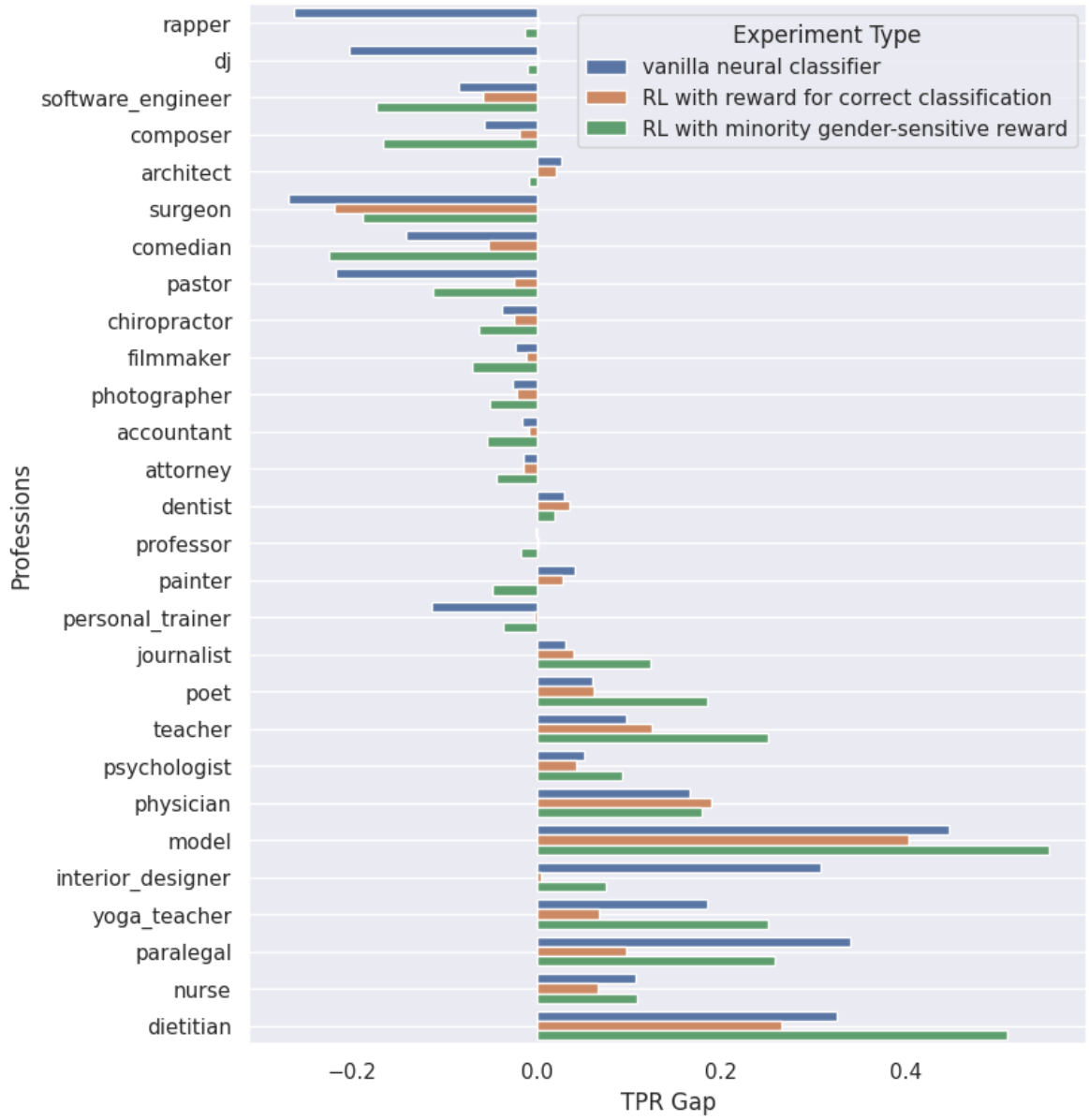

Figure 5.1: TPR gender gap of every occupation in the three different experiment types, sorted by the gender distribution (represented as % female) of the gold data.

Figure 5.1 demonstrates the TPR gender gap per experiment type and occupation. The occupations in this figure are sorted by the gender distribution as found in the gold data, i.e. the "true" gender distribution. It is visible from this figure that RL with reward for correct classification reduces the TPR gender gap in almost every occupation, except for *dentist, journalist, poet, teacher* and *physician*. In contrast, the RL system with a reward function that is minority-gender sensitive increases the TPR gaps of the classifier in most occupations. A notable observation that can be done from Figure 5.1 is that the bias in most occupations went down slightly as a result of the RL with reward for correct classification, but there is a great difference in

bias between the non-RL system and RL with reward for correct classification in the occupations *DJ, rapper, paralegal, pastor* and *interior designer* and *yoga teacher*. An explanation for this finding can be found in Figure 5.2. This figure shows the percentage of females in each occupation, as propagated by each of the classifiers. As is visible from Figure 5.2, the RL system with the minority-sensitive reward generally underclassifies females in occupations where the gold gender distribution has a male majority. It generally overclassifies females in occupations where the gold majority gender is female (from *poet* onward). This can also be seen in Figure 5.3, which shows the trend line of each of the gender distributions as propagated by the different classifiers. The opposite of this trend can be seen in the RL system with reward for correct classification, which generally overrepresents the minority gender, i.e. it overclassifies females in male-majority occupations, and overrepresents males in female-majority occupations. In the occupations where the difference in bias between the vanilla classifier and the RL with reward for correct classification was the largest, we can see that the gender distribution is skewed towards the minority gender the most (i.e. in *rapper, DJ, pastor, interior designer, paralegal, yoga teacher*). These occupations all have relatively small sample sizes in the data set, with extreme gender distributions. This suggests that the RL system with reward for correct classification was able to effectively reduce the gender bias in small sample sizes, by changing the gender distribution to include more biographies of the minority gender. The RL system with minority gender-sensitive reward, in contrast to the other RL system, was unable to reduce gender bias in these small sample sizes, and only compounded the existing skewed gender distribution.

5.2 Comparison to Other Debiasing Methods

In order to investigate whether reinforcement learning is a suitable method for debiasing occupation classification, the results of this thesis need to be compared to similar studies. Firstly, it is to be noted that the approach taken in this thesis is a training-based debiasing strategy, meaning that the debiasing method is applied during the training process of the model. Other strategies include input-based methods, where debiasing is done on the biographies or embeddings that are then fed into a classifier, and post-hoc strategies, where models are debiased after they are already trained. As will become evident in the following section, there is almost always a trade-off between the accuracy of a classifier and its bias, i.e. the less biased, the less accurate a model is. The results found in this thesis, only the debiasing strategy of RL with reward for correct classification versus the vanilla neural classification will be compared to other debiasing studies that were conducted on the same data set, since it performs better in terms of TPR gap reduction than the RL with minority gender-sensitive reward.

