Social Biases in Language Models
Gender Stereotypes in GPT-3 Generated Stories

Beatrice Lorentzen
Abstract

The Generative Pre-Trained Transformer 3 (GPT-3) is a language prediction model developed by OpenAI, which can interpret and can generate human language and code. The aim of this study was to assess whether GPT-3 reproduce gender biases in generated short stories. 900 stories were generated using GPT-3’s API with the engine “Davinci”. Gender stereotypes about female, male, and gender-neutral characters were looked into using lists of gender connoted traits, professions, and hobbies, as well as physical features. The results indicate that GPT-3 reproduces some gender biases that are seemingly benign in generated short stories.
Populärvetenskaplig sammanfattning

Generative Pre-Trained Transformer 3 (GPT-3) är en språkprediktionsmodell utvecklad av OpenAI som kan utföra flera uppgifter inom språkteknologi, exempelvis sammanfattningar. Syftet med denna studie var att bedöma om GPT-3 återger könsfordomar i korta genererade berättelser. 900 berättelser genererades med hjälp av GPT-3:s API med verktyget Davinci. Könsstereotyper om kvinnliga, manliga och könsneutrala karaktärer undersöktes med hjälp av listor innehållande könsbetonade egenskaper, yrken och hobbyer, såväl som fysiska attribut. Resultaten indikerar att GPT-3 reproducerar vissa könsfordomar i genererade berättelser.
Contents

List of Figures ........................................................................................................ iv

Part I: Introduction ................................................................................................ 1
  1 Background ...................................................................................................... 2
  2 Scope and Research Problem ........................................................................... 3
    2.1 Main Problem ............................................................................................. 3
    2.2 Sub-Problem .............................................................................................. 3
    2.3 Assumptions and Limitations ..................................................................... 4
  3 Methods ........................................................................................................... 5
  4 Major Findings ................................................................................................. 6
  5 Outline ............................................................................................................ 7

Part II: Theoretical Framework ......................................................................... 8
  6 Machine Learning ............................................................................................ 9
    6.1 Model Training .......................................................................................... 9
    6.2 Deep Learning ......................................................................................... 9
    6.3 Artificial Neural Networks ....................................................................... 9
    6.4 Natural Language Processing .................................................................. 10
    6.5 Word Embedding .................................................................................... 10
    6.6 Transformer .............................................................................................. 11
      6.6.1 Architecture ....................................................................................... 11
  7 Language Models ............................................................................................ 14
    7.1 Generative Pre-Trained Transformer 3 ..................................................... 14
      7.1.1 Engines .............................................................................................. 14
  8 Bias in Artificial Intelligence .......................................................................... 15
    8.1 Causes of Bias ......................................................................................... 15
    8.2 Measuring Bias ....................................................................................... 15
    8.3 Gender Bias ............................................................................................. 16
  9 State-of-the-Art ............................................................................................... 18
    9.1 Gender and Representation Bias in Stories ............................................. 18
    9.2 Fairness, Bias, and Representation ......................................................... 18
    9.3 Anti-Muslim Bias ..................................................................................... 19

Part III: Methods ............................................................................................... 20
  10 Literature Review .......................................................................................... 21
    10.1 Data Sources and Search Strategy ......................................................... 21
    10.2 Inclusion and Exclusion Criteria ......................................................... 21
    10.3 Process .................................................................................................... 21
  11 Implementation ............................................................................................... 23
    11.1 Generated Stories .................................................................................... 23
B.2 Traits ............................................................................................................................... 51
B.3 Professions .................................................................................................................... 52
B.4 Hobbies ........................................................................................................................ 52
B.5 Physical Features ........................................................................................................ 52

References .......................................................................................................................... 56
List of Figures

Figure 1.1: An example of gender stereotypes reproduced in Google Translate. .......................... 2
Figure 6.1: The structure of a neural network [6]. ................................................................. 10
Figure 6.2: Visualization of word embeddings [16]. ............................................................... 11
Figure 6.3: The Transformer model architecture [48]. ............................................................. 12
Figure 10.1: PRISMA flow diagram. ..................................................................................... 22
Figure 11.1: Illustration of the implementation flow. ............................................................... 23
Figure 11.2: Sentence patterns discovered by *spaCy*. ............................................................ 26
Figure 13.1: The results of the main characters’ traits in GPT-3 generated stories. ............... 30
Figure 13.2: Frequency of most used male connoted traits in GPT-3 generated stories. ........... 31
Figure 13.3: Frequency of most used female connoted traits in GPT-3 generated stories. ....... 31
Figure 14.1: The results of the main characters’ professions in GPT-3 generated stories. ....... 32
Figure 14.2: Frequency of main characters’ female-dominated professions in GPT-3 generated stories. ............................................................................................................. 33
Figure 14.3: Frequency of main characters’ male-dominated professions in GPT-3 generated stories. ............................................................................................................. 33
Figure 15.1: The results of mentioned hobbies in GPT-3 generated stories. ......................... 34
Figure 15.2: The results of hard hobbies in GPT-3 generated stories. ..................................... 35
Figure 15.3: The results of soft hobbies in GPT-3 generated stories. ....................................... 35
Figure 16.1: The results of mentioned body parts for female and male main characters in GPT-3 generated stories. ....................................................................................... 36
Figure 16.2: The adjectives used describing main characters’ hair. ......................................... 37
Figure 16.3: The adjectives used describing main characters’ hair. ......................................... 37
Figure 16.4: The adjectives used describing main characters’ eyes. ........................................ 37
Figure 16.5: The adjectives used describing main characters’ smile. ..................................... 38
Part I:
Introduction

This part serves as an introduction to the thesis containing the background of the study, a description of the scope and problem, and the limitations and assumptions made. An introduction to the methods used during the study and a summary of the significant findings are presented. This part also outlines the structure of the rest of this thesis.
1. Background

Today’s language models can be useful for a variety of problems in computational linguistic. They can summarize, parse, classify, and translate texts, among other things. The cornerstone of these algorithms is probabilities created by observations of a large number of texts generated by humans.

The third generation Generative Pre-Trained Transformer 3 (GPT-3), developed by OpenAI, is a language prediction model trained to generate human-like text [4]. As of early 2021, GPT-3 was the largest neural network ever produced with over 175 billion machine learning parameters [4]. By using a neural network machine learning model, GPT-3 can use text as input and transform it into a result predicted to be the most useful. The model can be used to create realistic human texts and is used today by companies to, for instance, get summaries of insights from customer feedback, create interactive stories, and do semantic searches [37].

Due to social biases existing in humans and in texts that the models are trained on, the language models that do not actively address these biases are more likely to reproduce them. Social biases reproduced by language models can lead to unfair and harmful or useless outputs.

There have been several cases of language models giving gender biased outputs. For instance, when Google Translate translates a gender-neutral language like Finnish to English, it gives biased results based on gender stereotypes (see Fig 1.1).

![Figure 1.1](image.png)

Figure 1.1. An example of gender stereotypes reproduced in Google Translate.

Another example is when Amazon Inc had to dispose their recruiting algorithm in 2018 after it was evident that it identified word patterns, rather than skill sets, inadvertently penalizing résumés containing certain words and as a consequence favored male candidates over female candidates by discounting women’s résumés [11].
2. Scope and Research Problem

This study aimed to assess whether gender bias in the form of stereotypes is reproduced in GPT-3 through the generation of short stories. The reproduction of gender bias in GPT-3 generated stories will be quantified, and the technical reasons for its existence and impact on society will be discussed. Finally, how the reproduction of social biases, like gender bias, can be avoided will be addressed.

2.1 Main Problem

Owing to the aim of this study, the main problem is defined as follows:

<table>
<thead>
<tr>
<th>Main Research Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do GPT-3 generated short stories contain gender biases in the form of stereotypes?</td>
</tr>
</tbody>
</table>

2.2 Sub-Problems

In order to solve the main research problem of this study, the following sub-problems will be investigated. First, addressing how the main characters’ personalities are described. The personal traits are divided into female and male connoted and are categorized by GPT-3. The first sub-problem is stated below:

<table>
<thead>
<tr>
<th>Sub-Problem 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>To what extent are female, male, and gender-neutral main characters described using female versus male connoted traits in GPT-3 generated stories?</td>
</tr>
</tbody>
</table>

The second sub-problem addresses the physical appearances of the main characters. The physical features are described using body parts and adjectives. The second sub-problem is stated as follows.

<table>
<thead>
<tr>
<th>Sub-Problem 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>To what extent are female, male, and gender-neutral main characters described using their physical features?</td>
</tr>
</tbody>
</table>

The third sub-problem addresses the main characters’ professions. The professions were categorized as female- or male-dominated. The third sub-problem reads as follows.

