EasyChair Preprint

№ 10954

Does AI Reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation Tools

Marco Smacchia, Stefano Za and Álvaro Arenas

September 23, 2023

Does AI reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation tools

Marco Smacchia[1], Stefano Za[1] and Alvaro Arenas[2]

[1] University "G. D'Annunzio" of Chieti-Pescara. Pescara, Italy
[2] IE Business School. Madrid, Spain
marco.smacchia@unich.it
stefano.za@unich.it
alvaro.arenas@ie.edu

Abstract. Natural language processing tools are becoming more and more important in our daily life, enabling us to perform many tasks in a timely and efficient manner. However, as the utilisation of these tools growth, so does the risk of unexpected consequences due to the presence of bias. This study investigates the presence of gender bias within the most popular neural machine translation and large language model tools. We defined a set of Italian sentences concerning ten specific jobs, where the gender of the subjects is not explicitly mentioned. Employing those AI tools, we translated the sentences from Italian to English, requiring the gender to be explicitly mentioned. Afterwards, we developed a survey to obtain human translations for the same sentences, allowing us to compare the differences between the responses generated by the tools and those from individuals. Results show a high presence of gender bias especially for the jobs associated with a male gender and demonstrate a consistency between the outcome obtained by the tools and the results of the survey. These findings serve as a starting point for exploring the origins of gender bias within natural language processing tools and how they reflect gender distributions in our society and human behaviour regarding job occupations.

Keywords: Natural Language Processing, Neural Machine Translation, Large Language Models, Gender Bias, Artificial Intelligence Bias.

1 Introduction

In the last few years, artificial intelligence (AI) has gained much attention from both scholars and practitioners, with a particular focus on societal aspects [1]. Within this context, natural language processing (NLP) tools, such as chatbots, translators and speech recognition, have witnessed a surge in popularity because they allow users to interact without using machine language [2]. NLP, a subfield of AI, encompasses the development of machine algorithms capable of understanding human language, consisting of two main branches: natural language understanding and natural language generation [2,3]. The first branch is related to the study of linguistics and to the understanding of human language, while the second is focused on the process of producing such

language [2]. NLP finds application in various tasks, such as automatic summarisation, discourse analysis, speech recognition, machine translations, and many others (ibid). In the last few years, this technology gained popularity due to the introduction of neural networks and large language models (LLM) that are able to solve many NLP tasks without having specific training data for a given job (zero-shot tasks) [4–6]. Machine translation (MT) is one of the NLP fields significantly impacted by this transition [6].

MT allows a machine to translate natural language sentences, providing an instant translation of words, sentences, documents and even of speeches [7,8]. In the past, MT primarily relied on a set of predetermined translation rules and linguistic knowledge. However, nowadays state-of-the-art MT systems implement neural networks that use a training dataset of parallel texts to translate an input sentence to a target output by using the information acquired during its training [9]. MT serve as a useful tool that simplifies the way we communicate with people that speak other languages by providing large volumes of translated text or speech in a few seconds [10]. As the industry value continues to growth, so does the number of tools available in the market that are used worldwide by companies and individuals for many tasks[1,2]. Since this technology is broadly used, unexpected results such as mistranslations or biases could have severe implications. Vieira et al. [8] conducted a review in which they explore MT bias across legal and medical fields. They reported interesting examples such as the low accuracy of Google Translate for translating medical instructions to patients suffering from diabetes. Additionally, they discuss an incident in which Google Translate was used by a police officer to obtain consent during inspections, which later faced challenges during the court audience due to concerns about overcoming language barriers. In both cases, the translation tools outcomes can cause serious damage to the individuals involved, underscoring the paramount importance of addressing these kinds of issues.

In this study, we focus on examining the presence of gender bias within AI translation tools. Specifically, we selected five jobs mostly done by women and five jobs mostly done by men (on the basis of relevant reports). We then we constructed sentences involving these jobs and translated them from Italian to English using the API of five different AI translation tools (Google Translate, Deepl, Microsoft Azure, GPT 3.5 and Text DaVinci). All the original sentences in Italian did not include a subject (the gender was not made explicit). Since in Italian it is possible omitting the subject in a sentence, our experiment aims at measuring the gender bias in English translations where, on the other hand, the subject must always be made explicit. To compare the results obtained from an AI with those of human beings, we also developed a questionnaire that we submitted to a sample of Italian individuals.

The purpose of this study is to address the limitations identified in the literature by investigating the presence of gender bias in the outcome of five different AI tools when

---

[1] Global Market Insights. (April 6, 2017). Machine translation market size worldwide, from 2016 to 2024 (in million U.S. dollars) [Graph]. In Statista. Retrieved May 14, 2023, from https://www.statista.com/statistics/748358/worldwide-machine-translation-market-size/.

[2] Stanford University. (April 15, 2023). Number of commercial machine translation (MT) programs available globally from 2017 to 2022 [Graph]. In Statista. Retrieved May 14, 2023, from https://www.statista.com/statistics/1378793/mt-translation-programs-number/.

sentences without subject are presented. More specifically, we aim to answer the following questions:

1. *To what extent does gender bias occur inside the most common AI translation tools?*
2. *Related to gender bias, are there differences in the output of a translation depending on how the query is formulated?*
3. *Concerning the translation output, which jobs are most likely to be biased?*
4. *What are the results of a comparison between the results obtained from an AI with those obtained by Human beings?*

The contribution of this study is structured as follows. In the next section, we provide a theoretical background about the gender bias inside MT along with the limitations detected and outline the methodology we used to perform our research. We subsequently present the results obtained together with the discussion. Finally, in the conclusion section, we state the theoretical and practical contribution together with the limitations and avenues for future research.