An input-based debiasing approach is De-Arteaga et al. (2019), who scrubbed the input biographies of gender indicators such as pronouns. Their approach reduced the correlation between existing TPR gender gaps and those found in their approach between 0.15 and 0.06, depending on the input representation. The intuition behind measuring this correlation is to quantify to what degree a classifier amplifies bias compared to the real distribution of gender in a profession. In this thesis, the correlation between existing gender gaps and those found in this thesis between the vanilla neural classification and RL with reward for correct classification is reduced by 0.196. The accuracy of the debiased models in De-Arteaga et al. (2019) is higher than the accuracy of those in this thesis (between roughly 78-82% in De-Arteaga et al. (2019) versus

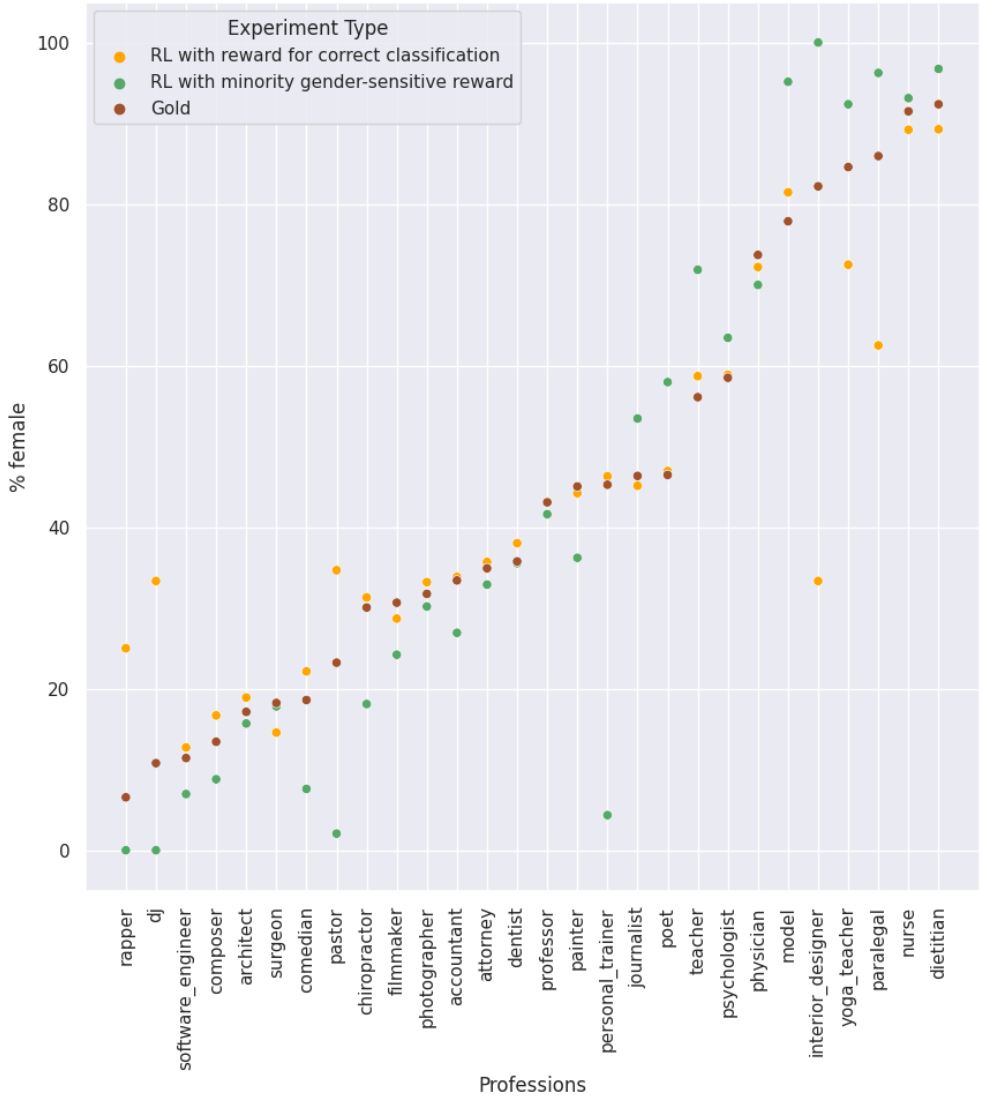

Figure 5.2: Gender distribution in predictions made by the three different classifiers, compared to the gender distribution in the gold data, sorted by the gender distribution (% female) of the gold data. The results presented in this figure are based on the models that were trained with seed 2.

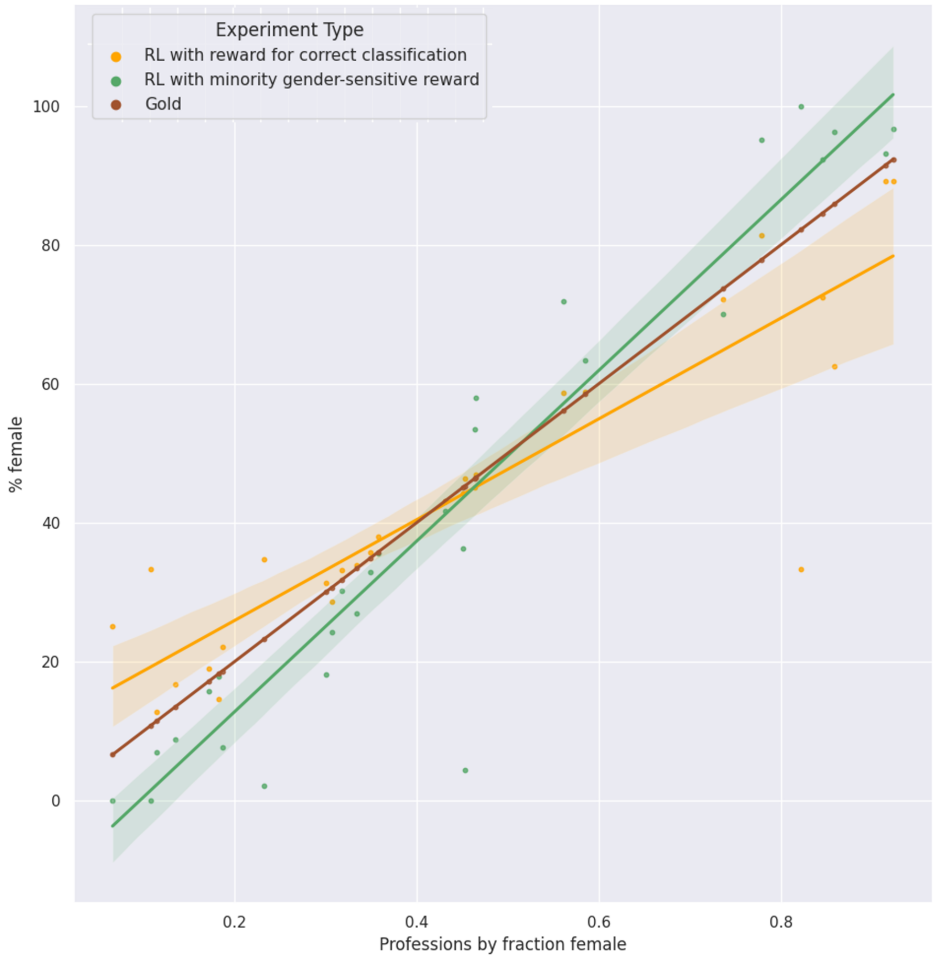

Figure 5.3: Trend line through gender distribution in predictions made by the three different classifiers, compared to the gender distribution in the gold data, sorted by the gender distribution (fraction female) of the gold data. The results presented in this figure are based on the models that were trained with seed 2.

