Can Chatbots Truly Be 'Unbiased'?

Joshua Keith Reed

UP847988

School of Computing

Abstract

It is well-known that bias is a big issue within the field of AI, and chatbots are no exception. Numerous reports of negatively biased chatbots have appeared in the media. Perhaps the most famous case is Microsoft's Tay, taken down less than 24 hours after its launch after offensive 'tweets' were generated. This bias can be likened to biases found in people, who are usually the source of data used to train chatbots. Many studies have been conducted to understand bias in chatbots, and more importantly, how this can be removed – with mitigation algorithms being the result. However, there is another hidden side to human bias that research into chatbot bias fails to consider: *implicit* bias. This project aims to be one of the first to look at implicit bias within chatbots, firstly by looking at how it affects people, especially everyday speech, and what implications this has for chatbots. A way to test for implicit bias within people is explored, and an adaptation is made to test chatbots similarly. This adaptation reveals the complexities surrounding implicit bias within chatbots, and inconclusive results highlight the need for further research. This project acts as one of the first preliminary studies linking the psychological concepts of implicit bias with the technical application of chatbots, laying the foundations for future research to fully identify the effects had on a rapidly growing area of technology, with the ultimate aim of answering the fundamental question: Can chatbots truly be 'unbiased'?

Introduction

The effects of bias in Artificial Intelligence (AI) are widely spread. Countless news articles cover this, including: The Guardian's "South Korean AI chatbot pulled from Facebook after hate speech towards minorities" (McCurry, 2021); PinkNews' "Google's new artificial intelligence bot things gay people are bad" (Jackman, 2017); and The Verge's "This girls-only app uses AI to screen a user's gender — what could go wrong?" (Schiffer, 2020). The issue with these is that they focus on the explicit biases shown by AI, little/no coverage is given to the hidden biases that arguably could have more severe consequences.

When a piece of AI shows explicit bias, it can easily be taken down, improved, and re-released, with little consequence. The real danger arises with hidden biases, where the consequences are felt, but the bias is not obvious. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software tool was found to be racially biased several months after it incorrectly labeled a Black individual as high risk and a White individual as low risk of re-offending for very similar crimes, despite the White individual's troubled past, and the Black individual's relatively crime-free past (Angwin et al., 2016).

The issue with bias in AI is not just a technical one – bias can be seen in people in everyday life. Bias is formed from a young age (Baron et al., 2014; Cameron et al., 2001), and influences almost everything people do and say,

yet people are not aware it exists (Ruhl, 2020). The danger of training AI using human-driven data is the possibility of turning these 'manual' biases into 'automatic' biases. The issue this research aims to tackle is whether it is possible to remove these hidden biases that are revealed to be intrinsically linked to people's day-to-day lives. Can a truly 'unbiased' machine be developed, or does bias simply need to be managed?

This project focuses primarily on bias within chatbots, where it argues bias can not be removed at all. As stated by P. Henderson et al. (2018), the primary purpose of a chatbot is to mimic human behaviour. With the inevitability of implicit bias in humans, is it not then inevitable chatbots will 'mimic' this bias? Is it possible to detect, and by extension remove this hidden bias?

Literature Review

2.1 The Chatbot

2.1.1 Definitions

A chatbot is a computer, algorithm, or AI that communicates with a person, intending to make users feel they are talking with another living person (Neff & Nagy, 2016; Zemčík, 2020). The simplest approach to chatbot creation is pattern-matching, where the chatbot selects a predefined answer from a corpus of responses using pattern-matching algorithms (Adamopoulou & Moussiades, 2020). Machine learning approaches use Natural Language Processing (NLP) to extract content from user input, considering the entire dialogue context, and respond without need for a predefined list of responses (Adamopoulou & Moussiades, 2020). Typical machine learning approaches use Artificial Neural Networks (ANNs) or Recursive Neural Networks (RNNs) for the chatbot's implementation and are generally more complex. A generative chatbot is an open-domain chatbot that uses seq2seq (Sequence-to-Sequence) models to generate its very own responses to input (Adamopoulou & Moussiades, 2020; "Generative Chatbots", n.d.) – making it ideal to mimic human speech.

2.2 Bias in AI

There are many examples of bias in real-world applications of AI, including: Google Photos automatically labeling selfies of Black people as “Gorillas” (Miller, 2017); and an AI powered photo filter app, FaceApp, whitening a user’s face when they apply the “hot” filter (Miller, 2017).

There are also many cases of bias in chatbots, potentially the most infamous being Microsoft’s Tay.

2.2.1 Bias Formation

Microsoft’s Tay was taken down within 24 hours of its creation after it started ‘tweeting’ offensive messages (Larson, 2016; Lee, 2016; Neff & Nagy, 2016; Victor, 2017; Wolf et al., 2017; Zemčík, 2020). Users started ‘attacking’ the chatbot, by sending offensive ’tweets’ and messages (Larson, 2016; Lee, 2016; Victor, 2017). Tay, designed to learn from Twitter users, started creating its own offensive ’tweets’, learning from the messages it received (Larson, 2016; Victor, 2017; Wolf et al., 2017).

It is natural through learning via data observation, machine learning algorithms will develop biases towards certain types of input (Fuchs, 2018). P. Henderson et al. (2018) state chatbots’ susceptibility to bias is due to their subjective nature, and their overall goal to mimic human behaviour.

Leavy et al. (2020) highlights every dataset used to mimic humans do so per a world-view/ideology reflected in the humans the data was collected from. Miller (2017) states bias in AI is usually unintentional and often creeps in via unintentionally biased training sets. Data used to train models is usually obtained from online chat platforms like Reddit and Twitter, which due to their sheer volume are impossible to hand-filter (P. Henderson et al., 2018). These datasets can include subtle biases, which, if not removed/filtered, are then encoded (P. Henderson et al., 2018).

2.2.2 Mitigating Bias

To detect/mitigate bias, Bellamy et al. (2019) developed an open-source Python toolkit, AI Fairness 360 (AIF360).

When combating bias, algorithms used can be split into three categories: pre-processing; in-processing; and post-processing.

2.2.2.1 Pre-processing Algorithms

Reweighing, proposed by Calders et al. (2009), attaches different weights to objects found in the dataset(s) used, the higher the weight of an object, the more expensive it is to get wrong. This algorithm assigns weights to remove dependencies between a predetermined 'sensitive' attribute and the Class attribute (positive, $+$, or negative, $-$) (Calders et al., 2009).

