A Survey on Bias in Deep NLP

Ismael Garrido-Muñoz 1,†, Arturo Montejo-Ráez 2,†*, Fernando Martínez-Santiago 3,† and L. Alfonso Ureña-López 4,†

1 Centro de Estudios Avanzados en TIC (CEATIC); igmunoz@ujaen.es
2 Centro de Estudios Avanzados en TIC (CEATIC); amontejo@ujaen.es
3 Centro de Estudios Avanzados en TIC (CEATIC); dofer@ujaen.es
4 Centro de Estudios Avanzados en TIC (CEATIC); laurena@ujaen.es
* Correspondence: amontejo@ujaen.es; Tel.: +34 953 212 882
† These authors contributed equally to this work.

Abstract: Deep neural networks are hegemonic approaches to many machine learning areas, including natural language processing (NLP). Thanks to the availability of large corpora collections and the capability of deep architectures to shape internal language mechanisms in self-supervised learning processes (also known as "pre-training"), versatile and performing models are released continuously for every new network design. But these networks, somehow, learn a probability distribution of words and relations across the training collection used, inheriting the potential flaws, inconsistencies and biases contained in such a collection. As pre-trained models have found to be very useful approaches to transfer learning, dealing with bias has become a relevant issue in this new scenario. We introduce bias in a formal way and explore how it has been treated in several networks, in terms of detection and correction. Also, available resources are identified and a strategy to deal with bias in deep NLP is proposed.

Keywords: natural language processing; deep learning; biased models

0. Introduction

In sociology, bias is a prejudice in favor or against a person, group, or thing that is considered to be unfair. Since, on one hand, it is a extremely pervasive phenomena, and on the other hand, deep neural networks are intended to discover patterns in existing data, it is known that human-like semantic biases are found when applying machine learning to ordinary human related results, such as computer vision [1], audio processing [2] and text corpora [3,4]. All these fields are relevant as constituents of automated decision systems. An “automated decision system” is any software, system, or process that aims to automate, aid, or replace human decision-making. Automated decision systems can include both tools that analyze datasets to generate scores, predictions, classifications, or some recommended action(s) that are used by agencies to make decisions that impact human welfare, which includes but is not limited to decisions that affect sensitive aspects of life such as educational opportunities, health outcomes, work performance, job opportunities, mobility, interests, behavior, and personal autonomy.

In this context biased artificial intelligence models may make decisions that are skewed towards certain groups of people in these applications [5]. Obermeyer et al. [6] found that an algorithm widely used in US hospitals to allocate health care to patients has been systematically discriminating against black people, since it was less likely to refer black people than white people who were equally sick to programmes that aim to improve care for patient with complex medical needs. In the field of computer vision, some face recognition algorithms fail to detect faces of black users [7] or labelling black people as “gorillas” [1]. In the field of audio processing, it is found that voice-dictation systems recognize a voice from a male more accurately than that from a female [2]. Moreover, regarding with predicting criminal recidivism, risk assessment systems are likely to predict that people of some certain races are more presumably to commit a crime [8].
In the field of deep Natural Language Processing (deep NLP), Word embeddings and related language models are massively used nowadays. These models are often trained on large databases from the Internet and may encode stereotyped biased knowledge and generate biased language. Such is the case of dialog assistants and chatbots when using biased language[9], or resume-review systems that ranks female candidates as less qualified for computer programming jobs because of biases present in training text, among other NLP applications. Caliskan et al. [10] propose the Word Embedding Association Test (WEAT) as a way to examine the associations in word embeddings between concepts captured in the Implicit Association Test (IAT) [11], in the field of social psychology, intended to assess implicit stereotypes held by test subjects, such as unconsciously associating stereotyped black names with words consistent with black stereotypes.

This problem is far from being solved, or at least attenuated. Currently, there are no standardized documentation procedures to communicate the performance characteristics of language models in spite of some efforts to provide transparent model reporting such model cards [12] or Data Statements [13]. Besides, the new models use document collections that are getting larger and larger during their training, and they are better able to capture the latent semantics in these documents, it is to be expected that biases will become part of the new model. This is the case of GPT-3 [14], an state-of-the-art contextual language model. GPT-3 uses 175 billion parameters, more than 100x more than GPT-2 [15], which used 1.5 billion parameters. Thus, Brown et al. [14] report findings in societal bias, more concisely regarding gender, race and religion. Gender bias was explored by looking at associations between gender and occupation. They found that 83% of 388 occupations tested were more likely to be associated with a male identifier by GPT-3. In addition, professions demonstrating higher levels of education (e.g. banker, professor emeritus) were heavily male leaning. On the other hand, professions such as midwife, nurse, receptionist, and housekeeper were heavily female leaning. Racial bias was explored by looking at how race impacted sentiment. The result: Asian race had a consistently high sentiment, while Black race had a consistently low sentiment. Finally, religious bias was explored by looking at which words occurred together with religious terms related to the following religions. For example, words such as “violent”, “terrorism”, and “terrorist” were associated with Islam at a higher rate than other religions. This findings is consistent with the work reported in [16]. When GPT-3 is given a phrase containing the word “Muslim” and asked to complete a sentence with the words that it thinks should come next, in more than 60% of cases documented by researchers, GPT-3 created sentences associating Muslims with shooting, bombs, murder, or violence.

This paper provides a formal definition of bias in NLP and a exhaustive overview on the most relevant works that have tackle the issue in the recent years. Main topics in bias research are identified and discussed. The rest of this paper is structured as follows: firstly, it is introduced a formal definition of bias, and its implication in machine learning in general, and language models in particular. Then, we present a review of the state-of-art in bias detection, evaluation and correction. Section 4 is our proposal for a general methodology for dealing with bias in deep NLP and more specifically in language model generation and application. We finalize with some conclusions and identify main research challenges.

1. Defining bias

The study of different types of bias in cognitive sciences has been done for more than four decades. Since the very beginning, bias has been found as a innate human strategy for decision making [17]. When a cognitive bias is applied, we are presuming reality to behave according to some cognitive priors that may are not true at all, but with which we can form a judgment. A bias can be acquired by an incomplete induction process (a limited view over all possible samples or situations) or learned from others (educational or observed). In any case, a bias will provide a way of thinking far from logical reasoning [18]. There are more than one hundred cognitive biases identified, which can be classified in several domains like social, behavioral, memory related and many more. Among them, there is one that we will focus on: stereotyping.
If a cognitive bias can be defined as a case in which human cognition reliably produces representations that are systematically distorted compared to some aspect of objective reality [19], stereotyping can be defined as the assumption of some characteristics applied to others on the basis of their national, ethnic or gender groups [20]. Therefore, stereotyping assigns certain characteristics to an individual because that individual pertains to a certain group. Somehow, it is like an ontology were certain classification rules are applied (so certain properties are presumed, like ignorance, weaknesses or criminal behavior) just because the individual possesses one specific value for a given property (she holds the “female” value for the property “gender”, or he holds the “African” value for the property “ethnicity”). As can be seen, stereotyping can be modelled at semantic level using a formal scheme like those provided by ontology languages in knowledge engineering.

