GRASE

GENDER AND RACE STEREOTYPES ERADICATION · IN LABOR MARKET ACCESS ·

FUNDED BY THE EUROPEAN UNION'S
RIGHTS, EQUALITY AND CITIZENSHIP
PROGRAMME (2014-2020)

AI-based gender and race/origin bias detection Toolkit

GRASE Toolkit

FONDAZIONE ISMU INIZIATIVE E STUDI SULLA MULTIETNICITÀ
THE ADECCO GROUP
Amiga
aproximar COOPERATIVA DE SOLIDARIEDADE SOCIAL CRL
FGB Fondazione Giacomo Brodolini
FUNDACIÓN ADECCO

GRASE

GENDER AND RACE STEREOTYPES ERADICATION · IN LABOR MARKET ACCESS ·

FUNDED BY THE EUROPEAN UNION'S
RIGHTS, EQUALITY AND CITIZENSHIP
PROGRAMME (2014-2020)

Authorship & Disclaimer

The Toolkit on *AI-based gender and race/origin bias detection toolkit* was realized within the GRASE project (2021-2022), funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020). The project was coordinated by ISMU Foundation (Italy) in partnership with Fondazione Giacomo Brodolini (Italy), the Adecco Group (Italy), Fundación Adecco (Spain), Asociación AMIGA por los Derechos Humanos de las Mujeres (Spain), APROXIMAR Cooperativa de Solidarieda de Social (Portugal).

The Toolkit is the result of a shared effort of the project's Steering Group and Board of Experts and has been edited by Mylia in collaboration with the University of Pisa and Erre Quadro srl. The content of the present Toolkit represents the views of the authors and is their sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains. The GRASE Project's Steering Group and board of Experts are the only responsible for the final content, words and phrases used.

Fondazione Iniziative e Studi
sulla Multietnicità - ISMU
Via Copernico, 1 - 20125 Milano
Tel. 02.6787791
www.ismu.org

Graphic Project
Alexandra Gredler

Suggested citation
GRASE project partners (2022),
AI-based gender and race/origin bias detection toolkit
ISMU, Milano.

Table of Contents

PART 1

PART 2

PART 3

PART 4

The GRASE Project

GRASE stands for "Gender and Race Stereotypes Eradication in Labour Market Access" and is a 2-year project (2021-2022), funded under the European Union's Rights, Equality & Citizenship program which addresses the double discrimination faced by migrant women in their access to employment, through the adoption of a truly intersectional approach. GRASE focuses on facilitating the access of women with a migratory background to the labour market by reducing the barriers they may find in career counselling services systems.

To reach this objective, GRASE combined **expert knowledge**, through the involvement of antidiscrimination specialists and researchers working on gender, race and migration, as well as **practical knowledge**, by activating three Communities of Practice with career counselling professionals in Italy, Spain and Portugal – three countries where women and migrants varyingly face barriers in the access to the labour market.

The final goal of the project is to contribute to **reduce gender and race gaps** in migrant women's participation in the labour market, with a view to provide full implementation of the principle of non-discrimination, heralded in the European legislative framework as one of its core elements, and enshrined also in the Constitutions and the laws of Italy, Spain and Portugal – the three countries where the project was implemented.

The project produced three Toolkits to fight the reproduction of bias and stereotypes against women with a migratory background: *"Effective strategies to fight race and gender stereotypes in career counselling services"* **(Toolkit 1)**, *"Raising awareness against gender and race stereotypes in recruitment: training for career counselling professionals"* **(Toolkit 2)**; *"AI-based gender and race/origin bias detection toolkit"* **(Toolkit 3)**.

For a detailed description of GRASE's approach and products, please refer to the project's Website.

Go!

451.91
8,904
2,053
YU-Qi-pO-AW
451.91
743.06
11.08
841.52
536.85

Objective of the Toolkit

The present technical specification describes **a tool able to identify gender and race (or origin) biases hidden behind the recruitment process**. Since recruitment is ever more conducted through the assistance of AI algorithms, the evaluation of likely biases embedded in machine learning processes is becoming extremely relevant.

As the literature shows, the recruitment process can be affected by distortions both in job advertising and in the software used to simplify this procedure, which may contain distortions that are then reflected in the assessments they make.

This tool is flexible and easily adaptable to the investigation of racial – or country of origin - bias in the area of interest. The toolkit can be used successfully by employment companies (by their IT professionals), verifying the absence of bias in job vacancies drawn up before their publication and testing the software used in selected applications. For this reason, as an example of application, a set-up of job applications for hostesses and stewards was analysed, with particular focus on the characteristics that refer to the gender and provenance of the candidates; subsequently, the application of a tool able to explore and evaluate biases in a sample of job advertisements is presented; finally, in the last section, some final evaluations are discussed.

Glossary

This Glossary provides an overview of the terminology and of the approach used by GRASE. It aims at helping the reader navigate the contents of the present Toolkit. Terms are listed under four main topics: “stereotypes”, “discrimination”, “bias”, “diversity, race, ethnicity and gender”.

STEREOTYPES

Compounded stereotypes	Generalised view or preconception about groups that results from the ascription of attributes, characteristics or roles based on one or more grounds.
Gender stereotypes	Preconceived ideas whereby females and males are arbitrarily assigned characteristics and roles determined and limited by their gender. Gender stereotyping can limit the development of the natural talents and abilities of girls and boys, women and men, as well as their educational and professional experiences and life opportunities in general. Stereotypes about women both result from, and are the cause of, deeply engrained attitudes, values, norms and prejudices against women. Stereotypes can be both hostile and explicitly negative (e.g. women are irrational) or seemingly benign (e.g. women are nurturing) – both kinds, though, can produce harmful effects, which justify and maintain the historical relations of power of men over women as well as sexist attitudes that hold back the advancement of women.
Judicial stereotyping	Practice of judges ascribing to an individual specific attributes, characteristics or roles on the sole basis of her or his membership of a particular social group. It also refers to the practice of judges perpetuating harmful stereotypes through their failure to challenge stereotypes.
Racial / ethnic stereotypes	Stereotype is a generalized perception ascribing particular traits, characteristics, values, aspect, appearance or behaviour to a group or a member of a group without regard to accuracy or applicability **(Corsini, 2016)**. Racial / ethnic stereotypes are reflexive and exaggerated mental pictures that we hold about all members of a particular racial / ethnic group. These stereotypes are so rigid, we tend to ignore or discard any information that is not consistent with the stereotype that we have developed about the racial / ethnic group **(University of Notre Dame, 2020)**.