2.2.2.2 In-processing Algorithms

Kamishima et al. (2012) introduced a prejudice remover, which is a regulariser that attempts to reduce the 'prejudice index'. Prejudice index is calculated from a dataset, D, such that $D = (y, x, s)$, where y are random variables corresponding to a class, x are 'non-sensitive' features, and s is a 'sensitive' feature. The bias is calculated using a Calders-Verwer (CV) score, which must be supplied a 'sensitive attribute' (Calders & Verwer, 2010; Kamishima et al., 2012).

2.2.2.3 Post-processing Algorithms

Discrimination-Aware Ensemble (DAE) is an ensemble of classifiers made discrimination-aware by exploiting the disagreement region amongst the classifiers (Kamiran et al., 2012). Traditionally, a classifier ensemble will classify instances by assigning the class label held by the majority. With DAE, if all classifiers predict the same label, then it is assigned, otherwise, instances belonging to a predefined deprived group are given a boosting label (C^+), those

in a favoured group are penalised (C^-) (Kamiran et al., 2012).

2.2.3 Critiques of Bias Mitigation

The presented approaches attempt to mitigate bias by focusing on developing fairness-aware machine learning algorithms (Leavy et al., 2020). All of these algorithms have displayed an ability to remove bias (Bellamy et al., 2019; Calders et al., 2009; Kamiran et al., 2012; Kamishima et al., 2012). However, none of them cover hidden biases – all algorithms above require predefined 'sensitive' attributes and/or 'deprived/favoured' groups, thus focusing on explicit biases found in everyday language.

2.2.4 Conclusions

It is clear objectivity in data-driven AI is unrealistic (Leavy et al., 2020); Dignum (2019) states "Therefore, bias is inherent in human thinking and an unavoidable characteristic of data collected from human processes" (p. 60). Bias mitigation algorithms exist, however they require predefined labels for 'sensitive' attributes. However, all humans possess hidden, 'implicit' bias, which cannot be defined in singular terms. With no way to define these implicitly biased attributes, how can bias mitigation algorithms remove this bias?

2.3 Implicit Human Bias

2.3.1 Introduction of Concepts

While explicit bias is bias people are fully aware of, implicit bias is automatic associations of certain stereotypes residing outside of conscious awareness/control ("Bias", n.d.; Jakeman & Clark, 2019; Ruhl, 2020). Implicit bias can sometimes contradict a person's conscious thoughts without them knowing (Ruhl, 2020), for example, an employer implicitly biased against pink hair. This person will likely say they have nothing against those with pink hair, and may say they

like it. However, when choosing between two similar applicants, one with pink hair, and one with 'normal' hair, the employer will choose the 'normal' hair applicant, perhaps (incorrectly) associating the pink-haired applicant with being too 'immature' for the role, despite the applicant showing no other signs of immaturity. A study by Greenwald and Banaji (1995) explores implicit modes of attitudes, stereotypes, and bias, providing the definition: "Implicit attitudes are introspectively unidentified (or inaccurately identified) traces of past experience that mediate favourable or unfavourable feeling, thought, or action toward social objects." (p. 8).

2.3.2 Causes of Implicit Bias

From a young age, children need to reason about people's social memberships (Baron et al., 2014). When doing so, there are two main identifying factors, 'noun labels' and 'visual cues'. In their study, where they conducted four experiments with children and adult participants, Baron et al. (2014) found that with a shared noun label but a lack of visual cue, children as young as four could generalise bad behaviour with new people with the same noun label as those previously associated with bad actions.

While growing up, children as young as three start associating things and people similar to themselves as positive, and those different from them as negative (Cameron et al., 2001). A study by Sinclair et al. (2005) showed the more a child identified with their parent(s), the more the child's biases corresponded to their parents'.

TV's portrayal of individuals can also ingrain biases (Ruhl, 2020). For example, the popular portrayal of Black people as criminals, or females as teachers can cause automatic associations later relied upon in everyday life (Ruhl, 2020).

An important point to make is children are not born with any bias/prejudices, children are born unable to form such prejudices. Instead, children learn them, as discussed above.

2.3.3 Identifying Implicit Bias

An Implicit Association Test (IAT) can be used to measure implicit bias (Greenwald et al., 1998). Two target concepts appear as choices for a task, with each concept being accompanied by an evaluation attribute (e.g., pleasant vs unpleasant words). When highly associated categories (e.g., “flower” + “pleasant”) share a response key, users perform faster than if less associated categories (e.g., “insect” + “pleasant”) share a response key. Therefore, users’ response times can be measured to indicate differences in evaluative associations (Greenwald et al., 1998).

A 2007 study (700,000+ participants), found 70% of White subjects more quickly associated White faces with positive keys, and Black faces with negative keys, showing implicit racial bias (Nosek et al., 2007; Ruhl, 2020).

The IAT is not perfect. A survey conducted by Motzkus et al. (2019) showed 84% of 250 pre-clinical medical students believed bias needs to be acknowledged or recognised, but only 27% believed the IAT was a beneficial tool in acknowledging racial bias. Selmi (2018) suggests recent findings demonstrate a modest connection between the IAT and behaviour, making the value returned useless. In an article written by Bartlett (2017), it was stated that Greenwald (one of the IAT’s creators) does not think the IAT is reliable enough to diagnose something that inevitably results in racism or prejudicial behaviour.

2.3.4 Implications for Chatbots

It has been shown implicit bias is an intricate part of human speech, therefore, it is fair to assume it would be an intricate part of all datasets based on human speech too. Unlike explicit bias, it is hard to detect implicit bias, meaning it would be hard to gather ‘sensitive attributes’ necessary for bias mitigation algorithms to work.

2.4 Conclusion

Research into how implicit biases develop revealed implicit biases can be seen in children as young as three years old, showing this bias is an integral part of human life. Playing such a big role in human speech means traces of implicit bias will be present in the datasets used to train chatbots. Can this implicit bias be detected? And can it be completely removed? Or is implicit bias an integral part of how chatbots make decisions, similar to how it is in humans?