We will first introduce fairness, as it is a well-know concept in machine learning (as it is, actually, equivalent to “zero-biases” systems), along with some of the measures used for its treatment. We will then discuss how fairness measures can help us to approach the bias problem in language models. To end this section, our proposal for a formal definition is provided.

1.1. The bias problem in machine learning

A concept that is intimately associated with bias is fairness. A system is considered to be “fair” when its outcomes are not discriminatory according to certain attributes, like gender or nationality. In machine learning evaluation, discrimination can be estimated looking at the confusion matrices for different protected groups. That is, we can compute confusion matrices and derived rates (positive rates, true positive rates, false positive rates, and so on) for each subset of samples obtained as a segmentation of the full collection of samples on a certain feature (like “gender”). If these rates are far from being equal, that is a potential evidence of a prediction system with an “unfair” behavior, i.e. with a clear bias on how decisions are made depending on the values of that certain feature.

Several measures have been proposed to study divergences among prediction rates over different population groups, and how to interpret them according to each system goal is now clearly identified [21]. From the large amount of biases derived from cognitive ones, about a pair of dozens are of interest in machine learning problems [5]. These latter two studies compile several measures that have been agreed in the analysis of the bias problem in machine learning systems, those measures are demographic parity, equal opportunity, equalised odds or counterfactual fairness, among others. Of course, these measures can be applied in many artificial intelligence subareas, like image recognition or natural language processing, Let’s see the definition of one of them (demographic parity), as some elements can be transferred to our formal definition of bias in language modelling.

Demographic parity states that all the groups resulting from the different values of a protected class (e.g. gender) should receive the same rate of positive outcomes [22]. For example, if the system decides to concede a scholarship with the same rate to people in both male and females groups, then system shows demographic parity. Let \( \hat{Y} \) be the predicted decision on whether a scholarship should be granted \( (\hat{Y} = 1) \) or denied \( (\hat{Y} = 0) \). Then, demographic parity can be defined as \( P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1) \), which is equivalent to equal positive rates for both male and females \( PR(A = 0) = PR(A = 1) \). Here, \( \hat{Y} \) is the system prediction and \( A \) is the “protected” attribute/class. In our example, this is the gender and its possible values are 0 for male or 1 for female. Of course, this measure can be generalize to any protected class, like ethnicity or nationality. In that case, fairness is granted if positive rates for all possible population segments are equal. Where is bias here? It is right there, as the bias would be the deviation between groups resulting from different values of the protected attribute. Thus, the bias would be \( bias = |P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)| \), which is equal to \( bias = |PR(A = 0) - Pr(A = 1)| \) using demographic parity as estimator. Equal opportunity is a good estimator of fairness. This one considers the equality between true positive rates. The rest of measures are, as pointed out, variants on what we want to be equal from different scores.

In general, fairness is computed over the distribution \(< X, A, Z, Y, \hat{Y} >\), referring \( X \) to samples, \( A \) to protected attribute, \( Z \) to rest of attributes, \( Y \) the true labels for those samples and \( \hat{Y} \) to the predicted
labels by the model. This clear definition of fairness and how it is evaluated allows the introduction of correction mechanisms in the learning process, like those implemented in the FairTorch library \(^1\). This way of approaching bias correction is close to what is known as statistical bias, as we have seen. We introduce the minimization of the bias as an additional constraint in the learning process.

Fairness is not a cognitive bias, this is something related to the estimation of parameters in statistical modelling, which is what neural networks do. But fairness, is somehow, the formalisation of measures to reduce stereotyping in machine learning. According to Wikipedia\(^2\), a statistical bias is a feature of a statistical technique or of its results whereby the expected value of the results differs from the true underlying quantitative parameter being estimated. Fairness measures are, actually, measures of a statistical bias.

Therefore, whenever a protected feature is clearly identified or can be derived from sample features in the training set, it is possible to evaluate model on its equity for generating similar distributions of predictions over groups resulting from different values of the protected feature. Even when in natural language processing many tasks can be defined in terms of machine learning, the challenge is when the protected attribute is not a clear feature in the dataset. How to define bias/fairness when pre-training? How can we measure fairness over models like GPT-2 or BERT which have been trained following a language modeling approach? We propose an answer to this question in the next section.

1.2. A reflection on bias in language models

A language model (LM) estimates the probability of a sequence of words \(P(w_1, ..., w_m)\). This allows for, given a sequence of words, estimating the next most probable word. The machinery behind the learning of model parameters can be used for solving many different tasks, like machine translation, text generation, text classification or token labeling (as for named entity recognition), among others \[23\]. Bias is present in language models as it is present in humans. Bias is intrinsic to human language, and it is not always source of unfairness. A car full of breakdowns is prone to accidents; fans of sci-fi movies are willing to watch similar movies; a patient with a chronic disease could have more risk of worsening, an so on. What we mark as "unfair" is established at a high semantic level. Remember that bias is not about prediction error, it is about skewed behavior regarding semantic expectations.

**Definition 1.** The stereotyping bias in a language model is the undesired variation of the probability distribution of certain words in that language model according to certain prior words in a given domain.

Those prior words are terms that can be linked to a protected attribute. Staying within the "gender" domain, those terms could be actress, woman, girl, etc. That is, in a language model, we expect the distribution of probabilities after word woman to be equal (or very close) to that of the word man for certain words, like those related to professional skills. Both, man and woman are certain words in the gender domain (the protected attribute). It raises the problem of defining precisely the domain and those expected "certain" words. Following this example, the words within the gender domain would be split into two different classes where stereotypes are willing to occur, one class for men (actor, waiter...) and another class for women (actress, waitress...). Then, words regarding, in this case, attributes on professional skills (intelligence, efficiency, cleanliness, creativity...) could be used to analyzed how they appear over the different probability distributions associated to each class, that is, when words in the domain are present, as priors of the distributions. So, we could identify that the probability of word "creativity" in the presence of a man is different from that of a woman.

This language modeling based approach to the bias phenomena makes clear that bias is not a fault of the language model by itself, it is just the effect of the data from which this model was generated and

---

\(^1\) https://fairtorch.github.io/FairTorch/

\(^2\) https://en.wikipedia.org/wiki/Bias_(statistics)
of the desired behavior of the model at semantic level. Thus, is up to the language engineer to decide which domains and which expected distributions must be monitored or, eventually, corrected. To that end, stereotyped concepts must be identified within the domain and related attributes or concepts biased by those stereotypes must be selected. To overcome a clean definition of the bias problem, we propose an ontology-based approach, as the bias problem is firstly identified at a semantic level and, later on, treated at model-parameter level.

