
| | |
|---|---|
| **Document Title** | **Literature Review** |
| **Project Title and acronym** | DEtecting Stereotypes in human ComputAtioN Tasks (DESCANT) |
| **Pillar** | II. Sustainable RTDI System |
| **Programme** | Excellence Hubs |
| **Grant Agreement** | EXCELLENCE/0918/0086 |
| **Deliverable No.** | D3.1 |
| **Work package No.** | WP3 |
| **Work package title** | Conceptual Framework |
| **Authors (Name and Partner Institution)** | E. Christoforou (CYENS) |
| **Contributors (Name and Partner Institution)** | J. Otterbacher (CYENS)<br>A. Lanitis (CYENS)<br>G. Demartini (UQ) |
| **Reviewers** | S. Kleanthous (OUC) |
| **Status (D: draft; RD: revised draft; F: final)** | F |
| **File Name** | D3.1_LiteratureReview.docx |
| **Date** | 30 May 2020 |

ΙΔΡΥΜΑ ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ
This project has received funding from Research and Innovation Foundation (RIF) under grant agreement EXCELLENCE/0918/0086.

| **Draft Versions - History of Document** | | | | |
|---|---|---|---|---|
| **Version** | **Date** | **Authors / contributors** | **e-mail address** | **Notes / changes** |
| 1.0 | 20/1/20 | E. Christoforou | e.christoforou@cyens.org.cy | initial document |
| 1.2 | 16/5/20 | E. Christoforou | e.christoforou@cyens.org.cy | review version |
| 1.3 | 30/5/20 | E. Christoforou | e.christoforou@cyens.org.cy | final version |

**Abstract**

This deliverable provides a comprehensive review of the state-of-the-art regarding social biases emerging through intelligent systems in general. The deliverable documents in a methodical way the impact of stereotypes and biases, which are created during the crowdsourcing process in the datasets developed for intelligent systems. Through this deliverable the different parameters of the crowdsourcing process involved in the creation or propagation of stereotypes and biases are identified. In this respect, this deliverable relates to deliverable D3.2, the Conceptual Framework Description.

| **Keyword(s):** | Literature Review, Bias, Stereotypes, Intelligent Systems, Artificial Intelligence (AI), Machine Learning (ML), crowdsourcing, datasets. |
|---|---|

# Contents

# 1. About this Deliverable

This deliverable reviews a number of academic and scientific resources from the various disciplines involved in the project, thus paving the way and setting the (theoretical) foundations for the development of each WP in the project, and in particular, WP4. All the scientific resources gathered throughout the bibliographical study that took place during WP3 are collected and presented in the Zotero reference management system:
https://www.zotero.org/groups/2441143/descant/library/

In this deliverable, we lay down the background for the DESCANT project, reviewing initially what motivated this work. Furthermore, we review all the components upon which this work is based on. We give particular emphasis to the crowdsourcing component, reviewing the works referring to the possible parameters affecting the creation and propagation of biases through the crowdsourcing process.

# 2. Background of the Study

## 2.1 Motivation

There is a strong tendency for people to believe that decision-making facilitated by algorithms is relatively objective -- free from the social bias that often plagues human decision-making. For instance, it has been suggested that algorithmic decision-making in hiring can help to promote diversity in the workplace, mitigating the subconscious biases of human decision-makers (Houser 2019). The justice system is another context in which algorithmic decision-making is increasingly used. In particular, algorithmic support has been applied to decisions on whether or not a defendant can make bail or will be detained (Kleinberg, et al. 2018). In this setting, a machine-learned algorithm can make use of a huge volume of historical data to make the optimal predictions of defendant outcomes based on a variety of factors. Judges' unassisted decisions, on the other hand, may be inadvertently sensitive to particular factors such as the nature of the crime and/or the defendant's race.

Even in “everyday" matters (e.g., online shopping or news consumption) algorithmic judgements are delegated a good deal of authority to take decisions on our behalf, operating in a largely autonomous fashion and without human oversight (Willson 2017). Scholars have long noted that once such mediating technologies have worked their way into our routines, they tend to become a “transparent” part of our lives; we often forget that they are there (Van Den Eede 2011, Verbeek 2005). Another issue is that algorithms are seen as not having agency, being technical artifacts (Klinger and Svensson 2018). Thus, as Artificial Intelligence (AI) increasingly takes over traditional processes and everyday tasks, we may become less likely to challenge their outputs. However, it is important to recognize that human judgement and agency are still a part of the equation, at least in the form of *data*.

