1EB9ẹ

Assessing Gender Bias in Machine Translation – A Case Study with Google Translate

April 18, 2018

Abstract

Recently there has been a growing concern in academia, industrial research labs and the mainstream commercial media about the phenomenon dubbed as *machine bias*, where trained statistical models – unbeknownst to their creators – grow to reflect controversial societal asymmetries, such as gender or racial bias. A significant number of Artificial Intelligence tools have recently been suggested to be harmfully biased towards some minority, with reports of racist criminal behavior predictors, Apple's Iphone X failing to differentiate between two distinct Asian people and the now infamous case of Google photos' mistakenly classifying black people as gorillas. Although a systematic study of such biases can be difficult, we believe that automated translation tools can be exploited through gender neutral languages to yield a window into the phenomenon of gender bias in AI.

In this paper, we start with a comprehensive list of job positions from the U.S. Bureau of Labor Statistics (BLS) and used it in order to build sentences in constructions like "He/She is an Engineer" (where "Engineer" is replaced by the job position of interest) in 11 different gender neutral languages such as Hungarian, Chinese, Yoruba, and several others. We translate these sentences into English using the Google Translate API, and collect statistics about the frequency of female, male and gender-neutral pronouns in the translated output. We then show that Google Translate exhibits a strong tendency towards male defaults, in particular for fields typically associated to unbalanced gender distribution or stereotypes such as STEM (Science, Technology, Engineering and Mathematics) jobs. We ran these statistics against BLS' data for the frequency of female participation in each job position, in which we show that Google Translate fails to reproduce a real-world distribution of female workers. In summary, we provide experimental evidence that even if one does not expect in principle a 50:50 pronominal gender distribution, Google Translate yields male defaults much more frequently than what would be expected from demographic data alone.

We believe that our study can shed further light on the phenomenon of machine bias and are hopeful that it will ignite a debate about the need to augment current statistical translation tools with debiasing techniques – which can already be found in the scientific literature.

1 Introduction

Although the idea of automated translation can in principle be traced back to as long as the 17th century with René Descartes proposal of an "universal language" (**?**), machine translation has only existed as a technological field since the 1950s, with a pioneering memorandum by Warren Weaver (**??**) discussing the possibility of employing digital computers to perform automated translation. The now famous Georgetown-IBM experiment followed not long after, providing the first experimental demonstration of the prospects of automating translation by the means of successfully converting more than sixty Russian sentences into English (**?**). Early systems improved upon the results of the Georgetown-IBM experiment by exploiting Noam Chomsky's theory of generative linguistics, and the field experienced a sense of optimism about the prospects of fully automating natural language translation. As is customary with artificial intelligence, the initial optimistic stage was followed by an extended period of strong disillusionment with the field, of which the catalyst was the influential 1966 ALPAC (Automatic Language Processing Advisory Committee) report((**?**). Such research was then disfavoured in the United States, making a re-entrance in the 1970s before the 1980s surge in statistical methods for machine translation (**??**). Statistical and example-based machine translation have been on the rise ever since (**???**), with highly successful applications such as Google Translate (recently ported to a neural translation technology (**?**)) amounting to over 200 million users daily.

In spite of the recent commercial success of automated translation tools (or perhaps stemming directly from it), machine translation has amounted a significant deal of criticism. Noted philosopher and founding father of generative linguistics Noam Chomsky has argued that the achievements of machine translation, while successes in a particular sense, are *not successes in the sense that science has ever been interested in*: they merely provide effective ways, according to Chomsky, of approximating unanalyzed data (**??**). Chomsky argues that the faith of the MT community in statistical methods is absurd by analogy with a standard scientific field such as physics (**?**):

> *I mean actually you could do physics this way, instead of studying things like balls rolling down frictionless planes, which can't happen in nature, if you took a ton of video tapes of what's happening outside my office window, let's say, you know, leaves flying and various things, and you did an extensive analysis of them, you would get some kind of prediction of what's likely to happen next, certainly way better than anybody in the physics department could do. Well that's a notion of success which is I think novel, I don't know of anything like it in the history of science.*

Leading AI researcher and Google's Director of Research Peter Norvig responds to these arguments by suggesting that even standard physical theories such as the Newtonian model of gravitation are, in a sense, *trained* (**?**):