2 Theoretical Background

Inside the AI debate, research focusing on identifying, investigating, and addressing biases has become increasingly relevant [11]. Although AI and its subfields bring enormous potentialities and advantages in many domains, the presence of biases could bring dangerous and unpleasant consequences [12]. Since AI is used worldwide by private, public and governmental organisations to support the decision-making process of many activities, its outcome influences many people simultaneously [13]. Bias inside data and resources used to train the AI model or inside the algorithm could cause and amplify issues such as discrimination, inequality, and lack of privacy and trust [12,14–17].

Among the possible biased outcomes concerning AI technology, the presence of gender bias inside neural MT tools has a significant relevance [18–20]. Gender bias occurs when one gender is preferred or prejudiced with respect to the others. NLP algorithms or training datasets, on which machine translation tools are based, could reinforce or amplify damaging gender stereotypes [17]. There are many real-world cases related to this issue, from resume filtering giving preference to male candidates, to credit applications refused to women even if they have a higher score than a male applicant [17,21].

Gender bias in MT tools based on AI could be attributed to different factors. For example, behaviours or habits embedded in society described by the training data could affect the outcome of a translator. Technical issues in the data or the algorithm used to develop the artefact could lead to some forms of gender bias. Finally, since AI tools are capable of learning, different use in various contexts could influence the behaviour of the MT [22].

The presence of gender bias inside neural MT tools is not a new topic. There are many press articles that underlined sexist behaviour in online translators[3-4]. In addition, several articles also tried to inspect and also address this issue. Stanovsky et al., [18] propose an evaluation method for gender bias inside MT tools based on morphological analysis. They analysed six MT tools both industrial and academic and detected the presence of translation errors related to gender. Another study focuses on reducing gender bias by training a neural MT tool and applying several automated techniques to address the issue. Results indicate that when these techniques are used to reduce the gender bias in a neural MT tool, the performances are worse with respect to the baseline system. Among the methods used only domain adaptation provides better results after a model is trained [19]. Bernagozzi et al., [20], used a method proposed by Srivastava et al., [23] to evaluate the presence of gender bias in MT tools that could be used by third parties. Moreover, through a questionnaire, they inspect the impact that such biases have on people and how this issue is perceived. Schiebinger [24] underlines the necessity to analyse gender bias inside MT tools as she notes their trend to translate with a default masculine form.

Although this phenomenon has been studied several times, there are some aspects that still need to be addressed. The language source of multiple datasets used to perform the research is English [18,20]. Many studies have not considered the fact that in some languages sentences without a subject are valid while, in other languages, the subject must be outlined [23]. Moreover, some contributions use a limited number of tools to make comparisons and simple sentences [20,23]. To carry out the research we adapted the framework of Srivastava et al., [23] to assess the presence of gender bias inside MT tools without having access to their source code or training data.

In the next section, we describe the process adopted to conduct our research. Starting from studies on gendered language concerning the reinforcement of gender stereotypes in job titles [25,26] and from statistics about the occupations mostly performed by a specific gender, we developed a set of sentences to translate. Every phrase is related to a specific job that could be used to identify both men and women (gendered-language free). We avoided referring to occupations that could be related specifically to masculine or feminine forms to avoid possible bias due to gendered language.

3 Methodology

This contribution adopts an exploratory design as the main aim is to *seek new insights* into the evaluation of gender bias inside AI translation tools [27]. Our primary goal was to assess the presence of bias in translators when a sentence with no subject (a common practice in some languages) is translated into a sentence in which the subject must be

[3] Parmy Olson, The Algorithm That Helped Google Translate Become Sexist, Forbes. Availabel at: https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/.

[4] Shivali Best, Is Google Translate SEXIST? Users report biased results when translating gender-neutral languages into English, Daily Mail. Available at: https://www.dailymail.co.uk/sciencetech/article-5136607/Is-Google-Translate-SEXIST.html

mandatorily made explicit. To this end, we have developed a set of sentences in Italian in which the subject has not been made explicit and which can be attributed to either a male or a female and then we translated them in English. Each sentence focuses on ten particular jobs, which we chose by searching within Eurostat databases[5-6], in World Bank Group[7] and U.S. Bureau of Labor Statistics[8] reports. More specifically, we selected the five jobs mostly done by women and just as many held mostly by men. In order to be selected, the name of the job had to be impersonal in Italian (not attributable to a male or female audience). We developed three sentences for each job, the first one quite simple, while the other two more elaborate. Since all sentences have been translated in English, we arranged for the first sentence to contain only personal pronouns, while the other two contained both personal pronouns and possessive adjectives. As an example, the following three sentences are those already translated in English and referring to carer (in bold the words concerning the gender):

- ***She*** *said* ***she*** *works as a carer in the home of an elderly person.*
- ***She*** *started* ***her*** *job as a carer after an elderly person requested assistance.*
- *After losing* ***her*** *job as a carer,* ***she*** *thought of changing life.*

In the next step, we selected the tools used to perform our study. We chose them according to two parameters: the popularity of the translator and the availability of the APIs. After a screening, we chose three of the most used neural MT tools in the market: Google Translate, DeepL, and Microsoft Azure. Since LLM models are becoming more and more relevant inside the debate and are also trained on much bigger datasets than translators, we decided to add to the current list, the models GPT 3.5 turbo and Text-DaVinci 003 on which Chat GPT is based [5,28]. We referred to the API documentation of these MT tools to write a Python script in which we translated the 30 sentences in three different ways, first, we translated the entire text corpus (all sentences at once), then we translated all the corpus from an external document and finally, we translated the sentences one by one. The results, consisting of 90 translations per tool, were listed in an Excel sheet and labelled according to the gender that was made explicit in the English translation (M for male, F for female and N for neutral) and job. Depending on the output of the translation we classified the results as *biased* when the output reflects the underlying data used to develop the sentences, *converging biased* when the output is influenced by the previous translations (underlying data) and converges towards the same result and *unbiased* when the output is translated using in a neutral way or the translation path doesn't reflect a biased behaviour.