Acknowledging that AI and consequently "intelligent systems" are not free from human bias is the only way to guarantee that progress against algorithmic social bias – including discrimination and unfair treatment – will continue to happen.

Fortunately, researchers and practitioners are increasingly questioning the perception of objectivity in AI applications and systems. Different forms of algorithmic bias have been documented in applications used across domains of society, including the fields of education (Boratto, Fenu and Marras 2019), health (Kadija and Pitcan 2018), social services (Chouldechova, et al. 2018), and, as previously mentioned, the justice system (Dressel and Farid 2018). One of the most widely-discussed examples of algorithmic racial bias is found in the U.S. justice system, where the COMPAS risk assessment tool is extensively used during sentencing. In 2016, a group of data journalists demonstrated the system's tendency to discriminate against African-American defendants, exhibiting nearly twice the error rate in recidivism prediction for this group, as compared to their white counterparts (Angwin, et al. 2016).

Beyond sentencing, predictive policing -- a set of techniques used to identify potential criminal activity -- is another area in which algorithmic bias can result in significant harm (Richardson, Schultz and Crawford 2019). These systems are usually composed of various algorithms spanning in the fields of computer vision, natural language processing (NLP), speech recognition and expert systems. Consider for example a system used for predictive policing employing some face recognition algorithm and thus spanning in the field of computer vision. Recently it was made apparent that face recognition algorithms can exhibit a discriminatory behavior towards certain races due to racial characteristics such as the tone of the skin (Buolamwini and Gebru 2018). Face recognition algorithms employing machine learning techniques are trained based on image data categorized manually either in house or through a crowdsourcing process (Deng, et al. 2009). Data such as this one, originating from human annotators suffer from stereotypes and efforts to make them more balanced and consequently fair are in action (Kaiyu, et al. 2020).

## 2.2 Algorithmic stereotyping and crowdsourcing

Even in more "everyday" contexts, algorithmic bias can affect multiple aspects of a person's life, online and offline; extending from advertising (Speicher, et al. 2018), to hiring (Sánchez-Monedero, Dencik and Edwards 2020, Raghavan, et al. 2020), even to dating (Hutson, et al. 2018).

Furthermore, AI systems aiming at exhibiting human-like behaviors tend to perpetuate social stereotypes. On the one hand, this is arguably expected, given the important role that social stereotypes play in human cognition (e.g., serving as heuristics to judge others when ample information is not available (Bodenhausen 1993)). However, in an AI, this can have the effect of sustaining and sometimes amplifying inequalities. For example, AI virtual assistants, like Siri and Alexa have a woman's name, voice, and personality (to some extent) as a way to give users a more pleasant experience according to opinion polling studies. Similarly, a UNESCO white paper report contrasted the "unmistakably male" voice of the antagonist *HAL in 2001: A Space Odyssey*, to the pleasant female voices of modern digital assistants, questioning the proliferation

of feminization of their speech and other features (pp. 89--92) (West, Kraut and Chew. 2019). It is noteworthy how this gender stereotype, that women are ideal assistants, found its way into these AI systems and from there, back to society enriching a biased belief and perpetuating inequalities.

Artificial intelligence (AI) systems are used to achieve specific goals and they are characterized by their ability to adapt by quickly and correctly interpreting external data and learning from such data. An AI system can demonstrate biased behavior by systematically discriminating against a social group that is characterized by a distinguishable characteristic such as race, gender, age, economic status among others, leading to an unfair or harmful outcome. Furthermore, an AI system can perpetuate social stereotypes amplifying inequalities aiding at the prolong of unfair treatment of social groups in the society.

*Tracing the reasons* for these biased behaviors by AI systems is an active topic of discussion and it has its root in the technical characteristics of some of the AI algorithms, like optimization functions and criteria choices; however, for many AI systems the *data* upon which these algorithms are trained is fundamental. A vast amount of these data represent our perceptions of the world, the society and in general, our surrounding environment, thus they are inevitably biased. It is understood that many of these data arrive from expert communities or "big data" (typically collected from Web and social media) but when classification is the main objective of the algorithm, then these sources of data many times can fail to comply (Deng, et al. 2009). Expert communities might fail to provide the required amount of data or the necessary heterogeneity, while big data are collected in such a fashion that they might be missing many of the required attributes for the process of classification. A straightforward solution for obtaining a massive amount of data capturing a human's intelligence is the act of *crowdsourcing* (Howe 2006).