> *As another example, consider the Newtonian model of gravitational attraction, which says that the force between two objects of mass m_1 and m_2 a distance r apart is given by*
>
> $$F = Gm_1m_2/r^2$$
>
> *where G is the universal gravitational constant. This is a trained model because the gravitational constant G is determined by statistical inference over the results of a series of experiments that contain stochastic experimental error. It is also a deterministic (non-probabilistic) model because it states an exact functional relationship. I believe that Chomsky has no objection to this kind of statistical model. Rather, he seems to reserve his criticism for statistical models like Shannon's that have quadrillions of parameters, not just one or two.*

Chomsky and Norvig's debate (**?**) is a microcosm of the two leading standpoints about the future of science in the face of increasingly sophisticated statistical models. Are we, as Chomsky seems to argue, jeopardizing science by relying on statistical tools to perform predictions instead of perfecting traditional science models, or are these tools, as Norvig argues, components of the scientific standard since its conception? Currently there are no satisfactory resolutions to this conundrum, but perhaps statistical models pose an even greater and more urgent threat to our society.

On a 2014 article, Londa Schiebinger suggested that scientific research fails to take gender issues into account, arguing that the phenomenon of male defaults on new technologies such as Google Translate provides a window into this asymmetry (**?**). Since then, recent worrisome results in machine learning have somewhat supported Schiebinger's view. Not only Google photos' statistical image labeling algorithm has been found to classify dark-skinned people as gorillas (**?**) and purportedly intelligent programs have been suggested to be negatively biased against black prisoners when predicting criminal behavior (**?**) but the machine learning revolution has also indirectly revived heated debates about the controversial field of physiognomy, with proposals of AI systems capable of identifying the sexual orientation of an individual through its facial characteristics (**?**). Similar concerns are growing at an unprecedented rate in the media, with reports of Apple's Iphone X face unlock feature failing to differentiate between two different Asian people (**?**) and automatic soap dispensers which reportedly do not recognize black hands (**?**). *Machine bias*, the phenomenon by which trained statistical models unbeknownst to their creators grow to reflect controversial societal asymmetries, is growing into a pressing concern for the modern times, invites us to ask ourselves whether there are limits to our dependence on these techniques – and more importantly, whether some of these limits have already been traversed. In the wave of algorithmic bias, some have argued for the creation of some kind of agency in the likes of the Food and Drug Administration, with the sole purpose of regulating algorithmic discrimination (**?**).

With this in mind, we propose a quantitative analysis of the phenomenon of gender bias in machine translation. We illustrate how this can be done by simply exploiting Google Translate to map sentences from a gender neutral language into English. As Figure 1 exemplifies, this approach produces results consistent with the hypothesis that sentences about stereotypical gender roles are translated accordingly with high probability: *nurse* and *baker* are translated with female pronouns while *engineer* and *CEO* are translated with male ones.

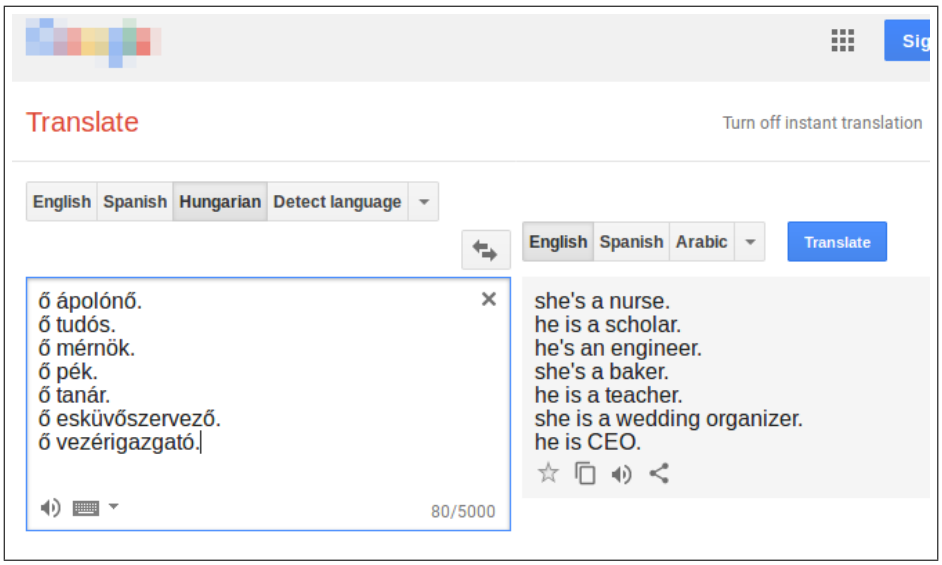

Figure 1: Translating sentences from a gender neutral language such as Hungarian to English provides a glimpse into the phenomenon of gender bias in machine translation. This screenshot from Google Translate shows how occupations from traditionally male-dominated fields (**?**) such as scholar, engineer and CEO are interpreted as male, while occupations such as nurse, baker and wedding organizer are interpreted as female.

