# Synthetic CVs To Build and Test Fairness-Aware Hiring Tools

JORGE SALDIVAR, Universitat Pompeu Fabra, Spain
ANNA GATZIOURA, Universitat Pompeu Fabra, Spain
CARLOS CASTILLO, Universitat Pompeu Fabra, Spain

Algorithmic hiring has become increasingly necessary in some sectors as it promises to deal with hundreds or even thousands of applicants. At the heart of these systems are algorithms designed to retrieve and rank candidate profiles, which are usually represented by Curricula Vitae (CVs). Research has shown, however, that such technologies can inadvertently introduce bias, leading to discrimination based on factors such as candidates' age, gender, or national origin. Developing methods to measure, mitigate, and explain bias in algorithmic hiring, as well as to evaluate and compare fairness techniques before deployment, requires sets of CVs that reflect the characteristics of people from diverse backgrounds.

However, datasets of these characteristics that can be used to conduct this research do not exist. To address this limitation, this paper introduces an approach for building a synthetic dataset of CVs with features modeled on real materials collected through a data donation campaign. Additionally, the resulting dataset of 1,730 CVs is presented, which we envision as a potential benchmarking standard for research on algorithmic hiring discrimination.

CCS Concepts: • **Information systems** → *Information systems applications*; **Decision support systems**;

Additional Key Words and Phrases: Synthetic Data, Algorithmic Hiring, Fairness, Benchmarking

**ACM Reference Format:**
Jorge Saldivar, Anna Gatzioura, and Carlos Castillo. 2025. Synthetic CVs To Build and Test Fairness-Aware Hiring Tools. *ACM Trans. Intell. Syst. Technol.* 37, 4, Article 111 (August 2025), 25 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

Although the European Union (EU) has long recognized equal employment as a force of social cohesion, dignity, and equality, its ambition remains unfulfilled as structural, institutional, and individual forms of discrimination continue to persist in the workplace [5]. The participation of women in the workforce continues to be unequal, and COVID-19 has derailed gender equality gains [18]. Similarly, ethnic minorities, people of African descent, women of color, and LGBTQ+ people have been shown to still be discriminated against.

In this context, algorithmic hiring is on the rise and rapidly becoming necessary in some sectors, as job postings that used to attract about 120 applicants in 2010 now attract over 250 [20]. AI technologies promise to deal with hundreds or thousands of applicants at high speeds. Moreover, their uptake in European HR teams and Public Employment Services is growing faster than the global average [21].

Authors' Contact Information: Jorge Saldivar, jorge.saldivar@upf.edu, Universitat Pompeu Fabra, Barcelona, Spain; Anna Gatzioura, anna.gkatzioura@upf.edu, Universitat Pompeu Fabra, Barcelona, Spain; Carlos Castillo, carlos.castillo@upf.edu, Universitat Pompeu Fabra, Barcelona, Spain.


ACM 2157-6912/2025/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

A key component in the realm of algorithmic hiring, i.e., the use of software technology to automate or assist employers through the different stages of recruitment (sourcing, screening, interviewing) [33], is the algorithms to retrieve and rank candidate profiles, which are usually materialized in a curriculum vitae (CV). They are at the core of job search engines and applicant tracking systems (ATS), which are the tools that can lead to algorithmic discrimination. It is known that discrimination in hiring can be due to personal data such as age [23] and gender [10], and due to sensitive attributes like national origin [32], or race [6].

Research on measuring, mitigating, and explaining bias in hiring processes assisted by algorithms, needs primarily real-world CVs that reflect the characteristics of people from diverse demographic backgrounds. However, the availability of sensitive information (i.e., sexual orientation, religion/-belief, or ethnicity) in the published datasets is scarce or simply nonexistent [19]. Also, the vast majority of datasets used in algorithmic hiring research are not publicly available due to the sensitive information typically contained in CVs, or they are composed of synthetic materials fabricated using aggregated, incomplete, or artificially annotated demographic data.

A way to advance the field can come from curating and publishing resources comprising synthetic CVs with i) representative job experience data; ii) diverse sensitive and demographic attributes; and iii) privacy guarantees for the data subjects involved. Moreover, the EU Artificial Intelligence Act (AI Act), in its article 10 (5.a), explicitly recommends the use of synthetic data as a primary option to avoid processing personal data when aiming to detect and mitigate algorithmic bias [7].

In this context, the contributions of this paper are twofold: i) an approach to build synthetic CVs with characteristics resembling actual CVs; and ii) a dataset of 1,730 CVs, created using this approach, from a set of materials collected through a data donation campaign [51]. We propose that this dataset could be used for benchmarking algorithmic hiring applications to promote increased fairness in ranking methods, especially when it comes to the evaluation and comparison against the state of the art of systems developed with relevance as a goal. Furthermore, we expect this dataset to become a standard reference for detecting biased ranking functions, in both academia and industry, while supporting increased explainability and research reproducibility.

The rest of the paper is structured as follows: Section 2 presents the theoretical foundation of our work and a review of the literature, while Section 3 introduces the proposed method, from the data collection to the generation of the dataset. In Section 4, we present the results, starting with the characteristics of the synthetic dataset, followed by its validation. In Section 5, we discuss our results, the main challenges faced and the elements we plan to improve in the future, to conclude our work in Section 6.

## 2 Related Works

In continuation, we present the theoretical background related to the main pillars of our work, namely: the most common application domain of synthetic data, their generation techniques, as well as some of the datasets that have been previously used in algorithmic hiring research.

### 2.1 Synthetic data application domains

Computer vision [42], and especially face recognition [3], audio, natural language processing and health, have been presented as some of the main domains where synthetic data has had a significant impact [22]. Furthermore, various possible uses of synthetic data have been identified in human analysis [27], financial applications [45], as well as in human behavior and activity applications [9].

Synthetic data has become a common approach for augmenting existing datasets or creating new ones, when real records are impossible to use due to logistic or ethical constraints [57]. In healthcare, for example, access to high quality data is particularly important, but also challenging due to privacy restrictions related to sensitive personal information and data sparsity, especially

in cases of less common conditions [37]. To this direction, synthetic data has been claimed to contribute to open research and innovation while enabling an improved diagnosis, treatment, and monitoring of diseases, being especially important for rare diseases or highly aggressive and complex conditions, to support more accurate, timely, and personalized healthcare solutions. The authors in [50] present a comprehensive literature review and the uses of synthetic data in multiples healthcare domains (oncology, neurology, cardiology, etc.) and their subdomains.

Synthetic data has been also widely recognized for its potential as a method to reduce bias in AI-based algorithms by removing imbalances and suppressing disparate impact, while maintaining data privacy, especially in high-risk applications, as in algorithmic hiring, referring to the use of systems that perform inferences for categorizing, classifying, ranking or recommending people [31]. More details on the usage of synthetic data in algorithmic hiring are presented next in 2.3.

### 2.2 Synthetic data generation techniques

Synthetic data refers to artificially generated data that mimics the characteristics and patterns of real-world data, but is created through algorithms or generative models rather than being directly collected or annotated by humans [34]. It relates to, but differs from, synthetic content generation (text, images, video, media, etc.), since rather than being created to be consumed by end-users, synthetic data are generated with the aim to be used as inputs to other data processing systems, where the original data cannot be used due to size or privacy concerns [53].

The first approaches to synthetic data generation and data imputation were based on non-parametric machine learning models, such as classification and regression trees, support vector machines and random forests, aiming to generate data based on the statistical distribution of attributes in real data [39]. However, the recent advances in generative AI have caused a shift in the computational paradigms used in many applications, including data generation techniques. Generative AI models refer to AI methods that learn the underlying distributions in existing data and generate novel data according to them. Among those, variational autoencoders (VAEs) and generative adversarial networks (GANs) have been widely used to generate synthetic data [26]. As the generation process depends on the type of data to be generated, for text data also NLP, BERT and transformer models, like GPT, have been applied, while convolutional neural networks (CNNs) have been found to better learn image representations. Finally, in domains where the data sequence is important, recurrent neural networks (RNNs) are used to generate meaningful synthetic sequences [14, 35].

To be considered as "good", synthetic data must be representative of the original data, while also providing guarantees about privacy [24]. Furthermore, the quality of synthetic data is highly dependent on both the quality of the original data and the generative model used: if the original data contains errors, biases, or under-represents certain features, the synthetic dataset will also reflect those flaws. Similarly, if the model is not able to properly capture the key characteristics of the original data, the generated data may not provide accurate insights. To this direction, Endres et al. highlight the importance of proper data collection and cleaning, as prior steps to training models to generate synthetic data [16].

For a synthetic dataset to be useful in a given application domain, it should have multivariate distributional properties similar to the dataset based on which it was made, ensuring that it could substitute the original dataset with very similar analytic results. Furthermore, the intended use has to be defined in advance, to properly select the most adequate generation technique and take into account domain specific characteristics and limitations [56].

An important aspect when generating synthetic data, and one of the main decisions to be made when building a model, is the desired trade-off between control and flexibility. Although some black-box approaches manage to produce high-dimensionality data and at a large scale, they are

particularly opaque, making it hard to have control over the process and estimate their privacy. Therefore, their use is not adequate for applications domains like algorithmic hiring [29]. As our research scope has been to generate a data set of synthetic CVs that could be used as a benchmarking dataset, under the premise that it would reflect the distribution of the original characteristics, while preserving anonymity, we have opted to develop a hybrid technique that allows us to have more control over the process and, as a consequence, a better understanding of the results.

### 2.3 Datasets and synthetic data in algorithmic hiring research

The study of fairness in algorithmic hiring requires access to personal data about job applicants, including their work experience and sensitive attributes. Unfortunately, there is a lack of publicly available resources comprising the CVs of candidates and their protected attributes [19]. Previous work on fairness in CV-based hiring has used private databases [8, 41] or short text snippets with limited sensitive attribute information [11, 59].