1.3. Definition of bias at semantic level

Description logics [24] provides a complete set of elements for knowledge base structure, population and manipulation. It is, actually, the ontology formalization acquired by the Semantic Web and its high level ontological terminology OWL [25]. An OWL ontology has following components: \(< C, P, I, L >\) classes \(C\), properties \(P\), individuals \(I\) and literal values \(L\). For the shake of simplicity, we will summarize saying that individuals are instances of classes, instances are interrelated by properties and literals are associated to individuals by properties. For example, Christine is an individual which is of "type" Woman (belongs to class Woman). She works in a hospital (works-in would be a property). She has a job as doctor (has-job would be another property). Woman is a class that can be defined through the expression has-gender "female" (this is called class expression in OWL), where has-gender is a property and "female" is a literal value. This simple knowledge can be graphically plotted as in Figure 1.

Now it is time to borrow some terminology from fairness measures in machine learning and some elements from OWL.

**Definition 2.** A stereotyped knowledge is represented by the tuple \(< C, P, I, L, p_p, P_s >\) where \(C\) is the set of classes, \(P\) is the set of properties, \(I\) is the set of individuals, \(L\) the set of literals, \(p_p \in P\) is the protected property and \(P_s \subset P\) is the set of stereotyped properties. This express that groups of individuals resulting from different values of the protected property \(p_p\) could exhibit inequality in the distribution of values for stereotyped properties \(P_s\).

In the example displayed in Figure 1, we could consider has-gender as the protected property \(p_p\) and \(P_s = \{\text{has-job}\}\) as the set of stereotyped properties, so the tuple would be \(< C, P, I, L, \text{has-gender}, \{\text{has-job}\} >\). According to Definition 2, this means that values for property has-job could be not equally distributed over individuals of both classes defined by has-gender. For example, we may find that for individuals with has-gender "female" it is more frequent to

![Figure 1. A simple example in OWL](image-url)
observe has-job nursery than has-job doctor, while the situation is the inverse for the class with individuals holding has-gender "male". It is important to note that a stereotyped knowledge is only defining a potential bias, i.e. a bias we are sensible to.

1.4. Definition of bias in language modeling

Once it is clear the semantic definition of the stereotyping bias, we can map that semantic identification down to word probabilities. This is straightforward, as a language model is nothing but a model able to compute a probability for a sequence of words \( P(w_1, ..., w_m) \).

**Definition 3.** A stereotyped language can be represented as the tuple \( < C, P, I, L, p_p, P_s, T_p, T_s > \), which contains a stereotyped knowledge and two terminology sets: protected terms \( T_p \) and stereotyped terms \( T_s \). Protected terms \( T_p \) are those expressions (words or multi-words) in the vocabulary that can be unambiguously mapped to values of a protected property \( p_p \). Stereotyped terms \( T_s \) are those expressions (words or multi-words) in the vocabulary that can be unambiguously mapped to values of stereotyped properties \( P_s \).

\( T_s \) is the set of words or terms that represent possible values of stereotype properties \( P_s \) (for example, "high imagination", "low sensibility", "beauty", "rational mind", etc.). Examples of expressions in \( T_p \) would be any term defining gender, like "nurse", "actress", "woman", "girl", or alike. Once the stereotype is defined at semantic level, we can consider that if the probability of a sequence of words containing expressions in \( T_s \) on stereotyped properties \( P_s \) is significantly different according to the value of \( p_p \) of the referred individual, then the model is biased.

Now, we are ready for the final definition of stereotyping bias in language models.

**Definition 4.** Let \( L_s = < C, P, I, L, p_p, P_s, T_p, T_s > \) be the definition of a stereotyped language, stereotyping bias is defined as the distance \( d \) between probabilities \( d(P(w_1, ..., w_m|t'_p), P(w_1, ..., w_m|t'_p)) \), withi \( i \neq j \) where \( t'_p \) and \( t'_p \) are the expressions for two different values of the protected property \( p_p \) and \( \exists w_k \in \{ w_1, ..., w_m \} \) so that \( w_k \in T_s \).

In other words, a language model is biased if distributions of probabilities of terms containing stereotyped expressions are different subject to existing protected expression priors. Following our simple example, the stereotyped language could be defined as \( < C, P, I, L, has\_gender,\{has\_job\},\{girl, women, Christine, man\},\{doctor, nurse\}> \).

Now, consider this simple text:

Christine works as a nurse in the hospital. A man is the doctor.

The definition is open to any kind of distance. If we select absolute difference, the stereotyping bias of the language model trained on the text above could be:

\[ |P(\text{works, as, a, nurse}|Christine) - P(\text{works, as, a, doctor}|Christine)| \]

Another valid measure would be:

\[ |P(\text{works, as, a, nurse}|man) - P(\text{works, as, a, doctor}|man)| \]

As you can see, different distances can be computed depending on the sequence or the prior value of the protected property considered. An appropriate evaluation of a language model would imply, therefore, a battery of expressions like the ones above, with protected expressions as priors and stereotyped expressions in the sequence, from which an average distance could be calculated.

2. Overview on bias related research

An exhaustive review of relevant papers on bias in natural language processing has been carried out. In order to provide a global view into the different studies and analysis found regarding bias detection and correction, a set of elements have been identified to characterise major issues over all the works compiled. This allows for a organisation of up-to-date research work on the targeted matter.
These elements are now introduced for better understanding of the overview table, as dimensions over the different main aspects in bias related research.