Model	Accuracy (in %)	aggregated TPR gap (reduction in %)	Correlation
RL (CC)	75.4	0.115 (35.4)	0.670
MP (FastText)	75.6	0.092 (46.5)	0.522
MP (BERT)	75.2	0.087 (43.1)	0.395
INLP (FastText)	73.6	0.103 (40.1)	0.493
INLP (BERT)	74.7	0.065 (43.1)	0.353
RL (CC) + scrubbed inputs	75.4	0.102 (42.7)	0.625

Table 5.2: Comparison of best performing model to related work, on accuracy, aggregated TPR gap bias measure , how much it was reduced compared to the non-debiased counterpart, and correlation of TPR_g and gold percentage of women for each occupation. The reduction in aggregated TPR gender gap is calculated compared the baselines used in the respective studies. RL (CC) means RL with reward for correct classification.

75.4% in this thesis). De-Arteaga et al. (2019) did not employ the aggregated TPR score as developed by Romanov et al. (2019), which is the main measure of bias used in this thesis. No raw scores from the study done by De-Arteaga et al. (2019) were available, so I was unable to compute the aggregated TPR gap for this method used in this study. The results of De-Arteaga et al. (2019) are thus difficult to compare to the results found in this thesis, and to other debiasing methods.

Ravfogel et al. (2020) leverage a post-hoc debiasing strategy, by changing the embedding space of the models employed in the study, using linear projections. This method is called iterative null-space projection (INLP). Haghighatkhah et al. (2022) employed a similar method to Ravfogel et al. (2020). Unlike INLP, which affects the entire embedding space, their approach (called Mean Projection or MP) targets a single attribute in the embedding space. Haghighatkhah et al. (2022) apply both methods to the BiasBios data set, with different input representations. The results used for comparison with the results of this thesis are those where embeddings were used as input. Two types of embeddings were used in Haghighatkhah et al. (2022), namely FastText and BERT embeddings, and both are compared to the results of this thesis (see Table 5.2). As is visible from Table 5.2, the best performing model in terms of reduction in aggregated TPR gender gap, namely RL with reward for correct classification, does not reduce gender gaps to the same extent as the other methods shown in this table. However, the RL system has higher accuracy than most of the other approaches, with the exception of MP with FastText embeddings.

5.3 Additional Experiment: RL with scrubbed biographies

This section presents the results of an additional experiment that was done to investigate the potential of combining RL with other debiasing methods. The results of this experiment, in comparison to other debiasing methods, can be found in Table 5.2 (see RL(CC) + scrubbed inputs).

In Cheng et al. (2021), the authors recommend to combine multiple debiasing approaches, such as debiased input representations, with their reinforcement learning bias mitigation method. Since the debiasing approach taken in this thesis, like in Cheng et al. (2021), is training-based, there is a possibility to combine it with other methods for bias mitigation. In the approach implemented in De-Arteaga et al. (2019), the authors scrubbed gender indicators from the input biographies, and found that this was a suitable approach for mitigating gender bias. For this thesis, in an additional experiment, it was tested whether combining the scrubbed input biographies with reinforcement learning decreases the TPR gap of the classifier more than reinforcement learning alone. To do so, the same setup as in the RL with reward for correct classification was employed. The biographies without gender indicators were present in the dataset as I received it, so the scrubbed texts were processed in the same manner as the regular biographies in the setup as described in Section 4.2.1. The reinforcement learning system was trained with the same amount of steps, and the same hyperparameters as in RL with reward for correct classification, the only difference pertains to the input biographies not containing gender indicators. Only one run was done with seed 14, as this is only an exploratory experiment to investigate the combination of multiple debiasing strategies, and there was little variation in the performance of the different seeds of RL with reward for correct classification in the original experiment.

For an overview of the results of this experiment, compared to other debiasing methods, see Table 5.2. The results of this preliminary experiment demonstrate that the classifier preserve the same accuracy as the classifier that was trained on the non-scrubbed texts (75.4%). The combination with the scrubbed texts resulted in a correlation with existing gender gaps of 0.625 ($p < 0.005$), and a aggregated TPR gap of 0.102, which is a reduction of 42.7% compared to the vanilla neural classifier. Comparing this to the version of RL with reward for correct classification which was not trained on scrubbed biographies, we can see that the gender bias was reduced on all metrics. Training the RL algorithm on texts that are scrubbed of gender indicators makes this debiasing strategy comparable to the debiasing methods shown in Table 5.2, in terms of percentage reduction of aggregated TPR gap. It outperforms most methods shown in Table 5.2 in terms of accuracy and aggregated TPR gender gap reduction in percentage compared to their respective non-debiased baselines. The only method that performs better on all metrics is the Mean Projection with FastText embeddings as developed by Haghighatkhah et al. (2022).

Chapter 6

Discussion and Conclusion

The results in Chapter 5 demonstrate that RL can effectively be applied to mitigate gender bias in the task of occupation prediction. The aim of the method developed in this thesis was to investigate how an NLP task, namely occupation prediction, could be formalized in order to apply RL to it. Moreover, I hypothesized that reinforcement learning would be a suitable approach for the mitigation of gender bias in this task, as I could specify desired (debiased) behaviour to the agent. As was seen in Cheng et al. (2021), reinforcement learning was able to reduce bias in hate speech classification, but the question as to whether this would be the case in a specific type of bias and in a different task was investigated in this thesis.

6.1 Discussion of Results

The BiasBios data set was employed to train two different reinforcement learning classifiers, which were compared to a non-RL classifier. The main finding of this thesis is that reinforcement learning reduces the gender bias in a system, compared to the non-RL baseline, on all metrics used to measure bias. It was found that the RL system with a simple reward function for correct classification yielded better results than the system with the minority gender-sensitive reward function. Additionally, it was found that combining a RL system with another debiasing strategy, namely input texts that were scrubbed of gender indicators, resulted in a larger reduction of gender bias across all metrics than the individual methods; it also resulted in a reduction of gender bias that was comparable to the debiasing methods as developed in related work. In this section, the interpretation of the results is elaborated upon.

The expectation that reinforcement learning would be a suitable method for mitigating bias in an NLP task, which was also found in Cheng et al. (2021) was met in this thesis: Reinforcement learning was able to reduce the bias as measured in the TPR gender gap, compared to the non-RL baseline, without harming the performance too much. With developing debiasing methods, there is often a trade-off between the efficiency of a system and its fairness (Bertsimas et al., 2012). This trade-off entails that the accuracy is often lowered when a system is debiased. However, in Cheng et al. (2021), it was found that applying RL as a debiasing method did not harm the accuracy of the model; it increased it. This finding was not echoed in this thesis: The accuracy dropped by about four percent, which is comparable to the drop in accuracy found in Haghighatkhah et al. (2022) and Ravfogel et al. (2020). Thus, reinforcement learning may be seen as a suitable gender debiasing strategy, because it is able to alleviate bias

in a classifier, without harming the accuracy too much. There is the possibility that the RL system can perform better in terms of accuracy, because it is not fully fine-tuned and a more elaborate neural network may be used as the agent. By doing so, the finding that RL was able to reduce bias without harming the performance (and even improving it) in Cheng et al. (2021) may be replicated.