To highlight the importance and challenging aspects of algorithmic tools for job discovery and hiring, the annual RecSys challenge[1] focused on job recommendations both in 2016 [46] and 2017 [47]. A data set from the XING platform[2] was used. This dataset was formed on a semi-synthetic resumes, enriched with noise to anonymize and abstract from real user profiles. However, the aim of both years' tasks has been different from our purpose, as this dataset served for building models to determine the relevance and potential interest of users to job postings under different use cases.

Our research scope is quite different from the majority of works with synthetic data in the recruitment domain, as rather than exploring the linguistic characteristics appearing in online CVs [12], our goal is the creation of a dataset of synthetic CVs, following the distributions of the characteristics in donated data, to enable algorithmic fairness benchmarking. In contrast, the majority of CV or biography datasets are compiled as supportive elements for classification or linguistic exploration tasks rather than to improve research related to fairness in algorithmic hiring.

For example, Skondras et al. explore the potential of large language models to augment a dataset of real resumes crawled from Indeed[3] with artificially generated resumes created with chatGPT. Resumes in dataset belong to 15 different job categories, ranging from technical roles, like software developers, to professional roles, such as lawyers. The authors' main goal was to test whether the augmented dataset helps to improve resume classification tasks [52].

In [25] the authors also used Indeed to construct a corpus based on approximately 28k anonymous IT resumes being available on this platform. Their objective was the identification of skills relevant to job roles based on the unstructured textual information found in resumes and the classification and relevance calculation of those, rather than the use of this corpus to support the generation of synthetic curricula vitae. One of the few open datasets found is the Kaggle "Resume dataset" [1] which includes around 24k resumes taken from LiveCareer[4], in text format, and the most suitable profession for each, from 25 distinct profession sectors (e.g., Designer, IT, Teacher, Business, Healthcare, Agriculture), aiming to support an improved categorization of resumes into defined labels. Another open dataset is presented in [13], where the authors introduce a collection of 230k real CVs extracted from the Ukranian platform Djinni[5], primarily in English and from the IT sector, which also includes within the content of the documents explicit mentions to some demographic attributes such as age, gender, marital and military status, and religion.

[1]Please refer to RecSys 2016: http://2016.recsyschallenge.com; RecSys 2017: https://www.recsyschallenge.com/2017
[2]Please refer to https://www.xing.com
[3]Please refer to https://www.indeed.com
[4]Please refer to https://www.livecareer.com
[5]Please refer to https://djinni.co

Peña et al. [44] study how current multi-modal algorithms for algorithmic hiring are affected by sensitive attributes and preexisting biases in the data. They use a dataset of 24k synthetic resumes covering 4 job sectors, with 12 features including education, availability, previous experience, occupation, name, and language, 2 demographic attributes (gender and ethnicity) and a face photograph from the DiveFace database [36] correlated with the demographic attributes. They further extended this dataset [43] by including short biographies to the CVs using the Common Crawl Bios dataset [11] which contains online biographies related to 28 different occupations. The Common Crawl Bios dataset [11] was created with the aim to study gender bias in occupations' classification and consists of almost 400k biographies. Although in both works bias in algorithmic hiring is considered an important aspect, their focus is on its detection, rather than on the generated datasets. In addition, in the synthetic CV dataset that we propose, only structured information related to education, work experience and obtained skills, is included. We do not assign images or names to the generated CVs.

Bruera et al. [4], starting from the Indeed dataset used by [25], propose a method that combines Bayesian networks and natural language processing techniques to resemble candidate attributes as found in a dataset of real CVs. They first get the initial structure of a CV and then complete it using prompts to query a generative model. In contrast to our work, they include artificial personal information in their synthetic CVs. This work is probably the closest to ours; however, the authors aim to generate a synthetic dataset permitting to train machine learning models and do not focus on fairness issues or the benchmarking of ranking algorithmic in terms of fairness. Table 1 summarizes the information from the CV datasets presented in this section.

Table 1. Datasets of CVs published in the literature

| Ref. | Type | Source(s) | N | Job sectors | Main contents | Demographics | Availability |
|---|---|---|---|---|---|---|---|
| [46] | Semi-synthetic | Xing | Unknown | Unknown | Education, experience, career level, discipline, industry, country | No | Unavailable |
| [47] | Semi-synthetic | Xing | Unknown | Unknown | Education, experience, career level, discipline, industry, country | No | Unavailable |
| [52] | Hybrid (Real, synthetic) | Indeed, ChatGPT | 2K | 15 | Unknown | No | Upon request |
| [25] | Real | Indeed | 28K | 1 | Unknown | No | Unavailable |
| [1] | Real | Livecareer | 24K | 25 | Education, experience, skills | No | Public |
| [13] | Real | Djinni | 230K | 1 | Education, experience, skills, english level, years of experience, name | Age, gender, religion, marital and military status | Public |
| [44] | Synthetic | DiveFace, US Census Bureau 2018 | 24K | 4 | Education, occupation, availability, experience, languages, name, photo | Gender, ethnicity | Unavailable |
| [11] | Real | Common Crawl | 400K | 28 | Short biography, name | Gender | Unavailable |
| [4] | Synthetic | [25] | 4K | 1 | Skills, education, experience, hobbies | No | Unavailable |

**Note.** *Ref.*: reference; *N*: number of CVs in the dataset; *Jobs sectors*: number of job sectors represented in the dataset. The dataset [43] reviewed above was not considered in the table, as it is an extension of [44], which is already included.

## 3 Method

Our proposed approach is to generate a dataset of synthetic CVs based on a reference dataset built from data donated by consenting individuals whose identities were safeguarded by anonymizing the donated data.

### 3.1 Data Donation Campaign

We proposed collecting real CVs and attributes potentially leading to discrimination (i.e., age, gender, religion, origin, and disability condition) through a data collection campaign. Residents of the European Economic Area and Switzerland who are part of the labor force (i.e., employed or seeking employment) were invited to voluntarily donate up to two anonymous CVs by completing an online survey[6]. Donors were recruited online and provided with an information sheet and a consent form [51].

The campaign started in June 2023 and remained open until the end of May 2024. In total 1,143 donations were received, four of which were discarded because of incomplete fields. The remaining 1,139 completed submissions included 1,211 CVs, considering that about 15% of them contained two CVs (up to two CVs could be attached to the submission). Most donations were in Spanish (78%, 895 out of 1,139), impacting the language of CVs, which are primarily written in Spanish (69%, 836 out of 1,211).

Half of the donors declared themselves as professionals (567 out 1,139) from various sectors, including science and engineering, business and administration, ICT, legal, social, cultural, and health. An interesting balance between junior and senior workers is available in the data. Fifty percent of the donors are between 26 and 45 years of age (581 out of 1,139), and half of donations were submitted by women (576 out of 1,139). Almost 20% of donors reported belonging to the LGBTQ+ community (192 out of 1,139), a similar proportion declared being part of a minority group (229 out of 1,139), and 15% perceived themselves as foreign in the country where they live (179 out of 1,139). About 45% of the campaign participants (491 out of 1,139) reported being either secular or not religious; the rest are mostly Christians, while Muslims, Buddhists, Hinduists, and Jews are marginally represented in the data. Less than 10% of donors (84 out of 1,139) declared having a disability condition.

Although most donors are Spanish nationals, the donation dataset is representative of the European workforce in terms of gender, age, and foreign condition (i.e., whether donors feel foreign in the country where they live) [51]. On the other hand, it overrepresents people from the LGBTQ+ collective, ethnic minorities, and disabled communities, which is somewhat expected given the evidence of employment discrimination suffered by these groups [2, 55, 58]. Regarding religion, Europe is majority Christians [17], while donations were primarily submitted by agnostic/secular people. The group of Muslims is, however, fairly well represented, whereas other religions, like Buddhists, Jews, and Hinduists, are represented in both the general population and donations with less than 1%.

Some professional sectors, such as ICT, engineering, clerical support, and business administration, are overrepresented in the donations. We understand that this is related to the digital means used to run and advertise the campaign, which might facilitate donations from these sectors, i.e., those using computers to support their daily tasks, but complicates participation from sectors that typically involve manual labor. Another explanation might be that overrepresented sectors align with professionals familiar with using CVs and job search platforms to access the labor market.

The collected sample with the above characteristics is used as a reference for constructing the synthetic data. We aim to generate synthetic CVs with statistical properties resembling those of people who participated in the donation campaign.

### 3.2 Synthetic CV Generation Approach

We propose an approach that combines automatic and manual tasks to generate the synthetic data. Figure 1 shows the pipeline to collect, process, store, and produce the synthetic CVs. For our purpose,

[6]Please refer to https://findhr.eu/datadonation to access the website used to run the campaign.