Design

3.1 Design

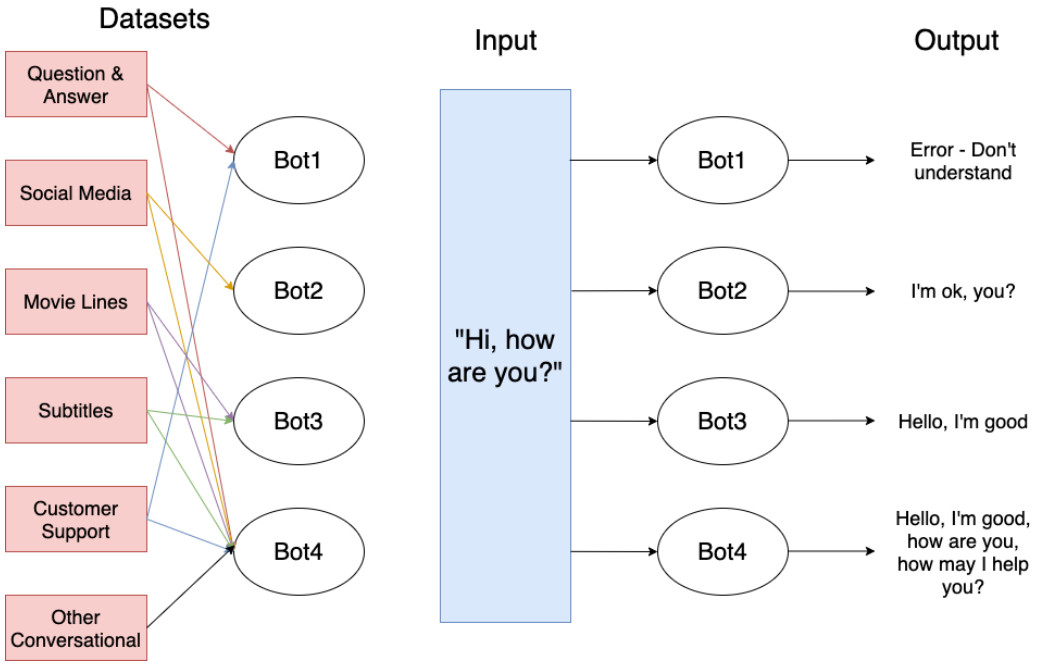

Figure 3.1: The chatbot design. A set input will likely yield different outputs from different chatbots. *Note*: outputs here are dummy values, not actual data.

For this research, four chatbots were created, each trained over a different subset of datasets (Figure 3.1). Generative chatbots were created to allow

them to reveal their own bias, rather than bias present in a corpus of responses.

3.2 Datasets

There are six main groups of datasets detailed above: question and answer; customer support; social media; movie lines; subtitles; and other. Each of these represents different types of data likely encoded into different chatbots (with 'other' being a 'catch-all'). The hope of using different types of datasets was to highlight implicit bias in all four of the chatbots.

The first chatbot was trained over a mixture of question and answer, and customer support datasets, the second chatbot was trained exclusively on datasets built around social media, the third chatbot was trained over datasets containing movie lines and subtitles, and the fourth chatbot was trained over all of the datasets used for this project, including any that did not fit into the groups above.

Each dataset contained human-created data. The primary aim of this research is to show implicit bias found in human-generated data is encoded into chatbots; not using human-generated data would not fulfill this aim.

Implementation

4.1 Chatbot Creation

To create chatbots, a tutorial by Inkawhich (2017) was followed. This tutorial provided code examples capable of creating a trained chatbot that could be interacted with.

4.1.1 Gathering Datasets

Before a chatbot is created, datasets must be obtained. The datasets collected for this project were:

- Amazon [Customer Support] (M. Henderson et al., 2019)
- Convai [Other] (Aliannejadi et al., 2020)
- Twitter Customer Support [Social Media, Customer Support] (Axelbrooke, 2017)
- SQUAD [Question and Answer] (Rajpurkar et al., 2018)
- OpenSubtitles [Subtitles] (M. Henderson et al., 2019)
- Cornell [Movie Lines] (Danescu-Niculescu-Mizil & Lee, 2011)
- QA [Question and Answer] (Smith et al., 2008)

- Reddit [Social Media] (M. Henderson et al., 2019)

Per Section 3, the datasets were assigned to each chatbot as follows:

Chatbot 1:

- Amazon, SQUAD, QA

Chatbot 2:

- Twitter Customer Support

Chatbot 3:

- Cornell

Chatbot 4:

- Amazon, Convai, Twitter Customer Support, SQuAD, Cornell, QA

4.1.1.1 Unused Datasets

From the datasets above, two were unused. The Opensubtitles and Reddit datasets were too large for training, being 13GB+ each. Whilst training, the training script would cause too much data to be processed at once, making the chatbot untrainable.

Being a customer support dataset, Twitter Customer Support fits into both chatbot 1 and chatbot 2. However, due to complications with the Reddit dataset, the Twitter Customer Support dataset was removed from chatbot 1, to allow more variation.

4.1.2 Loading Datasets

Next, the datasets were loaded. The tutorial only loaded the Cornell dataset. To overcome this, some 'general' functions were written to handle processes needed by all datasets, namely: reading the dataset file and writing the formatted pairs to `formatted_lines.txt`. Reading the file was largely the same

```python
def load_files(*filepaths, open_func=open, line_eval_func=None):
    """Loads in dataset files, given filepaths, and optional open and evaluation functions.

    Args:
        filepaths (str): relative filepaths to the datafiles to be loaded.
        open_func (func, optional): Function to open the file, should 'open' not be sufficient. Defaults to open.
        line_eval_func (func, optional): Function to further process the data loaded before it is yielded. Defaults to None.

    Yields:
        iterator: Iterator of the line loaded.
    """
    # open_func allows for different open funcitons, in case the built-in open() funciton is not enough
    # line_eval_func is optional, and allows some evaluation before return. Mostly used for JSON files, where json.loads() is needed
    for file in filepaths:
        print(f"    Loading {file.split('/')[-1]}...")
        with open_func(file) as f:
            for line in f:
                yield line if line_eval_func is None else line_eval_func(line)


def load_csv_files(*filepaths, delimiter=','):
    """Loads in csv files, given filepaths and an optional delimiter.

    Args:
        filepaths (str): relative filepaths to the datafiles to be loaded.
        delimiter (str, optional): Delimiter to use to load csv file. Defaults to ','.

    Yields:
        iterator: Iterator containing the lines
    """
    for file in filepaths:
        print(f"    Loading {file.split('/')[-1]}...")
        with open(file, mode="rb") as f:
            lines = []
            for line in f:
                try:
                    line = line.decode("utf-8")
                except UnicodeDecodeError:
                    continue  # Ignore any lines with non-decodable strings in
                lines.append(line)
            csv_reader = csv.DictReader(lines, delimiter=delimiter)
            for row in csv_reader:
                yield row
```

Figure 4.1: Two functions to load data from generic files, and csv files respectively.

for most datasets, though there were differences between Comma Separated Variable (csv) files and other files, requiring two different functions (Figure 4.1).