---

[5] European Jobs Monitor 2021: Gender gaps and the employment structure. Available at: https://www.aranagenzia.it/attachments/article/12440/European%20Jobs%20Monitor%202021-Gender%20gaps%20and%20the%20employment%20structure.pdf

[6] Jobs still split along gender lines. Available at: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20180307-1

[7] Gender-based Employment Segregation: Understanding Causes and Policy Interventions. Available at: https://documents1.worldbank.org/curated/en/483621554129720460/pdf/Gender-Based-Employment-Segregation-Understanding-Causes-and-Policy-Interventions.pdf

[8] Labor Force Statistics from the Current Population Survey. Available at: https://www.bls.gov/cps/cpsaat39.htm.

We then analysed the results measuring the presence of bias in the translation considering each tool both individually and in aggregated form. Then we compared the results with those obtained from the questionnaire for each job. Lastly, we consider the translation trend between the different responses to the last question concerning the aim of the study.

To better explain the processes on which the research is based we graphically represented our research protocol (Fig. 1) that we adapted from the model of Srivastava et al., [23] to assess the presence of bias inside MT tools by third parties. Moreover, to assess the presence of bias, we take into account the statistics concerning the gender gap inside the selected jobs. To make the model clearer and more reproducible we used the business process modelling notation (BPMN).

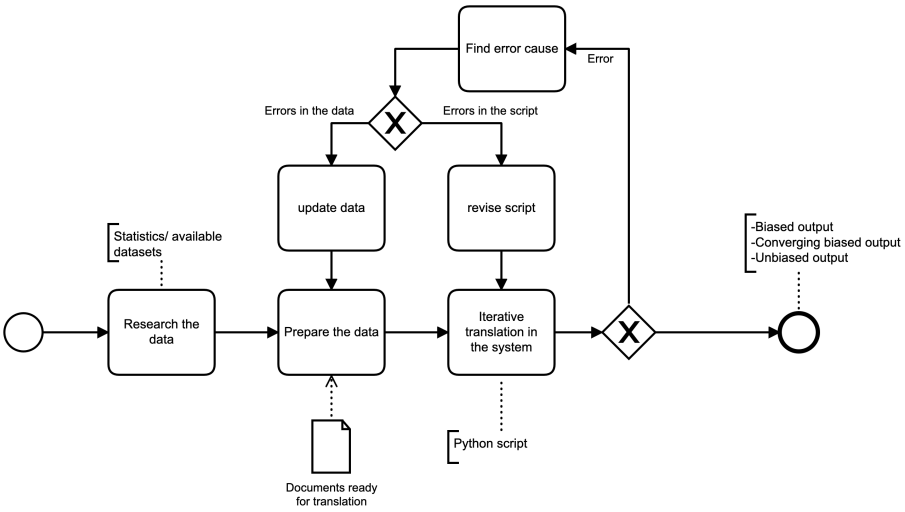

Fig. 1. Research protocol

In this study, we also developed a questionnaire to compare the results obtained by AI tools with those obtained by humans. More specifically, our aim is to understand whether gender bias is still present in the way people think in our society.

The questionnaire reported the 30 sentences in Italian translated into English, to which all references regarding gender (pronouns and adjectives) were appropriately removed, by asking the respondent to complete the sentences giving them a proper meaning (fill in the blank questions). All the sentences were presented in the same order they were submitted to the AI tools. At the beginning and at the end of the form were also asked demographic questions. The first question was related to the English level to assess the familiarity of the respondents with the language. After completing the 30 questions related to the translated sentences, respondents were asked: the gender they identify with, their Italian region of origin, their age, and their educational qualification. All the demographic questions, although mandatory had the option “I prefer not to answer” or “other”. Finally, we added a further question asking the respondent to explain

in a few words what was in her/his opinion the main aim of this study. After the questionnaire was completed, we provided information about the objective of the research.

We decided to use the fill in the blank structure for various reasons. First, letting people translate all the sentences would have gone beyond our purposes and would have led to possible errors and a difficult analysis of the outcomes, caused by an excessive amount of noise within the dataset. Then, we wanted to make the questionnaire available for the people that are not proficient in the English language, by reducing the possibility of mistakes due to their lack of expertise.

We submitted the questionnaire to a random sample of 100 Italian speakers. The respondents were 53. Of these responses, 11 were discarded because they were inconsistent with what was required within the questionnaire. The data concerning the 42 questionnaires, was collected from late April to early June 2023. To develop the survey, we used the tool Jotform. Further information about the sample is discussed in paragraph 4.2.

4 Results

In this section we present the results obtained by the analysis conducted on the translation tools first and then the outcome of the questionnaire.

4.1 Analysis of AI translation tools

We analyse the results given by the MT tools considering first the outcome of every single tool, concerning the presence of bias, taking into account the statistics that we used to select the jobs to develop the sentences (Fig. 2). For each AI tool, we provide the percentage of female, male and neutral translations both in aggregate (statistics for the five jobs done mostly by men and by women) and individual form (statistics for every single job selected). Afterwards, we summarise those results in a combined diagram.

Microsoft Azure. The tool is consistent between the three methods of translation. We obtained the same results by translating the 30 sentences from an external document, the whole text at once and one sentence at a time. Concerning the distinction between groups' jobs (male vs female according to the statistics) results point out a balance regarding the translations of the sentences related to the occupations held mostly by women (53% female translations vs 47% male translations). While the tool obtains good results regarding the previous group, when it comes to the one populated by jobs done mostly by men the outcome becomes highly biased. In this group, all the translations are in the masculine form (0% female translations vs 100% male translations). By looking at single jobs, almost all of them are always translated using the masculine or the feminine form. The only exception is represented by kindergarten teacher (67% female translations vs 33% male translations). Finally, the only two occupations translated with the feminine form are carer (translated by the tool as caregiver) and secretarial assistant. All the other jobs are translated always using male adjectives and pronouns.