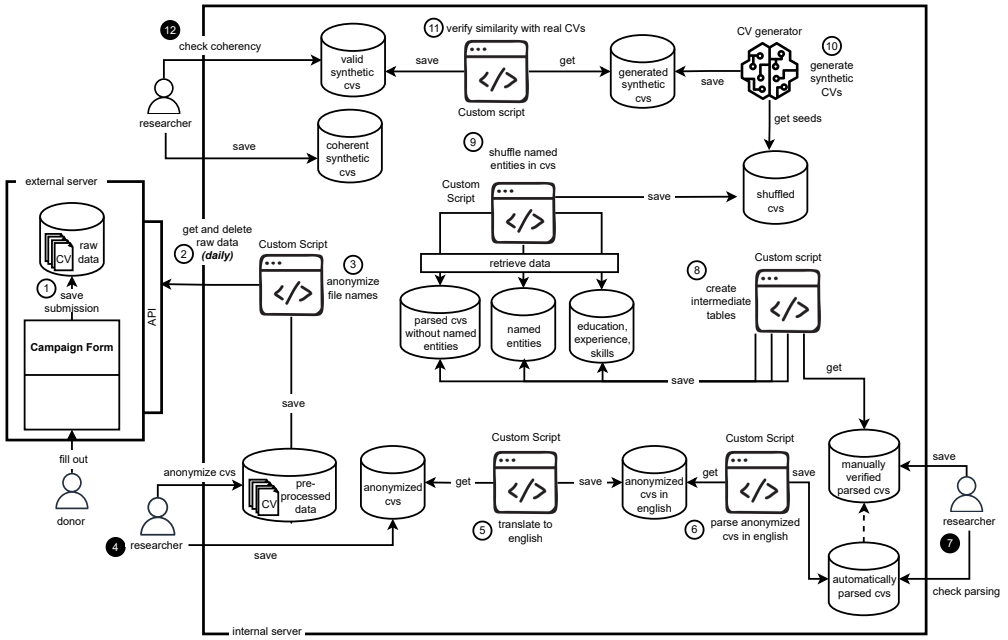



Fig. 1. Pipeline proposed to generate the synthetic dataset. Black circles indicate manual tasks.

a synthetic CV is defined as a structured text document organized in three sections: *educational background*, *professional experience*, and *skills*, each containing a number of items representing information related to the section.

The pipeline starts by processing the donations as they arrive (step 1 in Figure 1). This involves anonymizing the CV file name(s), creating a directory to allocate the CVs, and generating a JSON file containing all submission data as is, i.e., without any further processing (steps 2 and 3). In step 4, donated CVs are manually anonymized, removing all personally identifiable information (names, addresses, pictures, emails, phone numbers). Later, CVs written in languages different from English are automatically translated into English (step 5).

Anonymized CVs in English and formatted as PDFs are parsed by hitting the RESTful API of the GDPR-compliant resume parsing service EdenAI (step 6). EdenAI[7] was chosen as the parser due to the time, cost, and human resources limitations of building an in-house parsing tool. The parsing is automatically processed, generating the JSON structure shown in Listing 1. Later, the created JSON is manually verified (step 7), filling out missing information or fixing incorrect data. At this step, the processing of CVs finishes, and what remains is related to the generation of synthetic CVs.

*3.2.1 Intermediate dataset.* In preparing the generation and as a way to reduce potential re-identification risks, information in the parsed CVs is extracted from JSON files and split into three unrelated tables (step 8). One of the tables, *anonymized-cvs*, contains, in random order, one record per anonymous, parsed, and processed CV in JSON format, as well as the job sector, years of professional experience, and demographic data (age, gender, origin, etc.) associated with the CV (see Table 2). CVs included in this table not only miss personally identifiable information, but

[7]Please refer to https://www.edenai.co

Listing 1. Illustrative example of a JSON that resulted from processing the parsing output

```
{
    "education_background": [
        {
            "degree": "BSc. Computer Science",
            "start_date": "2019-09-15",
            "end_date": "2023-06-30",
            "institution": "Pompeu Fabra University"
        },
        { ... }
    ],
    "professional_experience": [
        {
            "role": "Junior Software Developer",
            "start_date": "2023-01-01",
            "end_date": "Present",
            "institution": "AI company",
            "description": "..."
        }
    ],
    "skills": [
        "hard": ["Python", "MySQL", ...],
        "soft": ["Leadership", "Teamwork", ...],
        "languages": ["English", "Spanish", ...],
        "others": ["Photography", "Public speaking", ....]
    ]
}
```

Table 2. Table anonymized-cvs

| Anonymized and parsed CV | Job Sector | Year of professional experience | Age | Gender | Ethnicity | … |
|---|---|---|---|---|---|---|
| {...} | IT | 12 | 38 | Woman | … | … |
| {...} | Engineering | 1 | 25 | Woman | … | … |
| {...} | Business | 7 | 31 | Man | … | … |
| ... | ... | ... | ... | ... | ... | ... |

**Note.** Illustrative example of the table containing, in random order, one record per anonymized and parsed CV plus the job sector, years of professional experience, and additional data (regular, sensitive).

also names of institutions (companies, universities, NGOs) are removed from the education and professional items. Due to parsing inconsistencies, education dates may not reflect typical study durations. To address this, rules–such as a minimum of three years for Ph.D. programs or nine months for master's–are applied before adding entries to the anonymized table. The other table, *education-experience-skills-combinations*, involves, in random order, one record per CV, including the job sector and experience, together with the list of educational institutions, workplaces, and skills contained in the donated CV (see Table 3). The last table, *named-entities*, contains a job sector, a value of a sensitive variable (e.g., "Woman" or "Muslim"), the number of CVs in the job sector, including that value, which must be larger than or equal to 5 to be included, and an alphabetic list of entities found in CVs with that value, i.e., educational institutions and workplaces (see Table 4). The latter rule—included in the approved protocol for data generation (see section 3.4)—was arbitrarily decided to protect less-represented groups of donors.

These tables are created to protect donors by separating sensitive information from their CVs, feeding the generator with this intermediate dataset. Donated materials are never directly used to generate synthetic CVs. Also, the content of these tables is employed to add variability to the seed data by shuffling the education institutions and workplaces of donors with similar demographic

Table 3. Table education-experience-skills-combinations

| Job Sector | Years of professional experience | Educational institutions | Workplaces | Skills |
|---|---|---|---|---|
| IT | 12 | Pompeu Fabra University, ... | AI Company, ... | Python, MySQL, Leadership, Teamwork, English, Spanish, Photography, Public speaking |
| Engineering | 1 | Complutense University of Madrid, ... | Sakyr SA, ... | Calculus, Excel, AutoCAD, Teamwork, Portuguese, Italian, Running, Communication |
| Business | 7 | INCAE Costa Rica, ... | Inditex, ... | Excel, PowerBI, Finance, Leadership, Spanish, Italian, Guitar |
| ... | ... | ... | ... | ... |

**Note.** Illustrative example of the table containing, in random order, one record per donated CV with job sectors and years of professional experience, list of educational institutions of an individual, list of workplaces of the individual, and words/phrases describing the individual's skills.

Table 4. Table named-entities

| Job Sector | Year of professional experience | Variable | Variable Value | Num. CVs | Education Institutions | Workplaces |
|---|---|---|---|---|---|---|
| IT | 12 | Age | 22 | 37 | Pompeu Fabra University, ... | AI Company, ... |
| IT | 12 | Gender | Woman | 18 | Pompeu Fabra University, ... | AI Company, ... |
| Engineering | 1 | Age | 25 | 31 | Complutense University of Madrid, ... | Sakyr SA, ... |
| Engineering | 1 | Gender | Woman | 12 | Complutense University of Madrid, ... | Sakyr SA, ... |
| Business | 7 | Age | 31 | 43 | INCAE Costa Rica, ... | Inditex, ... |
| Business | 7 | Gender | Man | 23 | INCAE Costa Rica | Inditex, ... |
| ... | ... | ... | ... | ... | ... | ... |

**Note.** Illustrative example of the table containing a job sector, a value in a regular or sensitive category data, the number of CVs in the job sector including that value, which must be larger than or equal to 5 to be included, and an alphabetic list of entities found in CVs with that value, i.e., locations, educational institutions, and workplaces.

characteristics (step 9). As part of this step, mappings between companies and professional roles and between education degrees and academic institutions are generated using Sentence-BERT [48], a variant of BERT [30] that uses siamese neural networks for large-scale text comparison. A manual inspection of the generated mappings demonstrates useful and reliable results.

Similarly, the probability distribution of skills associated with the degrees and roles is calculated. In the process, a higher abstraction level is introduced to group together degrees, or roles, that refer to the same concept while different names have been used in the donated CVs (e.g., a bachelor's degree that has been introduced as "bachelor in science" and "BSc"). This classification permits the calculation of probability distributions in relation to more general, educational, and working experience categories, leading to reduced data sparsity. The relevance of degrees and roles to skills is then stored into separate tables that contain, respectively, the educational degrees and their relevance to each of the observed skills and the job roles and their relevance to each of the skills.

*3.2.2 Final steps of the process.* After the generation (step 10), synthetic CVs are automatically verified (step 11) to ensure they do not closely resemble the donated materials of any individual donor. The final step (step 12) involved evaluating the generated CVs. A research assistant —trained by the authors— manually reviewed each synthetic CV, assessing its coherence with the target professional sector. The guiding question was whether the education, work experience, and skills described in the CV reflect a plausible career trajectory within that sector. For example, a full-stack

developer with a bachelor degree in software engineering would not be coherent with the *clerical support* sector, aligning instead with *ICT*.

In addition, the coherence within and between sections (education, experience, skills) is evaluated, allowing flexibility for non-traditional career paths. This means that a political science graduate who worked in customer support at a telecommunications company while studying would be considered coherent, whereas a graduate in web design and marketing working as a motorcycle mechanic would not. Similarly, it is implausible for a seller with a high-school diploma to be presented as an expert in schizophrenia therapies. Moreover, highly divergent professional or academic trajectories, such as completing a master's degree in microprocessors followed by a PhD in law, or serving as a CEO before working as an accounting assistant, while possible, are not included in our resulting dataset. Each criterion was rated on a scale from 1 (poor) to 5 (excellent). After this process, only CVs with an overall rating of at least 4 (very good) out of 5 were retained; the rest were discarded.

### 3.3 Synthetic CV Generator

The primary goal of the synthetic CV generator (step 10 in Figure 1) is to resemble the characteristics of real donated CVs. Generated documents do not contain personal information (e.g., name, email address, phone number, or picture) and are structured into three sections, namely educational background, professional experience, and skills, each section containing a list of items.