Writing data to `formatted_lines.txt` was simple, the pairs were written to the file, with a tab separating the two pair items, and a new line separating different sets of pairs. Each dataset was written to a file named after itself, to avoid confusion.

Finally, new datasets were loaded using unique functions and the above general functions. Since every dataset is initially formatted differently, each needs different logic to format them.

4.1.3 Training over Multiple Datasets

Next, the chatbot needed to be able to train over multiple datasets at once. To achieve this, the `formatted_lines_X.txt` files were combined into a single `formatted_lines_combined.txt` file, which was then used to train the chatbot.

4.1.4 Data Preparation

The data needs to be prepared so the chatbot can understand it. First, a `Voc` (Vocabulary) class is created, responsible for mapping each discrete word to an index.

Then, functions are defined to normalise strings: one function converts Unicode strings to plain ASCII; the other trims the string, removes non-letter characters (except '.', '!' and '?'), and sets the entire string to lowercase. Both functions aim to make it easier for the chatbot to train.

The chatbot expects numerical tensors instead of sentence pairs (a tensor is a multidimensional list). For this, 'mini-batches' were used to store input sentences as tensors. To turn each word into a number, the words are converted to their index, via the `Voc` class. Mini-batches must have a set size, thus `max_length` and `batch_size` variables are made, where `max_length` dictates the maximum length of input sentences, and `batch_size` indicates the number of sentences stored.

One of the key features of a seq2seq model is the ability to take varying input. Therefore, to ensure the input sentences are exactly `max_length` long, they are cut down if too long, and 'zero-padded' if too short. Zero-padding means all elements after the End Of String (EOS) token are set to 0 until the `max_length` is reached.

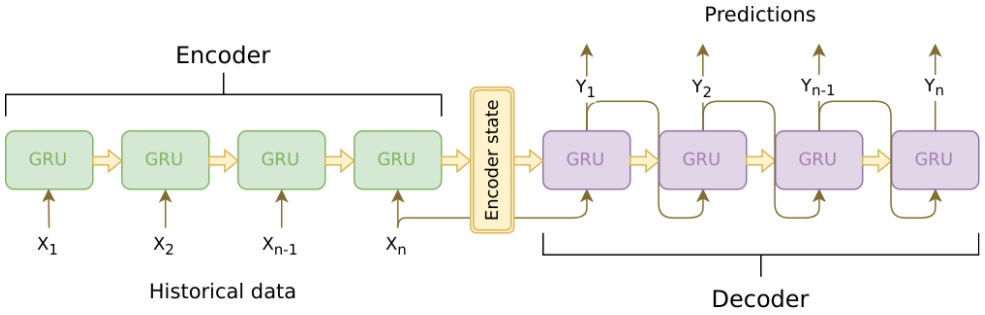

Figure 4.2: Seq2Seq model (Eddy, 2018). Two RNNs are used to form the Seq2Seq, an encoder and a decoder.

4.1.5 Model Definition/Creation

4.1.5.1 Encoder

The encoder RNN iterates through input sentences word-by-word, outputting an 'output' vector and a 'hidden state' vector after each step. The output vector is recorded, while the hidden state vector is passed to the next step of the encoder. The encoder transforms the context it sees at each point into a set of points in high-dimensional space.

The encoder is essentially a multi-layered Gated Recurrent Unit (GRU), first proposed by Cho et al. (2014). A bidirectional variant is used, meaning there are two independent RNNs, one fed the sentence in normal sequential order, and the other the sentence backwards (Figure 4.3). The outputs of each RNN are summed together at each step.

4.1.5.2 Decoder

The decoder RNN will generate the response in a token-by-token fashion. It takes the encoder's context vectors, along with its internal hidden states to generate the next word in the sequence. It will continue generating new words until it outputs an `EOS` token, meaning it has reached the end of the sentence.

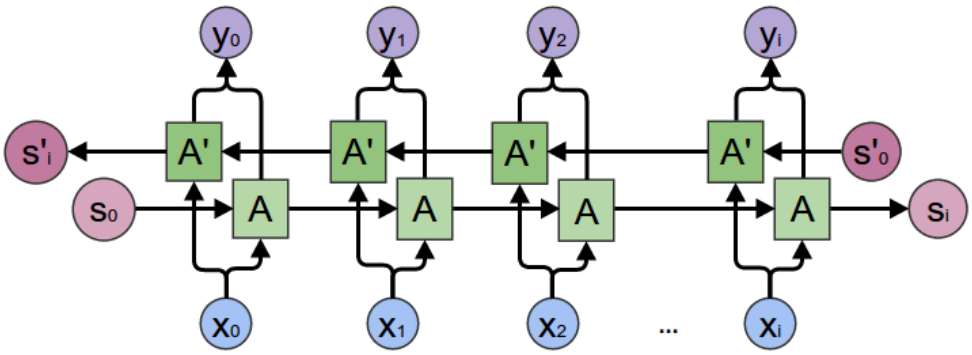

Figure 4.3: A bidirectional Recursive Neural Network (Olah, 2015).

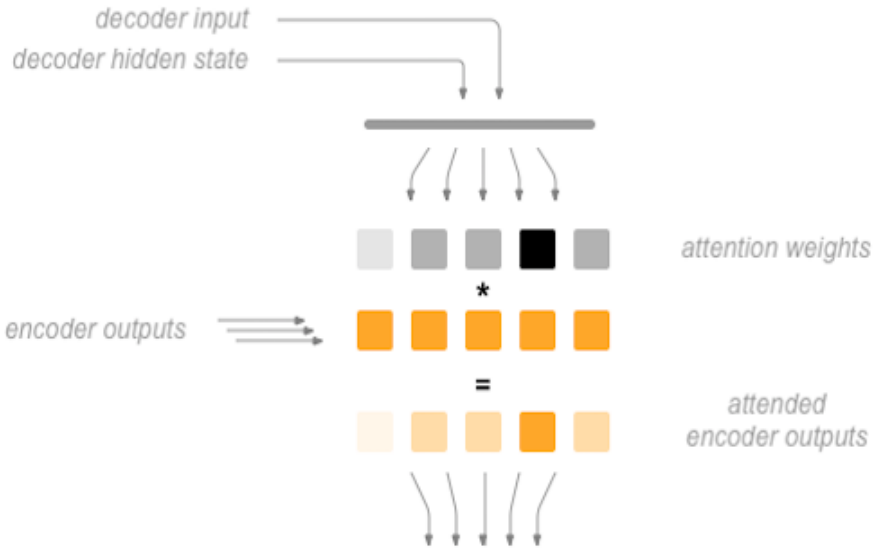

Figure 4.4: A decoder with a built in "attention mechanism" (Inkawhich, 2017).