<table>
<thead>
<tr>
<th>Sub-Problem 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>To what extent are female, male, and gender-neutral main characters occupied with female-versus male-dominated professions in GPT-3 generated stories?</td>
</tr>
</tbody>
</table>
Finally, the fourth sub-problem addresses the main characters’ hobbies. To find out what type of hobbies the main characters like, the hobbies were separated into *hard* (physically demanding or challenging) and *soft* (for the purpose of pleasure of relaxation).

### Sub-Problem 4

| To what extent are female, male, and gender-neutral main characters performing soft versus hard hobbies in GPT-3 generated stories? |

---

### 2.3 Assumptions and Limitations

This section addresses the assumptions made prior to or during the study and the limitations of the study are brought up. How the results were affected by the assumptions and limitations is discussed in Part V.

Names do not necessarily have universal gender connotations. Countries, cultures, and social norms have over time associated particular sexes and/or gender identities with names. For instance, Andrea is an Italian male name but is commonly understood as a female name in other countries, such as Sweden, Spain, Germany, and England. Nevertheless, in this study the assumption that names are associated with a gender is made to establish whether GPT-3 reproduces any gender stereotypes based on traditional female, male, and gender-neutral names in the United States.

Furthermore, the assumption that GPT-3 interprets gender-neutral names as gender-neutral characters is made. However, GPT-3 can conceivably associate genders to gender-neutral names, which will affect the results.

The content of the categories are limited to the traits, professions, and hobbies that GPT-3 suggested for each category. There are most likely more components that should have qualified in those lists to make them complete.

Additionally, the lists are limited to the form of their contents. That is, for example, ”humorous” can also be described as ”making people laugh”, which is not taken into account. There are a lot of different ways words and sentences can be structured and all could not be taken into consideration. Hence, it is viewed as a limitation and should be considered when reading the results.
3. Methods

A literature review was conducted to find papers on gender bias in GPT-3. Two data sources were used, ACM (Association for Computer Machinery) Digital Library and IEEE (Institute of Electrical and Electronics Engineers) Xplore. The literature review was done according to the PRISMA approach and resulted in 55 results which were screened. Finally, the literature review and recommendations received from this thesis’ supervisor resulted in three papers.

In total, 900 stories were generated using GPT-3’s API from OpenAI with the engine Davinci. The stories were about the main characters, their traits, professions, hobbies, and physical features. To assess whether the stories contain gender stereotypes, the traits, professions, and hobbies were divided into lists representing feminine and masculine stereotypes. Traits were divided into female and male connoted, professions into female- and male-dominated, and hobbies into hard and soft. Body parts were listed and the adjectives describing the body parts were found using an open-source library for Natural Language Processing (NLP) in Python called spaCy.
4. Major Findings

The results of this study indicate that GPT-3 contains gender bias seeing that the output reproduced gender stereotypes in the generated short stories. Female characters were mostly described using female connoted traits. However, male characters and gender-neutral characters were also described mostly using female connoted traits. Male characters were described using male connoted traits more frequently than female and gender-neutral characters. Female characters were almost exclusively given female-dominated professions and were associated more often than male characters with professions requiring no higher level of education, such as stay-at-home mom, whereas male characters were mostly given male-dominated professions. Hobbies that are physically demanding or challenging were heavily male leaning, while hobbies for the purpose of pleasure or relaxation were heavily female leaning. Female characters’ physical appearances were described more frequently compared to male characters. Additionally, females were more associated with appearance associated words, such as "pretty", "sweet", and "attractive".
5. Outline

The remaining parts of this thesis are structured as follows: Part II presents the Theoretical Framework, which includes descriptions of topics that serve as the underlying basis and principles of this thesis, such as machine learning, including model training, neural networks, deep learning, transformer neural networks, and word embeddings; language models and more specifically GPT-3; and the causes of bias in artificial intelligence, how to measure it, and what gender bias is.

Then, Part III presents the methods used to conduct the literature review and the implementation to address the problem statements of this thesis.

Afterwards, Part IV presents the results of the study and Part V discusses the results presented in Part IV, critique to the methods used, and discusses the results from a technical and societal perspective.

Finally, the concluding Part VI closes the thesis by adding several remarks concerning its coverage.
Part II:
Theoretical Framework

This part supports the theory of the thesis. The subjects addressed in this part serves as the basis of this study. Topics such as machine learning, language models, and bias in AI are covered.
6. Machine Learning

This chapter presents a definition of machine learning, including an overview of some relevant concepts to create a common understanding for the rest of this thesis.

*Machine learning* is defined by Mohri et al. as "computational methods using experience to improve performance or to make accurate predictions" [30]. The purpose of machine learning is to design efficient Artificial Intelligence (AI) algorithms that can make accurate predictions and classifications based on input data [30]. The focus is learning, that is, acquiring skills or knowledge from previous experiences through data [30].

6.1 Model Training

*Modeling* is the process of creating an AI model, which makes decisions based on its interpretation of the data [9]. When training a machine learning model, the machine learning algorithm takes training data as input and trains itself to recognize the patterns in the training data [30]. The aim is that the algorithm can predict the output for a new set of similar data that it has not been trained on [45]. The success of a learning algorithm is dependent on the data used [30]. There are several learning techniques. *Supervised learning* is when the learning algorithm uses labeled data, iteratively makes predictions, and adjusts for the correct answer for all unseen points by learning from the training data set [30]. The predicted output from the training data is compared to the actual output [45]. Based on its accuracy, it can optimize itself to fit better to the data points in the training set [45]. By using labeled inputs and outputs, the algorithm can learn over time and measure its accuracy [30].

*Unsupervised learning* clusters and analyzes data sets by using unlabeled training data and making predictions [30]. The technique is called "unsupervised" because the algorithm discovers patterns in the data sets without human intervention [30]. The goal of unsupervised learning is to get insights from large volumes of new data [30].

6.2 Deep Learning

*Deep learning* is a subset of machine learning that attempts to mimic the human brain by accurately clustering data and making predictions [8]. It is a neural network with at least three layers that learns from a large amount of data [8]. More layers can optimize and refine the accuracy of the algorithm [8]. In contrast to machine learning, deep learning is not dependent on humans and does not pre-process the data to the same extent as machine learning does [8]. This means that deep learning algorithms can process unstructured data, like text, to adjust and fit itself for accuracy and increase its precision [8].

6.3 Artificial Neural Networks

*Artificial neural networks*, also known as *neural networks*, allow computer programs to recognize patterns and do problem solving within AI, machine learning, and deep learning [6]. Data is
passed through the different layers of a neural network [6]. As seen in Fig 6.1, a neural network contains one input layer, one or more hidden layers, and one output layer [6]. There are different types of hidden layers that perform different kinds of transformations on their inputs and can be better suited for specific tasks [6]. The connections between the layers are assigned a weight [6]. The weighted sum is computed and passed to an activation function that transforms the number and passes it on to the next neuron in the next layer [6]. This process is repeated until the result reaches the output layer and the weight is optimized by the optimization model as it is learning [6].

![Deep neural network](image)

*Figure 6.1. The structure of a neural network [6].*

6.4 Natural Language Processing

Neural networks are the foundation of Natural Language Processing (NLP), which is a branch in AI that gives computers the ability to understand and respond to text and voice data in a human-like way [7]. NLP is a combination of computational linguistics and statistical, machine learning, and deep learning models [7]. Using these technologies, computers are able to process human language in text and voice data and understand the meaning of it [7]. Human language can contain ambiguities that make it difficult to create software that accurately determines the meaning of the text or voice data [7].

6.5 Word Embedding

Word embedding is a term used in NLP referring to the representation of words in the form of real-valued vectors in a predefined vector space that encodes the meaning of words [21]. Each word is mapped to one vector and associated with a point in a vector space [3]. The closer the points are in the vector space, the more similar the words’ semantic or contextual meanings are [21]. The distributed representation is learned based on the usage of words, which makes it possible to capture the intended meaning of a word that has several representations and ways to be used [3].

---

1The scientific and engineering discipline concerned with understanding human language from a computational perspective [43].
As seen in the simplified example in Fig 6.2, each word is mapped to a word vector. The numbers in the word vectors constitute the words’ distributed weights across dimensions [21]. The words’ numerical weights on that dimension capture the closeness of their associations with and to that meaning [21]. The points for the words “cat” and “kitten” are closer in the vector space due to being more similar since their numbers in the words vectors have lower differentials (see Fig 6.2).

![Figure 6.2. Visualization of word embeddings [16].](image)

### 6.6 Transformer

The *Transformer* model was first introduced in the paper ”Attention Is All You Need” by Vaswani et al. in 2017 [48]. The deep learning model was the first sequence transduction model\(^2\) entirely based on attention mechanisms [48].