DISCRIMINATION

Discrimination against women	Any distinction, exclusion or restriction made on the basis of sex and gender that has the effect or purpose of impairing or nullifying the recognition, enjoyment or exercise by women, irrespective of their marital status, and on a basis of equality between women and men, of human rights and fundamental freedoms in the political, economic, social, cultural, civil or any other field. Discrimination can stem from law (de jure) or from practice (de facto). The CEDAW Convention recognises and addresses both forms of discrimination, whether contained in laws, policies, procedures or practice.
Direct discrimination	Discrimination occurring where one person is treated less favourably on grounds such as sex and gender, age, nationality, race, ethnicity, religion or belief, health, disability, sexual orientation or gender identity, than another person is, has been or would be treated in a comparable situation.
Indirect discrimination	Discrimination occurring where an apparently neutral provision, criterion or practice would put persons of one sex at a particular disadvantage compared with persons of the other sex, unless that provision, criterion or practice is objectively justified by a legitimate aim, and the means for achieving that aim are appropriate and necessary.
Intersectional discrimination	Discrimination that takes place on the basis of several personal grounds or characteristics / identities (sex, racial or ethnic origin, religion or belief, disability, age, sexual orientation, gender identity, etc.) which operate and interact with each other at the same time in such a way as to be inseparable.
Sex- and gender-based discrimination	Discrimination occurring due to interaction between sex (as the biological characteristics of women and men) and their socially constructed identities, attributes and roles and society's social and cultural meaning for biological differences between women and men. Such interactions result in hierarchical and unequal relations and roles between and among women and men, and a disadvantaged social positioning of women. The social positioning of women and men is affected by political, economic, cultural, social, religious, ideological and environmental factors, and can be changed over time.
Racial / ethnic discrimination	Any distinction, exclusion, restriction or preference based on race, colour, descent, or national or ethnic origin which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life. (Art. 1.1 of the United Nations International Convention on the Elimination of All Forms of Racial Discrimination).

BIAS

Implicit bias	Behaviours by which people act on the basis of prejudice and stereotypes without intending to do so and without consciously recognizing their bias. These behaviours display a bias – i.e. rather than being neutral, they show a preference for (or an aversion to) a person or a group of people. However, this bias is present but not consciously held or recognized, meaning we are unaware of them or mistaken about their nature. For instance, a host of studies have demonstrated that white people tend to associate criminality with black people. The present definition is based on the Merriam Webster dictionary and on the definitions provided by **The Perception Institute.**
Invisible barriers	Attitudes and the underlying traditional assumptions, norms and values that prevent women's and migrants' empowerment / full participation in society.
Gender bias	Prejudiced actions or thoughts based on the gender-based perception that women are not equal to men. Bias represents the enactment" of stereotypes and prejudices: through preconceived ideas, females and males are arbitrarily assigned characteristics and roles determined and limited by their gender. For example, this may lead a career counselling professional to avoid proposing certain kinds of jobs to women, assuming that, because they are women, they are not "suitable" for those kinds of jobs.
Racial / ethnic bias	Prejudiced actions or thoughts based on reflexive and exaggerated mental pictures that we hold about all members of a particular racial / ethnic group. Bias represents the enactment" of stereotypes and prejudices: through preconceived ideas, members of specific racial or ethnic groups are arbitrarily assigned characteristics and roles determined and limited by their belonging to that group. For example, this may lead a career counselling professional to avoid proposing certain kinds of jobs to people of colour, assuming that, because they are people of colour, they are not "suitable" for those kinds of jobs.

DIVERSITY, RACE, ETHNICITY AND GENDER

Diversity	Differences in the values, attitudes, cultural perspective, beliefs, ethnic background, sexual orientation, gender identity, skills, knowledge and life experiences of each individual in any group of people.
Gender awareness raising	Process that aims at showing how existing values and norms influence our picture of reality, perpetuate stereotypes and support mechanisms (re)producing inequality. It challenges values and gender norms by explaining how they influence and limit opinions taken into consideration and decision-making. In addition, awareness raising aims at stimulating a general sensitivity to gender issues.
Gender roles	Social and behavioural norms which, within a specific culture, are widely considered to be socially appropriate for individuals of a specific sex. Collectively, gender roles often determine the traditional responsibilities and tasks assigned to women, men, girls and boys (see gender division of labour). Gender-specific roles are often conditioned by household structure, access to resources, specific impacts of the global economy, occurrence of conflict or disaster, and other locally relevant factors such as ecological conditions. Like gender itself, gender roles can evolve over time, in particular through the empowerment of women and transformation of masculinities.
Gender segregation	Differences in patterns of representation of women and men in the labour market, public and political life, unpaid domestic work and caring, and in young women's and men's choice of education.
Racial segregation	The practice of restricting people to certain circumscribed areas of residence or to separate institutions (e.g., schools, churches) and facilities (parks, playgrounds, restaurants, restrooms) on the basis of **race** or **alleged** race. Racial **segregation** provides a means of maintaining the economic advantages and superior **social status** of the politically dominant group, and in recent times it has been employed primarily by white populations to maintain their ascendancy over other groups by means of legal and social colour bars (Britannica 2022).

Two approches for job offers analysis

This section presents the main features of the toolkit through its application on two job advertisements, in order to determine whether they implicitly contain gender distortions that would penalize certain categories of people at the selection phase. The toolkit is based on a lexicon of distorted terms, which in the case of use are related to gender, making it easily adaptable to the recognition of other types of bias.

The toolkit is declined according to two logics: in the first, defined top-down, the classification of terms is entrusted to the human being who builds the lexicon by choosing the terms; the second, called **bottom-up**, uses artificial intelligence models to detect new biased words. As mentioned above, the structure of the tool makes it flexible and suitable for application to identify other kinds of bias.

Top-down approach:
using expert knowledge to develop the toolkit

The **top-down approach** allows to identify the words present in a text that can be discriminating with respect to one gender than to another. The analysis focuses on several Italian job vacancy texts, explored separately, in which the terms that make the ad biased are searched.

One of the main outputs of the gender bias literature map is represented by Gaucher et al. (2011), where the authors analyze the use of words in job advertisements in order to understand whether the presence of male words deters women from applying. The result of this work is a list of lemmas, assigned to the gender towards which they are distorted. This list, integrated with a further shorter lexicon extracted from Hoyle et al. (2019), is used as the basis for bias detection in the proposed method.