4.1.6 Training

Because of the variable-length property of the input (and the output), not all of the elements of the tensors can be used when calculating loss. Therefore, a loss function is defined to calculate the loss based on the decoder's output tensor, the target tensor, and a binary mask tensor, which describes the padding of the target tensor. The loss function is used to test how the chatbot is doing, a higher loss value means further away from its target response.

Next, functions for training the chatbot are created. Whilst training, the aim is to get the chatbot's responses to 'converge' from an initial response with a high loss value, down to a response with a much lower loss value.

4.1.7 Evaluation

To talk to the chatbot, a `GreedySearchDecoder` class is defined, whose objects take an input, a scalar input length tensor, and a `max_length`, to choose a word with the lowest loss.

Next, an evaluate function is defined to process the input. First, the sentence is formatted into an input batch of word indexes. A lengths tensor is made, containing the length of the input. Next, the `GreedySearchDecoder` object is used to return a decoded response tensor, containing the index values of the words in the output. Finally, these indexes are converted back into words and returned.

A final function, `evaluateInput` is created. This function creates a Textual User Interface (TUI) for the user to interact with the chatbot. After typing an input sentence and hitting *Enter*, the text is evaluated, and a response generated by the chatbot.

Verification and Validation

5.1 Implicit Association Test Adaptation

5.1.1 Current Form

Shown in Section 2, the IAT is used to detect implicit bias in people. Many bias categories can be tested, though some rely on users' ability to perceive images. Since chatbots cannot 'see' images, these categories cannot be used. Therefore, the biases tested by the IAT that can be used are: gender (both gender-science and gender-career); sexuality; Arab-Muslim; and religion.

For simplicity's sake, only the gender-careers category was adapted for use by the chatbots. The average bias found in IAT users can be found in Figure 5.1, whilst the researcher's result can be found in Figure 5.2.

There are three sections to the IAT. The first is a demographic section, which has no impact on the result. The second is a questionnaire with multiple-choice questions, specific to the category taken. The final section is the association section. The introduction screen shows the user four tags that are used, in this case: 'male', 'female', 'career', and 'family'. Next to each of these are some related words/names. This screen is followed by 7 sub-sections; for each sub-section, two categories contain 1-2 of the four tags. Words from

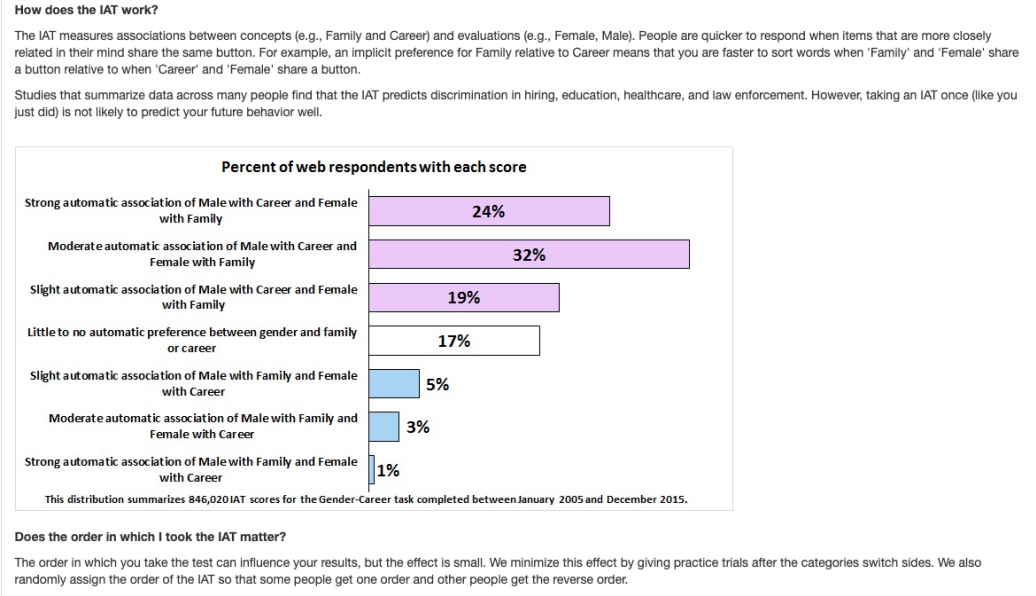

Figure 5.1: The average IAT results for users taking the Gender-Career test (as of May 14, 2021).

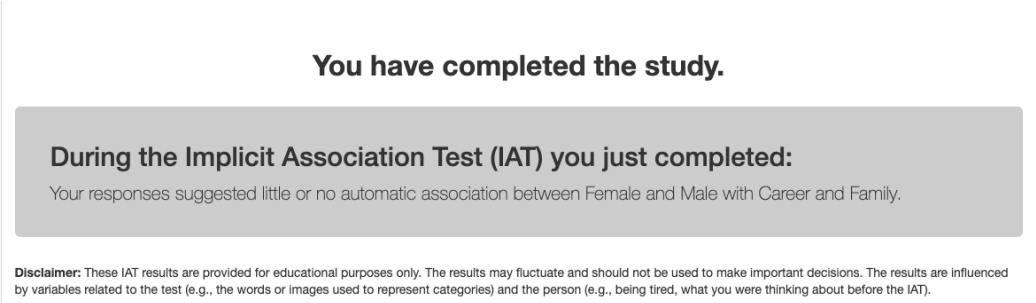

Figure 5.2: The researcher's personal results from taking the Gender-Career IAT on May 14, 2021.

the introduction screen appear in the middle of the screen, and the user must categorise them as quickly as possible.

5.1.2 Adaptation

5.1.2.1 Demographics and Questionnaire Section

The demographics and questionnaire sections are a set of predefined questions, thus not needing much adaptation. The demographics questions do not affect the result of the IAT and offer a 'Decline to Answer' option, so will be ignored. The questionnaire section will be adapted for chatbots. There are a few issues that arise from taking the questions out and directly asking the chatbot.

The first is the fixed nature of the multiple response options. Since the chatbots are generative, there is no easy way to restrict the response the chatbots can give. However, it is very feasible the chatbots will give a similar answer or one that fits only one answer.