Previous to the Transformer, Recurrent Neural Network (RNN) was commonly used in NLP systems [36]. Both models are designed to handle sequential input data, such as natural language [36][48]. The Transformer model is built on the technology and structure of RNNs but with added attention mechanisms [48]. Unlike RNNs, Transformers use attention mechanisms that provide context for any position in the input sequence and do not need to process data in order [36][48]. This allows the input sequence of a Transformer encoder to be passed in parallel [48].

#### 6.6.1 Architecture

The Transformer model has an encoder-decoder structure, which consists of two components, an *encoder*, which takes a variable-length sequence as the input and transforms it into a fixed shape state; and a *decoder*, which maps the encoded state to a variable-length sequence [48]. This type of structure is commonly found in neural sequence transduction models [48]. In Transformers, the encoder maps an input sequence to a sequence of continuous representations [48]. Using the result of the mapping, the decoder then generates an output sequence [48]. The model is *autoregressive*, meaning that it predicts the next symbol based on the previous value [48]. As seen in Fig 6.3, the Transformer model contains an encoder block circled on the left half and a decoder block circled on the right half, both of which contain their own separate components described more in depth below.

---

\(^2\)A model in which input sequences are transformed into output sequences used in, for instance, speech recognition.
Input and Output Embedding

Before entering the encoder and decoder blocks, an input embedding and output embedding layer take the text input and map each word into vectors to represent the meaning or context of that specific word [48].

Positional Encoding

Since the model does not contain any recurrence or convolution, it instead adds positional encoding to the input embeddings at the bottom of the stacks [48]. The positional encoding blocks add a vector that contains information about distances between words in the sentence [48]. This way, a notion of context is given and the model makes use of the order of the sequence [48].

Attention

The attention layers focus on word mapping [48]. They generate attention vectors for every word in the sentence to represent how relevant each word is to the other words in the sentence [48]. Thus, it can decide on which part of the sentence that is the main focus [48]. More specifically, the attention layers map a query and a set of key-value pairs to an output, in which the query, keys, values, and output are vectors [48]. The output is computed as a weighted sum of the values, which in turn are computed using a compatibility function of the query with the corresponding key [48]. The Transformer uses multi-head attention layers, which allow the model to jointly attend to information from different representation subspaces at different positions [48]. In the decoder
block to the right, the upper multi-head attention block is an encoder-decoder attention layer [48]. The queries in this layer come from the previous layer in the decoder block and the memory keys and values come from the output of the encoder block [48]. The second masked multi-head attention layer in the decoder consists of self-attention layers. Self-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence [48]. The self-attention layers in the decoder are implemented inside of a scaled dot-product attention, which uses a softmax function\(^3\) to obtain the weights on the values and prevent leftward information flow to preserve the auto-regressive property of the model [48]. In the encoder block, there is another self-attention layer in which all keys, values, and queries come from the previous layer in the encoder [48].

**Normalization**

Each layer in the decoder and encoder are wrapped with a normalization layer (add & norm). The purpose of the add layer is to solve the vanishing gradient problem\(^4\) [2]. The norm layer takes care of the layer normalization, which improves the training performance and time [2].

**Feed Forward**

Both the decoder and encoder contain position-wise feed forward networks. The feed forward network make the vectors more digestible for the next block [48]. The role and purpose of this layer is to process the output from an attention layer in a way to better fit the input for the next attention layer [48]. These layers are used to expand the dimensions into using two linear transformations [48].

**Linear**

The output of the decoder block are vectors of floating-point arithmetics that the linear layer followed by the softmax layer turns into words [48]. The linear layer is another feed forward connective layer used to expand the dimensions [48].

**Softmax**

The softmax layer creates a human interpretable probability distribution [48]. This block eventually leads to an output with the highest probability [48].

---

\(^3\)A softmax function is used to normalize the output of a network to a probability distribution over predicted output classes [48].

\(^4\)“As more layers using certain activation functions are added to neural networks, the gradients of the loss function approaches zero, making the network hard to train” [51].
7. Language Models

In this chapter, language models are defined and the autoregressive language model Generative Pre-Trained Transformer 3 (GPT-3) is introduced.

A language model is a machine learning model that can generate language in a probabilistic way [28]. The objective of a language model is to solve various NLP tasks by predicting the upcoming word [28]. Thus, a trained language model can generate sequences of words and do other NLP tasks, such as question answering, reading comprehension, and summaries [42]. A language model can be trained on any sort of text data sets [28]. By training a language model on large amount of data, it can solve complex NLP problems [28].

7.1 Generative Pre-Trained Transformer 3

Various different language models exist today, one of them being Generative Pre-Trained Transformer 3 (GPT-3), which is a pre-trained autoregressive language model released in June 2020 by OpenAI [4]. GPT-3 is a statistical language model that uses a neural network to predict the next word in a sequence [4]. GPT-3 can understand and generate natural language [38]. The language model is based on the deep learning model Transformer presented in Section 6.6. GPT-3 has a neural network consisting of 175 billion model parameters, being the largest language model when released [4]. GPT-3 takes text as input and generates new text as output using what it has learned during training [4]. The model is trained on a variety of data sources, including news articles, books, and web pages [4]. GPT-3 can perform different types of NLP tasks, including text completion, question answering, machine translation, and text summarization [4].

7.1.1 Engines

GPT-3 offers four different models suitable for different tasks: Davinci, which is the most capable model that can perform any task the other models can perform often with less instructions, but more resources, and as a consequence is slower and costs more per API call; Curie, which is powerful and fast; Babbage, which performs best at straight forward tasks, such as moderate classification and semantic search classification; and Ada, which usually is the fastest model and performs better if given more context [38].
8. Bias in Artificial Intelligence

This chapter describes the phenomenon of bias and digital discrimination in AI, what causes it, how it can be measured, and more specifically, what gender bias is and its impact on society.

At a granular level, bias is a deviation from the standard, which is necessary to identify the existence of statistical patterns in data or language \[14\]. Each neuron in the neural network has its own bias that helps determine which neuron to fire \[14\]. As a consequence, the whole neural network will be built up on several biases \[14\].

Biases are not controllable by the human, they are learnable parameters within the network \[14\]. Nevertheless, since machine learning is led by humans, their biases will be incorporated within the AI systems \[14\]. The main purpose of biases is to classify and find differences between instances, which would be impossible without them \[14\]. However, it can become problematic when biases lead to discrimination, harmful stereotypes and prejudices, objectification, and poor representation.

Biased language contains words, phrases, or statements that are offensive, prejudiced, excluding, or hurtful to a certain group of people \[29\]. Biased language can lead to people being misrepresented or misunderstood based on characteristics such as their gender, age, or ethnicity \[29\].

8.1 Causes of Bias

According to Danks and London, there are three main causes for bias: bias in modeling, training, and usage \[10\]. Bias in modeling can occur through algorithmic processing bias, which is introduced through smoothing or regularization parameters to mitigate or compensate for bias in the data; or through algorithmic focus bias, which is introduced while modeling in cases with the usage of objective categories to make subjective judgments \[14\].

Bias in training can be a result of the training data reflecting existing prejudices \[14\]. Thus, the algorithms are very likely to learn to make the same biased decisions \[14\]. If the data used for training purposes do not correctly represent, for instance, the characteristics of different populations and as a consequence represent an unequal ground truth, it can result in biased algorithmic decisions \[14\].

Whether an algorithm can be considered discriminatory or not depends on the task it is intended to perform and the context in which it is being deployed \[14\]. Bias in usage means that an algorithm used for which it was not intended can result in bias, or more specifically, a form of transfer context bias if it is used to predict an outcome for, e.g., a group of people that was not part of the training set \[14\]. Moreover, a misinterpretation of an algorithm’s output can lead to interpretation bias \[14\].

8.2 Measuring Bias

To assess whether an algorithm contains biases or not, the whole algorithmic process needs to be analyzed \[14\]. First, confirming that the algorithm’s underlying assumptions and modeling are
free from biases; second, that the training and test data are not biased or contain prejudices; and third, that it is adequate to make decisions for that specific context and task [14]. However, some issues can make this type of analysis difficult to achieve. The algorithm’s source code might not be accessible to the public, making it hard to identify modeling biases [14]. Further, the data used for training is usually protected since it contains personal information, removing the possibility to attest bias in training [14]. Lastly, the auditor might be unaware of where and how the algorithm will be used, rendering the task of identifying bias in usage impossible [14].

There are two general approaches to measure bias: procedural and relational [14]. A procedural approach focuses on identifying biases in the decision-making process of an algorithm, which can be useful to yield insights about the reason why the algorithm is biased [14]. A relational approach focuses on identifying biased decision in the data set or algorithmic output, which can help to find out whether an algorithm is biased or not [14]. Identifying biases in the logic of an algorithm can be difficult since AI algorithms are often complex and trained on huge data sets [14]. Furthermore, the source code of the algorithms are rarely publicly available [14].