Crowdsourcing in a systematic, on-demand fashion it is of great importance to developers as they can obtain easy access to human intelligence. In this study, we are interested in understanding how the stereotypes and prejudices of the humans involved in the crowdsourcing process find their way into AI algorithms, resulting in unfair treatment by these systems. An essential element of crowdsourcing is that it takes place over the Internet; thus, harnessing the power of the crowd happens either via dedicated platforms[1,2,3,4] or through social media sites (Gao, Barbier and Goolsby 2011, Sun, Chen and Viboud 2020). In this study, we will focus on crowdsourcing that is facilitated through dedicated platforms where a *requesting entity* (in this case, the developer/owner of the AI system) publishes a task on the platform, which acts as a mediator between the requester and a crowd of humans willing to execute the task. These sets of humans are usually referred to as *workers* when they receive a financial reward for their work while when they provide their services on a volunteer basis, they are usually referred to as volunteer workers. In the case of crowdwork for pay, the task posted by the requester often takes the form of a specific and well-defined task such as image tagging, image annotation, text annotation and so on. By the end of the task campaign, the requester is left with a set of responses from the workers that will serve as training data.

[1] https://www.mturk.com/
[2] https://appen.com/
[3] https://www.worldcommunitygrid.org/discover.action
[4] https://www.zooniverse.org/projects/zookeeper/galaxy-zoo

### 2.3 Structure and Methodology of the Study

Considering the background of the study described above and the goals of the DESCANT project, our study will focus on documenting the sources of bias in the produced crowdsourced data and the way stereotypes are created and propagated through crowdsourced data. As a first step, we lay down the framework of the study reviewing the general setting of crowdsourcing and formally define the concepts of bias and stereotypes after surveying the literature.

Furthermore, we outline the connection of crowdsourcing to intelligent systems and we briefly review how bias is propagated from the collected data to the final output of the intelligent system, possibly causing discriminatory phenomena. Once all the parameters of our framework are laid down, we use those concepts as a pillar in our search for each of the identified sources of bias in the crowdsourcing process.

## 3. Framework

### 3.1 Crowdsourcing and Human Intelligence

The term *crowdsourcing* was first introduced by Jeff Howe, writing in *Wired* Magazine in 2006 (Howe 2006). Compared to outsourcing, which involves the delegation of work to a specific individual or group, crowdsourcing involves harnessing the resources of a more public, loosely defined group of participants. As noted by Howe, while "offline" crowdsourcing existed before the digital age, today the term is primarily associated with the use of Internet technologies to facilitate the process. Following the definition of Estellés-Arolas and González-Ladrón-de-Guevara (Estellés-Arolas and González-Ladrón-de-Guevara 2012) *crowdsourcing* refers to a "*type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task.[...] The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken*."

The above definition encompasses many forms of crowdsourcing; for example, contributing to online communities like Wikipedia, producing physical work like TaskRabbit, or developing software in an open fashion involving an undefined group (Stol and Fitzgerald 2014). In contrast, the current work focuses on a specific category, *micro-task crowdsourcing*, that can harvest an enormous amount of data produced by human, intelligent workers on demand by connecting to platforms like Amazon Mechanical Turk, Appen and Microworkers.

We will formally refer to the *micro-task crowdsourcing process*, which consists of the following components: (1) the crowd, that is the workers, (2) the micro-task, (3) the crowdsourcing

platform, (4) the requester of the micro-task and (5) the crowdsourced data harvested and built from the workers' replies to the micro-task.

## 3.2 Bias, Stereotypes and Discrimination

The goal of this study is to explore how the micro-task crowdsourcing process may contribute to inflicting an unfair, discriminatory and/or morally wrong decision or behavior on humans on the receiving end of the output of an AI algorithm. Data-driven AI systems usually deploy machine learning algorithms that can exhibit bias and propagate stereotypes. To this end, like in (Friedman and Nissenbaum 1996), we use the term bias in an AI system to describe a systematic treatment of a person or a group that differs from the norm / the treatment that others receive. Furthermore, AI systems are considered biased when the above mentioned systematic discrimination is accompanied by an unfair/harmful outcome. Regarding propagation of stereotypes by AI, we consider the systematic promotion of stereotypes that can perpetuate harmful discriminatory behaviors.

The core of the crowdsourcing process are the worker participants. When these workers are asked to give a subjective judgement as part of the micro-task assignment, their beliefs can end up affecting the final outcome of an AI system and provoke a discriminatory behavior. Workers might suffer from cognitive bias (Eickhoff 2018) , creating their own “subjective social reality,” systematically deviating from the norm in judgement. For example, a set of workers that is asked to judge whether the human depicted in the photo might be suffering from a health condition, could possibly be negatively predisposed (consciously or unconsciously) against sumo athletes.