The generation is based on the mandatory parameters: job sector and years of professional experience. It is required to specify the job sector of the CVs to be generated, as well as the years of professional experience that the synthetic CV should reflect. Also, at least one personal and/or sensitive attribute, like age, gender, disability condition, or ethnicity, should be specified. Next, the generator finds out what a "typical" CV of people with the specified characteristics looks like, i.e., it needs to recognize the typical characteristics of CVs belonging, for example, non-European female lawyers in their 30s who have 5 years of professional experience. In other words, it determines how to fill in the sections on educational background, professional experience, and skills with respect to the number of items in each section and their content.

*3.3.1 Structure Computation.* Following the example above, say we want to compute the number of items in the skills sections of a "typical" CV of non-European female lawyers in their 30s with five years of professional experience. We extract all anonymous CVs of non-European donors (from Table 2) and check the number of items in their skill sections. Then, a Weibull distribution [49] that satisfies these numbers is computed. A similar procedure is applied to the rest of the input parameters, i.e., in the example, gender (female), professional sector (law), age (30), and professional experience (five years). Weibull distribution was chosen because sampling from this curve does not result in negative values, which is crucial in our case. Also, Weibull is typically used in survival analysis, and hiring can be seen as a process in which candidates kind of "strive to survive" until they reach the final stages. Moreover, Weibull is shown to be sufficiently flexible despite requiring only two parameters (shape and scale).

At the end of this process, we have as many distributions as the number of input parameters. Each distribution is sampled, and the results are combined using a randomly chosen strategy, including mean, median, min, and max. The number of items for the section educational background and professional experience is calculated using this approach. For the skills section, the number of items is capped at 12, as generation tests showed that the section coherence drops significantly beyond this limit. Algorithm 1 shows the pseudo-code proposed to compute the number of items to be generated for each section.

*3.3.2 Content Generation.* Once the number of items in each section is known, the next step involves deciding how to fill them, and here is where the tables created during the processing tasks

**Algorithm 1** Compute number of items per section

**Require:** $input\_parameters$ ▷ Array of input parameters
$items \leftarrow$ INIT_DICTIONARY() ▷ Initialize an empty dictionary
$num\_params \leftarrow$ LENGTH($input\_parameters$) ▷ Return the length of the array as integer
**for** $section \in \{education, experience, skills\}$ **do**
  $dists \leftarrow$ INIT_ARRAY($num\_params$) ▷ Initialize an empty array of n elements
  **for** $i \in \{1, \ldots, num\_params\}$ **do**
    $cvs \leftarrow$ EXTRACT_CVS($table\_2, input\_parameters_i$) ▷ Get items in Table 2
    $num\_cvs \leftarrow$ LENGTH($cvs$)
    $num\_items \leftarrow$ INIT_ARRAY($num\_cvs$)
    **for** $j \in \{1, \ldots, num\_cvs\}$ **do**
      $num\_items_j \leftarrow$ GET_NUM_ITEMS($cv_j, section$) ▷ Get number of items in section
    **end for**
    $dists_i \leftarrow$ WEIBULL($num\_items$) ▷ Draw Weibull distribution
  **end for**
  $samples \leftarrow$ INIT_ARRAY($num\_params$)
  **for** $k \in \{1, \ldots, num\_params\}$ **do**
    $samples_k \leftarrow$ SAMPLING($dists_k$) ▷ Sample Weilbull distribution
  **end for**
  $samples\_comb \leftarrow$ MEAN($samples$) ▷ Combine samples
  $items[section] \leftarrow$ ROUND($samples\_comb, 0$) ▷ Round value to 0 decimals
**end for**
**return** $items$

(step 8 in Figure 1) come in handy. The procedure proposed to create the content of synthetic CVs, is described next.

(1) Collect CVs that satisfy the input parameters by querying Table 2. Remember that CVs included in this table are in JSON format, do not contain personal information, and miss the names of all institutions reported in the document. Going back to the example, this means recovering all CVs of non-European female lawyers in their 30s with five years of professional experience. The number of CVs must be larger than 20 to continue; otherwise, the process is abandoned. This rule was decided based on a sufficiently large number of CVs (20) to avoid compromising donors' identity and is part of the safeguard protocol approved by the Institutional Committee for Ethical Review of Projects at Pompeu Fabra University, see section 3.4;
(2) Use Table 4 to extract institutions (universities and workplaces) belonging to CVs that satisfy the input parameters. That is to say, create a list with the institutions contained in the CVs of non-European lawyers with five years of professional experience plus the institutions available in CVs of female lawyers with five years of professional experience plus institutions in CVs of lawyers in their 30s with five years of professional experience;
(3) Fill in the missing university and workplace names of CVs collected before with those extracted in the previous step. The mappings created in step 9 are employed here to ensure coherent replacements. The chosen institutions are checked against the Table 3 to ensure they do not accidentally match or near-match (match, except for one parameter) a real combination for a donor;
(4) Separate CVs' items into three groups: education, experience, and skills. Compute embeddings for the items in each group and cluster the embeddings using Agglomerative Clustering [54];
(5) Generate sections education background, professional experience, and skills by selecting the required number of items according to the following generative procedures:
   - For the **education background** section, it is arbitrarily defined that it can have, at maximum, five items, namely, one Bachelor, one abroad experience (e.g., Erasmus, internships, visiting scholar periods), two Masters, and one PhD. The procedure for the education background section starts by randomly selecting a bachelor item—if, for the given combination of parameters, reference CVs do not contain bachelor items that can be selected,

a vocational experience (if any) is chosen, and the process stops. Next, at maximum, one abroad, two Masters, and one PhD experience that belong to the same cluster of the picked bachelor are randomly selected. Whether to include or not these items is guided by a stochastic mechanism that aims to increase diversity in the produced education sections. If an abroad experience is decided to be included, it should be within the period of the already included bachelor, Master, or PhD. If master experiences are picked, their start dates should be later than the end date of the included bachelor. A similar logic is applied to a PhD, meaning if a PhD experience is chosen, it cannot start before the Master's ends;

- For the **professional experience** section, items are added subsequently, ensuring that picked items follow a chronological order. In the worst case, if no professional experience comes later than the initially selected item, the section has only one item. But, on average, between one and the required number of professional experience items are included. Another restriction imposed on the procedure is that the total duration of the selected professional experiences cannot surpass the expected professional experience specified in the mandatory parameter *years of professional experience*. Like the procedure to generate the education section, the process starts by randomly selecting a professional experience. If the duration of the chosen experience is longer than the expected *years of professional experience*, the process stops. Otherwise, the process continues by randomly selecting subsequent items that belong to the cluster of the first chosen experience while the duration of the selected professional experiences altogether does not exceed the specified *years of professional experience*;
- For the **skills** section, we identify the skills being most relevant to the items that are included in the education background and professional experience sections. The selection is based on probability distributions (calculated during the pre-processing phase, see 3.2.1) that associate education backgrounds and working experiences with skills. As skills may be associated with multiple education and professional entries, the retrieved distributions are aggregated and ranked. Finally, the top $n$ skills are selected, where $n$ represents the number of items that were previously calculated to be included in the skill section. In cases where a smaller number of relevant skills are identified, all of them are included.

*3.3.3 Generation Rules.* A set of rules has complemented the generation procedure to guarantee a certain quality in the resulting fabricated CVs and minimize clumsy inconsistencies and errors. Generally, there cannot be two similar synthetic CVs in a generation attempt. Second, items either in the skills section as well as in the educational background or professional experience cannot be duplicated. Synthetic CVs with empty sections are rejected.

*3.3.4 Generation Parameters.* In generating the synthetic CVs, we aim to resemble the characteristics of the donated materials. To satisfy this requirement, we need to feed the generator with some parameters, including both professionally and personally related information.

Mandatory parameters correspond to the job sector and years of professional experience that CVs should reflect. Values that the mandatory parameters can take are presented in Table 5. Names of all job sectors, except Public officials, correspond to categories of the European classification of Skills, Competences and Occupations (ESCO), which were used in the donation form (see Section 3.1)[8]. Public officials, on the other hand, is a renaming of the ESCO category *Chief executives, senior officials and legislators* aiming to look for a more concise and representative term that covers all professional activities in this sector.

[8]Interested readers can check ESCO website (https://esco.ec.europa.eu/en/classification/occupation) for information about the meaning of each job sector

Table 5. Mandatory generation parameters and their corresponding values

| Parameter | Job sector | Years of professional experience |
|---|---|---|
| Values | Business and administration; Clerical support; ICT; Science and engineering; Sales; Legal, social and cultural; Construction, manufacturing and transport; Health; Public officials; Production and specialized services; Personal service; Teaching; Cleaning; Food preparation; Food Processing, Woodworking, Garment and Other Craft; Agricultural, forestry and fishery; Armed forces; Hospitality, retail and other services; Personal care; Handicraft and Printing; Protective | 4 years or less, 5-9 years, 10-14 years, 15+ years |

Table 6. Generation parameters related to demographic attributes and their values

| Parameter | Age | Disability condition | Gender | Minority | Perceived foreign | Religion | LGBTQ+ |
|---|---|---|---|---|---|---|---|
| Values | <= 30, 31-40, 41-50, > 50 | Yes, No | Woman, Man, Non-binary | Yes, No | Yes, No | Buddhism, Christianity, Hinduism, Muslim, Judaism, Other, Secular | Yes, No |

In addition, the generator requires one or more demographic attributes, such as age, gender, ethnicity, or religion. Unlike the others, the attribute ethnicity was not operationalized through direct options but implemented via the indirect parameters *Perceived foreign in the country of residence* and *Belong to an ethnic minority in the country of residence*. This corresponds to how information about donors' ethnicity was collected. Table 6 lists the parameters related to the demographic attributes.