The toolkit takes as input the job vacancy, analyzing and splitting the text into words, which are saved into rows of a database. At the same time, the biased terms resulting from the paper mentioned above are analysed and, since the job vacancies are written in Italian, these words are translated into the same language of the job offers. Finally, the model identifies within the job vacancy the words present in the reference list, and the labels with the genre towards which they are distorted. The output of the toolkit is the list of biased terms present and the percentage that quantifies the total amount of text distortion.

Top-down approach testing

The approach presented is tested on two examples of job vacancy for hostesses or stewards on the locations of Venice and Milan, very close to each other. The result shows that in the text words such as “trust” and “empathy” are biased towards the female gender, while “activity”, “values”, “care”, “good” and “hierarchy” for male ones. Both texts have about **0.02%** biased-terms, pointing out that they are **not discriminating** against gender.

Bottom-up approach:
improving the toolkit with AI

The **bottom-up approach** aims to broaden the list of distorted terms on which the instrument previously proposed is based, enlarged with a list of terms defined as neutral, and thus improve the results obtained. To do this, the toolkit performs word embedding, a technique used in natural language processing to represent a word as a vector of real numbers, allowing to extract semantic and syntactic internal information (Li et al., 2018).
In this case, word embedding was used to calculate the distance between vectors (Mikolov et al., 2013).

To identify gender biased words, two terms of reference for gender are identified, "male" and "female", whose vectors are used to calculate cosine distance with other words present in the paper mentioned above. The calculated distance is used to identify a possible threshold value to label words with one genre rather than another. As we see in **Figure 1** using the distances from the masculine and feminine terms as coordinates for the representation of words it is possible to notice that their distribution is almost random, thus making it difficult to identify a threshold distance to label words according to the gender to which they tend.

Alternatively to the use of a single word as a reference for a gender, the vector derived from the average of all word vectors in the reference list and labeled with "M" or "F" has been used. Subsequently, the two average vectors were used to calculate the cosine distance with the vectors of the other words in the list. Again, these distances were used as coordinates to display the distribution of words between the two reference vectors. As shown in **Figure 2**, also in this case such distribution seems random thus preventing the determination of a threshold value to label the words with respect to the gender towards which they are distorted.

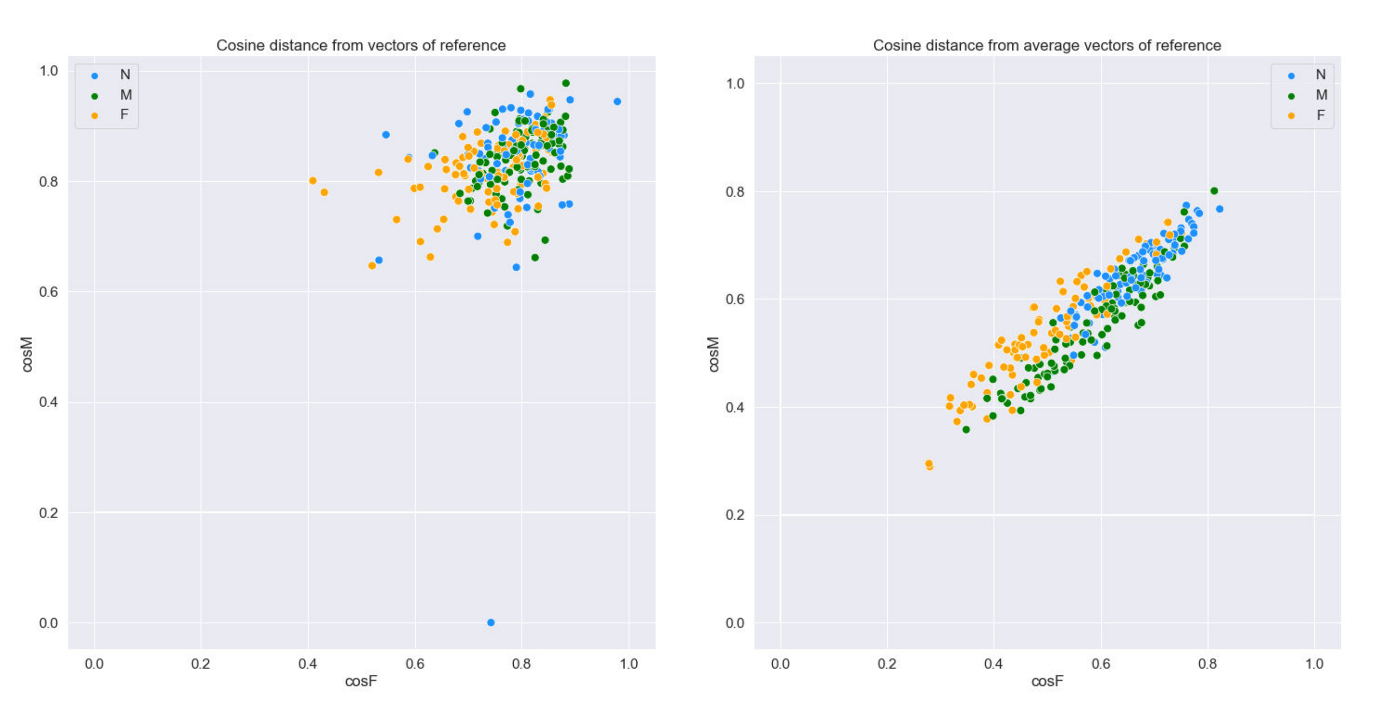

Cosine distance from vectors of reference
N
M
F
cosM
cosF
0.0
0.2
0.4
0.6
0.8
1.0
Cosine distance from average vectors of reference
N
M
F
cosM
cosF
0.0
0.2
0.4
0.6
0.8
1.0

Figure 1.
Representation of terms
with male and female vectors

Figure 2.
Representation of terms
with average vectors

Bottom-up approach testing

The bottom-up approach presented is tested on two examples of job vacancy for hostesses or stewards on the locations of Venice and Milan, very close to each other. The result shows that in the text words such as “good” are biased towards the female gender, while “activity”, “care”, “degree”, “work”, “operative” and “attention” for male ones. Both texts have about **0.04%** biased-terms, pointing out that they are **not discriminating** against gender.

Final considerations

The above presented applications show the toolkit use for detecting gender or race/origin biases hidden behind the recruitment process. More specifically, the toolkit may be used to identify gender bias in job vacancies or to test the software used in the selection of candidates, but it can also be easily adapted to find out racial bias.