6.1.1 Comparison to Other Debiasing Methods

In Chapter 5, the results of the experiments in this thesis were compared to related studies. It is difficult to accurately compare the results of this thesis to other work, as not all studies make use of the same metrics to evaluate how well a debiasing strategy work. For example Ravfogel et al. (2020) and Haghighatkhah et al. (2022) quantify overall gender bias with the root square TPR gender gap, in order to capture the bias of a classifier on all occupations. In contrast, De-Arteaga et al. (2019) does not employ this metric but rather only displays the gender gaps in terms of a figure, and the correlation with existing gender gaps. Thus, when comparing my results to other studies, it is difficult to gauge whether the system as used in this thesis performs comparably to other methods, since it coincides with the performance on some metrics, but not on others. Moreover, since the initial TPR gender gap of the non-debiased baselines differed between the studies, it was difficult to compare how well my strategy worked in terms of absolute reduction. I attempted to overcome this issue by calculating the reduction in percentages, which makes it relative to the initial bias of the classifier. In addition, the studies that this thesis is compared to make use of different input representations, namely Bag-of-Words, FastText embeddings and BERT embeddings, so the initial difference in TPR gaps may also be due to the variance in bias in the input representations. In order to be able to fully compare how well reinforcement learning mitigates gender bias in the BiasBios data set, compared to similar studies, it should be trained with the same input representation, and the same classifier. By doing so, it reduces the amount of factors that limit the full comparison between the studies.

6.1.2 Combining Debiasing Methods

As was seen in Chapter 5, reinforcement learning was successful in mitigating gender bias in occupation prediction, since the TPR gender gaps were reduced. De-Arteaga et al. (2019) note that as long as there is still some degree of bias present in the classifier, this will be harmful to individuals or groups: A debiased system may still compound gender imbalances when used in a pipeline, though to a lesser extent than a non-debiased system. Following this reasoning, the ultimate goal is to completely remove bias from a system, without harming the performance too much. Since reinforcement learning can be used as a training-based debiasing strategy, as we saw in this thesis, it leaves room for the possibility to combine it with other bias mitigation methods. The additional experiment as explained in Section 5.3 demonstrates that combining an input-based debiasing method with reinforcement learning reduces the aggregated TPR gap to a larger extent than RL alone. I see a source of gain in combining multiple debiasing methods to reduce gender gaps as much as possible, to limit the compounding effect of using a biased system in a pipeline. This observation was also echoed in Cheng et al. (2021), who recommend combining multiple debiasing strategies with reinforcement learning. The finding in Cheng et al. (2021) that using RL as a debiasing strategy did not reduce the accuracy of the classifier, and may lead to the limiting of the

trade-off between efficiency and fairness. The alleviation of gender bias while retaining the accuracy of the non-biased classifier was not found in this thesis, but the reduction in accuracy was fairly small. Combining multiple debiasing methods could lead to a drop in accuracy, but using RL as one of the strategies may limit this drop, as it has the potential to alleviate bias without harming the performance.

6.2 Addressing Research Question and Hypotheses

This thesis is focused around the research question:

Can reinforcement learning be employed for the mitigation of gender bias in the task of occupation prediction?

A second research question was posed in order to answer the main research question, namely:

What is the influence of the reward function on the use of reinforcement learning for the mitigation of gender bias in occupation prediction?

In the following section, these questions will be addressed using the two hypotheses as posed in Section 4.

Hypothesis 1: A deep reinforcement learning classification algorithm will contain less gender bias, measured in true positive rate (TPR) gender gaps, compared to a classification algorithm that does not use reinforcement learning.

With regard to this hypothesis, the results of this thesis demonstrate that the RL-based classifier with a reward for correct classification only, contains less bias than the vanilla classifier. In Figure 5.1, we saw that RL with reward for correct classification was able to reduce the overall gender gaps, a finding also reflected in the aggregated TPR gap (see Table 5.1). It was not the case that both RL systems developed in this thesis contained less bias than the non-RL baseline; the RL with minority gender-sensitive reward can be argued to amplify the bias, and this finding is discussed in more detail in the next paragraph. However, it can be said that the reinforcement learning-based classifier with the reward for correct classification contains less bias than the vanilla classifier, and Hypothesis 1 can thus be accepted.

Hypothesis 2: A reward function based on higher reward for correct classification for minority genders in an occupation will have less gender bias, measured in TPR gender gaps, compared to a reward function based on correct classification only.

The second hypothesis pertains to the influence of the reward system. Before conducting the experiments, the minority-sensitive reward system was thought to result in lower TPR gender gaps than the reward for correct classification. The results of the experiments demonstrate that the opposite is the case: the simple reward system that only rewarded correct classification reduced gender bias to a higher extent than the minority gender-sensitive reward; the latter increased the bias in some occupations. A possible explanation for this may be that the minority-sensitive reward was overly sensitive to the minority class, and thus paid attention to the minority too much, as it received a higher reward for doing so. The reward system was designed as such that

the smaller the minority share, the higher the reward if the classifier predicted the occupation correctly. By doing so, it may have paid too much attention to the minority group which could have increased the bias. This was reflected in Figure 5.3, where it was visible that the RL system with the minority gender-sensitive reward generally underclassified females in male-dominated occupations, and overclassified females in female-dominated occupations. The opposite of this trend was seen in the RL system with reward for correct classification.

In Figure 5.1, it is evident that the RL system with minority gender-sensitive reward increased the bias in almost every occupation, and in the case of *architect, professor* and *painter* became biased towards the other gender.. For *professor*, it is noteworthy to observe that the vanilla classifier and the RL with reward for correct classification contained virtually no bias towards either gender, likely because of the large sample size and the nearly equal gender distribution. As a result of the minority-sensitive reward system, in which the agent receives a higher reward when it correctly classifies a female professor (since this is technically the minority gender with 45.9% female), it will pay more attention to classifying female professors. By doing so, a small bias towards males is created, which is reflected in Figure 5.1.

Regarding Hypothesis 2, it can thus be seen that this hypothesis cannot be accepted: My prior expectation was that the minority gender-sensitive reward system would result in lower TPR gaps than the reward for correct classification, but the opposite was shown to be the case in the results. Thus, we can say that changing the reward in an RL system has great impact on the outcomes and effectiveness of the system as a debiasing method. Which reward system is most useful depends on the use case: If the goal is simply to reduce or eliminate bias in an occupation prediction system, then looking at the TPR gaps indicates that the reward for correct classification performs the best. It could also be the case that you wish to create more visibility for minority groups, in order to change the real gender distribution. For example, in the professions *rapper* and *DJ*, the RL with reward for correct classification overclassifies females and thus shifts the gender distribution to include more females. If the goal is to have representation for female DJs or rappers, then this reward system can be said to be successful at this task. The RL system with minority-gender sensitive reward compounds the gender bias in most occupations, amplifying the stereotypical gender distribution that is found in the gold gender distribution. To answer the main research question, it can be stated that reinforcement learning can be applied as an effective debiasing method in occupation prediction. The reward function should be created and tested with the goal of the use case in mind, as it can greatly influence the outcome of RL as a debiasing method.

6.3 Limitations

Firstly, due to the limited time and computational resources that were available for this thesis, I do not expect that the RL algorithms performed to their full potential in the experiments of this thesis. The classifier used in the RL systems consists of a simple neural network, that is not state-of-the-art. This leaves room for a more elaborate neural network, which may improve the performance of the overall system. A more recent development in reinforcement learning, namely that of Decision Transformer (Chen et al., 2021), also shows promising results that may be used in the development of a more elaborate RL-based debiasing method. In a Decision Transformer, reinforcement learning as a sequence modeling problem is combined with Transformer architectures

such as GPT-x and BERT. This framework could be a solution for the trade-off between efficiency and fairness of an algorithm, as it draws on the advanced architectures of these large language models and may thus lead to higher accuracy of the model. Future work would need to investigate whether this approach is suitable as a debiasing method.