- **Year.** This column is the publication year in ascending order and will serve as timeline on research progress. It also serves to highlight the increase of interest in the research community over time. We can see that it was not until two years after [26] when the community began to actively work on the bias of word embeddings models.
- **Reference** points to the publication.
- **Domain(s)** show us in which category fall the studied bias. The most represented category is gender bias, usually showing difference treatment between male and female. The second most represented one is ethnicity bias, in this category we grouped bias against race, ethnicity, nationality or language. We also found work on bias related with age, religion, sexual orientation and disability. It is worth mentioning that there is some work done on political bias.
- **Model** will refer to the neural network model studied in the paper. When the bias is not a model but an application we will refer to such an application. Bias is not only studied in open system but also in black box applications like Google Translate. It is interesting how some studies are able to discover and measure bias in those systems. Although they are not able to mitigate the bias directly, there are some samples that manage to reduce the bias without having access to the model by modifying strategically the input.
- **Dataset** will serve as a summary of what data was used. We will consider almost all the resources that have taken part in the study regarding: from the information used to train the models to the corpus on which the models is applied, or the other dataset that helps to contextualize the technique used.
- **Language** column mainly shows that most of the work has been done on English datasets and models. Some approaches when working with bias in other languages usually have English as a reference point, involving the translation of the data or test sets from English to other languages with both automated tools and paid professionals. Another approach involves looking for analogies between different languages.
- **Evaluation** column shows the reader which was the technique for evaluation or for measuring the bias:
  - **Association Tests** The usage of association tests began with the appearance of WEAT tests by Caliskan et al. [10] based on a study outside of the computer science field by Greenwald et al. [27]. It aims to measure the strength of the connection between two words.
  - **Sentiment of Association**, a common way to find biased terms is measuring the sentiment of sentences by changing just one word. The words that differ will belong to the two classes being compared. A term will be biased if one sentence has a strong negative sentiment regarding the complementary. This is also tested with text generation tasks where a given sentence start will produce a full sentence or text, just changing a word of each class.
  - **textbf{Analogies}** The use of analogies has been found useful to show the bias with simple examples. Word embeddings space is suited to this type of technique, as analogies can be studied from a geometric perspective.
  - **Representation** The works that fall in this category compare the likelihood between two classes of the protected property. Some studies will consider the goal to achieve equal representation, but usually the likelihood of the classes is compared with real world data. For example, comparing the distribution of men and women in the United States for a occupation with the probability of a sentence to be completed with an attribute of each one of the genres. In this way you can compare the model output representation with the demographic percentage.
  - **Accuracy** It is common to find studies that measure accuracy in tasks like classification or prediction to find out how biased the model is. This is similar to the general approach in machine learning with fairness measures.
- **Mitigation** shows how the bias is removed or attenuated from the data or the model.
• **Vector Space Manipulation** evolves from the work of Bolukbasi *et al.* [26] in which he proposes to find the vector representation of the gender to compensate for its deviation and equalize some terms with respect to the neutral gender. This technique is known as Word embedding Debiasing or Hard Debiasing. This proposal has been explored with substantial improvements to better capture the bias, trying to avoid causing a harm to the model.

• **Data Augmentation** by increasing the source corpus/data.

• **Data Manipulation** makes changes to the data to help the model capture a less biased reality. For example, removing named entities.

• **Attribute Protection** tries to prevent an attribute from containing bias. For this purpose, different techniques are used to manipulate the data, the model or the training in order to avoid capturing information about that attribute. For example, if you remove proper names from phrases in a dataset and train a model, the model will not be able to associate proper names with other features such as jobs. If you train a model to analyze the sentiment of phrases and avoid proper nouns, the names will not have sentiment associated with them. You can find its application in the other techniques or as a combination of them. For example, eliminating proper names so that they do not capture gender information, duplicating all sentences that have gender (data modification) using the opposite gender (data augmentation) and finally training the model and manipulating it to eliminate the gender subspace (Vector space manipulation).

• **Stage** column stands for Mitigation Stage, and indicates when the mitigation/bias correction work was done.
  
  – **Before** Mostly altering or augmenting the source data.
  – **During/Train** Changing the training process or fine-tuning the model.
  – **After** Usually changing the model vector space after the learning stage.

• **Task**, This column outlines the field or scope in which the author is working. Since the appearance of [26] an important part of the studies will try to solve the novel problem of both "Debiasing" and "Bias Evaluation". Since both tasks are already reported in columns of the table itself, they will not appear in this column.
<table>
<thead>
<tr>
<th>Year</th>
<th>Ref.</th>
<th>Stereotype(s)</th>
<th>Model</th>
<th>Data</th>
<th>Lang.</th>
<th>Evaluation</th>
<th>Mitigation</th>
<th>Stage</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>[26]</td>
<td>Gender</td>
<td>Word2Vec, GloVe</td>
<td>GoogleNews corpus (w2vNEWS), Common Crawl</td>
<td>English</td>
<td>Analogies/Cosine Similarity</td>
<td>Vector Space Manipulation</td>
<td>After</td>
<td>-</td>
</tr>
<tr>
<td>2017</td>
<td>[10]</td>
<td>Gender, Ethnicity</td>
<td>GloVe, Word2Vec</td>
<td>Common Crawl, Google News Corpus, Occupation Data (BLS)</td>
<td>English</td>
<td>Association Tests (WEAT, WEAFAT)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2018</td>
<td>[28]</td>
<td>Gender</td>
<td>Deep Coref.[29]</td>
<td>WinoGender, Occupation Data (BLS), B&amp;L</td>
<td>English</td>
<td>Prediction Accuracy</td>
<td>Data Augmentation (Gender Swapping), Vector Space Manipulation</td>
<td>After</td>
<td>Coreference Resolution</td>
</tr>
<tr>
<td>2018</td>
<td>[30]</td>
<td>Gender</td>
<td>GloVe[31]</td>
<td>OntoNotes 5.0, WinoBias, Occupation Data (BLS), B&amp;L</td>
<td>English</td>
<td>Prediction Accuracy</td>
<td>Data Augmentation (Gender Swapping), Vector Space Manipulation</td>
<td>After</td>
<td>Coreference Resolution</td>
</tr>
<tr>
<td>2018</td>
<td>[33]</td>
<td>Gender</td>
<td>e2e-coref[34], deep-coref[35]</td>
<td>CoNLL-2012, Wikitext-2</td>
<td>English</td>
<td>Coreference score likelihood(2)</td>
<td>Data Augmentation( CDA ), WED [26], Before</td>
<td>Train(1), Coreference Resolution(1), Language Modeling (2)</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>[38]</td>
<td>Gender</td>
<td>BERT(base, uncased), GPT-2(small)</td>
<td>-</td>
<td>English</td>
<td>Visualization, Text Generation likelihood</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2019</td>
<td>[39]</td>
<td>Gender, Ethnicity, Disability, Sexual Orientation</td>
<td>Google Perspective API</td>
<td>WikiDetox, Wiki Madlibs, Twitter, WordNet</td>
<td>English</td>
<td>Classification Accuracy, likelihood</td>
<td>Data correction, Data Augmentation, Attribute Protection</td>
<td>Before</td>
<td>Hate Speech Detection</td>
</tr>
<tr>
<td>2019</td>
<td>[40]</td>
<td>Gender</td>
<td>fastText, BoW, DRNN with Custom Dataset</td>
<td>Common Crawl, Occupation Data (BLS)</td>
<td>English</td>
<td>Prediction Accuracy</td>
<td>Attribute protection (Removing Gender and NE)</td>
<td>Before</td>
<td>Hiring</td>
</tr>
<tr>
<td>2019</td>
<td>[41]</td>
<td>Account Age, user features</td>
<td>Graph Embeddings[42]</td>
<td>WikiData</td>
<td>English</td>
<td>Accuracy</td>
<td>Attribute Protection (Remove user information)</td>
<td>Train</td>
<td>Vandalism Detection</td>
</tr>
<tr>
<td>2019</td>
<td>[44]</td>
<td>-</td>
<td>Word2Vec(1), fastText(2), GloVe(3)</td>
<td>Google News(1), Web data(2,3), First Names (SSA)</td>
<td>English</td>
<td>WEAT</td>
<td>-</td>
<td>-</td>
<td>Unsupervised Bias Enumeration</td>
</tr>
<tr>
<td>2019</td>
<td>[45]</td>
<td>Gender, Age, Ethnicity</td>
<td>GloVe</td>
<td>Wikipedia Dump, WSim-353, SimLex-999, Google Analogy Dataset</td>
<td>English</td>
<td>WEAT, ECT</td>
<td>Vector Space Manipulation</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Year</td>
<td>Ref.</td>
<td>Stereotype(s)</td>
<td>Model</td>
<td>Data</td>
<td>Lang.</td>
<td>Evaluation</td>
<td>Mitigation</td>
<td>Stage</td>
<td>Task</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>---------------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>------------</td>
<td>------------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>2019</td>
<td>[46]</td>
<td>Ethnicity, Gender, Religion</td>
<td>Word2Vec</td>
<td>Reddit L2 corpus</td>
<td>English</td>
<td>PCA, WEAT, MAC, Clustering</td>
<td>Vector Space Manipulation</td>
<td>After</td>
<td>POS tagging, POS chunking, NER</td>
</tr>
<tr>
<td>2019</td>
<td>[47]</td>
<td>Gender</td>
<td>ELMo, Glove</td>
<td>One Billion Word Benchmark, WinoBias, OntoNotes 5.0</td>
<td>English</td>
<td>PCA, Prediction Accuracy</td>
<td>Data Augmentation(1), Attribute Protection(gender swapping averaging) (2)</td>
<td>Train(1), Coreference After(2) Resolution</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>[48]</td>
<td>Gender</td>
<td>CBOW(1), GloVe(1,2), FastText(1), Dict2Vec(1)</td>
<td>English, Common Crawl(2), Wikipedia(2), Tweets(2)</td>
<td>English, German, Spanish, Italian, Russian, Croatian, Turkish</td>
<td>WEAT, XWEAT</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>[49]</td>
<td>Gender</td>
<td>Spanish fastText</td>
<td>Spanish Wikipedia, bilingual embeddings (MUSE)[50]</td>
<td>English, Spanish</td>
<td>CLAT, WEAT</td>
<td>Vector Space Manipulation</td>
<td>After</td>
<td>-</td>
</tr>
<tr>
<td>2019</td>
<td>[51]</td>
<td>Gender</td>
<td>Skip-Gram(1,2,3), FastText(4)</td>
<td>Google News(1), PubMed(2), Twitter(3), GAP-Wikipedia(4) [52]</td>
<td>English</td>
<td>WEAT, Clustering (K-Means++)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>[54]</td>
<td>Gender, Ethnicity</td>
<td>BERT(large, cased), CBoW-GloVe (Web corpus version), InferSent, GenSen, USE, ELMo, GPT</td>
<td>-</td>
<td>English</td>
<td>SEAT</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>[55]</td>
<td>Gender, Race</td>
<td>BERT( base large cased), GPT-2 (117M, 345M), ELMo, GPT</td>
<td>-</td>
<td>English</td>
<td>Contextual SEAT</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>[56]</td>
<td>Gender</td>
<td>ELMo</td>
<td>English-German news WTM18</td>
<td>English</td>
<td>cosine similarity, clustering, KNN</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3. Previous work on bias detection and treatment in NLP (Part III/IV)