From an algorithmic point of view, the tool looks simple, in fact the value is in the creation of the dictionary, which can be top-down (by humans) or bottom-up (with AI models), through which the distorted terms are identified. When software is used in the recruitment process it is well known that risks exist, as studies have shown (Pedreschi et al., 2008; Hajian et al., 2013; Mehrabi et al. 2019), the algorithms themselves can be discriminatory with respect to gender, ethnicity and civil status.

These prejudices exist even when there is no real discriminatory intention in the development of the same: the data sources used can influence, software that consequently amplify some historical discrimination in the data, or a trained algorithm may discriminate based on sensitive attributes due to correlations between the data themselves (Caliskan et al., 2017). Biases can sometimes occur when an algorithm is trained on unbalanced data sets, which therefore do not represent a population well enough, and are then incorporated into the design of the algorithm (Zhang et al., 2018).

The toolkit presented can be used successfully by employment companies, allowing both to verify the absence of bias in job vacancies drawn up prior to their publication, and to test the software used in the selection of applications. In addition, since the literature also confirms that it is easy to discriminate in the selection of personnel, in order to ensure a bias-free recruitment process of any kind it is necessary to verify first the absence of distorted terms in the job advertisement to be published and, secondly, the absence of distortions in the software used for the evaluation of candidate.

Job-vacancies analysis: an example of distorsions detecting

This section presents the analyses carried out on a set of job applications in order to assess the presence or not of distortions in the software used for the selection of candidates.

In more detail, we analyse a dataset of candidates for hostesses and stewards, focusing on the main feature that represents the evaluation attributed to the candidate in order to detect the possible presence of gender and race biases. First of all, the dataset and the key variables are introduced, then a brief description of the data cleaning phase is presented. Finally, the relationship between the different features is presented, with particular attention to those expressing the gender and the country of origin of the candidates.

Dataset description

The dataset under analysis consists of **10’801 rows and 21 columns**, which summarise the responses received in the applications for the job offer as a hostess or steward on board.
Within the dataset there are 15 features with null values to manage. In addition, the dataset is composed of binary variables, corresponding to questions with closed answers (yes/no), and by features related to questions with open answers, which therefore need to be properly managed. The first type includes questions about the availability of shift work, on weekends or in Milan, while the second type includes questions on the city and State of origin, the last job covered and the discovery of the job offer.

Finally, there are features related to the system used to collect the data, such as:

- **“*status*” that is relative to the status of the application and can be inbox if completed and online, draft if not completed or not compliant if completed but not online;**
- **“*pipeline step*” that indicates the position in the pipeline and can be Sc (screening) or cv (sent cv to customer);**
- **“*trashed*” which is set to 1 if the application is removed, 0 otherwise.**

In **Table 1**, summarize the attributes present in the dataset, also indicating the amount of null values.

Data cleaning

The preprocessing phase is conducted with the aim of standardizing non-binary variables. Special characters such as brackets, bars and numbers are removed in the “*city*” feature. With a more in-depth analysis, typing errors are identified and corrected. Another characteristic to be standardized is represented by the last job position covered. Here the data are much more heterogeneous and for this reason several steps have been carried out. First, the special characters are removed, and several tasks are replaced with more general synonyms. In the execution of these passages, the gender is kept, to store those specifications. Also in the attribute “*English level*” changes are carried out, to gain a single level for every answer. Finally, all the categorical variables, except the one related to the last covered work, have been transformed into numerical ones, by mapping each value with a number starting from 0. In this way data can be treated as numerical and can be analyzed more deeply.

In **Table 1**, summarize the attributes present in the dataset, also indicating the amount of null values.

Feature	**Null-values**	**Length of the module**
Unique key	0	0%
Name	0	0%
Zip code	531	4,92%
City	495	4,58%
Province	2176	20,15%
State	489	4,53%
Last job	1433	13,27%
Start date	1675	15,51%
End date	2550	23,61%
Starting availability	489	4,53%
English level	500	4,63%
Availability of shift work	500	4,63%
Availability to work on holidays	500	4,63%
Availability to work anywhere	500	4,63%
How did you know about the selection	500	4,63%
Data application	0	0%
Score	0	0%
Status	0	0%
Pipeline step	5315	49,21%
Trashed	5815	53,84%
Gender	0	0%

Table 1. Dataset description

Data analysis and visualization

The analysis performed on the different features of the dataset is presented, studying their behavior jointly with the values of *"score"* that represents the main variable of interest, as it expresses the final evaluation of the applicant. First, the variables of interest regarding possible gender or origin biases will be explored. As regards the origin of the candidates, the variable *"city"* is analyzed, whose distribution turns out to be very heterogeneous. In particular, as shown in **Figure 3**, the most frequent city of origin among the candidates is Rome, followed by Naples, Milan, Reggio Calabria, Turin and Venice, also emphasizing a uniform distribution between cities in the north, in the center and in the south of Italy.