Additionally, the RL algorithms contain many parameters, which I was unable to fully fine-tune within the scope of this thesis. Therefore, it was decided to only tune the learning rate and the number of training steps, though there are many more parameters that may be altered for a more optimal performance. Thus, the system built in this thesis has the potential to obtain higher performance than was shown in the results, possibly when combined with more state-of-the-art models, and when fine-tuned completely.

6.4 Summary and Conclusion

This thesis presents the results of the experiments conducted to investigate the potential of reinforcement learning as a debiasing method for reducing gender gaps in occupation prediction. The study compares different debiasing strategies and evaluates their impact on bias reduction and model performance. The results show that reinforcement learning can effectively reduce gender bias in occupation prediction. The RL system with a reward for correct classification performs better than the system with a minority gender-sensitive reward function. The accuracy of the best-performing debiasing system drops by four percent compared to the non-RL classifier. Comparisons to other debiasing methods reveal that the RL-based approach in this study achieves similar results in bias reduction and accuracy. Combining RL with an input-based debiasing approach, i.e. biographies scrubbed of gender indicators, further improves bias reduction. Overall, the findings suggest that reinforcement learning is a suitable method for mitigating gender bias in occupation prediction tasks.

In conclusion, the research questions of this thesis can be answered as follows: A reinforcement learning algorithm may be applied to the task of occupation prediction which effectively reduces gender gaps, but the reward system plays an important role in the debiasing power and outcome of the method. The findings of this thesis can serve as a starting point for future work, for which recommendations will be discussed in the following section.

6.5 Implications and Future Work

The findings of this thesis contribute to the field of bias mitigation in NLP and highlight the potential of reinforcement learning as a tool for addressing gender gaps in occupation prediction. The implications of this research extend to various applications where reducing bias is crucial, such as fair hiring practices, job recommendation systems, and automated decision-making.

Reinforcement learning proved to be a suitable method for mitigating bias in the NLP task of occupation prediction. It reduced gender bias without significantly harming the accuracy of the model. However, there was still a trade-off between bias reduction and accuracy, as the accuracy dropped slightly compared to the non-debiased baseline. Further exploration can be done to optimize the RL algorithm, by fine-tuning it more, or combining reinforcement learning with a more state-of-the-art classifier. Moreover, to mitigate the finding that the RL systems had low performance on small

classes, a more balanced data set may be used to train the models. Another promising path is that of Decision Transformer (Chen et al., 2021), in which the power of large language models such as BERT may be leveraged in an RL environment. In addition, the debiasing method proposed in this thesis may be combined with other debiasing strategies, to further reduce gender bias. In order to fully compare the results of a RL debiasing strategy to other studies, RL can be trained with the same input representations as used in Haghighatkhah et al. (2022), Ravfogel et al. (2020) and De-Arteaga et al. (2019), namely Bag-of-Words, FastText embeddings and BERT embeddings. By doing so, the only variable that is different is reinforcement learning, so it can be ruled out that the debiasing effect is due to the input representation. The method used in this study may also be used in other tasks, and to mitigate other types of bias, as it shows potential of being a widely applicable debiasing strategy.

Appendix A

Appendix A: Results of Individual Experiments

Table A.1: Classification Report Vanilla Neural Classifier: Seed 2

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.76	0.68	0.72	1413
architect	0.73	0.53	0.61	2528
attorney	0.85	0.86	0.85	8151
chiropractor	0.78	0.32	0.46	649
comedian	0.79	0.60	0.68	699
composer	0.74	0.81	0.77	1400
dentist	0.91	0.93	0.92	3619
dietitian	0.84	0.81	0.82	994
dj	0.55	0.66	0.60	371
filmmaker	0.81	0.72	0.76	1752
interior_designer	0.79	0.43	0.56	365
journalist	0.64	0.76	0.69	4987
model	0.73	0.65	0.69	1875
nurse	0.86	0.82	0.84	4742
painter	0.80	0.72	0.75	1934
paralegal	0.89	0.40	0.55	441
pastor	0.48	0.35	0.40	633
personal_trainer	0.74	0.53	0.62	358
photographer	0.82	0.88	0.85	6081
physician	0.85	0.91	0.88	9641
poet	0.76	0.66	0.71	1752
professor	0.85	0.88	0.86	29527
psychologist	0.73	0.68	0.71	4573
rapper	0.69	0.51	0.59	351
software_engineer	0.60	0.77	0.67	1726
surgeon	0.84	0.60	0.70	3318
teacher	0.52	0.58	0.55	4049
yoga_teacher	0.76	0.60	0.67	415
accuracy			0.79	98344
macro avg	0.75	0.67	0.70	98344
weighted avg	0.80	0.79	0.79	98344

Table A.2: Classification Report Vanilla Neural Classifier: Seed 72

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.77	0.67	0.72	1413
architect	0.68	0.58	0.62	2528
attorney	0.85	0.86	0.85	8151
chiropractor	0.83	0.30	0.45	649
comedian	0.75	0.63	0.69	699
composer	0.78	0.81	0.79	1400
dentist	0.91	0.93	0.92	3619
dietitian	0.87	0.79	0.83	994
dj	0.65	0.54	0.59	371
filmmaker	0.82	0.71	0.76	1752
interior_designer	0.74	0.43	0.54	365
journalist	0.62	0.78	0.69	4987
model	0.74	0.65	0.69	1875
nurse	0.84	0.83	0.84	4742
painter	0.80	0.72	0.75	1934
paralegal	0.87	0.41	0.56	441
pastor	0.65	0.25	0.37	633
personal_trainer	0.79	0.51	0.62	358
photographer	0.81	0.88	0.84	6081
physician	0.84	0.91	0.88	9641
poet	0.75	0.64	0.69	1752
professor	0.85	0.87	0.86	29527
psychologist	0.76	0.67	0.71	4573
rapper	0.64	0.57	0.60	351
software_engineer	0.64	0.73	0.68	1726
surgeon	0.83	0.62	0.71	3318
teacher	0.49	0.59	0.54	4049
yoga_teacher	0.73	0.58	0.65	415
accuracy			0.79	98344
macro avg	0.76	0.66	0.69	98344
weighted avg	0.80	0.79	0.79	98344

Table A.3: Classification Report Vanilla Neural Classifier: Seed 14

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.83	0.66	0.73	1413
architect	0.66	0.59	0.63	2528
attorney	0.84	0.85	0.85	8151
chiropractor	0.83	0.32	0.46	649
comedian	0.70	0.66	0.68	699
composer	0.74	0.82	0.78	1400
dentist	0.92	0.92	0.92	3619
dietitian	0.90	0.78	0.83	994
dj	0.84	0.49	0.62	371
filmmaker	0.83	0.70	0.76	1752
interior_designer	0.80	0.42	0.55	365
journalist	0.57	0.81	0.67	4987
model	0.71	0.66	0.68	1875
nurse	0.88	0.80	0.84	4742
painter	0.75	0.75	0.75	1934
paralegal	0.85	0.42	0.57	441
pastor	0.56	0.36	0.44	633
personal_trainer	0.82	0.56	0.66	358
photographer	0.82	0.87	0.84	6081
physician	0.85	0.91	0.88	9641
poet	0.77	0.62	0.69	1752
professor	0.85	0.87	0.86	29527
psychologist	0.77	0.66	0.71	4573
rapper	0.72	0.51	0.60	351
software_engineer	0.67	0.66	0.67	1726
surgeon	0.79	0.64	0.71	3318
teacher	0.51	0.58	0.55	4049
yoga_teacher	0.73	0.59	0.65	415
accuracy			0.79	98344
macro avg	0.77	0.66	0.70	98344
weighted avg	0.80	0.79	0.79	98344