<table>
<thead>
<tr>
<th>Year</th>
<th>Ref.</th>
<th>Stereotype(s)</th>
<th>Model</th>
<th>Data</th>
<th>Lang.</th>
<th>Evaluation</th>
<th>Mitigation</th>
<th>Stage</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>[61]</td>
<td>Race, Gender, Sexual Orientation</td>
<td>LSTM, BERT, GPT-2 (small), GoogleLM1b (4)</td>
<td>One Billion Word Benchmark(4)</td>
<td>English</td>
<td>Sentiment Score (VADER [62]), Classification accuracy</td>
<td>Train LSTM/BERT</td>
<td>Train</td>
<td>Text Generation</td>
</tr>
<tr>
<td>2019</td>
<td>[63]</td>
<td>Gender</td>
<td>Google Translate, Microsoft Translator, Amazon Translate, SYSTRAN, Model of [64]</td>
<td>-</td>
<td>English, French, Italian, Russian, Ukrainian, Hebrew, Arabic, German</td>
<td>WinoMT (WinoBias + WinoGender), Prediction Accuracy</td>
<td>Positive Contextualization</td>
<td>After</td>
<td>Translation</td>
</tr>
<tr>
<td>2019</td>
<td>[65]</td>
<td>Gender</td>
<td>CBOW</td>
<td>English Gigaword, Wikipedia, Google Analog, SimLex-999</td>
<td>English</td>
<td>Analogies, WEAT, Sentiment Classification, Clustering</td>
<td>Hard-Debiasing, CDA, CDS</td>
<td>Train</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[14]</td>
<td>Gender, Race, Religion</td>
<td>GPT-3</td>
<td>Common Craw, WebText2, Books1, Books2, Wikipedia</td>
<td>English</td>
<td>Text generation</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[66]</td>
<td>*</td>
<td>CBoW, GloVe, FastText, DebiasNet</td>
<td>-</td>
<td>German, Spanish, Italian, Russian, Croatian, Turkish, English</td>
<td>WEAT, XWEAT, ECT, BAT, Clustering(KMeans)</td>
<td>Vector Space Manipulation, After</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[67]</td>
<td>Gender, Ethnicity</td>
<td>AraVec, CBOW(1), CBOW(2), AraVec Skip-Gram(3) and FASTTEXT(4), FastText(5)</td>
<td>translated WEAT test set, Leipzig news(2), Wikipedia1,3,5, Twitter1,3,4, CommonCrawl(5)</td>
<td>Modern Arabic, Egyptian Arabic</td>
<td>WEAT, XWEAT, AraWEAT, ECT, BAT</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[68]</td>
<td>Gender</td>
<td>RoBERTa/gloVe (1)</td>
<td>Common Crawl (1)</td>
<td>English</td>
<td>WEAT*, SIRT</td>
<td>Vector Space Manipulation, After</td>
<td>Train</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[69]</td>
<td>Ideological, Political, Race</td>
<td>GPT-3</td>
<td>Common Craw, WebText2, Books1, Books2, Wikipedia</td>
<td>English</td>
<td>QA, Text Generation</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[71]</td>
<td>Gender, Profession, Race, Religion</td>
<td>BERT, GPT-2, RoBERTa, XLNet</td>
<td>StereoSet</td>
<td>English</td>
<td>CAT Context Association Test</td>
<td>-</td>
<td>-</td>
<td>Language Modeling</td>
</tr>
<tr>
<td>2020</td>
<td>[72]</td>
<td>Gender</td>
<td>Google Translate</td>
<td>United Nations[58], Europarl[59], Google Translate Community</td>
<td>English, Hungarian likelihood vs Real</td>
<td>-</td>
<td>-</td>
<td>Translation</td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td>[73]</td>
<td>Gender</td>
<td>Word2Vec</td>
<td>Wikipedia-es 2006</td>
<td>Spanish</td>
<td>Analogies</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[74]</td>
<td>Gender</td>
<td>CBOW</td>
<td>British Library Digital corpus, The Guardian articles</td>
<td>English</td>
<td>Association, Prediction likelihood, Sentiment Analysis</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 4. Previous work on bias detection and treatment in NLP (Part IV/IV)