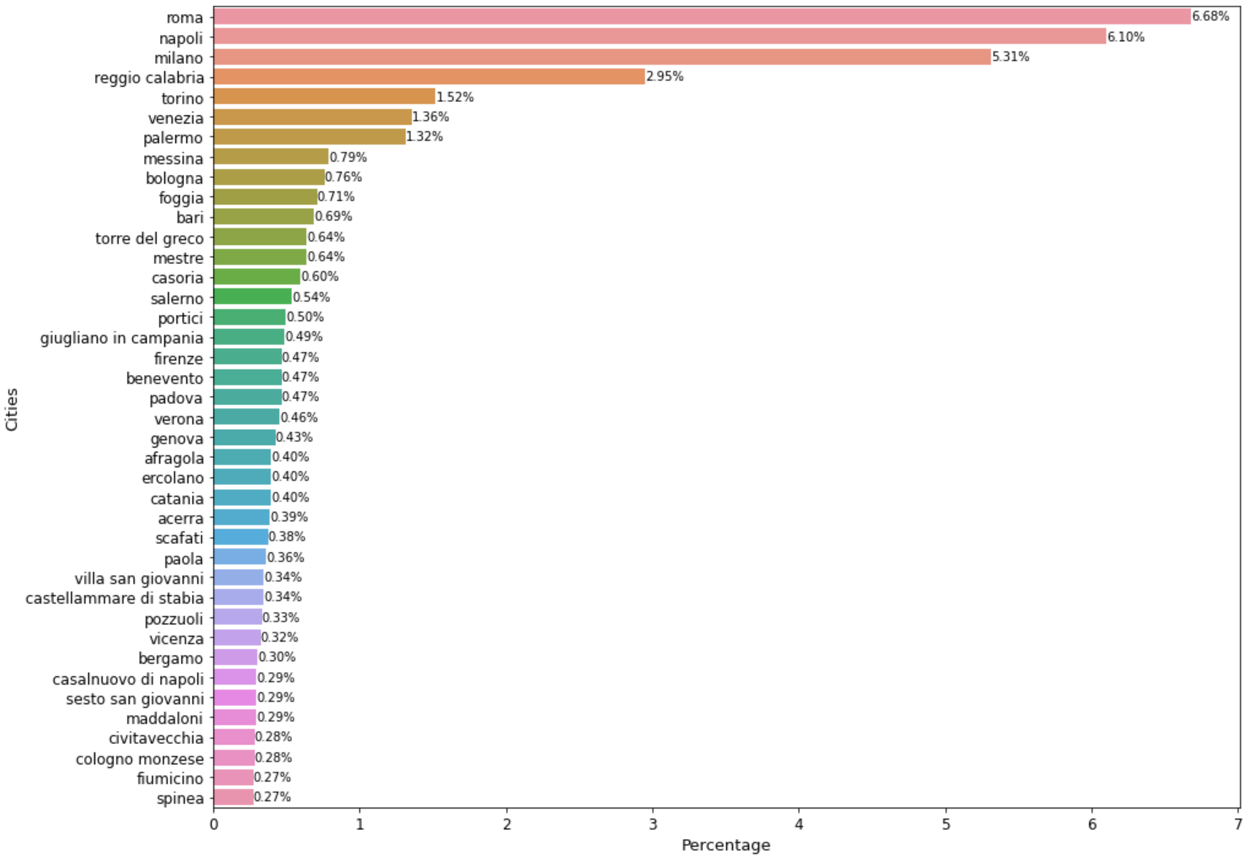


Figure 3. Distribution of cities' frequency

Similarly, the variable “*state*” is analyzed. In this case, the most frequent value is Italy, which represents the country of origin of more than 98% of the candidates, while the remaining 2% includes 21 other countries. In order to observe the distribution of these values, the State “IT” (Italy) is excluded from the analysis thus allowing the authors to focus on the other 21 states that appear with less frequency. The distribution of these values is shown in **Figure 4**, where the state “GB” (Great Britain) stands out as the most popular, while the others reach a frequency below 0.2%.

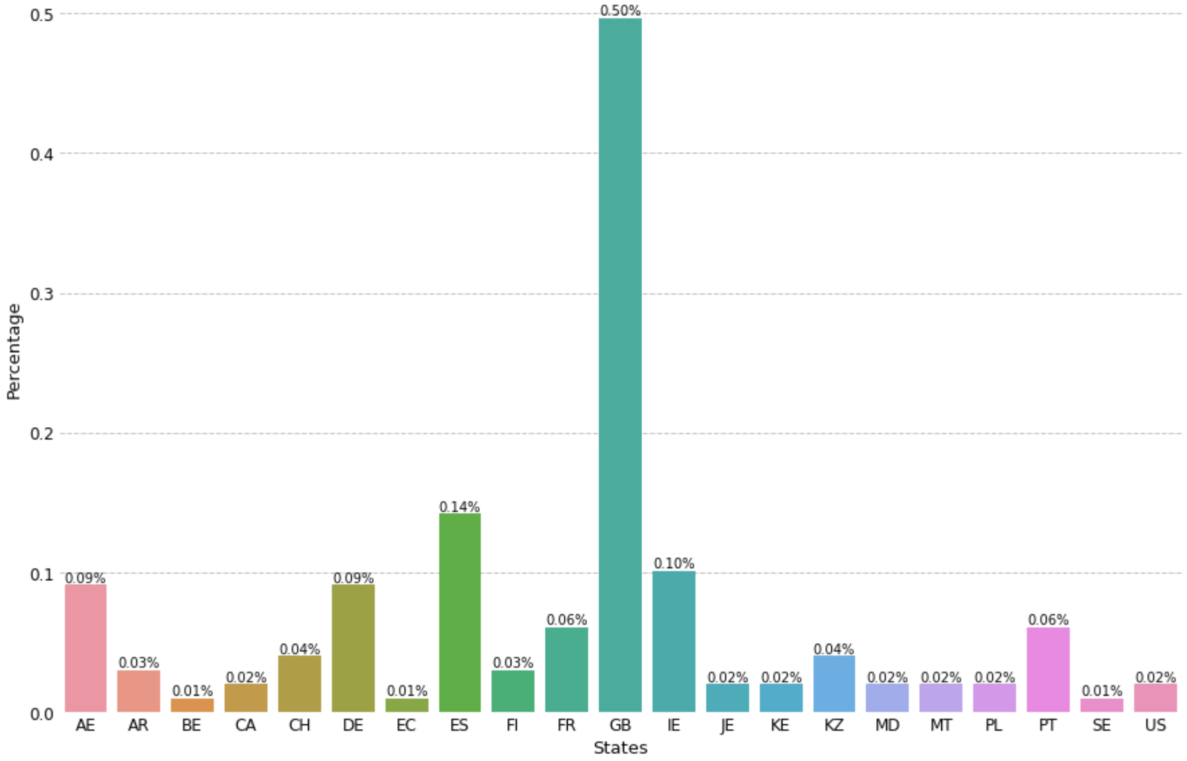
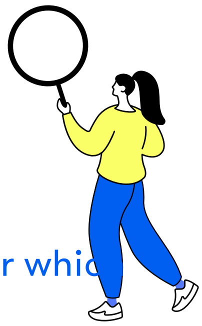


Figure 4. Distribution of States’ frequency



As for gender, the sample in analysis seems to be fairly balanced between the two sexes, but there is a majority of candidates (37%) for whic gender is unknown (**Figure 5**).