DeepL. This tool presents a few differences between the methods of translation. Taking into account the outcome obtained by translating all the text from an external document, there is a perfect distinction between the female group job (100% female translations vs 0% male translations) and the male group job (0% female translations vs 100% male translations). Considering the methods of translating the whole text at once and one sentence at a time, the results are consistent for the occupations held mostly by men (0% female translations vs 100% male translations). While for the female group job, there is a small component translated with the masculine form (80% female translations vs 20% male translations). Concerning the single occupations, we have a clear distinction for the jobs done mostly by women (carer, kindergarten teacher, secretarial assistant and cleaner, 100% female translations vs 0% male translations) and the jobs done mostly by men (truck driver, bricklayer, electrician, engineer and IT technician, (0% female translations vs 100% male translations). The only exception is represented by customer service (33% female translations vs 67% male translations).

Google Translate. This translator has substantial differences between the outcome obtained by translating all the text from an external document and the other two methods. By looking at the results of the first method, all the sentences are translated using the masculine form for both female group job (6% female translations vs 100% male translations) and male group job (0% female translations vs 100% male translations), the only exception is represented only by the first sentence (a simple phrase with the job carer). Concerning the other two methodologies, results indicate a degree of variation in translations related to works carried out by a female majority (66% female translations vs 33% male translations) and almost no variation regarding the male group job (6% female translations vs 94% male translations). Looking at the overall results of the three translation methods, there is almost an equality between masculine and feminine forms in the women's group of jobs (47% female translations vs 53% male translations), while in the male group, there is almost a total trend in translations using the masculine form (4% female translations vs 96% male translations). By analysing the single jobs, it can be noted a variation for all the occupations contained in the female group. Kindergarten teacher, secretarial assistant and cleaner are equal with 44% of translations using feminine forms and 56% using a masculine form. Also, carer (78% female translations vs 22% male translations) and customer service (22% female translations vs 78% male translations) are translated using different forms. Regarding the men group, the only job that has outcome translated with the feminine form is IT technician (22% female translations vs 78% male translations).

Text DaVinci-003. In the tool analysis, we have found consistency between the translation from an external document and the whole text at once. In particular, through these two methods, we obtained a one-way outcome, consisting of all sentences translated with the masculine form for both groups (0% female translations vs 100% male translations). Things change by translating the statements one by one only regarding the female group job (87% female translations vs 13% male translations). Overall, there is an overwhelming majority of the male form of translation, with the female counterpart only present in the third method of translation (14% female translations vs 86% male translations). Concerning the single jobs there is a variation only regarding the female group. Carer, kindergarten teacher and cleaner are translated 33% of the time

with a female form and 67% with a male form. While customer service and secretarial assistant sentences are translated using a feminine form (22%) and using a masculine form (78%).

GPT 3.5 Turbo. The last tool has a difference from the others as it is an LLM tool which the primary function is the generation of textual content and not only translation. Since it doesn't support the translation from an external document, we have opted instead for a continuous interaction (hit and run talk) where we submitted the sentences one by one without resetting the chat. Starting from the method just described results indicate the prevalence of feminine form for the translation of the sentences in both groups of jobs. Only the last sentences in the men group have been translated with a male form and the last one with a neutral form (73% female translations, 20% male translations, 7% neutral translations). Taking into consideration the method regarding the whole text translated at once, all the translations have been made using the male pronouns and adjectives (0% female translations vs 100% male translations). Finally, when the chat was reset for each sentence, we had as a result a small variation between male and female forms in the women group job (67% female translations vs 33% male translations) and a univocal translation concerning men group (0% female translations vs 100% male translations). By looking at the jobs, it could be noted that all the occupations present a variation (carer and secretarial assistant 67% female, 33% male, cleaner and kindergarten teacher 56% female, 44% male, customer service and truck driver 33% female, 67% male, bricklayer, electrician and engineer 22% female, 78% male and IT technician 22% female, 67% male and 11% neutral).

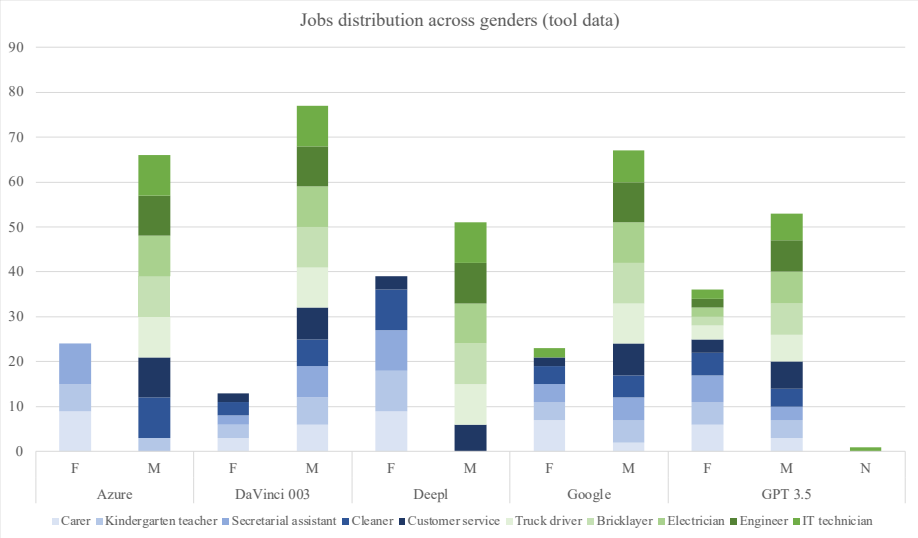

Fig. 2. Jobs distribution across genders (tools taken separately).

Overall results. After inspecting each tool, we decided to take into consideration the results in aggregated form. Results highlight a substantial difference between the female group job (54% female translations vs 46% male translations) and the male group

job (6% female translations, 94% male translations, 0.4% neutral translations). Concerning the jobs (Fig. 3), all occupations related to the men's group are translated almost exclusively using male adjectives and pronouns (truck driver 7% female, 93% male, bricklayer, electrician and engineer 4% female, 96% male and IT technician 9% female, 89% male and 2% neutral). As regards jobs done mainly by women, there is much more variation (carer 76% female, 24% male, kindergarten teacher 60% female, 40% male, secretarial assistant 67% female, 33% male, cleaner 47% female, 53% male and customer service 22% female, 78% male).