Additionally, crowd workers may suffer from stereotypical beliefs about the characteristics, the attributes and the behaviors present in a certain group of people (Hilton and Von Hippel 1996). Stereotypes serve as a way to provide cohesion among a group of people and separate them from other groups. They are used as a heuristics for a fast and effortless opinion about people belonging to a particular group, while they also provide a way for identifying our standing compared to other people (McGarty, Yzerbyt and Spears 2002). It thus becomes apparent that when a requester taps into a large pool of workers from around the globe, it is hard -- if not impossible -- to identify the workers’ stereotypical beliefs that can span from gender to ethnicity, sexual orientation and religion, to less apparent stereotypes that are regional or that are based on a person's socio-economic status, profession or educational background.

It is clear that there is a correlation between the human bias introduced by the crowdworker in the data creation process, and the potential bias that an AI system developed using that data can exhibit. This survey will try to address how and why bias is introduced into the AI system despite some efforts to mitigate it and highlight the paths that research has to take in order to develop better strategies at resolving this issue from different stakeholders' perspectives.

## 3.3 Crowdsourcing and AI-based systems

As we mentioned earlier, for AI systems, micro-task crowdsourcing is a primary source for obtaining valuable data. These data can be used by AI systems in a *static fashion,* as training data for a machine learning application (Vaughan 2017, Kovashka, et al. 2016). For example, crowdworkers can annotate images or classify images that are later used by computer vision algorithms (Deng, et al. 2009, Kaiyu, et al. 2020) , or label data for natural language processing purposes (Callison-Burch and Dredze. 2010). Data obtained through crowdsourcing are also being used in an online, on-demand *dynamic fashion* by some systems (Russakovsky, Li and Fei-Fei. 2015, Laws, Scheible and Schutze 2011), while crowdsourcing can even happen in a *real-time fashion* (Abbas, et al. 2020, Lasecki, Homan and Bigham 2015) by AI systems.

Other sources of data for AI-based systems are public datasets[5] (Deng, et al. 2009), proprietary data obtained and curated strictly in house by the institution deploying the AI system or can be extracted on demand online from social networks (Kumar, Morstatter and Liu. 2014). An AI system collects and curates these data, that might be arriving from different sources, and uses them to train or test the developed algorithms. Due to the human biases that are carried downstream and are usually mitigated after the collection or during the algorithmic process, the final system might still exhibit a discriminatory behavior (favorable or unfavorable) towards certain social groups, as is shown in Figure 1.

One of the main advantages of crowdsourcing is the relatively effortless creation of the huge amounts of data necessary to train machine learning algorithms. These crowdsourced data often embed the subjective beliefs of the crowd, e.g., portraying the existing, prevalent social stereotypes at a local and/or global level. The goal of this study is to outline the path leading to biased behavior in AI systems from the human input all the way up to the impact that an application of an AI system has on humans and the society in general, as shown in Figure 1.

While the recent survey by Olteanu and colleagues (Olteanu, et al. 2019) provides an overview of the sources and effects of social data biases (e.g., social media sources, which are oftentimes used in creating training data), our intent is to zone in on how the *data augmentation processes* via crowdsourcing, may introduce social biases. We will look in-depth at the elements of crowdsourcing that produce and promote biases and how the crowdsourced data are used by AI systems. In other words, we focus on data as the source of bias; it is out of the scope of this work to study how the learning algorithm(s) can actually inflict additional biases.

Machine learning algorithms (especially the ones focused on classification) and crowdsourcing have a teacher-student dynamic. As is mentioned in (Domingos 2012): " A classifier is a system that inputs (typically) a vector of discrete and/or continuous feature values and outputs a single discrete value, the class." Learning comes as a result of the way the classifier is represented by an algorithm, the way it is evaluated and the optimization method used to search for classifiers with the highest score (Domingos 2012).

[5] https://www.kaggle.com/

A large part of the research focusing on reducing machine learning bias are concerned with pre-processing the already established training data (Feldman, et al. 2015, Krasanakis, et al. 2018) or training by establishing some fairness constraints (Zafar, et al. 2017, Dwork, et al. 2012). Less emphasis is given however on the different aspects concerning the collection of the data used for training and testing the algorithm/learner. Some research does exist though exploring how biased data end up affecting machine learning models and ways to mitigate bias (Liu, Reyzin and Ziebart 2015, Chen, Johansson and Sontag 2018, Kallus and Zhou 2018).