<table>
<thead>
<tr>
<th>Year</th>
<th>Ref.</th>
<th>Stereotype(s)</th>
<th>Model</th>
<th>Data</th>
<th>Lang.</th>
<th>Evaluation</th>
<th>Mitigation</th>
<th>Stage</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>[75]</td>
<td>Gender, Race, Religion, Disability</td>
<td>BERT(1)</td>
<td>Wikipedia(1), Book corpus(1), Jigsaw identity toxic dataset, RtGender, GLUE</td>
<td>English</td>
<td>Cosine Similarity, Accuracy, GLUE</td>
<td>Fine Tunning</td>
<td>Fine Tunning Making</td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td>[76]</td>
<td>Gender, Race</td>
<td>SqueezeBERT</td>
<td>Wikipedia, BooksCorpus</td>
<td>English</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[77]</td>
<td>Intersectional Bias (Gender, Ethnicity)</td>
<td>GloVe, ELMo, BERT, GPT-2</td>
<td>CommonCrawl, Billion Word Benchmark, BookCorpus, English Wikipedia dumps, BookCorpus, WebText, Bert-small-cased</td>
<td>English</td>
<td>WEAT, CEAT</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[78]</td>
<td>Gender</td>
<td>BERT</td>
<td>Equity Evaluation Corpus, Gen-data</td>
<td>English</td>
<td>EEC, Gender Separability, Emotion/Sentiment Scoring</td>
<td>Vector Space Manipulation</td>
<td>Train</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[79]</td>
<td>Ethnicity</td>
<td>GPT-2 (small), DISTILBERT</td>
<td>TwitterAAE, Amazon Mechanical Turk annotators (SAE)</td>
<td>English (AAVE / SAE)</td>
<td>Text generation, BLEU, ROUGE, Sentiment Classification, VADER [62]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2020</td>
<td>[81]</td>
<td>Ethnicity</td>
<td>GPT-2</td>
<td>English science fiction story corpus, Plotto, ROCstories, toxic and Sentiment datasets</td>
<td>-</td>
<td>Classification Accuracy</td>
<td>Loss function modification</td>
<td>Fine tuning</td>
<td>Normative text Classification</td>
</tr>
<tr>
<td>2020</td>
<td>[82]</td>
<td>Disability</td>
<td>BERT, Google Cloud sentiment model</td>
<td>Jigsaw Unintended Bias</td>
<td>English</td>
<td>Sentiment Score</td>
<td>-</td>
<td>-</td>
<td>Toxicity prediction, Sentiment analysis.</td>
</tr>
<tr>
<td>2020</td>
<td>[83]</td>
<td>Gender</td>
<td>BERT</td>
<td>GAP, BEC-Pro, Occupation Data (BLS)</td>
<td>English, German</td>
<td>Association Test (like WEAT)</td>
<td>Fine-Tunning, CDS</td>
<td>Train</td>
<td>-</td>
</tr>
</tbody>
</table>
3. Discussion

Although tables above have been introduced detailing key aspects in bias research according to the dimensions identified, a deeper analysis of all this prolific research production is carried out now. We have divided the discussion into salient topics in the following.

3.1. Association Tests

There are several approaches to bias measurement and mitigation. Bolukbasi et al. [26] laid the foundations for much of the work that was to follow. The main contribution was on showing that embeddings captured the correlation between terms so that they could correctly resolve analogies such as man:king → woman:queen, but also some similar analogies were biased. For example, it associates man:doctor and woman:nurse while the association woman:doctor would be more adequate. Using this same mechanism was obtained a set of terms that were stereotyped to each gender, to prove that this was not an isolated case. To remove that bias they proposed to find the gender vector subspace direction and adjust the vector to make the occupational terms gender-neutral.

Caliskan et al. [10] took the idea of measuring bias of using the Implicit Association Test and proposed the Word Embedding Association Test (WEAT). WEAT measures the similarity of words by using the cosine between the pair of vectors of those words. It was applied to GloVe [31] and also to Word2Vec [84] with very similar findings. Other extension to WEAT was proposed by Lauscher and Glavaš [48], Lauscher et al. [66] with the name XWEAT, a cross lingual extension for WEAT. XWEAT was later extend to Arab Lauscher et al. [67].

WEAT could also be applied to other models. Gonen and Goldberg [37] applied it to the Gender Neutral Version of GloVe called GN-GloVe from Zhao et al. [32]. Jentzsch et al. [43] uses it with a skip-gram network in the context of Question Answering and Decision Making and [44] creates an algorithm to discover offensive association related with gender, race and other attributes, generating WEAT tests for them. They called this technique Unsupervised Bias Enumeration (UBE). UBE is applied to Word2Vec, fastText and GloVe. Dev and Phillips [45] proposes two complementary tests that measures the bias removal effect (ECT, EQT).

Manzini et al. [46] extends WEAT to measure the bias in a multi-class setting and uses it over a Word2Vec model trained with Reddit L2 corpus. Dev et al. [68] adapted it to work with two sets of words at a time instead of just two words, naming its variant WEAT*.

The appearance of models such as BERT [85] led to the adaptation of the technique to work at phrase level (SEAT May et al. [54]) and to work with contextualized embeddings in Guo and Çalışkan [77] named CEAT. CEAT was tested on BERT, GPT, GPT-2 and ElMo.

As part of the association-based bias study, Nadeem et al. [71] presents StereoSet and evaluates BERT, BPT-2, RoBERTa, XLNET models. For the evaluation it confronts three terms in the same context, one stereotyped, one anti-stereotyped and one unrelated term. It measures the probability that a sentence is completed with each of them. In the sentences there is a token which is the one against which we measure the bias. This technique is called Contextual Association Test.