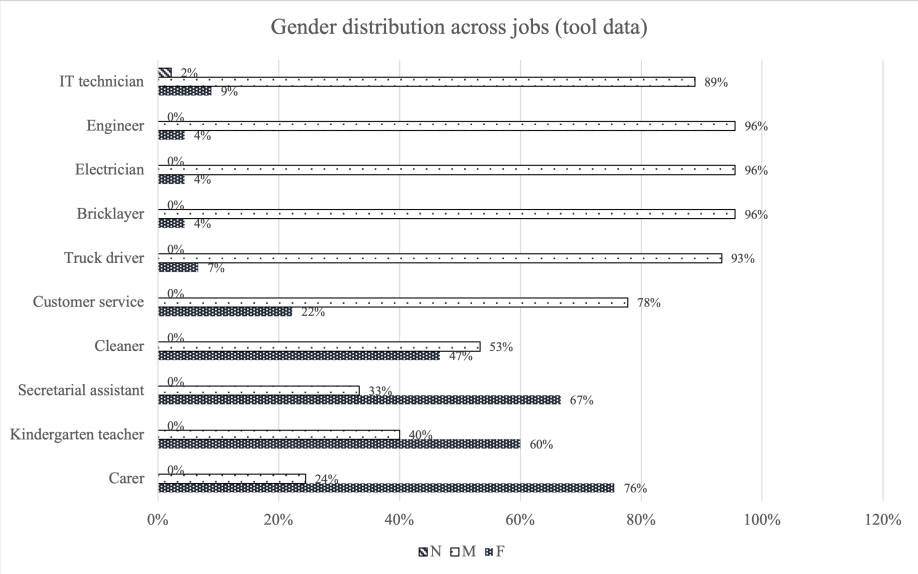

Fig. 3. Gender distribution across jobs taken singularly (tool aggregate data).

4.2 Analysis of the survey

On the basis of the demographic information gathered from the survey, 57% (24) of the respondents identify themselves as woman while 43% (18) as men. The majority of the respondents are between 26-35 years old (55%); the other group is represented by people whose age is between 18 and 25 (40%) and 55-65 (5%). Regarding the Italian region of origin, most of the people who answered the questionnaire came from Lazio (26%) and Abruzzo (43%) (central part of the country). As for the educational background, almost all the respondents have a bachelor's or a master's degree (74%), other participants have a PhD (12%) and a high school diploma (12%), only one preferred not to say his background. Finally, 12% of the participants have a C2 knowledge of the English language, while the 31%, 29%, 23% and 5% have respectively C1, B2, B1 and A2 knowledge. Considering the data in aggregate form it can be distinguished a similar pattern to that observed for data collected from the tools (female group job: 55% female

translations, 37% male translations, 8% neutral translations, male group job: 14% female translations, 79% male translations, 7% neutral translations). Concerning the single jobs there is also consistency between tool data and questionnaire data with slight differences regarding cleaner and engineer. Fig. 4 display the results concerning the differences in translation between the selected jobs. We have more variation concerning the occupations held mostly by women (carer 62% female, 32% male, 6% neutral, kindergarten teacher 56% female, 36% male, 8% neutral, secretarial assistant 55% female, 38% male, 7% neutral, cleaner 65% female, 29% male, 6% neutral and customer service 39% female, 49% male, 12% neutral) than the men's group of jobs (truck driver 4% female, 90% male, 6% neutral, bricklayer 8% female, 87% male, 6% neutral, electrician 11% female, 83% male, 6% neutral, engineer 24% female, 67% male, 9% neutral and IT technician 25% female, 66% male and 10% neutral).

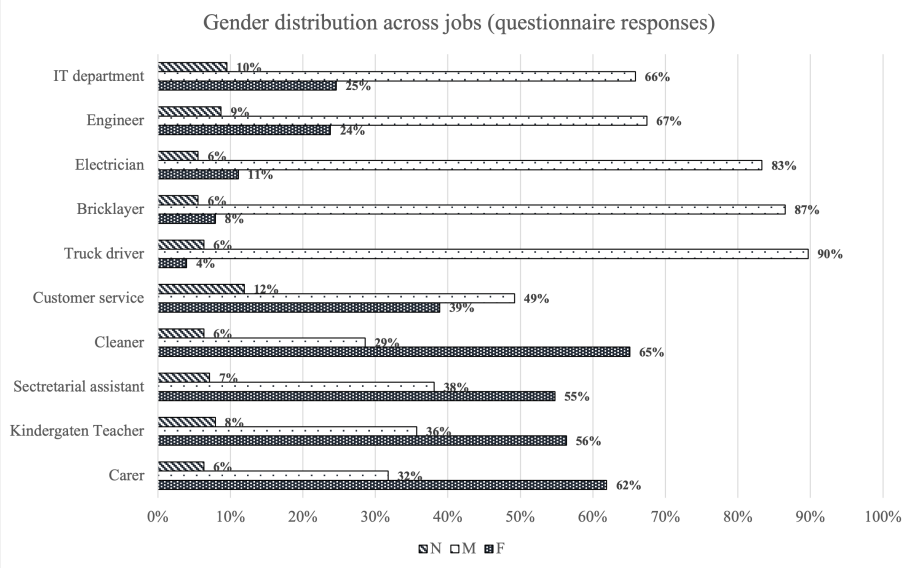

Fig. 4. Gender distribution across jobs taken singularly (questionnaire data).