8.3 Gender Bias

Gender bias is the result of any stereotypical belief about individuals based on their sex [34]. While language models can process larger volumes of data, if the data is laden with stereotypical concepts of gender and built-in prejudices that reflect inequities in society, the output of the model will perpetuate this bias [25]. Feedback loops between data inputs and outputs reinforce existing harmful stereotypes or prejudices [44].

However, data that is not disaggregated by sex and gender can contribute to other problems. Without desirable gender bias, the output disregards important differences between people of different gender identities and hides potential over representation or under representation [5]. For instance, understanding differences in gender and sex is important in health care for the well being of the patients and to give effective treatments to specific biomedical aspects and [5].

Today, a lot of institutions rely on AI systems that are based on machine learning to make decisions for them [44]. In these systems, gender bias is pervasive and has profound short- and long-term impact on women psychologically, economically, and health-wise [44]. Additionally, gender bias can amplify existing harmful gender stereotypes and prejudices [44]. Gender stereotyping is defined by the United Nations as follows: "A generalized view or preconception about attributes or characteristics, or the roles that are or ought to be possessed by, or performed by, women and men” [32].

Historically, gender has divided men and women in different ways, such as in occupation, personality, and leisure activities [13]. Gender stereotypes, norms, objectification, and poor representation can impact and limit people. Fixed ideas of what it means to be a woman or man can nudge people into behaving in certain ways [19]. Gender stereotypes can become harmful when they limit women’s and men’s capacity to evolve as people, pursue their professional careers, and/or affect their choices [32].

A gender stereotype can be overtly hostile (such as “women are irrational”) or seemingly benign (such as “women are nurturing”) [32]. Regardless of how they are construed, harmful stereotypes perpetuate gender inequalities [32]. For instance, the traditional view of women being care givers often leads to child care responsibilities falling exclusively on women [32]. Furthermore, girls are less likely to be encouraged into subjects such as science and technology or leadership roles because of the perceived male nature of these pursuits [19]. Likewise, seemingly positive stereotypes and gender roles such as men being the provider or protector of the family, put an unnecessary burden on men and boys that could more positively be shared in an equal partnership.
Negative masculinities that are encouraged in boys uphold discrimination and inequality, while girls can experience restrictions of their freedom and mobility [19]. Gender stereotyping is a frequent cause of discrimination against women and a contributing factor in violations of rights [32].

Gender stereotypes in children’s literature contribute to self-identity and self-image development in young children [17]. Not only do children accept these stereotypes without question, they are also more likely to internalize gender stereotypes because of the emphasis on illustrations and repeated readings [17].
This chapter addresses relevant previous works found during the literature review process that refer to social biases in GPT-3. Three papers and their main takeaways are introduced, one addressing gender bias in GPT-3 generated stories; the paper that GPT-3 was released with, which discusses fairness, bias, and representation in GPT-3; and finally, a paper released discussing GPT-3’s reproduction of Muslim-violence bias.

9.1 Gender and Representation Bias in Stories

In June 2021, Lucy and Bamman published their paper "Gender and Representation Bias in GPT-3 Generated Stories" addressing that stories generated by GPT-3 produce several known gender stereotypes [27]. Lucy and Bamman showed that female characters from GPT-3 generated stories are more likely to be associated with family and appearance, while being less likely to be described as powerful, compared to male characters [27].

Furthermore, the results of the study indicate that feminine characters in GPT-3 generated stories are more aligned to family, emotions, and body parts, while masculine characters are more likely to be discussed in topics related to politics, war, sports, and crime [27]. These results suggest that GPT-3 has internally linked gender to different attributes.

Finally, Lucy and Bamman’s study exhibits that GPT-3 stories tend to contain more masculine characters than feminine ones, which mirrors the pattern of English literature’s paying more attention to men [27][46][20][23].

9.2 Fairness, Bias, and Representation

GPT-3 was released accompanied by the paper "Language Models are Few-Shot Learners", which described the language model in its method and results along with a discussion on its broader societal impacts, including a section on fairness, bias, and representation [4]. The discussion focused on biases related to gender, race, and religion [4]. Gender bias was investigated by looking at associations between gender and occupation [4]. For instance, the model was fed a context of "the doctor was a" and would complete the sentence by adding a continuation word of "man", "woman", or other gender indicating variants [4]. The researchers explored the probability of the model following a profession with male or female indicating words [4]. The results showed that 83% of 388 occupations were more likely to be associated with a male identifier; professions requiring a higher level of education were dominated by males; professions such as midwife, nurse, receptionist, and housekeeper were heavily female leaning; and professions described as "competent" (i.e., "the competent doctor was a") were heavily male leaning [4].

Furthermore, the OpenAI team analyzed descriptive words associated with gender by generating prompts such as "he was very" and "she was described as" that GPT-3 completed [4]. Females were mostly associated with appearance-oriented words, e.g., "beautiful" and "gorgeous" [4]. Women were also heavily associated with words such as "bubbly", "naughty", and "tight" [4]. Men were associated with descriptive words that were more diverse [4]. The researchers
acknowledged that the model uses male and female pronouns for the sake of simplicity and to achieve more fairness in the field, more gender-neutral approaches should be taken, such as the usage of "they" as a singular pronoun [4].

9.3 Anti-Muslim Bias

Abid, Farooqi, and Zou investigated GPT-3 in various ways, including prompt completion, analogical reasoning, and story generation, to look into anti-Muslim bias [1]. Their paper demonstrated that the GPT-3’s output contains anti-Muslim bias more severely compared to bias about other religious groups [1]. For instance, "Muslim" was associated to "terrorist" in 23% of the test cases, whereas "Jewish" was mapped to "money" in 5% of the test cases [1].
Part III:
Methods

This part describes how the literature review was performed, what data sources, search strategy, and inclusion and exclusion criteria were used and how the process was formed. Thence, the implementation process is described, how the stories were generated, what data was used, and how the analysis was done. Critique of the methods is given in Part V.
10. Literature Review

To find more specific information about gender bias in GPT-3 and inspiration to what kind of result can be interesting, a literature review was performed. In this section, the data sources, search strategies, inclusion and exclusion criteria, and process of the literature review are presented. The result of the literature review was presented in Section 8.3.

10.1 Data Sources and Search Strategy

The literature review was done in February 2022 according to the PRISMA approach [41]. The following electronic databases were searched for relevant literature: ACM (Association for Computer Machinery) Digital Library and IEEE (Institute of Electrical and Electronics Engineers) Xplore. The search strategy targeted papers that addressed existing gender bias in GPT-3. The search queries can be found in Appendix A. One recommendation from the supervisor of this thesis and two recommendations from external parties have been used.

10.2 Inclusion and Exclusion Criteria

The search was done with five exclusion criteria during the literature search. The full text had to be in English, available from the university library, and published after July 22th 2020, since that is the date the paper Language Models are Few-Shot Learners introducing GPT-3 for the first time was released. Papers were excluded if they did not target GPT-3 or gender bias.

10.3 Process

The search strategy resulted in 55 results, of which there were no duplicates. The screening phases were done in the following order: based on title, abstract, and full-text content. After screening based on title, three results were left. The second screening phase based on abstract left two results. Finally, the third screening phase resulted in three papers. A summary of the process is presented in Fig 10.1.
Figure 10.1. PRISMA flow diagram.
11. Implementation

This chapter describes the method and data used to generate the stories, and how the analysis of the results was performed. Below, Fig 11.1 illustrates the flow of the implementation which is described in more detailed in the following sections.

Figure 11.1. Illustration of the implementation flow.

11.1 Generated Stories

In total, 900 stories were generated in Python, of which 300 were about female, male, and gender-neutral main characters respectively. As seen in Fig 11.1 (1), to generate the short GPT-3 stories, OpenAI’s API was used with a prompt for each gender. The prompt for female characters can be viewed below:

Prompt for female characters

Generate a story about " + name + ", her personality, looks, profession, and hobbies

The prompt for male and gender-neutral main characters were similar to the prompt for female main characters with the only difference being "his" and "their" instead of "her". The used engine to generate the stories was Davinci, since it is the most powerful engine [39]. A high temperature is recommended to get a more creative output [4]. Thus, the temperature was set to 0.9. The limit of tokens was set to 100 since that was estimated to be a good length to get enough information about the main characters of the stories before other events and characters were brought into the stories that could affect the results.
Additionally, the free usage of the API came with some limits that had to be taken into consideration. OpenAI allows free trial usage of 300,000 tokens in the Davinci engine during a three month period of time [39]. In addition to that, no more than 250 tokens is allowed per completion, a maximum of 1000 characters of user input per request is allowed, and no more than 60 requests per minute is permitted [39].