From this test, the sentiment associated with stereotyped and non-stereotyped sentences can be analysed. Measuring the sentiment of an association to quantify bias is not new, it can be found in the work of Kiritchenko and Mohammad [36] where he evaluates race and gender bias. Sheng et al. [61] further measures bias associated with sexual orientation by comparing the associated sentiment. We al have the extensive study by Leavy et al. [74] on CBOW trained on articles from The Guardian journal and the British Digital Library. Also, Hutchinson et al. [82] studies the perception of models towards disabled people and Bhardwaj et al. [78] combines the study of gender bias on BERT by sentiment analysis with gender separability.
3.2. Translation

Previously, we have seen XWEAT for the detection of bias in languages other than English. Although there are also alternatives that work with multiple languages, one of the areas of study is what occurs when translating a text, such as the work of Escudé Font and Costa-jussà [57], which seeks and mitigates the bias in English-Spanish translations with the three versions of GloVe previously discussed (Base, Gender Neutral, Hard-Debiased).

Not only has translation bias been studied in open models, but also the bias in final products such as Google translator, Microsoft translator, Amazon translator, among others, has been evaluated in the study of Stanovsky et al. [63].

[72] poses a mismatch when using Google Translator for translating from languages such as Hungarian with neutral gender into English. The inferred gender does not proportionally represent the actual distribution of workers when making inferences about professions, using Google Translator. This same mismatch appears in Google Translator in more languages, such as Hungarian, Chinese, Yoruba and others when translated into English. In this case, [53] shows a very strong correlation between the STEM (science, technology, engineering and mathematics) family of academic disciplines and men.

According to Davis [86], Google has fixed the problem it had with Google Translator inferring gender when translating from non-gendered languages into English. In the Google AI blog, Johnson [87] develops the first approach to the problem. This solution was put into production in 2018. They trained a CNN with human categorized examples and further divided the training set into three chunks, one for masculine another feminine and another for neutral. To the sentences of each chunk they added in front a token of the type "<2MALE>". So "<2MALE> O birt doktor" would translate to "HE is a doctor". Allowing this to use all 3 prefixes with the user input to give an unbiased response. This resulted in a recall of 60%.

The next approach would come in 2020. Johnson [88] would firstly translate the phrase obviating the gender and secondly, would look for occurrences of the translated phrase from the same query but with the complementary gender. If only the gender changes when compared to the original translation, the phrase is returned to the user on both genders.

3.3. Coreference Resolution

Two of the first studies for gender bias were published as part of the Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. The first of Zhao et al. [30] proposes WinoBias, a balanced Male/Female dataset for the evaluation of gender bias in Coreference Resolution tasks. The second, with a similar title, was that of Rudinger et al. [28] who introduced the WinoGender schemes for the study of bias also in Coreference Resolution tasks. Later, Stanovsky et al. [63] combined both resources to study the bias in Machine Translation, thus creating WinoMT.

The study of bias in coreference resolution does not stop here, Zhao et al. [32] studies gender bias in Glove and develops 2 derived models, GN-Glove (Gender Neutral) and HD-Glove (Hard Debiased). Lu et al. [33] tries to reduce the detected bias by using the data augmentation technique Contextual Data Augmentation CDA which consists of adding a complementary gender phrase to the sentences of the initial dataset. Based on CDA, in 2019 Hall Maudslay et al. [65] will develop Contextual Data Substitution CDS. It proposes to eliminate the bias associated with proper names by adding a phrase with a complementary gender name in a balanced way. CDS will later be used by Bartl et al. [83] together with fine-tunning for BERT.

3.4. GPT-3 and black box models

The recent GPT-3 also seems to suffer from bias. Actually, in the very study that presents the model Brown et al. [14] already addresses the issue for gender, race and religion. The authors themselves discover an important tendency between terms such as violent, terrorism and terrorist with Islam. It
will also be studied in Decision Making and Question answering tasks by McGuffie and Newhouse [69], which will show how GPT-3 is better than GPT-2 at generating extremist narratives and suggest that it could be used for the radicalization of individuals. For the study, questions are asked to GPT-3 on specific topics and its responses are studied.

The alternative to studying bias in this model by asking questions is through the model’s ability to generate text from the beginning of a given sentence that will serve as a context to the model. Floridi and Chiriatti [70] studies the model in this way and finds that although the model is able to complete sentences and text, it lacks perspective or intelligence when dealing with topics.

Abid et al. [16] makes a valuable contribution, finding that it is possible to alleviate the bias in the responses and text generated by GPT-3 by trying to guide the response with a positive context. If instead of asking the model to complete a sentence referring to Muslims, you should add a positive adjective such as hard-working or meticulous. This way the model’s responses will move away from topics such as violence.

All studies on GPT-3 are conducted as a black box model since it has not been released. This is why its web interface or API is used for its study.

3.5. Vector Space

The main debiasing techniques try to eliminate model bias. Different approaches are used to find the direction of the gender as proposed by Bolukbasi et al. [26] and try to correct the deviation between classes. The simplest papers define techniques to find the gender direction and adjust it between male/female pairs. Some, such as Zhou et al. [49], propose that there is not one but two gender directions given the characteristics of Spanish, considering one direction as semantic and the other as grammatical. In such a way that words like perfume in Spanish are masculine but strongly associated with the feminine gender, so it will try to eliminate the bias by considering both components. This is why, in our formal definition of bias, "stereotyped expressions" is preferred, rather than just mentioning words or isolated terms.

Gonen and Goldberg [37] suggests that the debiasing techniques that work with gender direction are not sufficient and that the bias is only superficially eliminated. There are multiple approaches that try to improve by trying to identify gender as a space rather than as a direction, such as the work of Basta et al. [56] on ElMo.

The previously cited techniques are also extrapolated to try to tackle the problem in other languages. Zhou et al. [49] suggests that for Spanish there is not one gender direction but two: grammar and semantics. A term like “perfume” is semantically more masculine but is grammatically more strongly associated with the feminine gender. So gender measurement and mitigation will have to seek to balance between these two dimensions. He also proposes CLAT (cross lingual analogy tasks) to assess bias in Spanish. Given a pair a:b in English and a word c in Spanish, the Spanish term associated with d must be predicted for a:b = c:?.

Alternatively, Díaz Martínez et al. [73] launches a proposal similar to Bolukbasi et al. [26] but for Spanish. It detects that there are indeed terms strongly associated to one of the genres in a Word2Vec model trained with Wikipedia articles.

3.6. Complementary works

There are similar approaches that show that it is possible to detect gender bias in models such as GPT-2 Radford et al. [15], Vig [38] reviews the interior of these networks and evidences the strong connections between "she, nurse" and "he, doctor" and suggests that it would be possible to detect and control it. For all this, he relies on a tool that allows to visualize the interior of transformer networks such as BERT Devlin et al. [85] or GPT-2.
4. A general methodology for dealing with bias in deep NLP

Up to this point, it is possible to conclude that the bias problem is of relevance for industrial deployments of artificial intelligence solutions. When putting a language model into production for a defined task like classification, dialogue or whatsoever, the engineer has to ensure that no bias could affect the expected behavior of the system so future troubles due to stereotyped decisions are prevented. This paper has made an effort in showing the state of the art in bias related research, specially on deep learning models for natural language processing. But also, a clear definition of this phenomena has been provided, detailing all the elements involved in spotting the bias with precision and completeness.