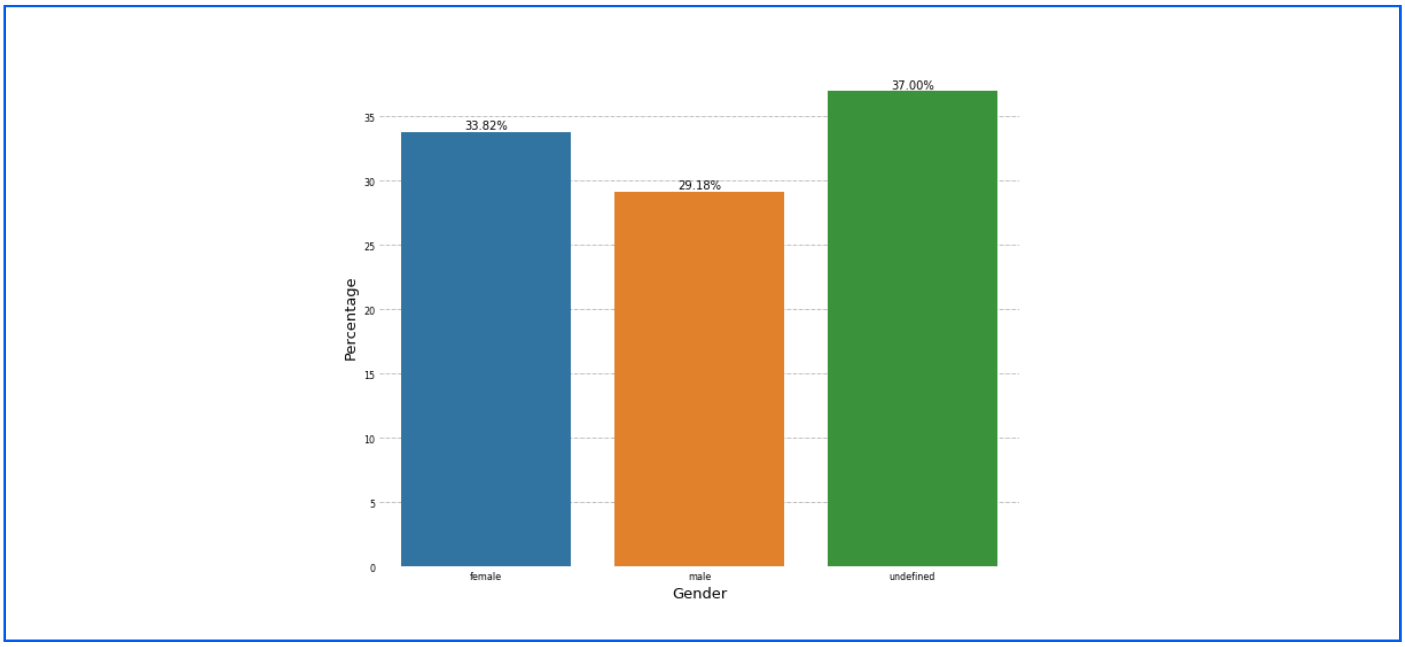
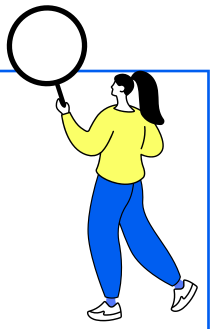


Figure 5. Gender distribution

The absence of this type of information we do not allow to conduct detailed analyses. To get around this, the gender is extracted by analysing the variable "name". As a result of this operation, the data obtained appear unbalanced compared to the two sexes, in fact 56.15% of the candidates are female, 39.18% male, while for 4.67% it is not possible to define the gender.

Subsequently, the feature relating to the candidate's last job is explored. Standardizing data makes it possible to identify the most frequent tasks among the candidates, pointing out that most of them have not carried out work recently. In addition, the last jobs held by most candidates are the flight attendant, the waiter (generic sex) and receptionist. Moreover, the variables indicating the availability of candidates to work on shifts, during holidays and elsewhere are analyzed, showing the candidates' positive propensity to do so. In a similar way, most candidates also show willingness to start working immediately or within 15 days. Also the distribution of the variable *"English level"* is analyzed, showing that most candidates have a B1 or B2 level, as represented in **Figure 6**.



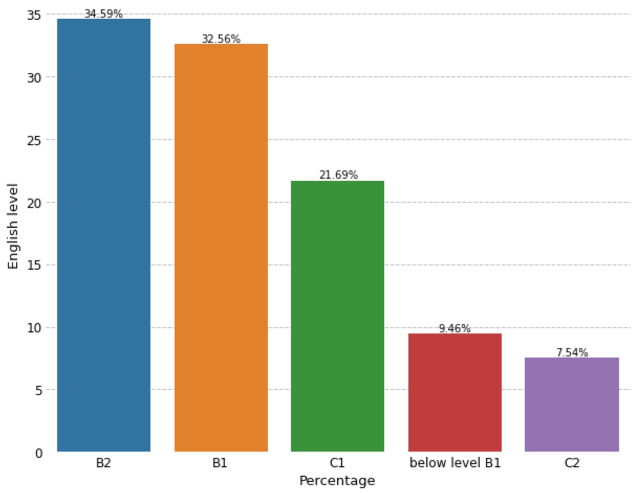
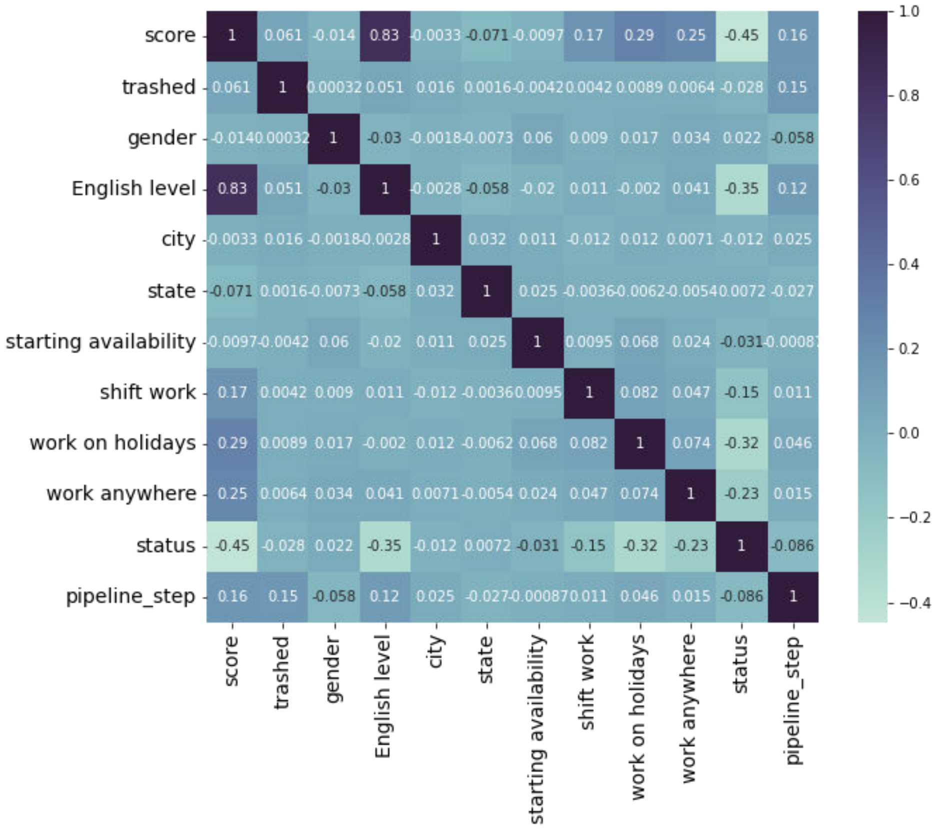