Table A.4: Classification Report Vanilla Neural Classifier: Seed 1344

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.82	0.64	0.72	1413
architect	0.66	0.60	0.63	2528
attorney	0.82	0.87	0.85	8151
chiropractor	0.82	0.32	0.46	649
comedian	0.76	0.63	0.69	699
composer	0.73	0.82	0.77	1400
dentist	0.91	0.93	0.92	3619
dietitian	0.86	0.80	0.83	994
dj	0.79	0.37	0.51	371
filmmaker	0.80	0.73	0.76	1752
interior_designer	0.83	0.39	0.53	365
journalist	0.63	0.76	0.69	4987
model	0.80	0.62	0.70	1875
nurse	0.84	0.82	0.83	4742
painter	0.83	0.69	0.75	1934
paralegal	0.89	0.42	0.57	441
pastor	0.62	0.32	0.42	633
personal_trainer	0.79	0.52	0.63	358
photographer	0.78	0.90	0.83	6081
physician	0.82	0.92	0.87	9641
poet	0.78	0.63	0.70	1752
professor	0.83	0.89	0.86	29527
psychologist	0.76	0.65	0.70	4573
rapper	0.60	0.63	0.61	351
software_engineer	0.65	0.66	0.65	1726
surgeon	0.86	0.56	0.68	3318
teacher	0.59	0.50	0.54	4049
yoga_teacher	0.75	0.58	0.65	415
accuracy			0.79	98344
macro avg	0.77	0.65	0.69	98344
weighted avg	0.79	0.79	0.79	98344

Table A.5: Classification Report Vanilla Neural Classifier: Seed 50

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.80	0.67	0.73	1413
architect	0.68	0.59	0.63	2528
attorney	0.86	0.85	0.86	8151
chiropractor	0.80	0.32	0.46	649
comedian	0.73	0.63	0.68	699
composer	0.77	0.79	0.78	1400
dentist	0.90	0.93	0.92	3619
dietitian	0.89	0.78	0.83	994
dj	0.63	0.54	0.59	371
filmmaker	0.79	0.73	0.76	1752
interior_designer	0.77	0.39	0.52	365
journalist	0.59	0.80	0.68	4987
model	0.73	0.66	0.69	1875
nurse	0.88	0.81	0.84	4742
painter	0.76	0.75	0.75	1934
paralegal	0.86	0.45	0.59	441
pastor	0.52	0.34	0.41	633
personal_trainer	0.74	0.54	0.62	358
photographer	0.79	0.89	0.84	6081
physician	0.83	0.92	0.87	9641
poet	0.74	0.67	0.70	1752
professor	0.85	0.87	0.86	29527
psychologist	0.79	0.64	0.70	4573
rapper	0.71	0.57	0.63	351
software_engineer	0.65	0.69	0.67	1726
surgeon	0.86	0.58	0.69	3318
teacher	0.53	0.57	0.55	4049
yoga_teacher	0.74	0.64	0.69	415
accuracy			0.79	98344
macro avg	0.76	0.66	0.70	98344
weighted avg	0.80	0.79	0.79	98344

Table A.6: Classification Report RL with Reward for Correct Classification: Seed 2

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.67	0.57	0.62	1413
architect	0.59	0.53	0.56	2528
attorney	0.79	0.84	0.81	8151
chiropractor	0.65	0.28	0.39	649
comedian	0.64	0.51	0.57	699
composer	0.64	0.72	0.68	1400
dentist	0.89	0.92	0.90	3619
dietitian	0.83	0.73	0.78	994
dj	0.08	0.02	0.03	371
filmmaker	0.62	0.70	0.65	1752
interior_designer	0.00	0.00	0.00	365
journalist	0.57	0.73	0.64	4987
model	0.68	0.57	0.62	1875
nurse	0.83	0.78	0.80	4742
painter	0.64	0.70	0.67	1934
paralegal	0.28	0.02	0.04	441
pastor	0.15	0.04	0.06	633
personal_trainer	0.28	0.04	0.07	358
photographer	0.78	0.85	0.81	6081
physician	0.83	0.90	0.87	9641
poet	0.59	0.61	0.60	1752
professor	0.82	0.88	0.85	29527
psychologist	0.73	0.61	0.67	4573
rapper	0.09	0.01	0.02	351
software_engineer	0.58	0.61	0.59	1726
surgeon	0.76	0.56	0.65	3318
teacher	0.54	0.42	0.48	4049
yoga_teacher	0.48	0.50	0.49	415
accuracy			0.75	98344
macro avg	0.57	0.52	0.53	98344
weighted avg	0.74	0.75	0.74	98344

Table A.7: Classification Report RL with Reward for Correct Classification: Seed 72

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.61	0.61	0.61	1413
architect	0.59	0.51	0.55	2528
attorney	0.76	0.86	0.81	8151
chiropractor	0.66	0.28	0.39	649
comedian	0.51	0.58	0.54	699
composer	0.63	0.75	0.68	1400
dentist	0.90	0.92	0.91	3619
dietitian	0.78	0.74	0.76	994
dj	0.19	0.01	0.02	371
filmmaker	0.63	0.71	0.67	1752
interior_designer	0.25	0.00	0.01	365
journalist	0.57	0.74	0.64	4987
model	0.72	0.56	0.63	1875
nurse	0.83	0.79	0.81	4742
painter	0.65	0.71	0.68	1934
paralegal	0.39	0.10	0.16	441
pastor	0.24	0.01	0.02	633
personal_trainer	0.33	0.06	0.10	358
photographer	0.81	0.84	0.82	6081
physician	0.83	0.90	0.86	9641
poet	0.60	0.56	0.58	1752
professor	0.82	0.87	0.85	29527
psychologist	0.71	0.62	0.66	4573
rapper	0.12	0.00	0.01	351
software_engineer	0.56	0.62	0.59	1726
surgeon	0.79	0.54	0.64	3318
teacher	0.52	0.42	0.47	4049
yoga_teacher	0.59	0.52	0.55	415
accuracy			0.75	98344
macro avg	0.59	0.53	0.54	98344
weighted avg	0.74	0.75	0.74	98344

Table A.8: Classification Repor RL with Reward for Correct Classification: Seed 14

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.69	0.59	0.64	1413
architect	0.59	0.51	0.55	2528
attorney	0.81	0.84	0.82	8151
chiropractor	0.51	0.31	0.38	649
comedian	0.48	0.53	0.50	699
composer	0.69	0.72	0.71	1400
dentist	0.89	0.92	0.91	3619
dietitian	0.76	0.73	0.75	994
dj	0.22	0.01	0.01	371
filmmaker	0.66	0.70	0.67	1752
interior_designer	0.07	0.01	0.01	365
journalist	0.58	0.71	0.64	4987
model	0.69	0.59	0.64	1875
nurse	0.87	0.78	0.82	4742
painter	0.64	0.71	0.68	1934
paralegal	0.40	0.04	0.07	441
pastor	0.06	0.00	0.01	633
personal_trainer	0.09	0.01	0.01	358
photographer	0.77	0.85	0.81	6081
physician	0.83	0.90	0.87	9641
poet	0.55	0.60	0.57	1752
professor	0.82	0.88	0.85	29527
psychologist	0.72	0.60	0.66	4573
rapper	0.11	0.01	0.02	351
software_engineer	0.53	0.63	0.58	1726
surgeon	0.79	0.57	0.66	3318
teacher	0.51	0.44	0.47	4049
yoga_teacher	0.40	0.60	0.48	415
accuracy			0.75	98344
macro avg	0.56	0.53	0.53	98344
weighted avg	0.74	0.75	0.74	98344