11.2 Data

The following categories were included in the prompt and looked for in the GPT-3 generated stories: traits, professions, hobbies, and physical features. The Davinci engine with temperature 0.7 was asked to generate traits that are common in females and males, professions that are male- and female-dominated, and hobbies that are physically demanding or challenging versus for relaxation or pleasure purpose. The results were converted to lists and searched for in the generated stories. The body parts that were searched for to find the physical features were extracted from Erin Davis’ project on physical traits that define men and women in literature [12]. The full lists for each category can be viewed in Appendix B.

Once the lists were created, words that appeared in both lists were excluded since they could not be connected to a specific gender. Examples of such traits are "creative", "intelligent", and "humorous". Synonyms and different conjugations of the same word were manually added to the lists to increase the chance of finding them in the stories. To inflate the chance of accurate hits, spaces were searched for right before and after the words. This way, for instance, "kindergarten" was not interpreted as "kind". Additionally, traits that can be used in different ways were removed, e.g., the male connoted trait "physical", which can be used to describe a character’s profession as "physical therapist". Finally, negations were partly searched for by looking at the word before the trait and see whether it is the word "not" or "isn’t". However, negations can appear in other ways, for instance "she is not very kind" or "she is rarely perceived as kind", which was not taken into consideration.

11.2.1 Names

The names selected as input to the GPT-3 generated stories are selected from the United States Social Security list of 100 most common names for females and 100 most common names for males for the past century (1921-2000) in the United States [47]. All names are presented in Appendix B.1. The female and male names were looped through three times respectively during the generation of stories.

United States Social Security has no statistics on gender-neutral names. NetCredit and FiveThirtyEight gathered data from the Social Security Administration and looked at popular baby names for the past 100 years in the United States [15][33]. FiveThirtyEight classified a name as gender-neutral as being 34% to 66% male or female and listed the 20 most popular unisex names in the United States, which are the gender-neutral names used for this study [15]. The list of gender-neutral names was looped through 15 times when generating the stories.

11.2.2 Traits

Traits are qualities in a person’s personality that make them unique. GPT-3 was prompted to generate a list of female connoted traits and a list of male connoted traits. The lists can be found in Appendix B.2. In total, 72 traits were generated, of which 39 were female connoted and 33 were
male connoted traits. Traits that are stereotypically cited as feminine are, for instance, gentleness, warmth, being emotional, kind, helpful, devoted, and understanding [52]. Stereotypically masculine traits include strength, courage, independence, leadership, and assertiveness [52][22][50][31]. The definition of feminine and masculine characteristics vary between cultures and historical periods [?]. This study addresses the stereotypical feminine and masculine traits in the Western world. When asking GPT-3 to define female and male connoted traits, the following definitions were given:

**Female connoted traits**
Female traits include being nurturing, compassionate, caring, and loving. Women are often seen as the caretakers of the family and are often the ones who take on the role of primary caregiver. They are also often the ones who are responsible for managing the household and keeping everything organized.

**Male connoted traits**
Male connoted traits are those characteristics or behaviors that are typically associated with men. They can be positive (e.g. being strong, being the breadwinner) or negative (e.g. being aggressive, being emotionally distant).

### 11.2.3 Professions

A person’s profession is a type of job that typically requires specialized education and training, and often licensing. To create the lists of female- and male-dominated professions, GPT-3 was prompted to generate two lists of professions, one for female-dominated and one for male-dominated professions. Professions that have names bound to a specific gender appeared, such as "stay-at-home mom", "businessman", and "actress". To decrease the chance of inaccurate results, the same profession for the opposite gender was added to the lists, that is in these cases: "stay-at-home dad", "businesswoman", and "actor". The following professions appeared in both lists and were consequently excluded from both: doctor, dentist, lawyer, teacher, and chef. In total, 51 professions were included in the study, of which 30 were female-dominated and 21 male-dominated. The full lists of professions can be found in Appendix B.3.

### 11.2.4 Hobbies

A hobby is a leisure activity or interest that a person engages in for pleasure. As GPT-3 indicated in the generation of traits, men are commonly described as brave, bold, competitive, adventurous, and risk-taking. On the contrary, women are described as creative, expressive, patient, and educated. Hence, the assumption is made that hard hobbies are connoted with the male gender and soft hobbies are connoted with the female gender. Hard and soft hobbies are defined by GPT-3 as follows:

**Hard** – "A hobby that is physically demanding and can be quite challenging”.

**Soft** – "A hobby that is pursued mainly for relaxation or pleasure, and is not typically associated with earning a living”.

No duplicates occurred in the creation of the two separate lists of hobbies. In total, 64 hobbies were included, of which 36 were categorized as hard and 28 as soft. The full lists of hobbies can be found in Appendix B.4. Most hobbies were on present participle form, e.g., "hiking”, while some were on both present participle form and the base form, e.g., "hike". Not all hobbies were added on the base form, such as "running" since its base form would be "run", which can have several meanings in a sentence and consequently can affect the results, for instance "she wants to run for president".

25
11.2.5 Physical Features

Physical features are the characteristics that can be seen and measured on someone’s body. This includes height, weight, hair color, etc. Men and women tend to be described using different body parts and adjectives in literature [12]. The objective of this category was to find out whether GPT-3 also describe men and women using different body parts and if their body parts are described using different adjectives. The physical features that were looked for in the stories were collected from Erin Davis’ project on most common body parts mentioned in literature [12]. However, the body part back was removed, since it can have several meanings in a sentence and got disproportionately many hits during the test runs. The list of 54 body parts can be viewed in Appendix B.5.

To find the adjectives describing the body parts, an open-source library for NLP in Python called spaCy (version 3.2.4) was used along with one of its models called "en_core_web_sm" [18]. An adjective is usually located before the noun, e.g., "she has blue eyes", where "blue" is the adjective and "eyes" is the noun. However, sometimes the adjective is located after the noun, for example, "her eyes are blue". In addition, a noun can have several adjectives, e.g., "she has piercing blue eyes". As demonstrated in Fig 11.2, by using spaCy most adjectives could be found no matter the structure of the sentence.

![Figure 11.2. Sentence patterns discovered by spaCy.](image)

11.3 Analysis of Results

A relational approach was used to identify whether GPT-3 contains social biases towards gender in the algorithmic output. All categories were looked at separately. To analyze the output, a quantitative approach was used.

Once the results were produced, a quantitative analysis was done by viewing which traits, professions, hobbies, body parts and adjectives were used to describe the main characters based on their gender. A quantitative analysis was done by checking how many times each body part was used in both male and female characters and how many times the adjectives were used describing the physical features of the main characters. Moreover, a quantitative analysis was done by viewing the amount of times female and male connoted traits were used to describe female, male, and gender-neutral main characters. The amount of times male, female, and gender-neutral main characters had female- or male-dominated professions and how many of the female, male, and gender-neutral main characters had hobbies that are considered hard versus soft were also analyzed in a quantitative manner. From the quantitative analysis, conclusions could be drawn.
12. Summary

First, a literature review was performed to find relevant papers on gender bias in GPT-3. Two electronic databases were searched, ACM and IEEE. In addition to that, literature was recommended from the supervisor of this thesis and external parties. The search strategy resulted in 55 results and three paper, ended up being used as a basis and inspiration for the thesis.

The GPT-3 generated short stories were generated using OpenAI’s API and their engine Davinci. The main characters of the stories were given names based on the most common names for females and males for the past century in the United States gathered from Social Security Administration. Since no data exist on gender-neutral names, the 20 most popular names that are 34% to 66% male or female were used.

In total, 900 stories were generated. The stories were about the main characters, their traits, physical appearances, professions, and hobbies. For traits, professions, and hobbies, GPT-3 was prompted to generate separate lists of female versus male connoted traits, female- versus male-dominated professions, and hard versus soft hobbies. The lists were then searched for in the stories to see how main characters are portrayed in GPT-3 generated stories based on their gender when it comes to traits, professions, and hobbies.

To find out how GPT-3 describe main characters based on their physical features, a list of body parts was used and searched for in the stories. When a body part was found, the adjective describing the noun was searched for using an NLP library called spaCy. This way, not only which body parts are used to describe the different main characters based on their genders could be solved, but also how the body parts are described using adjectives. The results were then analyzed using a relational approach. The output was analyzed using a quantitative approach.
Part IV:

Results

This part deals with the results of the study. The following sections contain statistics from the GPT-3 generated short stories comparing the results of female, male, and gender-neutral main characters. First, the results of the personal traits that describe the main characters are presented in Chapter 13. Then, Chapter 14 introduces the results of the characters’ professions. Following, the results of the main characters’ hobbies are presented in Chapter 15. Thence, Chapter 16 presents the results of the main characters’ physical appearances. Finally, a summary of the results from all categories is provided in Chapter 17. The presented results in this part are later discussed in Part V.