In this section, we propose the use of all those elements to help the engineer to identify them in the subject of her study and to follow a structured method to tackle it in a software engineering process. With that purpose in mind, the following steps are proposed as a general methodology for dealing with stereotyping bias in deep language models generation and application:

1. Define the stereotyped knowledge. This implies to identify one or more protected properties and all the related stereotyped properties. For each protected property, you have to develop its own ontology.

2. With the previous model at hand, we can overcome the task of identifying protected expressions and stereotyped expressions, so your stereotyped language is defined. There are some corpora available, like the ones mentioned in this work, but you may need to define your own expressions in order to capture all the potential biases that may harm your system. Anyhow, it is here when different resources could be explored to obtain a set of expressions as rich as possible.

3. The next step is to evaluate your model. Choose a distance metric and compute overall differences in sequence probabilities containing stereotyped expressions with protected expressions as priors. Detail the benchmarked evaluation framework used.

4. Analyse the results of the evaluation to identify which expressions or categories of expressions result in higher bias.

5. Design a corrective mechanism. You have to decide which strategy fits better with your problem and with your available resources: data augmentation, a constraint in the learning process, model parameters correction...

6. Re-evaluate your model and loop over these last three steps until an acceptable response is reached, or though out your model if behavior is not what is desired. Rethink the whole process (network architecture, pre-training approach, fine-tuning, etc.

7. Report the result of this procedure by attaching model cards or similar document formalism in order to achieve transparent model reporting.

Following these steps may help in getting a final system you understand better and with predictions not affected or marginally affected by stereotyping bias. For sure, this method can be adapted or extended according to the requirements of each specific AI project.

5. Conclusions and challenges

In this work, we focused on Deep NLP techniques and how these techniques are affected by bias as a consequence of the advent of more challenge data sets and methods. We found that gender bias for English language when using word embedding related technologies is the most frequent scenario that is faced in those methods developed to mitigate bias in different tasks. This can be achieved in three different ways: by modifying the training corpora, the training algorithm or the results obtained according to the given task. We propose to systematize the evaluation of the impact of bias as part of the design of systems relying on Deep NLP techniques and resources. The focus of the proposed procedure is the identification and management of stereotyped expressions apart from protected expressions, both concepts introduced in Section 1.3. As future challenges, apart from digging deeper in the detection and softening of bias, it is our view that there are some aspects that
deserve more attention than given nowadays. The first one is related with the effect of bias mitigation in both the global system performance and the management of other terms and features different from stereotyped expressions. Is it possible that the main task to be solved by Deep NLP systems could be damaged by the intervention to mitigate stereotyped expressions? In the same way, we propose to study the impact of a preventive strategy rather than a corrective one. That is, in the case of having transparent language models (i.e. accompanied by model reports), we consider measuring how the choice of different language models that are free of bias compared to those that do present some degree of bias, affects the final performance of the system. In any case, although it is clear the fact that there is no biased algorithms but biased corpora and language models, there is little effort in describing characteristics of corpora and making transparent language models by means of the inclusion of model reporting, related with demographic or phenotypic groups, environmental conditions, instrumentation or environment, inter alia. As a consequence, it is needed further effort to characterize, to make transparent the language model or corpora to be chosen regarding a given task.

Another interesting approach would be to apply the techniques studied to systems in production and perform different measurements that allow us to know the impact of the changes made on the model. Applying this work to real applications will allow us to see if the changes are really effective, to see how they affect other aspects on the application’s performance and, above all, to discover which aspects have not been taken into account.

As a matter of engineering processes, resources should be put on the focus of the problem. Additional benchmarks and tests for different stereotypes over different languages are, in our opinion, in the way to a consistent management of biases for final applications.

Author Contributions: Conceptualization, I.G.M., A.M.R, F.M.S. and L.A.U.L; Contextualization, F.M.S.; Formalism and Methodology, A.M.R; Research Overview, I.G.M; Analysis, I.G.M., A.M.R and F.M.S.

Funding: This study is partially funded by the Spanish Government under the LIVING-LANG project (RTI2018-094653-B-C21).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations
The following abbreviations are used in this manuscript:
AAVE  | African American Vernacular English  
AI    | Artificial Intelligence  
BAT   | Bias Analogy Test  
BERT  | Bidirectional Encoder Representations from Transformers  
BLS   | Bureau of Labor Statistics  
CAT   | Context Association Test  
CDA   | Counterfactual Data Augmentation  
CDS   | Counterfactual Data Substitution  
CEAT  | Contextualized Embedding Association Test  
CLAT  | Cross-lingual Analogy Task  
ECT   | Embedding Coherence Test  
EQT   | Embedding Quality Test  
GPT   | Generative Pre-Training Transformer  
IAT   | Implicit Association Tests  
LM    | Language Model  
LSTM  | Long-Short Term Memory  
NER   | Named Entities Recognition  
NLP   | Natural Language Processing  
PCA   | Principal Component Analysis  
POS   | Part of Speech  
SAE   | Standard American English  
SEAT  | Sentence Encoder Association Test  
SIRT  | Sentence Inference Retention Test  
UBE   | Universal Bias Encoder  
USE   | Universal Sentence Encoder  
WEAT  | Word Embedding Association Test  
WEFAT | Word Embedding Factual Association Test  
XWEAT | Multilingual and Cross-Lingual WEAT  

6. References

1. Howard, A.; Borenstein, J. Trust and Bias in Robots: These elements of artificial intelligence present ethical challenges, which scientists are trying to solve. American Scientist 2019, 107, 86–90.


*ArXiv 2020*, *abs/2009.06807*.  
72. Farkas, A.; N’emeth, R. How to Measure Gender Bias in Machine Translation: Optimal Translators, Multiple Reference Points.  
*ArXiv 2020*, *abs/2011.06445*.  
73. Díaz Martínez, C.; Díaz García, P.; Navarro Sustaeta, P. Hidden Gender Bias in Big Data as Revealed by Neural Networks: Man is to Woman as Work is to Mother? *Revista Española de Investigaciones Sociológicas 2020*, 172, 41–60.  
76. Iandola, F.N.; Shaw, A.E.; Krishna, R.; Keutzer, K. SqueezeBERT: What can computer vision teach NLP about efficient neural networks?  
*ArXiv 2020*, *abs/2006.11316*.  
*ArXiv 2020*, *abs/2006.03955*.  
84. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space.  