As the sample is still too small, we decided to compare only the answers of male and female respondents. There is no noticeable difference between the answers, although male individuals (female group job: 46% female translations, 47% male translations, 8% neutral translations and male group job: 18% female translations, 74% male translations, 8% neutral translations) are less biased than female ones (female group job: 63% female translations, 29% male translations, 8% neutral translations and male group job: 12% female translations, 82% male translations, 6% neutral translations). Afterwards, taking into consideration the further question concerning the perceived object of this research, it was possible to see how much awareness of the nature of the analysis had an impact on the responses.

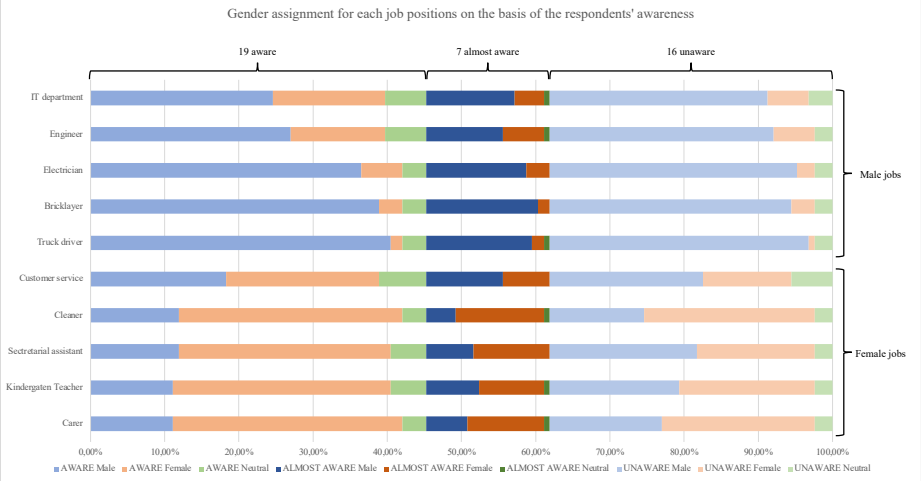

Fig. 5. Gender distribution across jobs, based on the respondent awareness of the research aim.

We divided the answers to the question by categorising them with “unaware” (16 participants) if there is no awareness from the respondent concerning the aim of the research, “almost aware” (7 participants) if there is partial awareness and “aware” (19 participants) if there is total awareness. Looking at the data (Fig. 5), the answers appear very similar regardless of the level of awareness the survey participant has. Respondents belonging to the category “unaware” (female group job: 44% female translations, 48% male translations, 8% neutral translations and male group job: 10% female translations, 83% male translations, 7% neutral translations) appear to be less biased on the female job group than respondents belonging to category “almost aware” (female group job: 57% female translations, 40% male translations, 3% neutral translations and male group job: 19% female translations, 78% male translations, 3% neutral translations) and “aware” (female group job: 62% female translations, 28% male translations, 10% neutral translations and male group job: 17% female translations, 74% male translations, 9% neutral translations), but more biased regarding the men’s job group.

5 Discussion

Based on the gathered data, our objective was to extract the necessary information to address our research questions. In examining the individual translation generated by the MT tools, we identified two primary biases. The first bias occurs mostly when the data is translated from an external document (DeepL, Google, DaVinci) or when the sentences are submitted all at once (DaVinci, GPT). We called this behaviour *converging bias* because the tool is influenced by the first translation output and remains consistent from the beginning until the end of the task. The second bias pertains to the inclination towards a specific gender when it becomes necessary to explicitly state the subject of a

sentence during translation. In instances where the occupation is predominantly associated with women, there exists some diversity in the form (masculine or feminine) used for translation. However, concerning occupations predominantly associated with men, the outputs reveal a presence of significant gender bias, as depicted in Figure 6.

Data used in AI, as it is the case of MT tools, is sourced from user-generated content or collected through other systems developed by humans. Consequently, any bias present in human behaviour become embedded in these systems, and their algorithms may reproduce, or even increase, existing biases [29]. Institutional bias refers to the tendency for the procedures and practices within institutions to operate in ways that favour certain social groups while disadvantaging others [30]. This bias does not necessarily stem from deliberate discrimination, but rather from the majority adhering to existing norms, with institutional sexism and gender discrimination being common examples of institutional bias. Therefore, a plausible explanation of the observed converging bias lies in institutional bias, where some discriminatory practices may have become ingrained in the data used by the MT tools.

In the context of LLM tools, GPT 3.5 gave more interesting results than classical MT tools, especially during the continuous interaction where we submitted the sentences one by one without resetting the chat. At first, the tool was hung up on the translation using only the feminine gender (converging bias), towards the last sentences to be translated it specified the fact that it had changed to the masculine pronoun as the gender of the worker was not specified in Italian. This behaviour represents a form of gender bias in the underlying data as the tool began to translate with male pronouns and adjectives coinciding with work mainly performed by males. During the last sentence, however, he used both the masculine and feminine forms for the translation. Intrigued by his behaviour, we continued to iterate sentences, and although the gender bias continued to be experienced, the tool started to use the neutral form more often. This can be interpreted as a sign of the superiority of LLM's ability to learn from a restricted data set.

Taking into consideration the single jobs, the most biased ones are related to the occupations mostly held by men. All of them have a masculine form translation rate of over 89%, determining the presence of a gender bias within the MT tools. This trend is not explainable by the converging bias, as the translation always began with the female job group. Regarding the occupations mostly detained by women, the most biased job seems to be the carer, followed by the secretarial assistant and the kindergarten teacher. The cleaner is the job that has almost gender equality among the translations, while customer service is biased more on the male side. Finally, IT technician is the only job that was translated using a neutral form (Fig. 3).

When comparing the results of the translation and the results of the questionnaire, no major differences emerge. Looking at Fig. 6, it is interesting to note that the ratio is similar between the two outcomes, although the results of the survey appear to be less biased, especially by looking at the fact that there are more neutral translations in which both male and female pronouns and adjectives are used. The same trend is noted by taking into account the single jobs. With the exception of cleaner, all the jobs appear to be less biased presenting a more equal distribution between genders. Truck driver, bricklayer and electrician, however, are still heavily biased.