How data are collected, repurposed and combined is less clear. This survey will try to shed some light and provide practitioners and stakeholders with useful recommendations when the data are generated through crowdsourcing.

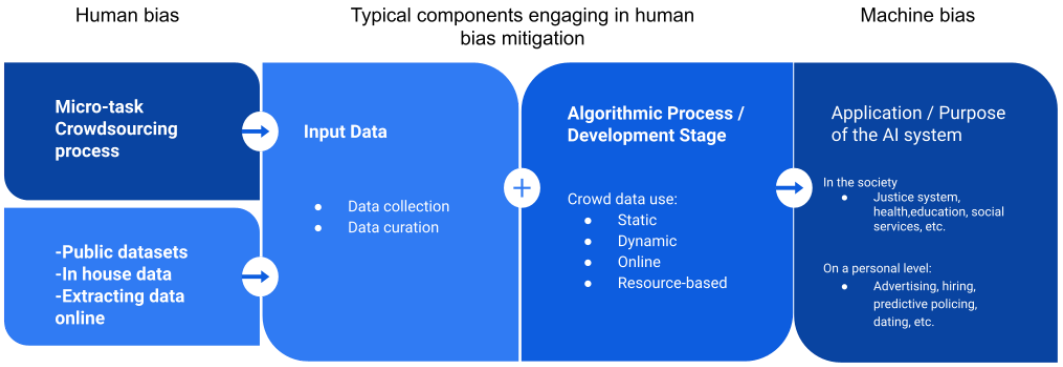


*Figure 1. AI-based system stages of development. A representation of where crowdsourced human bias is generated, where a mitigation process takes place and where disparate impact might appear.*

Many AI-based systems have a steady dependency on human input in order to operate and improve, following the human-in-the-loop approach. The crowdsourcing paradigm acts as a facilitator for the human input leveraging the machines' potentials (Kamar 2016, Demartini 2015). Crowdsourcing data are obtained in a dynamic fashion (Russakovsky, Li and Fei-Fei. 2015, Laws, Scheible and Schutze 2011). Active learning is the process of deciding which data should be labeled by the crowdworkers in order for the classifier to learn faster, and it is usually an iterative process. Through this survey we will try to check how the human-in-the-loop approach propagates and disseminates harmful stereotypes.

Capturing, processing, analysing and simulating human emotions plays a role in some AI systems. It can be part of intelligent advertising systems, chatbots, fake news detection, article creation, elderly care robots, etc. One of the primary sources for extracting data for this type of research is crowdsourcing (Breazeal, et al. 2013). Understanding how human stereotypes, like robot assistants having a gentle woman's voice, appear and become propagated in these types of systems is one of the objectives of this survey.

# 4. Micro-task crowdsourcing

We define **crowdsourcing** as an online participative activity in which a requesting entity (**requester**) proposes to a group of individuals (**crowdworkers** or **workers**) of varying knowledge and demographic background, via a dedicated **platform**, which connects requester with workers, the undertaking of a **task**. The worker agreeing to complete the task will receive a monetary reward and the requester will receive a response to the task. We refer to the set of processed task responses as **data**.

Crowdsourcing, as mentioned above, is a fast and inexpensive way of creating large datasets (Deng, et al. 2009) but the capabilities of this method are limited by its very nature. The nature of the task, the way the task is composed, the general and specific background of the workers, and the way data are aggregated to provide an answer to the task, are the main factors affecting the quality (Ipeirotis, Provost and Wang 2010) and accuracy (Frénay and Verleysen 2013) of the collected data, consequently affecting the reliability of the crowdsourcing process. Quality and accuracy of the crowdsourced data are interconnected topics that have largely captured the attention of researchers over the years (Ipeirotis, Provost and Wang 2010, V. C. Raykar, et al. 2010). Researchers focused on developing mechanisms for receiving the correct result usually by requesting multiple workers to provide an answer to the same task (redundancy) and devising a method for deciding on the correct/accurate or high-quality answer. In addition to the above approaches, a lot of research has also focused on the methods for identifying high quality/skilled workers by using techniques such as gold standard questions.