2 Motivation

As of 2018, Google translate is one of the largest publicly available machine translation tools in existence, amounting 200 million users daily(**?**). Initially relying on United Nations and European Parliament transcripts to gather data, since 2014 Google Translate has inputed content from its users through the Translate Community initiative(**?**). Recently however there has been a growing concern about gender asymmetries in the translation mechanism, with some heralding it as "sexist" (**?**). This concern has to at least some extent a scientific backup: A recent study has shown that word embeddings are particularly prone to yielding gender stereotypes(**?**). Fortunately, the researchers propose a rela-

tively simple *debiasing* algorithm with promising results: they were able to cut the proportion of stereotypical analogies from 19% to 6% without any significant compromise in the performance of the word embedding technique. They are not alone: there is a growing effort to systematically discover and resolve issues of algorithmic bias in black-box algorithms(**?**). The success of these results suggest that a similar technique could be used to remove gender bias from Google Translate outputs, should it exist. This paper intends to investigate whether it does. We are optimistic that our research endeavors can be used to argue that there is a positive payoff in redesigning modern statistical translation tools.

3 Assumptions and Preliminaries

In this paper we assume that a statistical translation tool should reflect at most the inequality existent in society – it is only logical that a translation tool will poll from examples that society produced and, as such, will inevitably retain some of that bias. It has been argued that one's language affects one's knowledge and cognition about the world (**?**), and this leads to the discussion that languages that distinguish between female and male genders grammatically may enforce a bias in the person's perception of the world, with some studies corroborating this, as shown in (**?**), as well some relating this with sexism (**?**) and gender inequalities (**?**).

With this in mind, one can argue that a move towards gender neutrality in language and communication should be striven as a means to promote improved gender equality. Thus, in languages where gender neutrality can be achieved – such as English – it would be a valid aim to create translation tools that keep the gender-neutrality of texts translated into such a language, instead of defaulting to male or female variants.

We will thus assume throughout this paper that although the distribution of translated gender pronouns may deviate from 50:50, it should not deviate to the extent of misrepresenting the demographics of job positions. That is to say we shall assume that Google Translate incorporates a negative gender bias if the frequency of male defaults overestimates the (possibly unequal) distribution of male employees per female employee in a given occupation.

4 Materials and Methods

We shall assume and then show that the phenomenon of gender bias in machine translation can be assessed by mapping sentences constructed in gender neutral languages to English by the means of an automated translation tool. Specifically, we can translate sentences such as the Hungarian " poln", where "poln" translates to "nurse" and "" is a gender-neutral pronoun meaning either he, she or it, to English, yielding in this example the result "she's a nurse" on Google Translate. As Figure 1 clearly shows, the same template yields a male pronoun when "nurse" is replaced by "engineer". The same basic template can be ported

to all other gender neutral languages, as depicted in Table **??**. Given the success of Google Translate, which amounts to 200 million users daily, we have chosen to exploit its API to obtain the desired thermometer of gender bias. Also, in order to solidify our results, we have decided to work with a fair amount of gender neutral languages, forming a list of these with help from the World Atlas of Language Structures (WALS) (**?**) and other sources. Table **??** compiles all languages we chose to use, with additional columns informing whether they (1) exhibit a pronominal gender system and (2) are supported by Google Translate. Because pronominal gender systems defy the purposes of our technique, such languages have been discarded.

There is a prohibitively large class of nouns and adjectives that could in principle be substituted in the templates of Table **??**. To simplify our dataset, we have decided to obtain a comprehensive list of professional occupations, which, we believe, are an interesting window into the nature of gender bias. Here, we resorted to using the Bureau of Labor Statistics' detailed occupations Table (**?**), from the United States Department of Labor, for providing such a list. The values inside, however, had to be expanded since each line contained multiple occupations and sometimes very specific ones. Fortunately this table also provided a percentage of women participation in the jobs shown, for those that had more than 50 thousand workers. We filtered some of these because they were too generic ("Computer occupations, all other", and others) or because they had gender specific words for the profession ("host/hostess", "waiter/waitress"). We then separated the curated jobs into broader categories (Artistic, Corporate, Theatre, etc.) as shown in Table **??**. Finally, Table **??** shows thirty examples of randomly selected occupations from our dataset. For the occupations that had less than 50 thousand workers, and thus no data about the participation of women, we assumed that its women participation was that of its upper category. Finally, we have selected a small list of 21 adjectives, presented in Table **??**.