Social constructionist theory posits that gender is a socially constructed concept shaped by cultural, societal, and historical factors [31], and may serve as a theoretical explanation for the similar results provided by the MT tools and the questionnaire. As mentioned before, bias in AI can arise when the training data used to develop AI systems reflects and perpetuates societal biases and stereotypes. Hence, if the training data contains gender imbalances or reinforces gender stereotypes, the AI system may learn and perpetuate those biases in its outputs.

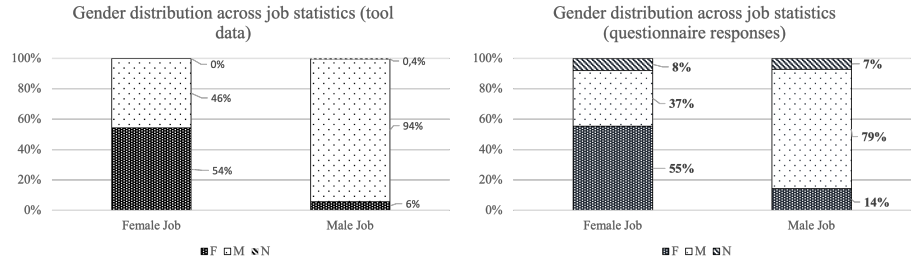

Fig. 6. Gender distribution across male and female group jobs (tools vs questionnaire).

After an aggregate analysis, we further inspected the data gathered from the survey. More specifically, we divided the responses according to gender and according to awareness about the research scope. From this analysis not emerged substantial differences, both in terms of gender difference and awareness of the purpose of the research (Fig. 5). However, the male respondents seem to be less biased with respect to the female respondents, while the participants that had no awareness about the aim of the study are less biased for what concerns the female job group.

Previous research has identified three approaches for bias mitigation [32]. Firstly, pre-processing methods concentrate on the data, with the aim of generating a "balanced" dataset that can be utilised by any learning algorithm. The underlying concept behind these approaches is that by ensuring fairness in the training data, the resulting model will exhibit reduced discrimination. Secondly, in-processing methods centre around the ML algorithm itself, where the classification problem is reformulated by explicitly incorporating the model's discriminatory behaviour into the objective function through regularization or constraints. Thirdly, post-processing methods focus on the ML model by either modifying its internal mechanisms (white-box approaches) or adjusting its predictions (black-box approaches).

6 Conclusion

This study aims to investigate gender bias within some of the most widely used neural machine translation and large language model tools. To conduct this research, we carefully selected a set of thirty Italian sentences related to ten specific gender-neutral occupations that can be expressed in both masculine and feminine forms. In our translations, we intentionally omitted the explicit mention of gender. These sentences were

then translated from Italian to English using various AI tools. Additionally, we designed a survey to gather translations of the same sentences from human participants. The purpose of this survey was to compare the responses obtained from the AI tools with those provided by human translators.

The results reveal the existence of two distinct biases. The first one, called converging bias, describes the tendency of the translation tool to utilise the same form of translation for all occurrences in the dataset, based on the initial translation. The second bias, known as gender bias, signifies a preference for a specific translation form (either male or female) over the other. In our dataset, gender bias is highly associated with occupations held mostly by men, while jobs which are statistically done by a female majority have more variation. We also determined the more biased jobs according to the group selected and investigated the discrepancies arising from different input modalities (full text, external document, one sentence at a time). Lastly, the results of the survey are consistent with the outcomes obtained from the translations of the tools. Notably, there is no significant difference in the outcomes based on the gender of the respondents or their awareness of the research objectives. This research serves as an initial exploration into the presence of gender-related biases within AI translation tools.

In the research, the observed gender bias could have different implications in society. For example, bias could reinforce the behaviour already embedded in the people by perpetuating the stereotypes about jobs and societal roles. Moreover, such issues could under-represent a determined gender leading to translation and non-recognition problems (women or men are not taken into consideration in determined translation concerning a specific job, for example). Under-representation issues in a specific tool could also affect gender representation across different cultures [22].

This paper has both theoretical and practical contributions. From a theoretical point of view, we contributed to the literature concerning AI bias derived from the data used to train the models by inspecting the occurrence of bias inside translations made by AI tools. Specifically, we described and implemented a method to assess the presence of gender bias inside MT and LLM tools and compare their performances with human behaviour and tendencies. We focused on translating sentences from a language that allows the omission of the subject in sentences to a language that requires the subject to be made explicit. To this end, we adapted the framework given by Srivastava et al., [23] to better represent the process of developing and translating the dataset occurrences and assessing the presence of bias. We discovered the presence of a converging bias in which the outcome is influenced by previous translations. Overall, our research methodology could be used to obtain preliminary insights into the phenomena under investigation.

From a practical point of view, this article can assist developers and designers to better identify the presence of gender bias inside translation tools. While the presence of gender bias within these tools has been discussed both in scientific literature and press articles [18], there is still considerable work to be done to address this issue and enhance the fairness of these AI tools. There is also the need to establish standardised and unbiased ways to measure and identify such biases inside MT tools to rigorously evaluate their outcome and give suggestions on which issues should be addressed. Moreover, as highlighted by Tomalin et al. [19], it is not always necessary to address

the bias in the underlying data before the training of the algorithm, as this may lead to a worse outcome in certain cases. Our results indicate that these kinds of bias are embedded inside the society and are subsequently transferred to AI tools. For this reason, it is crucial to address societal behaviour by raising awareness among individuals regarding these issues. In this scenario, governments, firms, and policymakers should play a central role in nudging people to give more attention to these aspects and, hence, in changing their behaviour. Additionally, it would be interesting to further explore how to evaluate a biased behaviour, particularly in the context of determining the right translation for each sentence.