Our goal in this work is to identify why and how the final crowdsourced data can propagate biases and stereotypes. Quality and accuracy are concepts usually associated with tasks that have a ground truth or that the ground truth can be estimated or that the reported response can be objectively verified. On the other hand, biases can be introduced in either high or low quality crowdsourced data; in fact, as we will review below, methods for improving quality (Hansen, et al. 2013) can introduce biases (Ipeirotis, Provost and Wang 2010). It is interesting to notice though, that many studies emphasising the improvement of the quality of crowdsourced data actually have as a goal the mitigation of bias in the data.

Many studies use the term *bias*, to describe the workers deviation from the true/expected answer (Snow, et al. 2008, Dawid and Skene 1979) and use probabilistic methods to estimate the correct answer. In this literature review, and in the project in general we will ***not use crowdworkers' bias as a term with a negative connotation.*** Thus, we will study the reason bias emerges from crowdsourced data, and we will evaluate approaches for mitigating bias from the final intelligent system output. However, we will not treat any sort of bias emerging from the crowdworkers as an *error* or source of *noise* that needs to be removed or smoothed out.

## 4.1 Shortcomings of crowdsourcing

Let us review some of the shortcomings of crowdsourcing that are directly related to the creation of biases in crowdwork. Table 1. provides a list with the known shortcomings of crowdsourcing, a brief note of how this shortcoming can introduce bias in the crowdwork (which we will assess extensively in the next subsection) and the works pointing out these shortcomings.

*Table 1. Shortcomings of crowdsourcing and its implications in bias creation in crowdwork.*

| **Shortcoming** | **Implications** | **Relevant papers** |
|---|---|---|
| Annotation sparsity, most crowdworkers provide a small number of annotations | Missing data when requiring multiple annotations from the same worker | (P. G. Ipeirotis 2010) |
| A large set of crowdworkers suffering from a specific bias relevant to the nature of the task | Aggregating the workers' replies, following multiple methods might give an erroneous crowdsourcing outcome for the task. | (Kamar, Kapoor and Horvitz, Identifying and accounting for task-dependent bias in crowdsourcing 2015) |
| Efforts to improve crowd data quality can lead to rejecting honest, rich but diverse replies. | Rejecting replies that are not in line with the ground truth or belong to a set of expected data can create a biased dataset or promote stereotypes. | (Chang, Amershi and Kamar 2017, Kairam and Heer 2016) |
| Workers belonging to certain demographic groups (i.e., specific countries) might require higher payments. | Many methods for recruiting crowdworkers are budget based, leading to underrepresentation of certain demographic groups. This can lead to the creation of biases against those demographic groups. | (Barbosa and Chen 2019, Joel, et al. 2010) |
| Workers across different cultures are shown the exact same task design. | Tagging images differs depending on the culture of the crowdworker, i.e., Americans tag the main object first, while Chinese provide tags about the overall properties of the image first. In a task design that only 3 open tag annotations are required it is clear that bias data might arise. | (Dong and Fu 2010) |

# 5. Biases in Crowdwork

## 5.1 Requesting crowdsourced data and data pre-processing

In their work, Kairam and Heer (Kairam and Heer 2016) recognize the weakness of "traditional" methods to improve quality and propose an alternative way for identifying high quality results while taking into consideration that a subset of workers might consistently produce predictable, diverging responses. The authors propose a method for analyzing responses from the crowdworkers by separating workers based on shared patterns in the way they respond to a text annotation task. In their work, Ipeirotis et al. (Ipeirotis, Provost and Wang 2010) also identify workers that have a predictable bias behavior but contrary to the work of Kairam and Heer, they propose methods for reducing the effect of these workers' replies.

Barbosa and Chen (Barbosa and Chen 2019) point out that the process of requesting a worker to do a crowdsourcing task strictly based on skill and experience (in an attempt to improve quality) might significantly skew the demographics of crowdworkers and as a result, introduce biases in the data. The authors propose the use of crowdworkers of different expertise at least for tasks that require no special skills.

The payment given to the workers is another factor that must be regulated by requesters in such a way that the minimum wage in the targeted crowdworkers' country is satisfied. In an opposite case, the crowdworker might feel underpaid (Hara,, et al. 2018) with consequences extending to the way they choose to reply to the task or not select the particular task (Barbosa and Chen 2019). Experienced workers might also fail to distance themselves from their own opinion, and as a result they might produce biased annotations (given the annotation task studied in their work). In other words, when it comes to bias induced by a worker's own opinion, the expertise level of the crowdworker will not guarantee less biased data, as opposed to high quality data. Additionally, when requesting workers with particular demographic characteristics their connection (emotional or other) with the task must be taken into consideration (Peesapati, Wang and Cosley 2010).