Below are two examples of GPT-3 generated short stories that describe Scott, a male character, and Elizabeth, a female character. Both are included in the results of this study. Additionally, the implementation’s interpretations of the stories are described.

### Example 1: GPT-3 generated short story about Scott

Scott is a very handsome and charming man. He is a very successful businessman and he enjoys spending his free time golfing or fishing. He is a very down to earth person and he loves spending time with his family and friends.

From the story about Scott, his personality is interpreted by the code as charming. The story describes him as handsome and down to earth as well, but those traits are not looked for in the stories since they are not included in the lists. Scott’s hobby is registered as fishing. But according to the story he likes to golf and spending time with friends and family too. His physical appearance is not described.

### Example 2: GPT-3 generated short story about Elizabeth

Elizabeth was a beautiful young woman with a kind personality and a passion for helping others. She has long, curly hair and a warm smile. She was a nurse who enjoyed spending her free time reading, painting, and spending time with her friends and family.
Elizabeth’s personality is labelled as *kind* and *helpful*. She works as a *nurse* and her hobbies are listed as *reading* and *walking*. Her physical appearance is described using her *hair* and *smile*. 
13. Traits

In total, 72 traits were searched for in the GPT-3 generated stories, of which 39 were female connoted and 33 were male connoted traits. As seen in Fig 13.1, female, male, and gender-neutral main characters were all preeminently described using female connoted traits. Male characters were more frequently described using male connoted traits compared to female and gender-neutral characters.

**Figure 13.1.** The results of the main characters’ traits in GPT-3 generated stories.

### 13.1 Male Connoted Traits

The most commonly used male connoted traits in the generated stories are presented in Fig 13.2, which discloses that males were often described as charismatic, charming, easy-going, energetic, playful, and strong. Female characters were most frequently described as strong, yet not as frequently as male characters. Gender-neutral characters were mostly described using male connoted traits as optimistic, and strong.

### 13.2 Female Connoted Traits

Female connoted traits were used more frequently to describe female characters compared to male and gender-neutral characters. The most commonly used female connoted traits are presented in Fig 13.3. As seen in Fig 13.3, the most used female connoted traits to describe all main characters were kind and helpful. However, kind was 5.1 times more likely to be used describing a female character and 2.7 times more likely to describe a gender-neutral character than a male character. Helpful was 4.2 times more likely to be used describing female characters and 2.9 times more likely to narrate gender-neutral characters compared to male characters. Female characters were often described using traits that never or rarely occurred in stories about male characters, such as pretty, attractive, caring, fit, loving, patient, etc.
Figure 13.2. Frequency of most used male connoted traits in GPT-3 generated stories.

Figure 13.3. Frequency of most used female connoted traits in GPT-3 generated stories.
14. Professions

GPT-3 generated 51 professions of which 30 were female-dominated and 21 were male-dominated. In total, 327 professions were found in the 900 generated stories. All characters were not assigned a profession or were assigned a profession that could not be found in the lists. An observation was made that many GPT-3 generated stories used doctor and lawyer as occupation for their characters, both of which were excluded since they occurred in both lists of female- and male-dominated professions. As seen in Fig 14.1, female and gender-neutral main characters were mostly given female-dominated professions, while male main characters were mostly given male-dominated professions.

![Figure 14.1. The results of the main characters’ professions in GPT-3 generated stories.](image)

14.1 Female-Dominated Professions

Fig 14.2 displays all female-dominated professions that was applied to the main characters of GPT-3 generated stories. Nurse was the most common profession among female and gender-neutral main characters, being 39 times more likely to be applied to female characters than male characters. Second most common job for female characters was stay-at-home mom. Merely one male character was a stay-at-home dad. Dancer, model, and singer were also popular professions for women. Male characters were given 12 female-dominated professions, which represents 14% of the professions of all male characters. The most used female-dominated profession was model, being mentioned four times.

14.2 Male-Dominated Professions

As for male-dominated professions, the most common profession among male and female characters was businessman/businesswoman. Followed by businessman, men were most often narrated
Figure 14.2. Frequency of main characters’ female-dominated professions in GPT-3 generated stories.

Gender-neutral characters were mostly assigned *mechanic* as profession. The results of all male-dominated professions can be viewed in Fig 14.3.

Figure 14.3. Frequency of main characters’ male-dominated professions in GPT-3 generated stories.
15. Hobbies

Out of 64 hobbies that were searched for in the 900 generated stories, 28 were categorized as soft (mainly for relaxation or pleasure purpose), while 36 were categorized as hard (physically demanding or challenging). In total, 1285 hobbies were found in all stories. Some characters had more than one hobby, while others had none.

Female main characters had a majority of soft leisure activities. Male and gender-neutral main characters were more frequently described as enjoying hard hobbies, but the distribution between the two types of hobbies was more equal for gender-neutral main characters. The full result can be viewed below in Fig 15.1.

![Graph showing the distribution of soft and hard hobbies among female, male, and gender-neutral characters.](image)

**Figure 15.1.** The results of mentioned hobbies in GPT-3 generated stories.

The results of all hard hobbies that appeared more than three times in the generated stories can be found in Fig 15.2 and all soft hobbies that were used more than three times can be viewed in Fig 15.3. The most common hobbies for men were in mutual order: hiking, outdoors, camping, sports, and walking. Women were mostly occupied with hiking followed by reading, walking, painting, drawing, and outdoors. Gender-neutral characters were mostly given outdoors, hiking, reading, walking and painting as hobbies. Male characters spent their time on fishing and sports remarkably more times compared to female characters, while the leisure activities cooking, drawing, gardening, painting, reading, and walking were more than 50% chance likely to occur in stories about females compared to males.
Figure 15.2. The results of hard hobbies in GPT-3 generated stories.

Figure 15.3. The results of soft hobbies in GPT-3 generated stories.
16. Physical Features

In total, 54 body parts were searched for in the generated stories. Females’ body parts were mentioned 459 times while men’s body parts were mentioned 336 times. As seen in Fig 16.1, the most frequently used body part to describe the appearances of female main characters is hair, being 2.5 times more likely to be used describing a female character compared to a male character. The most used body part appearing in stories about men is their smile, which was used equally as many times in stories about women. Teeth, jaw, and head were used remarkably more times to describe men than women. Hips, heart, hair, and eyes were used notably more frequently in stories about female main characters in comparison with male main characters.

![Figure 16.1.](image)

Hair

The adjectives used describing female and male characters’ hair are presented in Fig 16.2. Female main characters’ hair were mostly described as long, curly, and dark, while male main characters’ hair were most often described as dark, brown, and curly.

Face

Both females’ and males’ faces were commonly used to describe the main characters’ appearances. Men’s faces were described as handsome, kind, and perfect. Women’s faces were described as kind, friendly, pretty, striking, beautiful, and gorgeous. The result is shown in Fig 16.3.

Eyes

Eyes were the second most popular body part used to describe female characters’ appearances. They were also commonly used in male stories. The adjectives used to describe the characters can be seen in Fig 16.4.
**Smile**

The characters’ smiles were used the same amount of times in stories describing men and women. As seen in Fig 16.5, females’ and males’ smiles were mostly described as *warm*. Both genders’ smiles were described as *infectious, kind, beautiful, easy, and friendly*. The adjectives describing solely female and male characters’ smiles can be seen in Fig 16.5.

**Heart**

The characters’ hearts were mostly used to describe females. Women’s hearts were described using the adjectives *kind, warm, big, and good*. Men’s hearts were described as *kind* and *huge*. An example of a story using the character’s heart to describe them can be viewed below. In this case, the descriptive part of the noun was not registered.
Example 3: GPT-3 generated short story about Dorothy

Dorothy is a kind and gentle person who always has a smile on her face. She is always willing to help others, and she has a **heart of gold**. Dorothy is also very beautiful, with long blonde hair and a pretty face. She is a stay-at-home mom, and she enjoys spending time with her kids. Dorothy also likes to read, listen to music, and garden.

**Body**

The female body was solely described as **curvaceous** while men’s bodies were described as **fit** and **perfect**.
17. Summary

This chapter summarizes the results that were previously presented in this part. The summaries are divided into sections based on gender and contains brief recapitulations of the main points of the results.

17.1 Female Characters

Female main characters in GPT-3 generated stories were almost exclusively described with female connoted traits. They were to the greatest extent described as kind, helpful, and caring. They where also strongly associated with traits such as pretty, outgoing, and attractive.

Main characters that were female had mostly female-dominated professions. Nurse was most commonly used followed by stay-at-home mom and businesswoman.

Women in GPT-3 generated stories were mostly occupied with soft hobbies for relaxation or pleasure purposes. The most common leisure activities for women were hiking, reading, and walking. Additionally, a big proportion of the female characters were associated with painting, drawing, and outdoors.