Figure 6. Distribution of English level

In addition, the relationships between all the features are explored with particular focus on how these affect the final score. The correlation matrix, represented in **Figure 7**, shows a high positive correlation with the level of English: the higher the level , the higher the score.



	score	trashed	gender	English level	city	state	starting availability	shift work	work on holidays	work anywhere	status	pipeline_step
score	1	0.061	-0.014	0.83	-0.0033	-0.071	-0.0097	0.17	0.29	0.25	-0.45	0.16
trashed	0.061	1	0.00032	0.051	0.016	0.0016	-0.0042	0.0042	0.0089	0.0064	-0.028	0.15
gender	-0.014	0.00032	1	-0.03	-0.0018	-0.0073	0.06	0.009	0.017	0.034	0.022	-0.058
English level	0.83	0.051	-0.03	1	-0.0028	-0.058	-0.02	0.011	-0.002	0.041	-0.35	0.12
city	-0.0033	0.016	-0.0018	-0.0028	1	0.032	0.011	-0.012	0.012	0.0071	-0.012	0.025
state	-0.071	0.0016	-0.0073	-0.058	0.032	1	0.025	-0.0036	-0.0062	-0.0054	0.0072	-0.027
starting availability	-0.0097	-0.0042	0.06	-0.02	0.011	0.025	1	0.0095	0.068	0.024	-0.031	-0.00087
shift work	0.17	0.0042	0.009	0.011	-0.012	-0.0036	0.0095	1	0.082	0.047	-0.15	0.011
work on holidays	0.29	0.0089	0.017	-0.002	0.012	-0.0062	0.068	0.082	1	0.074	-0.32	0.046
work anywhere	0.25	0.0064	0.034	0.041	0.0071	-0.0054	0.024	0.047	0.074	1	-0.23	0.015
status	-0.45	-0.028	0.022	-0.35	-0.012	0.0072	-0.031	-0.15	-0.32	-0.23	1	-0.086
pipeline_step	0.16	0.15	-0.058	0.12	0.025	-0.027	-0.00087	0.011	0.046	0.015	-0.086	1

Figure 7. Correlation matrix

Although lower, further positive correlations could be detected analyzing the availability of working in shifts, on holidays, and everywhere. Analyzing more in detail the variables related to the gender and provenance of the candidates, the absence of a correlation emerges, proving that the system is **non-biased** with respect to those characteristics. With regard to gender, the distribution of scores is homogenous between the two values, thus confirming the absence of distortions.

Similarly, in **Figure 8** the variation in score relative to the state is explored. In this case it should be noticed that for the "IT" value the variability of the scores is wider, reaching both very high and very low scores, due to the imbalance of the dataset compared to this value. Looking at the other states, the score variance was not as wide, but in general the scores are higher. The absence of distortions within the system is confirmed, thus the system assigns the scores to the candidates without discriminating the applicants. This toolkit helps to respond well to the need.

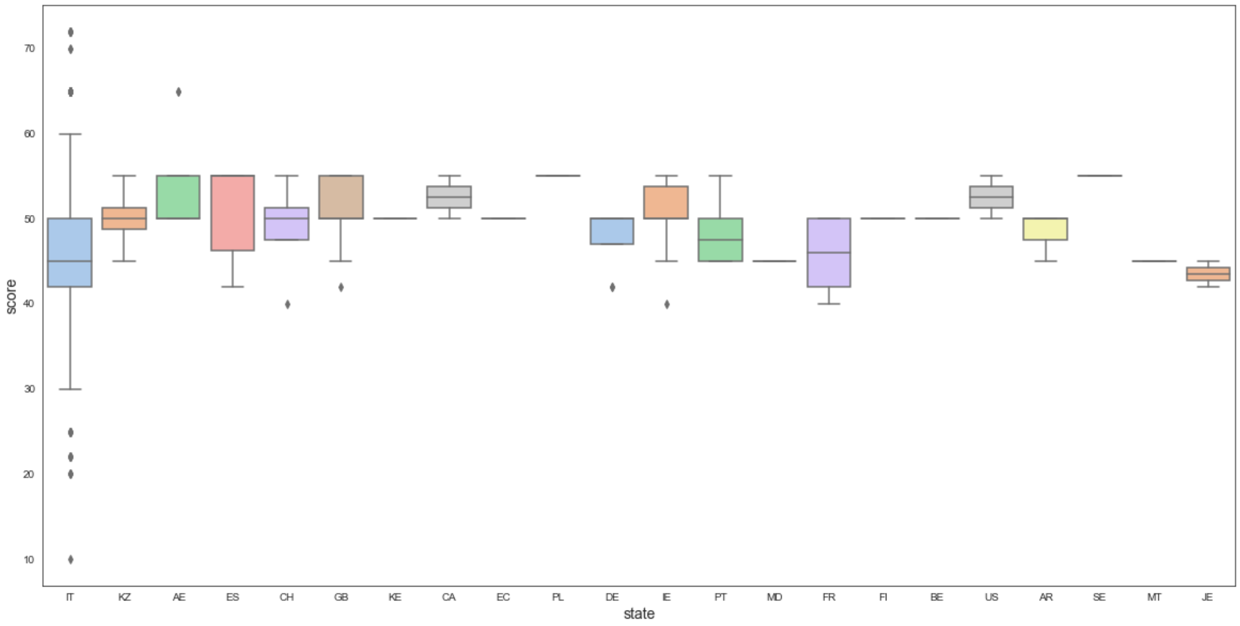
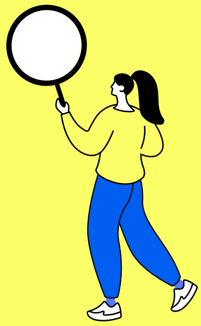


Figure 8. Variability of score by State

Towards a Gender Personnel Selection: Guiding tool to integrate AI and people in the selection process



Ensuring gender equality in the recruiting process, as well as in other business fields and processes, is ever more recognized as crucial. Despite the relevance of the topic, a lot of improvements should still be done and the biases retrieved from original approaches are quite findable also in digital recruitment processes.