Two other technical factors that must be taken into consideration when requesting crowdsourced data as possible sources of bias is the crowdsourcing platform and the time of execution of the crowdsourcing task. The selection of crowdsourcing platform dictates in a way the task design, the available crowd demographics and the perception of crowdwork (Kittur, et al. 2013). Launching the crowdsourcing task at a specific platform and at a specific hour of the day will attract first the active workers, which might all have common demographic characteristics (e.g., ethnicity, age).

## 5.2 The nature of the micro-tasks

The design of the micro-task is another important factor. The ethnicity of the targeted crowdworkers must be considered when designing the task (Dong, et al. 2012, Sen, Lam, et al. 2006, Peesapati, Wang and Cosley 2010). In the example of responses to an image tagging task provided by (Dong, et al. 2012), different ethnicities provided more factual tags compared to subjective tags; thus, the design of the tagging task must be adapted appropriately to encourage or discourage the emergence of factual tags given the worker audience. As an example, notice that in a later use of the tags, factual tags are more useful when searching an image, but subjective tags are more valuable when evaluating an image in an application of the created dataset.

Clarity of the task is a factor that affects the performance of the workers, and consequently how susceptible they are to systemic biases. Gadiraju et al. (Gadiraju, Yang and Bozzon 2017), showed that clarity is coherently perceived by workers and it varies according to the task type. Moreover, they found that the clarity of the task is affected by the keywords used and the readability of the task title. Dumitrache (Dumitrache 2015) in her work showed that the clarity of the annotation labels and the ambiguity of the text (when considering a text annotation task) presented to the workers can be two sources of disagreement among the workers.

Newell and Ruths (Newell and Ruths 2016) have shown that earlier tasks impact strongly the way workers respond to later tasks. Thus, this *intertask effect* can be a source of systematic bias when designing crowdsourcing tasks. The authors suggest that appropriate design options that help workers focus can reduce the influence of this effect and help the reproducibility of the tasks. Other works also explore the impact that the order of task executions has on the task outcome (Aipe and Gadiraju 2018, Cai, Iqbal and Teevan 2016).

## 5.3 Understanding the crowdworkers' behavior

In their work, Kamar et al. (Kamar, Kapoor and Horvitz, 2015) present the issue of bias going unnoticed and the correct answer not being discovered by methods that aggregate the collected results. The authors point out that a population of crowdworkers or a subgroup of them might suffer from a specific bias that leads to systematic errors (i.e., falling for a visual illusion) due to the nature of the task. Cognitive biases (Eickhoff, 2018) (ambiguity effect, anchoring, bandwagon and decoy effect) from the workers are a source of noise during the curation, annotation and generally the evaluation of the crowdsourced dataset.

Workers might also introduce biases in the dataset because of their beliefs (White 2013) and the stereotypes they might carry that are directly correlated with some demographic factors (i.e., age, gender, race, economic status, education etc.). Dong et al. (Dong and Fu 2010) mention that American and Chinese workers do not tag images in the same way and suggest that this is correlated to cultural differences among the two ethnicities. Peesapati et al. (Peesapati, Wang and Cosley 2010) in their work point out that the assumption that people will universally tag images in the same way does not hold and point a connection among the person's origin and the cultural origin of the image. Sen et al. (Sen, Giesel, et al. 2015) took their work a step further showing that

different cultural communities can produce different gold standard dataset from a judgement task and those datasets can directly affect the accuracy of the algorithms that use them. Ghai et al., (Ghai, et al. 2020) proposed a method based on counterfactual fairness for quantifying the social biases inherent in the crowdworkers. Nicholls et al., (Nicholls, Churches and Loetscher 2018) showed that in face perception tasks a crowdworkers own age subconconsciously biases the replies. Young crowdworkers were able to estimate better the age of young people and older crowdworkers the age of older people.

Personal traits of crowdworkers can affect the collected datasets. Otterbacher et al. (Otterbacher, et al. 2018) showed that sexist workers (as tested per a standard psychological measure) are less likely to report gender biases in image search results. Hube et al. (Hube, Fetahu and Gadiraju 2019) showed that workers with strong personal opinions tend to produce bias annotations, for text annotation tasks. In a similar flair, Roitero et al., (Roitero, et al. 2020) measured the political bias of crowdworkers on fact checking tasks. Leung et al., (Leung, et al. 2020) studied the crowdworker bias when it comes to hiring people and they notice that crowdworkers are biased based on gender and race stereotypical beliefs.