Another issue is the possibility words used in the questions do not exist in the chatbots' dataset. The way the chatbot is trained means any words not recognised cause an error message. To fix this, the questions will be paraphrased so the chatbot can answer. Paraphrasing the question may remove subtle meaning behind the question, which could affect the result. To minimise this, only words unrecognised by the chatbot will be changed, if possible.

5.1.2.2 Association Section

This section is a lot harder to adapt. There is a lot of context required inaccessible to the chatbot. There is no good way to ask the chatbot to categorise words that pop up in the middle of the screen into two different categories that also appear on separate sides of the screen. To mitigate this, the chatbots will be asked questions forcing them to categorise the word.

Another issue is the amount of time the user takes to categorise a word fac-

tors into their eventual bias score. Two separate issues arise from this. The first is chatbots usually reply instantaneously. The second is chatbots, being a textual interface, cannot itself take the IAT on a web-based platform. Therefore, the researcher must take the test for the chatbot. This means the test will gather information based on the chatbot's answers, but the researcher's timings, which could interfere with the results. To mitigate this, the researcher will ask the chatbot appropriate questions beforehand. Thus, the researcher can enter the chatbot's answers as quickly as possible, limiting interference.

5.1.2.3 Unrelated/Non-associable Responses

An issue faced within all three sections of the IAT is a chatbot returning a completely unrelated answer, or one not associated with a specific answer/category. In this case, more questions will be asked until a response can be categorised. Should the chatbot not give an appropriate answer after several questions, then a random response/category will be used.

Evaluation

6.1 Chatbot Implicit Association Test

6.1.1 General

Trying to get each chatbot to answer questions posed in a certain way was difficult, almost always requiring multiple follow-up questions to be asked. Some of the answers were completely unrelated, and even those related could not always be definitively categorised.

6.1.2 Chatbot 1

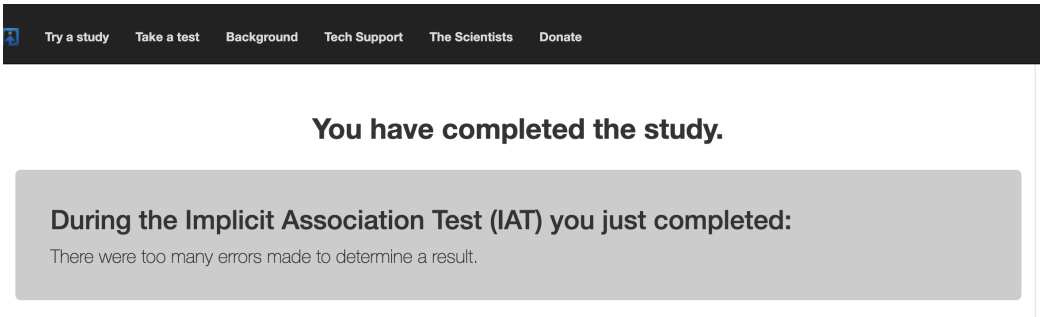

Figure 6.1: Chatbot 1 IAT result – inconclusive.

Throughout the questionnaire section, chatbot 1 was unimpressive. For seven out of twelve questions, further questions needed to be asked for a multiple-

choice answer to be picked. Some of the answers returned by the chatbot made no sense at all.

The association section was more difficult. Chatbot 1 failed to make associations for eighteen of the twenty-two words used by the IAT test, only managing to associate the words ‘parents’, ‘children’, ‘Paul’ and ‘Julia’. The category chosen for unassociated words was randomly chosen.

Due to lack of association provided by the chatbot, there were a lot of incorrectly categorised words, resulting in inconclusive results (Figure 6.1).

6.1.3 Chatbot 2

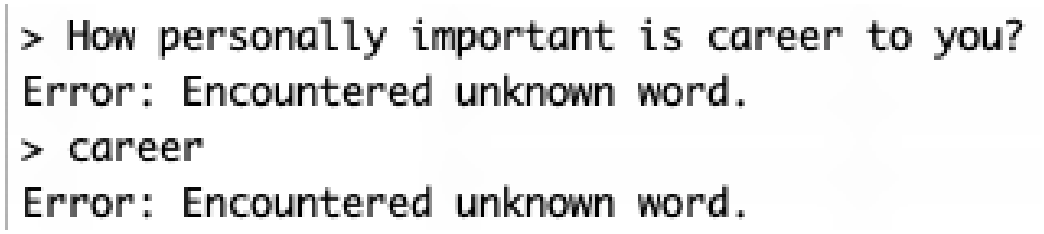

Figure 6.2: Chatbot 2 failure – it failed to understand the word ’career’.

Chatbot 2 was deemed invalid for this research. In Section 5 it was decided only the ’Gender-Career’ IAT would be used. After asking initial questions, it became clear the word ‘career’ was not part of chatbot 2’s training data, therefore it was unable to understand the word (Figure 6.2). Since a lot of interrogation would relate to career, the chatbot was unable to carry on.

6.1.4 Chatbot 3

Getting responses that could be associated with exactly one of the multiple-choice answers in the questionnaire section was quite difficult. The chatbot needed more questions for all but two of the eleven questions asked to get a valid response.

Though not the primary focus of this research, it is interesting to note chatbot

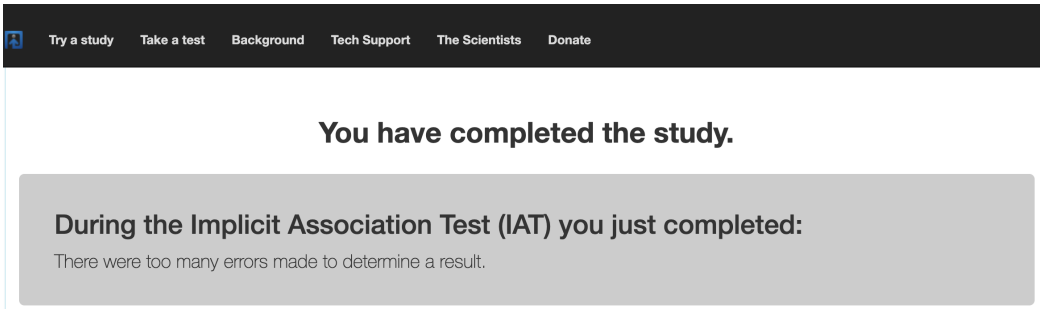

Figure 6.3: Chatbot 3 IAT result – inconclusive.

3 displayed some signs of explicit bias in this section. When asked the question “Do you associate career with men?”, the chatbot replied “i am . . . !” (assumed to be ‘I do’), yet when asked “Do you associate career with women”, the chatbot said, “no . . . . !”.