The study also presents some limitations. First, the dataset used in this research comprise only ten jobs and thirty sentences. Future studies could increase the dataset in order to explore gender bias across a broader range of occupations and also assess the presence of gender bias in more complex sentence structures. Moreover, concerning the usage of AI online tools, it would be interesting to investigate if the geographical location from which the translation request is performed (e.g., using a VPN connection) could affect the final outcome. Culture traits characterizing a specific geographical zone could influence the results as well as the iterative nature of translation requests. Additionally, the current study focused solely on Italian to English translations. Future studies could use more languages that allow subject omission in the construction, such as Spanish or Persian. Finally, concerning the survey, the sample used could be improved by increasing the number and demographic variety of the respondents in order to have more meaningful insights from the collected data.

References

1. Smacchia, M., Za, S.: Artificial Intelligence in Organisation and Managerial Studies: A Computational Literature Review. ICIS 2022 Proceedings. 6. 0–17 (2022).
2. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl. 82, 3713–3744 (2023). https://doi.org/10.1007/s11042-022-13428-4.
3. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent Trends in Deep Learning Based Natural Language Processing. IEEE Comput Intell Mag. 13, 55–75 (2017).
4. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: PaLM: Scaling Language Modeling with Pathways. (2022).
5. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D.: Is ChatGPT a General-Purpose Natural Language Processing Task Solver? arXiv preprint arXiv:2302.06476. (2023).

6. Stahlberg, F.: Neural Machine Translation: A Review. (2020).
7. Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., Liu, Y.: Neural machine translation: A review of methods, resources, and tools, (2020). https://doi.org/10.1016/j.aiopen.2020.11.001.
8. Vieira, L.N., O'Hagan, M., O'Sullivan, C.: Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases, (2021). https://doi.org/10.1080/1369118X.2020.1776370.
9. Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., Way, A.: Is Neural Machine Translation the New State of the Art? The Prague Bulletin of Mathematical Linguistics. 108, 109–120 (2017). https://doi.org/10.1515/pralin-2017-0013.
10. Doherty, S.: The Impact of Translation Technologies on the Process and Product of Translation. (2016).
11. Smacchia, M., Za, S.: Exploring Artificial Intelligence Bias, Fairness and Ethics in Organisation and Managerial Studies. In: ECIS 2023 Research Papers. p. 362 (2023).
12. van Giffen, B., Herhausen, D., Fahse, T.: Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. J Bus Res. 144, 93–106 (2022). https://doi.org/10.1016/j.jbusres.2022.01.076.
13. Zuiderwijk, A., Chen, Y.C., Salem, F.: Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Gov Inf Q. 38, (2021). https://doi.org/10.1016/j.giq.2021.101577.
14. Hoffmann, A.L.: Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Inf Commun Soc. 22, 900–915 (2019). https://doi.org/10.1080/1369118X.2019.1573912.
15. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv. 54, (2021). https://doi.org/10.1145/3457607.
16. Cathy O'Neil: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York, NY, Broadway Books. (2016).
17. Sun, T., Gaut, A., Tang, S., Huang, Y., Elsherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., Wang, W.Y.: Mitigating Gender Bias in Natural Language Processing: Literature Review. In: ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. pp. 1630–1640 (2020).
18. Stanovsky, G., Smith, N.A., Zettlemoyer, L.: Evaluating Gender Bias in Machine Translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1679–1684 (2019).
19. Tomalin, M., Byrne, B., Concannon, S., Saunders, D., Ullmann, S.: The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. Ethics Inf Technol. 23, 419–433 (2021). https://doi.org/10.1007/s10676-021-09583-1.
20. Bernagozzi, M., Srivastava, B., Rossi, F., Usmani, S.: Gender Bias in Online Language Translators: Visualization, Human Perception, and Bias/Accuracy Tradeoffs. IEEE Internet Comput. 25, 53–63 (2021). https://doi.org/10.1109/MIC.2021.3097604.
21. Kelley, S., Ovchinnikov, A., Hardoon, D.R., Heinrich, A.: Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending. Manufacturing and Service Operations Management. 24, 3039–3059 (2022). https://doi.org/10.1287/msom.2022.1108.

22. Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., Turchi, M.: Gender Bias in Machine Translation. Trans Assoc Comput Linguist. 9, 845–874 (2021). https://doi.org/10.1162/tacl.
23. Srivastava, B., Rossi, F.: Towards Composable Bias Rating of AI Services. In: AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 284–289. Association for Computing Machinery, Inc (2018). https://doi.org/10.1145/3278721.3278744.
24. Schiebinger, L.: Scientific research must take gender into account. Nature. 507, 9–9 (2014). https://doi.org/10.1038/507009a.
25. Liben, L.S., Bigler, R.S., Krogh, H.R.: Language at Work: Children's Gendered Interpretations of Occupational Titles. Child Dev. 73, 810–828 (2002). https://doi.org/10.1111/1467-8624.00440.
26. Bigler, R.S., Leaper, C.: Gendered Language: Psychological Principles, Evolving Practices, and Inclusive Policies. Policy Insights Behav Brain Sci. 2, 187–194 (2015). https://doi.org/10.1177/2372732215600452.
27. Makri, C., Neely, A.: Grounded Theory: A Guide for Exploratory Studies in Management Research. Int J Qual Methods. 20, (2021). https://doi.org/10.1177/16094069211013654.
28. King, M.R.: A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education, (2023). https://doi.org/10.1007/s12195-022-00754-8.
29. Karimi, F., Génois, M., Wagner, C., Singer, P., Strohmaier, M.: Homophily influences ranking of minorities in social networks. Sci Rep. 8, (2018). https://doi.org/10.1038/s41598-018-29405-7.
30. Henry, P.J.: Institutional Bias. In: The Sage Handbook of Prejudice, Stereotyping and Discrimination. pp. 426–440. SAGE Publications, London (2010).
31. Brickell, C.: The sociological construction of gender and sexuality. Sociol Rev. 54, 87–113 (2006).
32. Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., Staab, S.: Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdiscip Rev Data Min Knowl Discov. 10, 1–14 (2020). https://doi.org/10.1002/widm.1356.