Table A.9: Classification Report RL with Reward for Correct Classification: Seed 1344

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.69	0.59	0.63	1413
architect	0.60	0.50	0.55	2528
attorney	0.79	0.85	0.82	8151
chiropractor	0.66	0.28	0.39	649
comedian	0.38	0.38	0.38	699
composer	0.65	0.74	0.69	1400
dentist	0.90	0.90	0.90	3619
dietitian	0.66	0.76	0.70	994
dj	0.08	0.01	0.01	371
filmmaker	0.58	0.71	0.64	1752
interior_designer	0.12	0.01	0.01	365
journalist	0.59	0.72	0.65	4987
model	0.67	0.58	0.62	1875
nurse	0.85	0.79	0.82	4742
painter	0.68	0.70	0.69	1934
paralegal	0.64	0.20	0.30	441
pastor	0.17	0.05	0.08	633
personal_trainer	0.06	0.01	0.01	358
photographer	0.75	0.86	0.80	6081
physician	0.83	0.90	0.86	9641
poet	0.62	0.60	0.61	1752
professor	0.82	0.87	0.85	29527
psychologist	0.73	0.60	0.66	4573
rapper	0.03	0.00	0.01	351
software_engineer	0.59	0.61	0.60	1726
surgeon	0.76	0.57	0.66	3318
teacher	0.52	0.44	0.47	4049
yoga_teacher	0.44	0.59	0.51	415
accuracy			0.75	98344
macro avg	0.57	0.53	0.53	98344
weighted avg	0.74	0.75	0.74	98344

Table A.10: Classification Report RL with Reward for Correct Classification: Seed 50

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.69	0.59	0.64	1413
architect	0.61	0.50	0.55	2528
attorney	0.80	0.84	0.82	8151
chiropractor	0.70	0.29	0.41	649
comedian	0.39	0.43	0.41	699
composer	0.66	0.73	0.70	1400
dentist	0.89	0.91	0.90	3619
dietitian	0.80	0.75	0.78	994
dj	0.06	0.00	0.01	371
filmmaker	0.68	0.68	0.68	1752
interior_designer	0.00	0.00	0.00	365
journalist	0.59	0.71	0.64	4987
model	0.78	0.50	0.61	1875
nurse	0.85	0.79	0.82	4742
painter	0.61	0.71	0.66	1934
paralegal	0.50	0.27	0.35	441
pastor	0.23	0.08	0.12	633
personal_trainer	0.21	0.01	0.02	358
photographer	0.77	0.85	0.81	6081
physician	0.81	0.91	0.85	9641
poet	0.56	0.60	0.58	1752
professor	0.82	0.87	0.85	29527
psychologist	0.71	0.62	0.66	4573
rapper	0.10	0.01	0.01	351
software_engineer	0.57	0.63	0.60	1726
surgeon	0.80	0.54	0.65	3318
teacher	0.52	0.44	0.47	4049
yoga_teacher	0.52	0.53	0.53	415
accuracy			0.75	98344
macro avg	0.58	0.53	0.54	98344
weighted avg	0.74	0.75	0.74	98344

Table A.11: Classification Report RL with Minority Gender-Sensitive Reward: Seed 2

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.69	0.57	0.62	1413
architect	0.57	0.51	0.54	2528
attorney	0.79	0.84	0.81	8151
chiropractor	0.55	0.28	0.37	649
comedian	0.49	0.53	0.51	699
composer	0.61	0.71	0.66	1400
dentist	0.88	0.91	0.90	3619
dietitian	0.75	0.74	0.74	994
dj	0.09	0.01	0.01	371
filmmaker	0.67	0.67	0.67	1752
interior_designer	0.17	0.02	0.03	365
journalist	0.57	0.72	0.64	4987
model	0.72	0.55	0.62	1875
nurse	0.84	0.79	0.81	4742
painter	0.67	0.68	0.67	1934
paralegal	0.51	0.28	0.36	441
pastor	0.17	0.07	0.10	633
personal_trainer	0.13	0.01	0.02	358
photographer	0.77	0.85	0.81	6081
physician	0.82	0.90	0.86	9641
poet	0.62	0.57	0.59	1752
professor	0.82	0.87	0.84	29527
psychologist	0.72	0.61	0.66	4573
rapper	0.04	0.00	0.01	351
software_engineer	0.57	0.60	0.59	1726
surgeon	0.80	0.56	0.66	3318
teacher	0.49	0.44	0.46	4049
yoga_teacher	0.63	0.49	0.55	415
accuracy			0.75	98344
macro avg	0.58	0.53	0.54	98344
weighted avg	0.74	0.75	0.74	98344

Table A.12: Classification Report RL with Minority Gender-Sensitive Reward: Seed 72

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.72	0.56	0.63	1413
architect	0.60	0.50	0.54	2528
attorney	0.80	0.84	0.82	8151
chiropractor	0.58	0.28	0.37	649
comedian	0.55	0.49	0.52	699
composer	0.63	0.72	0.67	1400
dentist	0.90	0.90	0.90	3619
dietitian	0.81	0.71	0.76	994
dj	0.19	0.02	0.03	371
filmmaker	0.66	0.69	0.68	1752
interior_designer	0.14	0.02	0.03	365
journalist	0.58	0.71	0.64	4987
model	0.73	0.50	0.60	1875
nurse	0.79	0.80	0.80	4742
painter	0.68	0.65	0.67	1934
paralegal	0.54	0.27	0.36	441
pastor	0.25	0.16	0.19	633
personal_trainer	0.10	0.02	0.03	358
photographer	0.74	0.87	0.80	6081
physician	0.82	0.91	0.86	9641
poet	0.61	0.57	0.59	1752
professor	0.82	0.87	0.84	29527
psychologist	0.73	0.59	0.66	4573
rapper	0.09	0.02	0.03	351
software_engineer	0.56	0.60	0.58	1726
surgeon	0.78	0.58	0.66	3318
teacher	0.49	0.46	0.48	4049
yoga_teacher	0.57	0.48	0.52	415
accuracy			0.75	98344
macro avg	0.59	0.53	0.55	98344
weighted avg	0.74	0.75	0.74	98344

Table A.13: Classification Report RL with Minority Gender-Sensitive Reward: Seed 14