5 Distribution of translated gender pronouns per occupation category

A sensible way to group translation data is to coalesce occupations in the same category and collect statistics about how prominent male defaults are in each field. What we have found is that Google Translate does indeed translate sentences with male pronouns with greater probability than it does either female or gender-neutral pronouns. Furthermore, this bias is seemingly aggravated for fields suggested to be troubled by male stereotypes, such as life and physical sciences, architecture, engineering, computer science and mathematics (**?**). Table **??** summarizes these data, and Table **??** summarizes it even further by coalescing occupation categories into broader groups to ease interpretation. For instance, STEM (Science, Technology, Engineering and Mathematics) fields are grouped into a single category, which helps us compare the large asymmetry between gender pronouns in these fields (80.322% of male defaults) to that of

more evenly distributed fields such as healthcare (59.014%).

Plotting histograms for the number of gender pronouns per occupation category sheds further light on how female, male and gender-neutral pronouns are differently distributed. The histogram in Figure **??** (and its coalesced variant in Figure **??**) suggests that the number of female pronouns is inversely distributed – which is mirrored in the data for gender-neutral pronouns in Figures **??** and **??** –, while the same data for male pronouns (shown in Figures **??**, **??**) suggests a skew normal distribution. Furthermore we can see both on Figures **??** and **??** how STEM fields (labeled in red) exhibit predominantly male defaults – amounting predominantly near $X = 0$ in the female histogram although much to the right in the male histogram.

These values contrast with LBS' report of gender participation, which will be discussed in more detail in Section 8.

6 Distribution of translated gender pronouns per language

We have taken the care of experimenting with a fair amount of different gender neutral languages. Because of that, another sensible way of coalescing our data is by language groups, as shown in Table **??**. This can help us visualize the effect of different cultures in the genesis – or lack thereof – of gender bias. Nevertheless, the barplots in Figure 2 are perhaps most useful to identifying the difficulty of extracting a gender pronoun when translating from certain languages. Japanese and Basque are good examples of this difficulty, although the quality of Turkish, Chinese and Yoruba translations are also compromised.

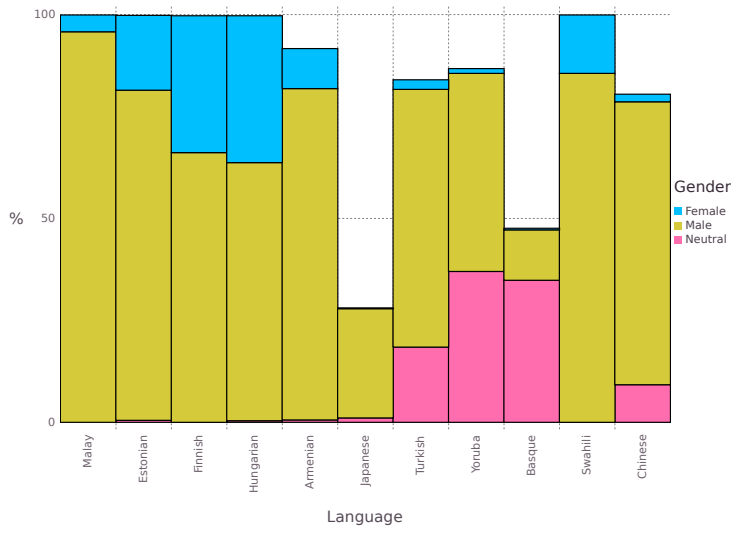

Figure 2: The distribution of pronominal genders per language also suggests a tendency towards male defaults, with female pronouns reaching as low as 0.196% and 1.865% for Japanese and Chinese respectively. Once again not all bars add up to 100% as Google Translate occasionally fails to translate sentences, particularly in Japanese and Basque. Among all tested languages, Basque was the only one to yield more gender neutral than male pronouns, with Yoruba and Turkish following after in this order.

7 Distribution of translated gender pronouns for varied adjectives

Apart from occupations, which we have exhaustively examined by collecting labor data from the U.S. Bureau of Labor Statistics, we have also selected a small and for all purposes not representative subset of adjectives, in an attempt to provide preliminary evidence that the phenomenon of gender bias may extend beyond the professional context examined in this paper.