### 3.4 Ethical Considerations

Generating synthetic CVs from donated real data is not exempt from risks. The main risk is the association between sensitive data (e.g., sexual orientation, ethnicity, religion) and donated CVs, which may lead to identifying a person as belonging to a minority group. In this sense, we consider internal and external risks. Regarding internal risks, researchers who have direct access to the data can associate CVs with sensitive data. The risk is mitigated by limiting access to the data on a need-to-know basis through the legal protections of a Non Disclosure Agreement. Also, data is kept encrypted and discarded after achieving the final version of the synthetic data.

Concerning the external risks, researchers accessing the synthetically generated CVs could use them to establish an association between a donated CV and sensible data. However, this risk could generally be considered minimal. First, this data intruder would need to find the original set of CVs to be able to reason about them based on the synthetic data, and neither the identity of the donors nor the corpus of CVs will be available online. Second, even if the original set of CVs were leaked, a data intruder will most likely not know which pieces of information of the synthetic CVs come from which original CVs. Hence, reasoning about the associations between identifying and sensitive information will be highly uncertain.

The procedure for collecting real CVs, the approach introduced to safeguard donors' identities, and the methods to process and generate the synthetic data have passed through a strict ethical review conducted by the Institutional Committee for Ethical Review of Projects (CIREP, by its Spanish acronym) at Pompeu Fabra University. Given the sensitive data collected during the campaign, CIREP was particularly concerned about protecting the identity of donors. After two rounds of reviews and revisions, the protocol was approved.

The dataset will be available to researchers at established European institutions, subject to a license agreement. This agreement prohibits re-identification attempts and outlines permitted uses of the corpus. Interested researchers should follow the instructions at https://findhr.eu/synthetic-cvs. According to the approved protocol, all donated materials will be destroyed following the publication of this research.

### 3.5 Technical Implementation

The pipeline is implemented with Python and open-source data processing libraries such as Pandas, Numpy, Streamlit, Sbert, and Spacy. Bash scripts automate the execution of a series of steps. Tables and databases were developed using comma-separated values (CSV) and JSON files.

The generator was implemented through a command-line interface (CLI) facilitates the automatic generation of CVs in batch. The CLI application operates in two steps. The first step computes a list of plausible generation parameters by trying all different combinations of professionally related attributes (job sector and years of professional experience) and demographic attributes, such as gender, age, ethnicity, or disability condition.

After checking whether the number of real CVs that satisfy a given combination is larger than the threshold established to safeguard donors' identity (20 according to the generation procedure, see Section 3.3), the combination is included in the list of plausible generation parameters. In the second step, the CLI application takes the combination of parameters from the list of plausible generation parameters and creates synthetic CVs. A generation execution is conducted per combination. That is to say, if, for example, the list contains 50 plausible combinations of parameters, the application executes 50 attempts to generate CVs. Apart from the generated CVs—saved as JSON files—a report file is produced describing the total number of CVs generated as well as the number of CVs produced per parameters.

### 3.6 Dataset Validation

We believe the donation dataset is valuable as a reference for studying bias in algorithmic hiring because it reflects the demographic characteristics of the European workforce. In this context, the synthetic dataset is considered valid if it demonstrates similar utility to the reference dataset.

Our utility validation is based on the indistinguishability between the synthetic dataset and the reference dataset. In this sense, two methods are proposed. On the one hand, we conduct univariate distribution comparisons, as proposed in [15]. This way, we can analyze whether variable distributions are similar in the reference and synthetic data. In particular, the distribution of the variables: job sector, years of professional experience, age, gender, ethnicity, religion/belief, and disability are computed for both datasets. Results are depicted on charts, whose shapes are visually inspected. Additionally, Jensen–Shannon divergence [38] is applied to analytically explore differences in distributions. Jensen–Shannon divergence is a technique from information theory that measures the divergence between two distributions.

On the other hand, a subjective evaluation [28] is conducted. In this sense, crowdsourcing workers are invited to look at 10 randomly selected CVs, 5 from the reference and 5 from the synthetic sets, and select whether they perceive the CV as real or artificially created. A representative random sample of 1,000 CVs, 500 from the reference and 500 from synthetic datasets, is used in this part of the evaluation, and three workers assess each CV. We make sure to include in the sample CVs corresponding to the different professional sectors available in the datasets.

After reading the information sheet and approving the consent form, the participants are instructed on the task with the following prompt: *You will be presented with 10 CVs and asked to determine whether each one appears to be real (i.e., belonging to an actual person) or artificially generated (i.e., created from information extracted from real CVs)* and the study can be started. The

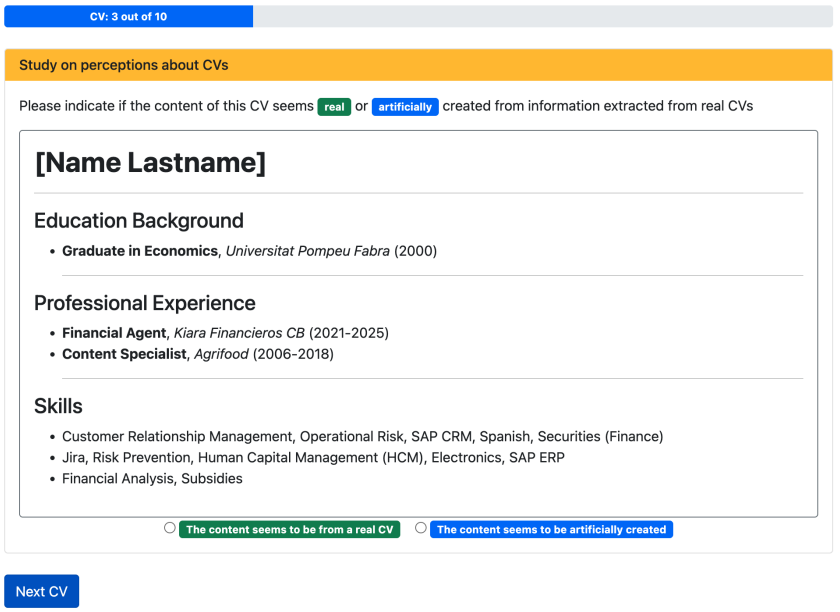



Fig. 2. Screenshot of the website used in the subjective evaluation of the synthetic CVs.

10 CVs (5 real, 5 synthetic) are presented one by one in random order and using Markdown format [40], as shown in Figure 2. For each CV, the following prompt is provided: *Please indicate if the content of this CV seems real or artificially created from information extracted from real CVs* and the participants are presented with the next options: *The content seems to be from a real CV* or *The content seems to be artificially created.*

The subjective evaluation was piloted with 28 crowdsourcing workers using a sample of 100 CVs in total (50 real, 50 synthetic). The goal was to ensure that the instructions were clear, the options unambiguous, and the interface seamless. These pretesting sessions allowed us to correct typos in the text and adjust the mechanism that ensures that each CV is evaluated by at least three workers.

## 4 Results

The execution of the implemented approach generated a dataset of synthetic CVs, which is described next. The generation was based on 948 of the 1,211 donated CVs. The rest were excluded for various reasons, including duplication, incompatible formats, incomplete content, and excessive length.

### 4.1 Dataset of Synthetic CVs

The initially generated synthetic dataset contained 2,218 CVs covering six job sectors, namely *science and engineering*, *ICT*, *business and administration*, *sales*, *clerical support*, and *legal, social, and cultural*. CVs from the initially generated dataset were manually analyzed by a research assistant (step 12 in Figure 1, see Section 3.2.2), who evaluated the coherence of the content on a scale of 1 (poor) to 5 (excellent). CVs with an overall coherence score of at least 4 out of 5 were included in the final dataset, which comprises 1,730 documents—almost 80% of those initially generated. The sectoral distribution of CVs is heterogeneous, with *business and administration* and *clerical support* accounting for over half of the dataset. Table 7 shows the distribution of CVs per sector.

Table 7. Distribution of synthetic CVs per job sector

| Job Sector | Number of CVs | Percentage |
|---|---|---|
| Clerical support | 468 | 27.05% |
| Business and administration | 451 | 26.07% |
| Science and engineering | 303 | 17.51% |
| ICT | 216 | 12.49% |
| Sales | 208 | 12.02% |
| Legal, social, and cultural | 84 | 4.86% |
| Total | 1,730 | 100% |

Table 8. Distribution of synthetic CVs per years of professional experience

| Years of Professional Experience | Number of CVs | Percentage |
|---|---|---|
| 4 years or less | 787 | 45.50% |
| 5-9 years | 209 | 12.16% |
| 10-14 years | 20 | 1.17% |
| 15+ years | 714 | 41.27% |
| Total | 1,730 | 100% |

The professional experiences reproduced in the CVs in the final dataset are predominantly clustered at the extremes of the experience spectrum. Specifically, 87% of CVs reflect either junior (with 4 years or less of experience) or senior profiles (with 15 or more years of experience). Table 8 outlines the distribution of CVs by professional experience.

The generated dataset reflects various demographic characteristics, including age, gender, religion, nationality, and ethnicity, but each CV highlights only one demographic attribute at a time. In other words, no synthetic CV combines multiple demographic traits, such as representing a female (gender attribute) engineer in her 40s (age attribute). Combinations involving more than one demographic parameter turned out to be impossible due to insufficient data to represent these combinations in accordance with the approved protocol, i.e., at least 20 real CVs must exist that satisfy the combination (see Section 3.3.2).