When it came to the association section, chatbot 3 was much better at associating words into the “female” and “career” categories than the “male” and “family” categories. Out of the five female names, the chatbot was able to correctly associate three, compared to just one of the male names. For the six career words, the chatbot was able to correctly associate five of them but was unable to associate any of the family words to the family category.

Due to the high number of categorisations chatbot 3 made, there were not many random categorisations made by the researcher. Furthermore, a number of the categorisations made by the chatbot were correct, especially when related to “female” and “career”. Despite this, there were still too many ‘mistakes’ made, meaning the IAT result was inconclusive (Figure 6.3).

6.1.5 Chatbot 4

It was difficult to get a valid response from chatbot 4 for the questionnaire section; of the eleven questions asked, further questions were needed for ten.

Out of twenty-two words in the association section of the IAT, chatbot 4 was

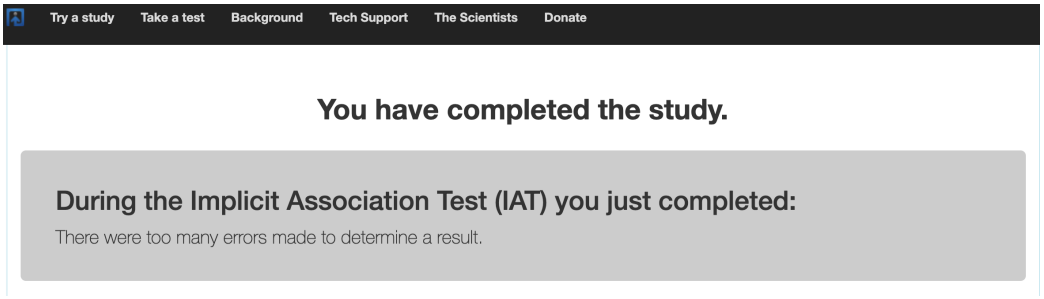

Figure 6.4: Chatbot 4 IAT result – inconclusive.

able to associate two of them to a category – both were wrong. This made this section of the IAT very difficult to fill out since random associations were made for twenty of twenty-two words.

Inevitably, the IAT results were inconclusive, due to too many errors (Figure 6.4).

6.2 Discussion of Results

Of the four chatbots used to answer the research question, three had inconclusive results, whilst the fourth was discounted entirely.

Discounting chatbot 2, all chatbots struggled to associate words into categories provided. Each required multiple questions, and the majority of word-category associations had to be randomly assigned after the chatbot was unable to form an association.

Whilst it seems the results are invalid, the opposite is true. One of the aims of this research was to determine whether it was possible to detect implicit bias within a chatbot. In answering this question, another question was posed: “is it possible to adapt the IAT test for use on chatbots?”. The results and findings from the four chatbots used show it is likely impossible to simply adapt the IAT for chatbots.

Conclusion

7.1 Overall Conclusions

Although the research question “Can Chatbots Truly Be ‘Unbiased’?” was not answered by the findings detailed in Section 6, the outcome of this project is not a failure. This project is seen as the first step into a new field of AI bias, linking psychological concepts of implicit biases to those in every human-data-driven chatbot.

Explorations into human and AI bias unearthed implicit bias. With this came the question, how can one show the existence of implicit bias? Implicit bias is shown to be difficult to spot, despite having massive effects. When it comes to identifying this bias in people, the IAT has been developed. However, there is very little research linking implicit bias with chatbots, meaning there is no test allowing its detection. This raised another question: is it possible to detect implicit bias within chatbots?

With this in mind, four generative chatbots were created and an adaptation was made to the IAT. Section 6 evaluated the results, showing all four chatbots came back inconclusive. This was due mostly to the inability of the chatbots to associate words to a predefined category.

As disappointing as the results were, they did not indicate the project as a failure. Instead, the entire project built a foundation for new research. Links

have been made between implicit bias and chatbots that had not been made before, and the data collected shows the difficulty in adapting the current IAT for use with chatbots. This project exposes complexities and subtleties associated with detecting implicit bias in chatbots, contributing to both the academic debates of implicit bias as a whole and bias within AI/Machine Learning.

7.2 Future Work and Recommendations

Firstly, an answer to the question 'can implicit bias be detected in chatbots?' must be answered. Without this, there is no hope of knowing if an 'unbiased' chatbot is possible. It is recommended other options besides the IAT test are explored. The idea of implicit bias is somewhat new in the field of psychology and has not been linked to the field of AI. New research could bring together the psychology and computer science fields to create a new test, capable of showing implicit bias in both people and chatbots.

To test chatbots for implicit bias, more competent chatbots would be required. All four chatbots in this research struggled to form associations. It is recommended more experimentation is done when developing chatbots, to create more advanced, realistic chatbots, which will be easier to test.

Described in Section 2, there are many ways to mitigate bias within chatbots. However, it is unknown whether these bias mitigation techniques could remove implicit bias. Research should be conducted to determine whether this is the case.

Links between human and chatbot implicit bias were formed due to the human data used when training a chatbot, but chatbots are not the only human-data-driven AI application. AI is being used in more and more fields, including the healthcare industry. Most healthcare research is conducted on male bodies and physiques (Allday, 2013; Zucker & Beery, 2010). Should a piece of AI be trained using the results from this biased research data, the AI would likely become just as biased. This bias could have severe consequences, including

misdiagnoses. Therefore, research into implicit bias in other applications is also necessary.

Bibliography

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*(2666–8270), 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., & Burtsev, M. (2020). ConvAI3: Generating Clarifying Questions for Open-Domain Dialogue Systems (ClariQ). https://arxiv.org/abs/2009.11352

Allday, E. (2013). Medical research has focused on males. Retrieved May 28, 2021, from https://www.sfgate.com/health/article/Medical-research-has-focused-on-males-4330773.php

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias*. Retrieved March 24, 2021, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Axelbrooke, S. (2017). *Customer Support on Twitter* (Version 10). https://www.kaggle.com/thoughtvector/customer-support-on-twitter/version/10

Baron, A., Dunham, Y., Banaji, M., & Carey, S. (2014). Constraints on the Acquisition of Social Category Concepts. *Journal Of Cognition And Development*, *15*(2), 238–268. https://doi.org/10.1080/15248372.2012.742902

Bartlett, T. (2017). Can We Really Measure Implicit Bias? Maybe Not. Retrieved March 24, 2021, from https://www.chronicle.com/article/can-we-really-measure-implicit-bias-maybe-not/