Once again the data points towards male defaults, but some variation can be observed throughout different adjectives. Sentences containing the words *Shy*, *Attractive*, *Happy*, *Kind* and *Ashamed* are predominantly female translated (*Attractive* is translated as female and gender-neutral in equal parts), while *Arrogant*, *Cruel* and *Guilty* are disproportionately translated with male pronouns (*Guilty* is in fact never translated with female or neutral pronouns).

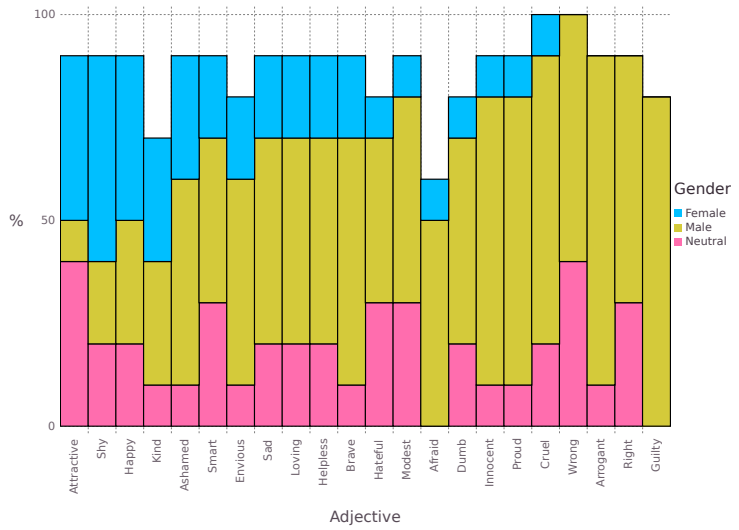

Figure 3: The distribution of pronominal genders for each word in Table **??** shows how stereotypical gender roles can play a part on the automatic translation of simple adjectives. One can see that adjectives such as *shy* and *attractive* are predominantly translated with female pronouns, while words like *guilty*, *innocent*, *wrong*, *right*, *arrogant* are almost exclusively translated with male pronouns. Objective statements have a tendency towards male defaults, while statements concerning emotional states (*shy*, *happy*, *kind*, *ashamed*) amass at the other extreme of the sex ratio spectrum.

8 Comparison with women participation data across job positions

A sensible objection to the conclusions we draw from our study is that the perceived gender bias in Google Translate results stems from the fact that (possibly) female participation in some job positions is itself low. We must account for the possibility that the statistics of gender pronouns in Google Translate outputs merely reflects the demographics of male-dominated fields (male-dominated fields can be considered those that have less than 25% of women participation(**?**), according to the U.S. Department of Labor Women's Bureau). In this context, the argument in favor of a critical revision of statistic translation algorithms weakens considerably, and possibly shifts the blame away from these tools.

The U.S. Bureau of Labor Statistics data summarized in Table **??** contains statistics about the percentage of women participation in each occupation category. These data is also available for each individual occupation, which allows us to compute the frequency of women participation for each 12-quantile. We carried the same computation in the context of frequencies of translated female

pronouns, and the resulting histograms are plotted side-by-side in Figure 4. The data shows us that Google Translate outputs fail to follow the real-world distribution of female workers across a comprehensive set of job positions. The distribution of translated female pronouns is consistently inversely distributed, with female pronouns accumulating in the first 12-quantile. By contrast, BLS data shows that female participation peaks in the fourth 12-quantile and remains significant throughout the next ones.