The synthetic CVs in final dataset generated using *gender* as the demographic attribute (270) are evenly split between male and female profiles. Most synthetic CVs created with *age* as the demographic factor (180) represent primarily profiles aged 40 or younger. A total of 242 CVs in the final dataset were produced using *LGBTQ+* as the defining characteristic, with 6% representing LGBTQ+ profiles. Among the 297 CVs generated with the *minority* parameter, 50% reflect profiles from minority groups. Additionally, 12% of the CVs created with the *foreign* attribute (301) represent profiles of people who feel foreign in the country where they live. Finally, synthetic CVs depict only Christian and non-disabled profiles, lacking CVs that reproduce data of individuals from other religions (e.g., Buddhism, Hinduism, Islam) and with disability conditions.

An example of a synthetic CV of a profile with less than four years of experience in the Sales sector is shown as JSON format in Listing 2. Apart from the JSON format, the Synthetic CVs are also available in Markdown to facilitate their inspections. The dataset is available for researchers in established academic institutions of the European Union (i.e., institutions having a participant number registration with the EU, known as a "PIC") under a free-of-charge license agreement that forbids donor re-identification attempts. Please contact the paper authors for instructions.

Listing 2. JSON example of a synthetic CV in the final dataset generated for the Sales sector


```
{
    "education_background": [
        {
            "institution": "UNED",
            "start_date": "April 2022",
            "end_date": "Ongoing",
            "degree": "Degree In Law"
        }
    ],
    "professional_experience": [
        {
            "institution": "Alcampo",
            "start_date": "January 2022",
            "end_date": "December 2023",
            "role": "Cashier  Stocker",
            "duration": "1 year, 11 months",
            "duration_months": 23
        }
    ],
    "skills": {
        "others": ["Literacy", "Informatics", "Social Integration", "Research"]
    }
}
```


### 4.2 Validation Results

Results of the validation are reported next. First, we discuss the findings of the univariate distribution comparisons. Later, we reflect on the subjective evaluation.

*4.2.1 Comparing Univariate Distributions.* Upon analyzing the distributions of variables in both the synthetic and reference datasets, we found that, overall, the distributions are fairly similar, although we do not aim for identical distributions, as this could signal potential privacy concerns.

As shown in Figure 3, the job sector distribution in the synthetic dataset is moderately close to the reference dataset, largely preserving the broad ordering of the sectors (i.e., most common to least common). Likewise, the distributions of years of professional experience in both datasets show to be pretty equivalent, as illustrated in Figure 4, suggesting a high level of data utility.

Regarding demographic variables, the analysis reveals that the gender distributions are identical in shape and closely balanced in terms of proportions. As shown in Figure 5, the categories *woman* and *man* are nearly evenly distributed across both datasets. When examining age distributions, we observe that, as in previous cases, the overall ordering is well preserved. The distribution of categories for LGBTQ+, ethnic minority status, and perceived foreignness in the reference and synthetic datasets exhibits a nearly identical pattern in which the general ordering is conserved, as shown in Figure 7. The ratio between the majority and minority classes is not always aligned because limitations in the reference dataset force the generator to under-/over represent them.

To complement the visual inspection of distributions across variables, we analytically explored how distributions in the reference dataset differ from the synthetic ones. The difference between these distributions is calculated through the Jensen-Shannon divergence technique, which account for missing categories as in some cases (e.g., job sector, gender, age) and provides a numerical score between 0 and 1 representing how similar the distributions are from one another. The closer the score closer to 0, the perfect align between the distributions.

Table 9 reports the Jensen-Shannon scores obtained for the distributions of the reference and synthetic datasets across variables. As shown, the scores are generally close to 0, with distribution similarities ranging from almost perfect—such as for *gender* and *minority*—to slightly different, as in *LGBTQ+* and *years of professional experience*.

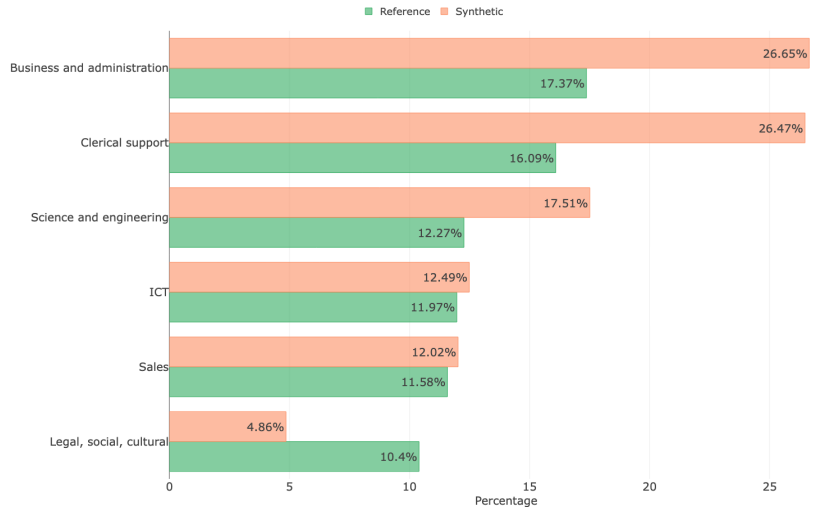



Fig. 3. Comparison of job sector distributions in the reference and synthetic datasets. Values in the reference dataset do not add up to 100% since sectors not included in the synthetic data were excluded.

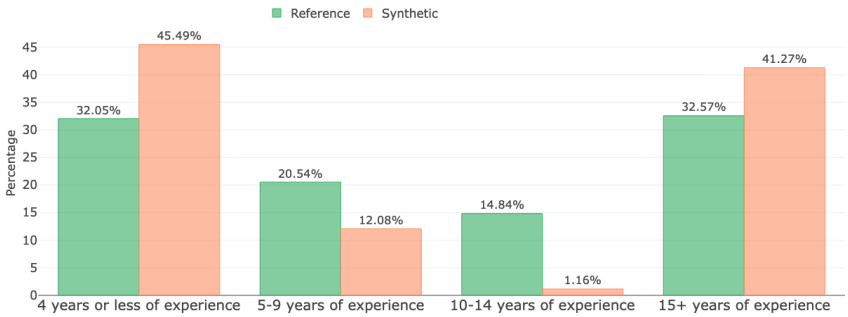



Fig. 4. Comparison of years of professional experience distributions in the reference and synthetic datasets.

Table 9. Jensen-Shannon (JS) scores for distribution comparison across variables.

| | Job sector | Professional Experience | Gender | Age | LGBTQ+ | Minority | Foreigness |
|---|---|---|---|---|---|---|---|
| JS score | 0.12 | 0.22 | 0.01 | 0.14 | 0.25 | 0.04 | 0.20 |

**Note.** The closer the score to 0, the similar the distributions.

In summary, our analysis reveals similar patterns between the reference and generated distributions across variables. These findings suggest that the synthetic CVs are likely to be as useful as the real documents for studying bias in algorithmic hiring, while not revealing personal information. As previously discussed, we understand that the collected CVs possess the necessary characteristics for this purpose.

*4.2.2 Subjective Evaluation.* In total, 300 crowdsourcing workers participated in the subjective evaluation of the synthetic CVs, each conducting a single run of the experiment, i.e., no worker

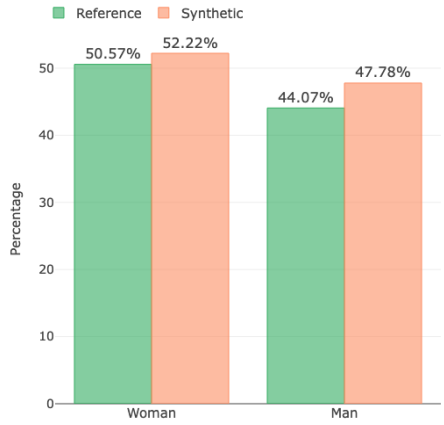



Fig. 5. Comparison of gender distributions in the reference and synthetic datasets. Percentages in the reference do not up to 100% since categories available in the reference dataset but no included in the synthetic data (e.g., non-binary) are excluded.

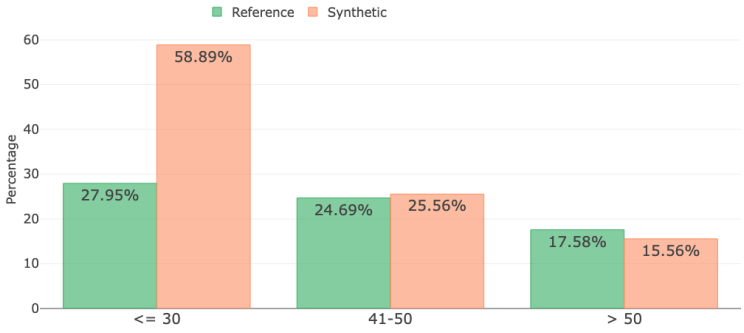
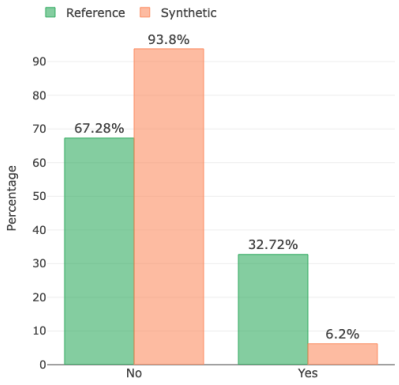


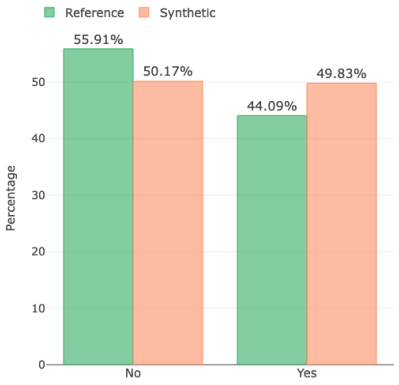

Fig. 6. Comparison of age distributions in the reference and synthetic datasets. Only categories available in both datasets are included.

repeated the study. All participants declared fluency in English, half identified themselves as women, and the large majority (70%) aged 20-40 years.