Bellamy, R. K. E., Dey, K., Hind, D., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovič, A., Nagar, S., Natesan Ra-

mamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal Of Research And Development*, *63*(4/5), 4:1–4:15. https://doi.org/10.1147/jrd.2019.2942287

Bias. (n.d.). Retrieved February 1, 2021, from https://www.psychologytoday.com/intl/basics/bias

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building Classifiers with Independency Constraints. *2009 IEEE International Conference On Data Mining Workshops*. https://doi.org/10.1109/icdmw.2009.83

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292. https://doi.org/10.1007/s10618-010-0190-x

Cameron, J., Alvarez, J., Ruble, D., & Fuligni, A. (2001). Children's Lay Theories About Ingroups and Outgroups: Reconceptualizing Research on Prejudice. *Personality And Social Psychology Review*, *5*(2), 118–128. https://doi.org/10.1207/s15327957pspr0502_3

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. http://arxiv.org/abs/1406.1078v3

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Dignum, V. (2019). *Responsible Artificial Intelligence*. Springer. https://doi.org/10.1007/978-3-030-30371-6

Eddy, J. (2018). Seq2Seq Model [Image]. Retrieved June 6, 2021, from https://jeddy92.github.io/ts_seq2seq_intro/

Fuchs, D. (2018). The Dangers of Human-Like Bias in Machine-Learning Algorithms. *Missouri S&T's Peer To Peer*, *2*(1). https://scholarsmine.mst.edu/peer2peer/vol2/iss1/1

Generative Chatbots. (n.d.). Retrieved May 28, 2021, from https://www.codecademy.com/learn/deep-learning-and-generative-chatbots/modules/generative-chatbots/cheatsheet

Greenwald, A. G., & Banaji, M. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295x.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Henderson, M., Budzianowski, P., Casanueva, I., Coope, S., Gerz, D., Kumar, G., Mrkšić, N., Spithourakis, G., Su, P.-H., Vulic, I., & Wen, T.-H. (2019). A repository of conversational datasets [Data available at github.com/PolyAI-LDN/conversational-datasets]. *Proceedings of the Workshop on NLP for Conversational AI*. https://arxiv.org/abs/1904.06472

Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., & Pineau, J. (2018). Ethical Challenges in Data-Driven Dialogue Systems. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 123–129. https://doi.org/10.1145/3278721.3278777

Inkawhich, M. (2017). *Chatbot Tutorial – PyTorch Tutorials 1.8.1+cu102 documentation*. Retrieved December 4, 2020, from https://pytorch.org/tutorials/beginner/chatbot_tutorial.html?highlight=chatbot

Jackman, J. (2017). *Google's new artificial intelligence bot thinks gay people are bad*. Retrieved March 24, 2021, from https://www.pinknews.co.uk/2017/10/26/googles-new-artificial-intelligence-bot-thinks-gay-people-are-bad/

Jakeman, K., & Clark, M. (2019). The neutral person: paradox-accepting and addressing unconscious bias in mediation. *Advocate (Vancouver Bar Association)*, *77*(5), 695–702.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision Theory for Discrimination-Aware Classification. *2012 IEEE 12Th International Conference On Data Mining*, 924–929. https://doi.org/10.1109/icdm.2012.45

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. *Machine Learning And Knowledge Discovery In Databases*, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

Larson, S. (2016). *Microsoft's racist robot and the problem with AI development*. Retrieved December 14, 2020, from https://www.dailydot.com/debug/tay-racist-microsoft-twitter/

Leavy, S., O'Sullivan, B., & Siapera, E. (2020). Data, Power and Bias in Artificial Intelligence. *CoRR*, *abs/2008.07341*. https://arxiv.org/abs/2008.07341

Lee, P. (2016). *Learning from Tay's introduction - The Official Microsoft Blog*. Retrieved December 14, 2020, from https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

McCurry, J. (2021). *South Korean AI chatbot pulled from Facebook after hate speech towards minorities*. Retrieved May 14, 2021, from https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook

Miller, D. (2017). *Design biases in Silicon Valley are making the tech we use toxic, expert says*. Retrieved January 17, 2021, from https://www.abc.net.au/news/2017-10-23/toxic-tech-bias-and-algorithmic-racism-in-our-technology/9042288

Motzkus, C., Wells, R. J., Wang, X., Chimienti, S., Plummer, D., Sabin, J., Allison, J., & Cashman, S. (2019). Pre-clinical medical student reflections on implicit bias: Implications for learning and teaching. *PLOS ONE*, *14*(11), e0225058. https://doi.org/10.1371/journal.pone.0225058

Neff, G., & Nagy, P. (2016). Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal Of Communication*, *10*(1932-8036), 4915–4931.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Psychology Press.

Olah, C. (2015). Bidirectional Recursive Neural Networks [Image]. Retrieved June 1, 2021, from https://colah.github.io/posts/2015-09-NN-Types-FP/

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *CoRR*, *abs/1806.03822*. http://arxiv.org/abs/1806.03822

Ruhl, C. (2020). *Implicit or Unconscious Bias*. Retrieved March 23, 2021, from https://www.simplypsychology.org/implicit-bias.html

Schiffer, Z. (2020). *This girls-only app uses AI to screen a user's gender – what could go wrong?* Retrieved May 28, 2021, from https://www.theverge.com/2020/2/7/21128236/gender-app-giggle-women-ai-screen-trans-social

Selmi, M. (2018). THE PARADOX OF IMPLICIT BIAS AND A PLEA FOR A NEW NARRATIVE. *Arizona State Law Journal*, *50*(1), 193–245.

Sinclair, S., Dunn, E., & Lowery, B. (2005). The relationship between parental racial attitudes and children's implicit prejudice. *Journal Of Experimental Social Psychology*, *41*(3), 283–289. https://doi.org/10.1016/j.jesp.2004.06.003

Smith, N. A., Heilman, M., & Hwa, R. (2008). Question generation as a competitive undergraduate course project. *Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, 4–6.

Victor, D. (2017). *Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk*. Retrieved December 14, 2020, from

https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html

Wolf, M., Miller, K., & Grodzinsky, F. (2017). Why We Should Have Seen That Coming. *The ORBIT Journal*, *1*(2), 1–12. https://doi.org/10.29297/orbit.v1i2.49

Zemčík, T. (2020). Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI & SOCIETY*, *36*(1), 361–367. https://doi.org/10.1007/s00146-020-01053-4

Zucker, I., & Beery, A. K. (2010). Males still dominate animal studies. *Nature*, *465*(7299), 690–690. https://doi.org/10.1038/465690a