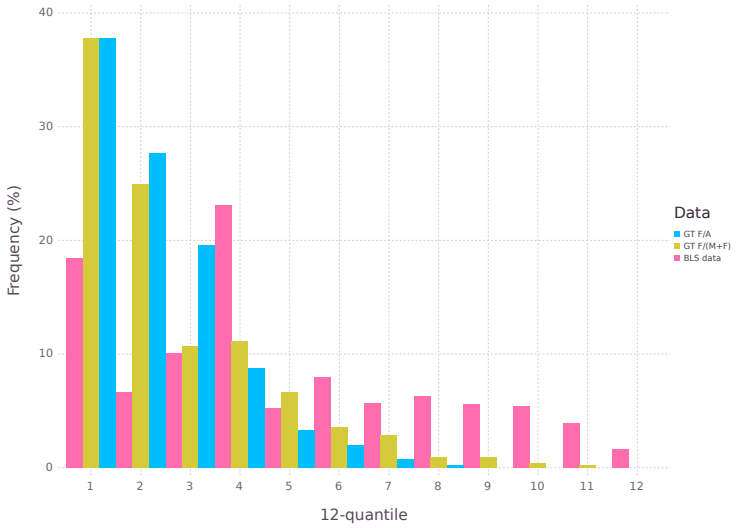

Figure 4: Women participation (%) data obtained from the U.S. Bureau of Labor Statistics allows us to assess whether the Google Translate bias towards male defaults is at least to some extent explained by small frequencies of female workers in some job positions. Our data does not make a very good case for that hypothesis: the total frequency of translated female pronouns (in blue) for each 12-quantile (which is consistently inversely distributed) does not seem to respond to the higher proportion of female workers (in red) in the last quantiles. To remove doubt we have also plotted in yellow the frequency of translated female pronouns among male and female pronouns (i.e., excluding eventual translation errors and gender-neutral pronouns).

Running a t-test on the percentage of female translations over all translations for every occupation against the percentage of female participation in an occupation (surrogating with the LBS class value for the profession, in case it was missing), we see a mean of 11.11% gender participation in the translations and a mean of 35.99% in the LBS report. The variance reported for the translation results is also lower, at ≈ 0.01451 in contrast with the report's ≈ 0.06668. Altogether, the Pearson correlation between the two values is ≈ 0.30356, with the p-values for one and two tail both being lower than the alpha value chosen

($\alpha = 0.05$). This result enforces the fact that the translation exhibits a tendency towards male defaults, even when the female participation in that job is higher than expected.

9 Conclusions

In this paper, we have provided preliminary evidence that statistical translation tools such as Google Translate can exhibit gender biases and a strong tendency towards male defaults. Although implicit, these biases possibly stem from the real world data which is used to train them, and in this context possibly provide a window into the way our society talks (and writes) about women in the workplace. In this paper, we suggest that and test the hypothesis that statistical translation tools can be probed to yield insights about stereotypical gender roles in our society – or at least in their training data. By translating professional-related sentences such as "He/She is an engineer" from gender neutral languages such as Hungarian and Chinese into English, we were able to collect statistics bout the asymmetry between female and male pronominal genders in the translation outputs. Our results show that male defaults are not only prominent but exaggerated in fields suggested to be troubled with gender stereotypes, such as STEM (Science, Technology, Engineering and Mathematics) jobs. And because Google Translate typically uses English as a *lingua franca* to translate between other languages (e.g. Chinese → English → Portuguese) (**??**), our findings possibly extend to translations between gender neutral languages and non-gender neutral languages (apart from English) in general, although we have not tested this hypothesis.

Although not conclusive, our results seem to suggest that this phenomenon extends beyond the scope of the workplace, with the proportion of female pronouns varying significantly according to adjectives used to describe a person. Adjectives such as *Shy* and *Attractive* are predominantly translated with female pronouns, while *Guilty* and *Cruel* are almost exclusively translated with male pronouns. Different languages also seemingly have a significant impact in machine gender bias, with Hungarian exhibiting a better equilibrium between male and female pronouns than for example Chinese. Some languages such as Yoruba and Basque were found to translate sentences with gender neutral pronouns very often, although this is the exception rather than the rule and these languages also exhibit a high frequency of translation errors.

To solidify our results, we ran our pronominal gender translation statistics against the U.S. Bureau of Labor Statistics data on the frequency of women participation for each job position. Although Google Translate exhibits male defaults, this phenomenon may merely reflect the unequal distribution of male and female workers in some job positions. To test this hypothesis, we compared the distribution of female workers with the frequency of female translations, finding no correlation between said variables. Our data shows that Google Translate outputs fail to reflect the real-world distribution of female workers, under-estimating the expected frequency. That is to say that even if we do

not expect a 50:50 distribution of translated gender pronouns, Google Translate exhibits male defaults in a greater frequency that job occupation data alone would suggest. The prominence of male defaults in Google Translate is therefore to the best of our knowledge yet lacking a clear justification.

We think this work shed new light on a pressing ethical difficulty arising from modern statistical machine translation, and hope that it will lead to discussions about the role of AI engineers on minimizing potential harmful effects of the current concerns about machine bias. We are optimistic that unbiased results can be obtained with relatively little effort and marginal cost to the performance of current methods, to which current *debiasing* algorithms in the scientific literature are a testament.