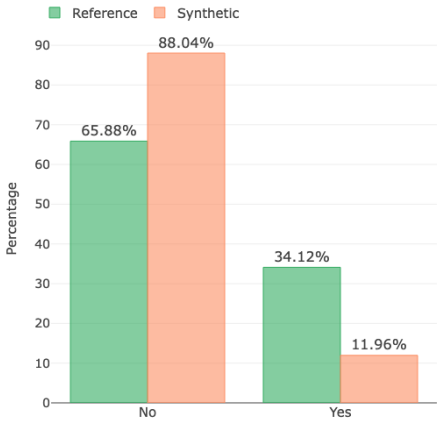
Of the 300 submissions, 14 were discarded due to low-quality answers or exceptionally high-speed execution (i.e., less than 180 seconds when the median completion time is almost 7 minutes). On average, real and synthetic CVs were evaluated 2.86 times. The participants' general accuracy in correctly identifying whether a CV was real or synthetic was 53%. The probability of correctly classifying a real CV was 0.52, while the probability of a synthetic CV being mistaken for a real one was 0.55. Moreover, almost 90% of the synthetic CVs (442 out of 500) were classified as real for at least one of the three evaluators, and more than half of them (260 out of 500) for two evaluators.




(a) LGBTQ+ status distributions in the reference and synthetic datasets.




(b) Ethnic minority distributions in the reference and synthetic datasets.




(c) Perceived foreignness distributions in the reference and synthetic datasets.

Fig. 7. Comparison of demographic attribute distributions in the reference and synthetic datasets

## 5 Discussion

The quantitative results reveal that distributions across variables in both datasets (reference and synthetic) show similar patterns, suggesting high utility of the synthetic CVs to study bias and benchmark mitigation techniques in algorithmic hiring.

Additionally, the results of the qualitative analysis of the synthetic CVs highlight their quality, as they appeared convincingly real more than half the time, demonstrating that they are not easily distinguishable as artificially created. Moreover, 90% of them were perceived as belonging to a real person by at least one of the evaluators, reinforcing their realistic appearance.

On the other hand, we can see that the synthetic dataset is in some way skewed towards sectors that are usually dominated by professionals with, at minimum, college studies (e.g., *business and administration*, *ICT*, *science and engineering*). Sectors—available in donations—whose workers learn occupation skills primarily from practical experiences rather than formal education (e.g., *construction and manufacturing*, *transport*, *personal service*), are not represented in the synthetic dataset due to their low representativeness in the reference dataset, which avoids the generator to apply specific combination of parameters that do not satisfy the requirements of the protocol set in place to safeguard donors' identities.

The same applies to some demographic characteristics missing in the synthetic dataset, which are a result of insufficient data in the reference dataset, such that the generator could operate while fulfilling the protection measures proposed to safeguard donors' identities. Furthermore, generating CVs with combinations involving more than one demographic parameter proved not plausible for the same reason.

Beyond these characteristics, our dataset shares certain similarities with, but also presents significant differences from, the collections of CVs described in Section 2 (see Table 1 for a summary). In terms of similarities, it includes synthetically generated CVs, as in [4, 44, 52], and, like [1, 4, 11, 52], it spans multiple professional sectors. Its content also resembles that of most reviewed datasets, with the exception of [11, 25, 52].

As for the differences, our dataset is, among the reviewed literature, the only synthetic collection created through a process explicitly designed to safeguard personal information and grounded in real CVs gathered specifically for this purpose. By contrast, previous efforts (e.g., [1, 11, 52]) have largely relied on web-scraped content. Moreover, our dataset introduces sensitive demographic attributes not previously available to researchers (e.g., disability condition, LGBTQ+ status). Finally, unlike earlier synthetic or semi-synthetic datasets (e.g., [4, 46, 47]), every generated CV underwent manual review. Apart from [1, 13], which are publicly available, our dataset is accessible for researchers in established academic institutions within the European Union, i.e., they are eligible to receive research funding from the European Commission.

### 5.1 Limitations and Future Work

The generation approach and the synthetic dataset have limitations.Next, we present details of these limitations and propose alternatives to address them in future work.

*5.1.1 Generation Approach.* Even when donations in formats different from PDF (i.e., DOC/DOCX, ODT) were received, the generator did not process them, and they were ignored. Additionally, due to limitations in the document parser, CV documents with more than 10 pages are not considered. These limitations prevented us to use the entire pool of donated CVs. Also, donations that do not include attached CV documents but the information about the donor's educational background, professional experiences, and skills are shared in the donation form, are omitted.

CVs not written in English are automatically translated into English with the risk of losing the nuances and particularities of the original language. The generator expects to operate on a fixed number and type of data fields, and it is compatible with a pre-defined structure of parsing output. Moreover, although it is built on free and open-source libraries and software, it depends on a third-party paid parsing service. Albeit replacing text processing dependencies, like the parser, does not represent a significant change, and can be done relatively easily.

As mentioned in Section 3, hand-made rules have been included in the generator to guarantee certain quality and consistency in the results. Through the various iterations of the CV generator, rules have been modified and added to improve the generation quality. However, we acknowledge that producing an exhaustive set of rules that covers all possible situations is unrealistic; therefore,

unexpected inconsistencies in the content of the synthetic CVs might be present. To overcome this challenge and ensure the quality of the final dataset, a manual validation step has been included in the pipeline.

Additionally, we aim to extend the generator to include a module that looks for plausible substitutes for academic institutions and workplaces. So, instead of using the academic institutions or workplaces originally found in the donated CVs, a plausible substitute will be identified and employed in the creation of the synthetic CV. For example, instead of using the University of Pisa in an education item of a synthetic CV (assuming that this university was included in some of the donated CVs), we could use the University of Trento, as those institutions come from the same country and are similar in terms of size and education offers. This feature will further enhance the measures adopted to protect donors' identities.

*5.1.2 Synthetic dataset.* As we saw, the generator reproduces the distribution of the collected data. Hence, if it operates on a biased source dataset those biases will be replicated in the synthetic dataset as no artificial corrective actions (e.g., up-sampling) have been performed to balance the seed dataset. Similarly, it was impossible to generate CVs for sectors without representation in the donated data. To address this issue, we propose to create CVs for some of the underrepresented sectors. In particular, we have already performed online research and evaluated the usual characteristics of CVs from the *education*, *tourism* and *arts & crafts* sectors, and plan to manually generate 100 artificial CVs for each of these sectors.

## 6 Conclusions

According to the AI Act, AI systems intended to be used for the recruitment or selection of natural persons, in particular to place targeted job advertisements, to analyze and filter job applications, and to evaluate candidates, are considered high-risk, as these systems could significantly affect individuals' fundamental rights. Therefore, the development of anti-discriminatory methods and algorithms contextualized within the actual technical, legal, and ethical context of algorithmic hiring is of utmost importance.

To advance research in this field, given the current lack of datasets reflecting the diverse characteristics found in real CVs of the actual workforce, in this work we present a dataset of 1,730 synthetic CVs representing six work sectors, along with the approach designed to generate them.

The developed dataset follows the AI Act recommendations on the use of synthetic data as a primary option to avoid processing sensitive personal data, when aiming to detect and mitigate bias [7]. Its envisioned use is as a benchmark during the pre-deployment phase of candidate ranking systems developed with relevance as a goal, to detect whether those systems result in discriminatory outcomes, based on attributes such as age, gender, or nationality, disadvantaging systematically some groups of applicants.

The dataset was developed using real CVs, collected through a data donation campaign, and found to follow the features' distribution in the donated data. Its delivery and use will permit an increased research reproducibility, while the proposed generation technique supports an increased explainability of the outcomes. Therefore, this work is expected to have an important impact on both the industrial and academic sectors.

Finally, this work is expected to be a significant contribution to current standardization efforts, especially in the direction of *Data Governance and Quality*. We envision that the presented synthetic dataset can become a standard for fairness benchmarking of algorithmic hiring systems, especially when implementing AI systems for candidate ranking based on CVs, as our dataset enables the comparison of multiple ranking algorithms, to ensure they do not introduce or reproduce discriminatory behaviors based on candidates' protected characteristics. In addition, in cases where the

specific dataset cannot be directly used, for instance, when a different working force population is evaluated, the proposed methodology allows the generation of an adequate benchmarking dataset.

## Acknowledgments

This work was partially supported by the project FINDHR (Horizon Europe research and innovation program, grant no.: 101070212). The authors would like to thank Marta Gracia, especially for her collaboration in processing the donated CVs and Tamara Vorobyeva for the careful evaluation of the generated CVs.

## References

[1] [n. d.]. Kaggle "Resume dataset". https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset. Accessed: 2025-04-04.

[2] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American economic review 94, 4 (2004), 991–1013.

[3] Fadi Boutros, Vitomir Struc, Julian Fierrez, and Naser Damer. 2023. Synthetic data for face recognition: Current state and future prospects. Image and Vision Computing 135 (2023), 104688. https://doi.org/10.1016/j.imavis.2023.104688

[4] Andrea Bruera, Francesco Alda, and Francesco Di Cerbo. 2022. Generating realistic synthetic curricula vitae for machine learning applications under differential privacy. In Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference. 53–63.

[5] Sarah Chander. 2017. Women of Color in the Workplace. European Network Against Racism.

[6] Zhisheng Chen. 2023. Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanities and social sciences communications 10, 1 (2023), 1–12.