To overcome these issues, a proper process for personnel selection that integrates AI or software based system, should consider a series of attention points, to guide the decisions of the recruiter and HR manager.

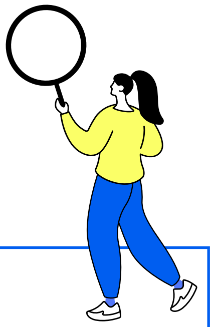
There thus exists the need to integrate computational tools and software for recruitment, with gender aware guidelines to lead the work of the developers and recruiters. This mixture of computational tools and qualitative methods can create a toolkit for the proper development of AI-based recruitment processes that are less prone to gender and ethnicity biases.

The present section goes exactly in this direction. We present here a series of guiding questions that can help the AI developer that works on an AI-based recruitment system. The questions and the attention points. The questions that compose the t INolkit, are divided into the main phases of data analysis when dealing with an AI process: data collection, training, evaluation, and implementation.

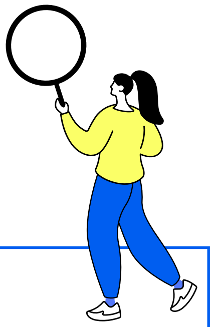
Checklist for ICT Specialists e AI developers to ensure gender equality in the design and development of recruiting tools

Table 2. Relevant questions for AI developers to implement gender equal tools for recruitment and the data analysis phase to which they belong to.

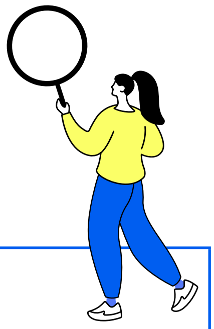
Questions	Measure to take	Phase
Is there any variables encoding the gender/ethnicity of the candidates? Is there any variables directly correlated to these dimensions? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **YES**, depurate data from gender info and from all the variables that could be related to it.	***Data Collection***
Is the dataset fairly balanced between men and women? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, integrate the dataset accordingly.	
Are the job posting channels equally used by men and women or not? How to measure the differences? Give option Y**ES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, change the platform to share the job post or find solutions (for example, right time for posting the vacancy) to increase the gender balance.	
Which are the best job posting platforms that ensure a gender equal application process? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	Detect the fairest job posting platforms.	
Is the data-collection team a diverse team (in terms of gender and etnicity)? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, integrate the team with new employees to increase the gender balance.	
Has the data collection process and result been revised by external (to the team) experts to audit for biases)? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, revise the results scouting external reviewers.	



Questions	Measure to take	Phase
Is the model relying on external data for the training (e.g. large textual models trained from social networks)? Has these external model been checked for gender and ethnicity biases? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.		***Training***
Is available, at the state-of-art, a lexicon of unbiased terms to be used as benchmark? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	Check literature and state-of-art and use it as benchmark.	
Has the model been audited using XAI (explainable AI) systems, to check for over-importance of gedner/etnicity realted features? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.		
Is the team working on the model engineering a diverse team (in terms of gender and etnicity)? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, integrate the team with new employees to increase the gender balance.	
Has the evaluation process been audited by external experts? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, revise the results scouting external reviewers.	



Questions	Measure to take	Phase
Are biases related metrics included in the performance metrics? Can a less accurate but also less biased model be preferred? Give option **YES/ NO**. If the answer could lead to bias, indicate the measure to take.		***Evaluation***
Is the team working on the model evaluation a diverse team (in terms of gender and ethnicity)? Give option **YES/ NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, integrate the team with new employees to increase the gender balance.	
Has the evaluation process been audited by external experts? Give option **YES/ NO**. If the answer could lead to bias, indicate the measure to take	If **NO**, revise the results scouting external reviewers.	



Questions	Measure to take	Phase
Are the users of the platform trained on the possible gender and ethnicity biases risks? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, prepare guidelines/ check list to align the team on the biases risks.	***Implementation***
If the systems is a continuous learning system (continually trained with new data) is the new data free of gender and ethnicity biases? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.		
If biases emerges during the implementation of the system, is the organization ready to re-implement the system and eliminate the biases? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, revise the results scouting external reviewers.	
Is the team working on the implementation of the system a diverse team (in terms of gender and ethnicity)? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, integrate the team with new employees to increase the gender balance.	
Is the system continuously audited by external experts process been audited by external experts? Give option **YES/NO**. If the answer could lead to bias, indicate the measure to take.	If **NO**, revise the results scouting external reviewers.	

References

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. *Semantics Derived Automatically from Language Corpora Contain Human-like Biases.* Science 356 (6334): 183–86. https://doi.org/10.1126/science.aal4230.

Gaucher, Danielle, Justin Friesen, and Aaron C. Kay. 2011. *Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality.* Journal of Personality and Social Psychology 101 (1): 109–28. https://doi.org/10.1037/a0022530.

Hajian, Sara, and Josep Domingo-Ferrer. 2013. *A Methodology for Direct and Indirect Discrimination Prevention in Data Mining.* IEEE Transactions on Knowledge and Data Engineering 25 (7): 1445–59. https://doi.org/10.1109/TKDE.2012.72.

Hoyle, Alexander Miserlis, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. *Unsupervised Discovery of Gendered Language through Latent-Variable Modeling.* In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 1706–16. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1167.

Li, Yang, and Tao Yang. 2018. *Word Embedding for Understanding Natural Language: A Survey.* In Guide to Big Data Applications, edited by S. Srinivasan, 26:83–104. Studies in Big Data. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-53817-4_4.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. *A Survey on Bias and Fairness in Machine Learning.* ArXiv:1908.09635 [Cs], September. http://arxiv.org/abs/1908.09635.

Mikolov, Tomas, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space.* Proceedings of Workshop at ICLR 2013 (January).

Pedreschi, Dino, Salvatore Ruggieri, and Franco Turini. n.d. *Discrimination-Aware Data Mining*, 9.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. *Mitigating Unwanted Biases with Adversarial Learning.* In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 335–40. New Orleans LA USA: ACM. https://doi.org/10.1145/3278721.3278779.

This Toolkit has been produced by the **GRASE project**. Visit our website to learn more about GRASE activities and products!