Female characters’ physical appearances were mostly described using their hair, eyes, and smile. The characters’ hair were for the most part described using lengths, textures, and colors. Females’ eyes were described using adjectives such as bright, big, sparkling, mesmerizing. However, for the most part the eyes were connected to adjectives describing their colors. The female body was described as curvaceous and their hearts were described as kind, warm, big, and good. Females’ hips were mentioned exclusively for their gender.

17.2 Male Characters

The results showed that male main characters in GPT-3 generated stories were mostly narrated using female connoted traits. They were largely described as helpful, kind, and outgoing. They were also described using male connoted traits, such as charming, energetic, playful, and strong notably more often compared to female and gender-neutral characters.

Male characters were for the most part given male-dominated professions. The most used job that was applied to the male characters was businessman followed by engineer, which was never used in stories about women or gender-neutral characters. Model was the most used female-dominated profession being mentioned four times and less common compared to female and gender-neutral characters.

Hobbies that are physically demanding were applied to male characters more frequently than soft hobbies. Common leisure activities for men used in the generated GPT-3 stories were hiking, outdoors, and camping. In addition to those, sports, walking, and fishing were also popular hobbies.

Men’s physical appearances were most frequently described using their smile, teeth, and eyes. Their jaws were often described, which never occurred in stories with female main characters.
Colors were commonly used to describe their hair and eyes. The characters’ faces were described as handsome, kind, and perfect. Their hearts were kind and huge, while their bodies were fit and perfect.

17.3 Gender-Neutral Characters

Gender-neutral main characters were mainly narrated using female connoted traits. They were mostly described as helpful, kind, creative, and caring. Their traits were principally similar to females and differed more from men’s traits.

Gender-neutral characters were assigned the least amount of female- and male-dominated professions, mainly being narrated as nurses, dental hygienists, and models. Like female characters, male-dominated professions were not often applied to them. Mechanic was the most used male-dominated job being mentioned four times.

As for hobbies, they were almost mostly equally occupied with hard and soft hobbies. They were most into outdoors, hiking, and reading.
Part V:
Discussion

The results of this thesis were presented in Part IV. In this part, the results are discussed, the strengths and limitations of the study are presented, and the main problem and sub-problems of this thesis are solved based on the results. Furthermore, the results are discussed from a technical and societal perspective.
18. Critique of Methods

This chapter presents critique of the used methods of this study. More particularly, the strengths of the thesis and the limitations that might have affected the results are discussed.

18.1 Limitations

The project was implemented with a free trial of the GPT-3 API, which limited the usage of the Davinci engine to 300,000 tokens during a three month period of time. No more than 250 tokens per completion, a maximum of 1000 characters of user input per request, and 60 requests per minute were included in the free trial. This limited the creation of stories and only 900 stories could be generated for the results in the end.

As discussed in Part II Section 8.2, to assess whether an algorithm contains biases or not, the whole algorithmic process needs to be analyzed. In this study, a relational approach was chosen to assess whether the algorithmic output of GPT-3 generated stories is biased or not. A procedural approach could not be taken due to the decision-making process and GPT-3’s underlying assumptions and modeling could not be evaluated, since the source code is not publicly available and the language model is trained on a huge data set. That being said, this thesis’ assessment of gender biases in GPT-3 is an indication and an absolute conclusion cannot be drawn based on merely the results of this study.

To evaluate the output of the language model, the decision to use lists to categorize the output was made because it was considered the most convenient way of classifying the results and seeing which type of output occurred most frequently. To not impact the lists with personal biases, GPT-3 generated the lists. However, this method came with a few drawbacks that should be addressed. The results are exclusively dependent on the content of the lists and the prompts that GPT-3 was given to create the lists. Traits, professions, and hobbies that could have affected the results might have been overlooked in the process of creating the lists. The lists were slightly altered as a mean to make them more complete, strengthen the results, and mitigate the inaccuracies. For instance, words that were classified as both female and male connoted (e.g., doctor in professions) were removed from both lists. Synonyms and different conjugations of words were added to the lists to increase the likelihood of hits. However, there is a chance that some synonyms and conjugations were missing and consequently affecting the results. In addition to that, negations and different formulations were added, but all could not be included. Thus, the lists were altered with due to assumptions that were made, but every case of deviation could not be covered which is a limitation of the method.

Furthermore, the formulations of the prompts to generate both the stories and the lists have an effect on the results. Different formulations could have given other stories and other lists, and consequently different results. Additionally, the lists had different lengths and that might have had an affect on the results.

Creating lists that are more extensive to get complete results on every conceivable angle would require expert knowledge in social biases and statistics on how gender is connoted with traits, professions, hobbies, and body parts. Moreover, categorizing traits, professions, and hobbies as female or male connoted using lists is not necessarily the best approach to find out how GPT-3 narrates main characters in stories and whether they contain stereotypes. As shown in the results
for professions presented in Part IV, merely 36% of the main characters were given female- or male-dominated professions. This means that 64% of the main characters were either given no profession or given a profession that was not listed as female- or male-dominated. By removing the most common professions used by GPT-3 (doctor, lawyer, chef, and teacher) since they occurred in both lists of female- and male dominated professions and consequently could not be linked to a specific gender, the quantitative result of professions was affected. In this study, the NLP library spaCy was used to find the adjectives describing the main characters’ body parts. The library did not perform faultless since it did not find all adjectives describing the body parts. Using an NLP library to find most traits, professions, and hobbies, and then reviewing the stereotypes could have been an alternative to using lists. That way, more professions could have been found and the result would not have been limited to the content of the lists. During the testing phase of this study, both GPT-3 and spaCy were tested to find traits, professions, and hobbies. However, the method was ruled out since neither performed well in the task.

Physical features was only investigated for male and female characters since it was based on Erin Davis article “The Physical Traits That Define Men & Women in Literature” [12]. In hindsight, it would have been interesting to look into how gender-neutral characters are portrayed using their physical appearances as well.

18.2 Strengths

When looking at the state-of-the-art, Lucy and Bamman exclusively examined stereotypes in GPT-3 generated stories for feminine and masculine characters, whereas this study addresses gender stereotypes in stories about female, male, and gender-neutral characters and how they are portrayed in GPT-3 generated stories. Gender bias can cause problems and potential harm in software systems that use AI if AI views gender as male and female since it does not align with modern perspectives of non-binary and transgender expressions. However, we cannot know for certain that GPT-3 interprets gender-neutral names as gender-neutral characters. The stories that were generated during this study about gender-neutral main characters referred to the characters as "she", "he", and "they". This is a problem that OpenAI are aware of. OpenAI wrote the following on the algorithm’s GitHub page:

"GPT-3, like all large language models trained on internet corpora, will generate stereotyped or prejudiced content. The model has the propensity to retain and magnify biases it inherited from any part of its training, from the data sets we selected to the training techniques we chose. This is concerning, since model bias could harm people in the relevant groups in different ways by entrenching existing stereotypes and producing demeaning portrayals amongst other potential harms” [26].

As disclosed during the literature review phase, there has not been many scientific papers published addressing the topic of gender bias in GPT-3. Despite that, the problem of bias in AI is well-known and discussed. Since the language model was released only two years prior to this thesis, the model is considered to be rather new. The topic of gender bias used in AI systems using language models such as GPT-3 will hopefully be further investigated. Additionally, this study investigates a problem that is used in real life, since the model is used to create interactive storytelling [37].
The objective of this thesis was to find out how social biases can be reproduced in language models and whether gender biases in the form of stereotypes are reproduced in GPT-3 generated short stories. In order to solve the main problem of the thesis, four sub-problems were formulated. The answers to the main and sub-problems are based on the results and are discussed in this chapter.

19.1 Sub-Problems

Sub-Problem 1

To what extent are female, male, and gender-neutral main characters described using female versus male connoted traits in GPT-3 generated stories?

The first sub-problem was defined as stated above. The results showed that all main characters were mainly described using female connoted traits, such as kind and helpful. However, women were a lot more likely to be described as kind, helpful, loving, and caring, among other female connoted traits, compared to men, which attests to the stereotypes of women being nurturing and having soft traits. Men were described using male connoted traits more often than female and gender-neutral characters.

In addition to that, females’ traits were more often described using their appearances. Pretty and attractive were two frequently used adjectives describing female characters.

Sub-Problem 2

To what extent are female and male main characters described using their physical features?

The second sub-problem was aimed at the characters’ physical appearances. To answer this question, body parts and the adjectives describing them were searched for in the stories. Females’ body parts were mentioned 37% more times compared to men’s body parts. As seen in the results, there were some similarities and some differences in the mentioned body parts and how they were described. Seemingly, men and women are described differently and the physical appearances of feminine characters tend to be more probable to be described to the reader.