[7] EP Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008,(EU) No 167/2013,(EU) No 168/2013,(EU) 2018/858,(EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU,(EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) Off. J. Eur. Union 50 (2024), 202.

[8] Bo Cowgill. 2018. Bias and productivity in humans and algorithms: Theory and evidence from resume screening. Columbia Business School, Columbia University 29 (2018).

[9] Jessamyn Dahmen and Diane Cook. 2019. SynSys: A synthetic data generation system for healthcare applications. Sensors 19, 5 (2019), 1181.

[10] Jeffrey Dastin. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of data and analytics. Auerbach Publications, 296–299.

[11] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In proceedings of the Conference on Fairness, Accountability, and Transparency. 120–128.

[12] Jens-Joris Decorte, Jeroen Van Hautte, Thomas Demeester, and Chris Develder. 2021. Jobbert: Understanding job titles through skills. arXiv preprint arXiv:2109.09605 (2021).

[13] Nazarii Drushchak and Mariana Romanyshyn. 2024. Introducing the Djinni Recruitment Dataset: A Corpus of Anonymized CVs and Job Postings. In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, Mariana Romanyshyn, Nataliia Romanyshyn, Andrii Hlybovets, and Oleksii Ignatenko (Eds.). ELRA and ICCL, Torino, Italia, 8–13. https://aclanthology.org/2024.unlp-1.2/

[14] Peter Eigenschink, Thomas Reutterer, Stefan Vamosi, Ralf Vamosi, Chang Sun, and Klaudius Kalcher. 2023. Deep generative models for synthetic data: A survey. IEEE Access 11 (2023), 47304–47320.

[15] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. 2020. Practical synthetic data generation: balancing privacy and the broad availability of data. O'Reilly Media.

[16] Markus Endres, Asha Mannarapotta Venugopal, and Tung Son Tran. 2022. Synthetic Data Generation: A Comparative Study. In Proceedings of the 26th International Database Engineered Applications Symposium (Budapest, Hungary) (IDEAS '22). Association for Computing Machinery, New York, NY, USA, 94–102. https://doi.org/10.1145/3548785.3548793

[17] European Commission. 2019. Final Report of the High-Level Expert Group on the Impact of the Digital Transformation on EU Labour Markets. European Commission website, https://digital-strategy.ec.europa.eu/en/news/final-report-high-level-expert-group-impact-digital-transformation-eu-labour-markets. Issued by the High-Level Expert Group created in May 2018, with recommendations for policy actions to shape a smooth, inclusive, and human-centric digital transformation of EU labour markets.

[18] European Institute for Gender Equality. 2021. Covid-19 derails gender equality gains. https://eige.europa.eu/newsroom/news/covid-19-derails-gender-equality-gains.
[19] Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. 2023. Fairness and Bias in Algorithmic Hiring. arXiv:2309.13933 [cs.CY]
[20] Joseph B Fuller, Manjari Raman, Eva Sage-Gavin, Kristen Hines, et al. 2021. Hidden workers: Untapped talent. Harvard Business School Project on Managing the Future of Work and Accenture 1 (2021).
[21] High-Level Expert Group et al. 2018. High-Level Expert Group on the Impact of the Digital Transformation on EU Labour Markets. Digital Single Market (2018).
[22] Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. 2024. Synthetic data in AI: Challenges, applications, and ethical implications. arXiv preprint arXiv:2401.01629 (2024).
[23] Christopher Harris. 2023. Mitigating age biases in resume screening AI models. In The International FLAIRS Conference Proceedings, Vol. 36.
[24] Bill Howe, Julia Stoyanovich, Haoyue Ping, Bernease Herman, and Matt Gee. 2017. Synthetic data for social good. arXiv preprint arXiv:1710.08874 (2017).
[25] Kameni Florentin Flambeau Jiechieu and Norbert Tsopze. 2021. Skills prediction based on multi-label resume classification using CNN with model predictions explanation. Neural Computing and Applications 33, 10 (2021), 5069–5087.
[26] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. 2022. Synthetic Data–what, why and how? arXiv preprint arXiv:2205.03257 (2022).
[27] Indu Joshi, Marcel Grimmer, Christian Rathgeb, Christoph Busch, Francois Bremond, and Antitza Dantcheva. 2024. Synthetic data in human analysis: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 7 (2024), 4957–4976.
[28] Ioannis Kaloskampis. 2019. Synthetic data for public good. https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/
[29] Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan. 2025. Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline. arXiv preprint arXiv:2501.18493 (2025).
[30] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT, Vol. 1. Minneapolis, Minnesota, 2.
[31] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. 2019. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 0–0.
[32] Siyka Kovacheva, Boris Popivanov, Radka Peeva, D Coletto, I Dimitriadis, G Fullin, M Fischer-Souan, and O Marcovici. 2018. Growth, Equal Opportunities, Migration and Markets (GEMM) Report on institutional and contextual factors.
[33] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 166–176.
[34] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data for language models. arXiv preprint arXiv:2404.07503 (2024).
[35] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2023. Machine learning for synthetic data generation: a review. arXiv preprint arXiv:2302.04062 (2023).
[36] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. 2020. SensitiveNets: Learning agnostic representations with application to face images. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 6 (2020), 2158–2164.
[37] Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. 2023. Synthetic data generation: State of the art in health care domain. Computer Science Review 48 (2023), 100546. https://doi.org/10.1016/j.cosrev.2023.100546
[38] Frank Nielsen. 2019. On the Jensen–Shannon symmetrization of distances relying on abstract means. Entropy 21, 5 (2019), 485.
[39] Beata Nowok, Gillian M Raab, and Chris Dibben. 2016. synthpop: Bespoke creation of synthetic data in R. Journal of statistical software 74 (2016), 1–26.
[40] Steven Ovadia. 2014. Markdown for librarians and academics. Behavioral & social sciences librarian 33, 2 (2014), 120–124.
[41] Prasanna Parasurama, João Sedoc, and Anindya Ghose. 2022. Gendered Information in Resumes and Hiring Bias: A Predictive Modeling Approach. Available at SSRN 4074976 (2022).
[42] Goran Paulin and Marina Ivasic-Kos. 2023. Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. Artificial intelligence review 56, 9 (2023), 9221–9265.

[43] Alejandro Peña, Ignacio Serna, Aythami Morales, Julian Fierrez, Alfonso Ortega, Ainhoa Herrarte, Manuel Alcantara, and Javier Ortega-Garcia. 2023. Human-centric multimodal machine learning: Recent advances and testbed on AI-based recruitment. SN Computer Science 4, 5 (2023), 434.

[44] Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fierrez. 2020. Bias in multimodal AI: Testbed for fair automatic recruitment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 28–29.

[45] Vamsi K Potluru, Daniel Borrajo, Andrea Coletta, Niccolò Dalmasso, Yousef El-Laham, Elizabeth Fons, Mohsen Ghassemi, Sriram Gopalakrishnan, Vikesh Gosai, Eleonora Kreačić, et al. 2023. Synthetic data applications in finance. arXiv preprint arXiv:2401.00081 (2023).

[46] RecSys Challenge 2016. 2016. Dataset page, ACM RecSys Challenge 2016. GitHub repository. https://github.com/recsyschallenge/2016/blob/master/TrainingDataset.md Training dataset specification for the ACM RecSys Challenge 2016.

[47] RecSys Challenge 2017. 2017. Dataset page, ACM RecSys Challenge 2017. Webpage, RecSys Challenge website. https://www.recsyschallenge.com/2017/#dataset Details on the dataset for ACM RecSys Challenge 2017 — job recommendation task by XING in a cold-start scenario.

[48] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084 (2019).

[49] Horst Rinne. 2008. The Weibull distribution: a handbook. Chapman and Hall/CRC.

[50] Miguel Rujas, Rodrigo Martín Gómez del Moral Herranz, Giuseppe Fico, and Beatriz Merino-Barbancho. 2025. Synthetic data generation in healthcare: A scoping review of reviews on domains, motivations, and future applications. International Journal of Medical Informatics 195 (2025), 105763. https://doi.org/10.1016/j.ijmedinf.2024.105763

[51] Jorge Saldivar, Jessica Wulf, and Carlos Castillo. 2025. Data Donation: A First Step Towards Improving Representativeness in Algorithmic Hiring Datasets. Available at SSRN 5160523 (2025).

[52] Panagiotis Skondras, Panagiotis Zervas, and Giannis Tzimas. 2023. Generating Synthetic Resume Data with Large Language Models for Enhanced Job Description Classification. Future Internet 15, 11 (2023). https://doi.org/10.3390/fi15110363

[53] Daniel Susser and Jeremy Seeman. 2024. Critical Provocations for Synthetic Data. Surveillance & Society 22, 4 (2024), 453–459.

[54] Kamal Taha. 2023. Semi-supervised and un-supervised clustering: A review and experimental evaluation. Information Systems 114 (2023), 102178.

[55] Nicholas Tilmes. 2022. Disability, fairness, and algorithmic bias in AI recruitment. Ethics and Information Technology 24, 2 (2022), 21.

[56] Pengda Wang, Andrew C Loignon, Sirish Shrestha, George C Banks, and Frederick L Oswald. [n. d.]. Leveraging Synthetic Data to Advance Organizational Science. ([n. d.]).

[57] Cedric Deslandes Whitney and Justin Norman. 2024. Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. In The 2024 ACM Conference on Fairness, Accountability, and Transparency. 1733–1744.

[58] Josephine Yam and Joshua August Skorburg. 2021. From human resources to human rights: Impact assessments for hiring algorithms. Ethics and Information Technology 23, 4 (2021), 611–623.

[59] Sijing Zhang, Ping Li, and Ziyan Cai. 2022. Are Male Candidates Better than Females? Debiasing BERT Resume Retrieval System. In 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 616–621.