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.65	0.56	0.60	1413
architect	0.58	0.52	0.55	2528
attorney	0.78	0.84	0.81	8151
chiropractor	0.58	0.27	0.37	649
comedian	0.50	0.53	0.52	699
composer	0.65	0.71	0.68	1400
dentist	0.89	0.91	0.90	3619
dietitian	0.74	0.74	0.74	994
dj	0.04	0.00	0.01	371
filmmaker	0.68	0.68	0.68	1752
interior_designer	0.28	0.19	0.23	365
journalist	0.58	0.71	0.64	4987
model	0.73	0.55	0.63	1875
nurse	0.90	0.77	0.83	4742
painter	0.66	0.66	0.66	1934
paralegal	0.48	0.32	0.38	441
pastor	0.17	0.08	0.11	633
personal_trainer	0.06	0.01	0.01	358
photographer	0.76	0.85	0.80	6081
physician	0.85	0.90	0.87	9641
poet	0.56	0.59	0.57	1752
professor	0.82	0.88	0.85	29527
psychologist	0.74	0.59	0.66	4573
rapper	0.05	0.01	0.01	351
software_engineer	0.56	0.59	0.57	1726
surgeon	0.78	0.59	0.67	3318
teacher	0.51	0.47	0.49	4049
yoga_teacher	0.69	0.42	0.52	415
accuracy			0.75	98344
macro avg	0.58	0.53	0.55	98344
weighted avg	0.74	0.75	0.74	98344

Table A.14: Classification Report RL with Minority Gender-Sensitive Reward: Seed 1344

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.72	0.56	0.63	1413
architect	0.53	0.53	0.53	2528
attorney	0.78	0.84	0.81	8151
chiropractor	0.61	0.27	0.37	649
comedian	0.48	0.55	0.51	699
composer	0.59	0.75	0.66	1400
dentist	0.90	0.90	0.90	3619
dietitian	0.79	0.71	0.74	994
dj	0.08	0.00	0.01	371
filmmaker	0.67	0.67	0.67	1752
interior_designer	0.26	0.10	0.14	365
journalist	0.59	0.70	0.64	4987
model	0.70	0.56	0.62	1875
nurse	0.83	0.79	0.81	4742
painter	0.62	0.70	0.66	1934
paralegal	0.48	0.17	0.25	441
pastor	0.18	0.07	0.10	633
personal_trainer	0.13	0.04	0.06	358
photographer	0.77	0.84	0.80	6081
physician	0.82	0.90	0.86	9641
poet	0.58	0.56	0.57	1752
professor	0.83	0.87	0.85	29527
psychologist	0.72	0.61	0.66	4573
rapper	0.10	0.01	0.02	351
software_engineer	0.55	0.63	0.59	1726
surgeon	0.79	0.57	0.66	3318
teacher	0.51	0.45	0.48	4049
yoga_teacher	0.52	0.52	0.52	415
accuracy			0.75	98344
macro avg	0.58	0.53	0.54	98344
weighted avg	0.74	0.75	0.74	98344

Table A.15: Classification Report RL with Minority Gender-Sensitive Reward: Seed 50

Class	**Precision**	**Recall**	**F1-score**	**Support**
accountant	0.62	0.59	0.60	1413
architect	0.56	0.53	0.55	2528
attorney	0.80	0.83	0.81	8151
chiropractor	0.60	0.29	0.39	649
comedian	0.60	0.54	0.57	699
composer	0.60	0.75	0.67	1400
dentist	0.88	0.92	0.90	3619
dietitian	0.73	0.71	0.72	994
dj	0.09	0.01	0.02	371
filmmaker	0.63	0.69	0.66	1752
interior_designer	0.13	0.01	0.02	365
journalist	0.61	0.69	0.64	4987
model	0.65	0.56	0.60	1875
nurse	0.82	0.80	0.81	4742
painter	0.64	0.65	0.65	1934
paralegal	0.51	0.21	0.30	441
pastor	0.16	0.11	0.13	633
personal_trainer	0.19	0.04	0.07	358
photographer	0.76	0.86	0.80	6081
physician	0.82	0.90	0.86	9641
poet	0.63	0.52	0.57	1752
professor	0.82	0.87	0.85	29527
psychologist	0.74	0.60	0.66	4573
rapper	0.10	0.02	0.03	351
software_engineer	0.56	0.57	0.57	1726
surgeon	0.77	0.54	0.64	3318
teacher	0.49	0.45	0.47	4049
yoga_teacher	0.56	0.50	0.53	415
accuracy			0.75	98344
macro avg	0.57	0.53	0.54	98344
weighted avg	0.74	0.75	0.74	98344

Bibliography

D. Bertsimas, V. F. Farias, and N. Trichakis. On the efficiency-fairness trade-off. *Management Science*, 58(12):2234–2250, 2012.

R. Bhardwaj, N. Majumder, and S. Poria. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018, 2021.

S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Y. T. Cao and H. Daumé III. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle. *Computational Linguistics*, 47(3):615–661, 2021.

L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

L. Cheng, A. Mosallanezhad, Y. Silva, D. Hall, and H. Liu. Mitigating bias in session-based cyberbullying detection: A non-compromising approach. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2158–2168, 2021.

K. Crawford. The trouble with bias. nips 2017 keynote, 2017.

M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

S. Dev, M. Monajatipoor, A. Ovalle, A. Subramonian, J. M. Phillips, and K.-W. Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084*, 2021.

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

P. Haghighatkhah, A. Fokkens, P. Sommerauer, B. Speckmann, and K. Verbeek. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection. *arXiv preprint arXiv:2212.04273*, 2022.

K. Hamberg. Gender bias in medicine. *Women's health*, 4(3):237–243, 2008.

L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018.

C. Isaac, B. Lee, and M. Carnes. Interventions that affect gender bias in hiring: A systematic review. *Academic medicine: journal of the Association of American Medical Colleges*, 84(10):1440, 2009.

S. Kiritchenko and S. M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

E. Lin, Q. Chen, and X. Qi. Deep reinforcement learning for imbalanced classification. *Applied Intelligence*, 50:2488–2502, 2020.

K. Lloyd. Bias amplification in artificial intelligence systems. *arXiv preprint arXiv:1809.07842*, 2018.

K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202, 2020.

N. Madaan, S. Mehta, T. Agrawaal, V. Malhotra, A. Aggarwal, Y. Gupta, and M. Saxena. Analyze, detect and remove gender stereotyping from bollywood movies. In *Conference on fairness, accountability and transparency*, pages 92–105. PMLR, 2018.

E. Matsuno and S. L. Budge. Non-binary/genderqueer identities: A critical review of the literature. *Current Sexual Health Reports*, 9(3):116–120, 2017.

C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*, 2019.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479, 2012.

A. Nelson. Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the national medical association*, 94(8):666, 2002.

M. Nissim, R. van Noord, and R. van der Goot. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497, 2020.

J. H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

M. O. Prates, P. H. Avelar, and L. C. Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32 (10):6363–6381, 2020.

S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL `https://aclanthology.org/2020.acl-main.647`.

A. Romanov, M. De-Arteaga, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. Rumshisky, and A. Kalai. What's in a name? Reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1424. URL `https://aclanthology.org/N19-1424`.

R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*, 2018.

D. E. Rupp, S. J. Vodanovich, and M. Crede. Age bias in the workplace: The impact of ageism and causal attributions 1. *Journal of Applied Social Psychology*, 36(6): 1337–1364, 2006.

G. Stanovsky, N. A. Smith, and L. Zettlemoyer. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*, 2019.

T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

R. Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59, 2017.

H. R. Tenenbaum and M. D. Ruck. Are teachers' expectations different for racial minority than for european american students? a meta-analysis. *Journal of educational psychology*, 99(2):253, 2007.

V. Uc-Cetina, N. Navarro-Guerrero, A. Martin-Gonzalez, C. Weber, and S. Wermter. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, pages 1–33, 2022.

E. Vanmassenhove, C. Hardmeier, and A. Way. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*, 2019.

M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018a.

J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, Oct.-Nov. 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL `https://aclanthology.org/D18-1521`.