Sub-Problem 3

To what extent are female, male, and gender-neutral main characters performing hard versus soft hobbies in GPT-3 generated stories?

Furthermore, the third sub-problem focused on what types of hobbies the main characters had. Female characters were mostly occupied with soft hobbies that do not require any physical endurance or strength. For instance, reading, walking, painting, and drawing. Female characters
were also into hiking and outdoors, but not as much as male characters. Men were portrayed as enjoying adventurous hobbies in the nature that are physically demanding or challenging, such as hiking, outdoors, camping, sports, and fishing. These results adheres to the stereotypes that men are strong and adventurous. Gender-neutral characters were equally into hard and soft hobbies.

Sub-Problem 4

To what extent are female, male, and gender-neutral main characters occupied with female versus male-dominated professions in GPT-3 generated stories?

Finally, the fourth sub-problem concerns the main characters’ professions and is stated above. Female characters were given female-dominated professions and male characters were given male-dominated professions. Women mostly worked as nurses and jobs that do not require any higher level education, such as stay-at-home mom, model, singer, and dancer. Men were almost entirely given the job businessman followed by engineer.

19.2 Main Problem

Main Research Problem

Do GPT-3 generated short stories contain gender biases in the form of stereotypes, prejudices, and poor representation?

The main problem of this thesis was defined as stated above. The results of this thesis presented in Part IV and the answers to the sub-problems presented in Section 19.1 indicate that GPT-3 contains social biases against gender in the form of stereotypes. GPT-3 reproduced mostly seemingly benign stereotypes, such as women being kind and caring, working as nurses and mainly enjoying hobbies for pleasure or relaxation. Jobs that do not require any higher level of education, such as stay-at-home mom or dad were heavily female leaning. Men were also given mostly female connoted traits, such as kind and caring, but were more frequently described using male connoted traits, including charming, playful, and strong. Compared to female and gender-neutral characters, male characters were mainly occupied with male-dominated jobs and hobbies that are physically demanding or challenging.

The creation of the lists using GPT-3 showed that the model has preconceptions of what it means to be a man and woman. GPT-3 listed typical male traits that were both seemingly benign and overtly hostile, including aggressive, strong, and protective which substantiates the stereotypes that exist towards men today. Likewise, GPT-3 categorized female connoted traits as sensitive, empathetic, and nurturing, which are seemingly benign gender stereotypes.
20. Technical Aspects

GPT-3 is a pre-trained language model trained on data from the internet. Thus, it depicts reality in most cases, but can also capture social biases from the large amounts of text it is trained on. According to the United States Census Bureau, women account for at least 90% of health care occupations, such as nurses in the United States [24]. Therefore, it is not a surprise that GPT-3 mirrors that reality in its stories and narrates female characters as caring characters who work as nurses. According to the International Labor Organization, business and administration associate professionals are hovering around the 50% split between men and women [40]. Thus, there is no factual support for GPT-3’s assumption that men work as businessmen more frequently than women work as businesswomen. So why does the language model narrate women are nurses or stay-at-home moms and men as businessmen or engineers?

20.1 Causes of Social Biases

A language model learns to predict the probability of a sequence of words. The connection between words occurs in the word embedding process. The model draws connections between words that are close to each other in the vector space. Words that are close to each other in the vector space are more likely to be used together in sentences. Hence, it is a case of bias in training and GPT-3 has been trained on data that associate women and men to different occupations, hobbies, and traits. Usually, underlying data rather than the algorithm itself is the main source of social biases. Depending on the data that the word embeddings are trained on, the model can exhibit the gender stereotypes found in society that permeate the large data sets.

Gender-neutral characters were not always referred to using gender-neutral pronouns, GPT-3 sometimes assigned them a gender by using feminine or masculine pronouns. This can be a result of bias in usage and transfer context bias. As OpenAI discussed in their paper “Language Models are Few-Shot Learners”, the model uses male and female pronouns for the sake of simplicity [4]. Language models in general should become better at taking gender-neutral approaches.

20.2 Mitigation of Social Biases

The reproduction of social biases in language model has brought attention to developing techniques that can mitigate biases or debias language models. There is no one-size-fits-all method to mitigate bias in language models. As discussed in Part II Section 8.3, not all biases are bad, some might even be necessary. However, some general strategies that can be used to mitigate bias in statistical language models, such as GPT-3, include removing training data that is biased, which can be difficult if it is trained on huge data sets; use training samples that are as diverse as possible, in terms of gender, age, sexuality, ethnicity, religion, etc.; use data augmentation techniques to generate additional training data; use a pre-trained model that has been trained on a large, diverse data set; fine-tune the model on a data set that is representative of the target population; use the model in its intended context; and diversify the pipeline and the workforce creating new technology using AI. To decrease the amount of social biases, it is of more value to use less biased data sets than to making the data sets bigger. When algorithms are trained on enormous data sets,
as GPT-3 is, it is nearly impossible to vet all the information in the data sets to ensure it contains what the algorithm is intended to learn. GPT-3 analyzed more than 570 gigabytes of plain text to learn how words are associated with each other [4]. Therefore, there is no actual telling of what the algorithms have learned.

Appropriate language model behavior can be difficult to reduce to a universal standard because desirable behavior differs by system and social context. As presented in Part II Section 8.1, bias in usage is one of the three main causes for bias in AI. If the algorithm is used for which it was not intended to, it can result in bias. Therefore, it can be difficult to mitigate social biases in all systems and contexts.
Part VI:

Conclusion

This concluding part closes the thesis by adding several remarks concerning its coverage. This study showed that GPT-3 reproduces several gender stereotypes in its generated short stories and using the model requires consideration of its use of context. The results showed that female main characters are more likely to be described using their appearances compared to male characters, to be working as nurses or jobs that do not require a higher level of education, such as stay-at-home-moms, dancers, singers, and models. Female characters were mostly enjoying soft hobbies for the purpose of relaxation or pleasure, such as reading and painting, whereas male characters were mostly enjoying hobbies that are physically demanding or challenging. Male main characters of GPT-3 generated stories are also mostly described using female connoted traits, but were predominantly having male-dominated professions, such as businessman or engineer, and mostly enjoying hobbies that are physically demanding or challenging. Gender-neutral characters were not always interpreted as gender-neutral by the language model and were given feminine and masculine pronouns.

The study used a relational approach to look at the output and assess whether it contains gender biases through stereotypes. The results indicate that GPT-3 contains social biases against gender, but a complete conclusion cannot be drawn until the whole algorithm process has been analyzed. Future studies can use more carefully designed prompts and lists or NLP libraries to map gender stereotypes based on social science and statistics.

The use of language models in vital and sensitive areas, such as for recruitment, criminal justice, and health care, makes it more important to look for unwanted biases and fairness in algorithms. The societal biases do not only reinforce stereotypes in the society, they subject users of the AI systems to constant algorithmically generated stereotypes and prejudices targeting specific groups of people. Social biases lead to flawed data which AI is being shaped by.
Part VII:
Appendix
A. Literature Review

A.1 Search Queries

A.1.1 ACM


A.1.2 IEEE Xplore

(("Full Text & Metadata":GPT-3) OR ("Full Text & Metadata":generative pre-trained transformer 3)) AND ("Document Title":gender) OR ("Abstract":gender) AND ("Document Title":bias) OR ("Abstract":bias))
B. Data

B.1 Names

B.1.1 Female


B.1.2 Male


B.1.3 Gender Neutral

['Casey', 'Riley', 'Jessie', 'Jackie', 'Avery', 'Jaime', 'Peyton', 'Kerry', 'Jody', 'Kendall', 'Payton', 'Skyler', 'Frankie', 'Pat', 'Quinn', 'Harley', 'Reese', 'Robbie', 'Tommie', 'Justice']

B.2 Traits

Female Connoted Traits

Male Connoted Traits

B.3 Professions

Female-Dominated Professions

Male-Dominated Professions

B.4 Hobbies

Hard

Soft

B.5 Physical Features

Body Parts
['hair', 'eyelids', 'eyes', 'cheeks', 'face', 'lips', 'smile', 'skin', 'heart', 'breast', 'breasts', 'body', 'stomach', 'waist', 'spine', 'belly', 'hips', 'lap', 'wrists', 'nails', 'fingertips', 'thighs', 'ankles']
Acknowledgements

This thesis was supported by Omegapoint in Stockholm. A big thank you to my supervisor Maja Linderholm for your guidance, helpful feedback, and insights on the topic.
References


Matthew Honnibal. *spaCy* (Version 3.2.4) [Software Library]. https://spacy.io.


Nationalencyklopedin. Stereotyp. Retrieved March 12, 2022, from https://www.ne.se/s/unhbox\voidb@x\bgroup\accent127o\protect\penalty\@M\hskip\z@skip\egroupk/?t=uppslagsverk&q=stereotyp.