## 5.4 Use of the crowdsourced data

As we mentioned earlier, the way crowdsourced data is used can perpetuate biases even more. Sen et al., (Sen, Giesel, et al. 2015) in their work showed how bias in gold standard datasets for a judgement task can directly affect the accuracy of the algorithms that use these gold standards. The use of transfer learning during the machine learning algorithm (especially deep learning algorithms) is shown to perpetuate the biases created through crowdsourcing (Khosla, et al. 2012, Tommasi, et al. 2017, Pan and Yang 2009).

Table 2. presents the most popular image dataset used currently by machine learning algorithms and that have been built entirely or partially through crowdsourcing. The table below points out an important issue that practitioners are faced with when using these datasets: ***the lack of documentation of the crowdsourcing process***. The additional obstacles faced by practitioners makes it even harder to eliminate biases arising from crowdsourcing since little information is known, for example, about the demographics of the crowd or the task design, etc.

*Table 2. Popular datasets, crowdsourcing platform used in the annotation process, and information provided.*

| ***Dataset Name*** | ***Link*** | ***Crowdsourcing platform*** | ***Crowdworker demographics provided*** | ***Annotation process recorded*** |
|---|---|---|---|---|
| Flickr-Faces-HQ Dataset (FFHQ) | https://github.com/NVlabs/ffhq-dataset | Partial use of MTurk | X | X |
| EMOTIC Dataset | http://sunai.uoc.edu/emotic/index.html | MTurk | X | √ |
| LAOFIW - Labeled Ancestral Origin Faces in the Wild | https://www.robots.ox.ac.uk/~vgg/data/laofiw/ | Proprietary platform (Microsoft Azure Face API) | X | partially |
| The 'Celebrity Together' Dataset | https://www.robots.ox.ac.uk/~vgg/data/celebrity_together/ | MTurk | X | X |
| UMDFaces Dataset | http://umdfaces.io/ | MTurk | X | partially |
| MPII Human Pose Dataset | http://human-pose.mpi-inf.mpg.de/ | MTurk | X | partially |
| FLIC - Frames Labeled In Cinema | https://bensapp.github.io/flic-dataset.html | MTurk | X | √ |
| AffectNet | http://mohammadmahoor.com/affectnet/ | MTurk | X | partially |
| Moments in Time Dataset | http://moments.csail.mit.edu/ | MTurk | X | √ |
| The 'Celebrity in Places' Dataset | https://www.robots.ox.ac.uk/~vgg/data/celebrity_in_places/ | MTurk | X | X |

## 5.5 Bias mitigation

An important objective of the DESCANT project is to identify the biases created in the crowdsourced data produced through the project and the sources of those biases. Addressing the above objective (i.e., recognizing the existence and source of bias) is an important step for mitigating bias. In this section we will review different approaches for mitigating bias, during and after data collection.

In the work of Barbosa and Chen (Barbosa and Chen 2019), an algorithm for assigning tasks to workers in order to mitigate demographic biases from the created dataset via fair compensation and compatible task assignments is designed. Faltings et al., (Faltings, et al. 2014) proposed a game theoretic scheme to remove bias among workers. Wauthier and Jordan (Wauthier and

Jordan 2011), proposed a model for capturing bias in the labelling process and predicting labels when consensus among labellers is missing.

Hube et al., (Hube, Fetahu and Gadiraju 2019) proposed and reviewed three mitigation methods for a text annotation task: i) Social projection, where the workers are asked to label a text according to what they believe the majority of workers would choose; ii) Awareness reminder, trying to make workers aware of the controversial and subjective nature of the task to influence their judgement when completing the task; iii) Personalized Nudges, giving personalized instructions to the workers according to the tendencies they exhibit in forming a judgement. The authors note that all three mitigation approaches reduce the total bias and average bias for workers with extreme opinions. Leung et al., (Leung, et al. 2020) showed that task design such as reducing the number of choices can reduce potential discriminations from the workers.

A set of works propose the use of different techniques for enabling workers to *interact* among themselves in an effort to mitigate biases. In (Duan., Ho and Yin 2020) the authors investigate whether the diversity in perspectives among the interacting workers can mitigate biases. Through their results it appears that task properties such as the difficulty of the task and the type of the task (objective v. subjective) play a role in the success of this approach. On the other hand, it is unclear whether the interacting workers must have a diverse or similar perspective for the mitigation technique to be successful. Chang et. al., (Chang, Amershi e Kamar 2017) in their work presented a collaborative approach for producing high quality labels that capture a rich interpretation of the data, harnessing on the diversity of crowdworkers.

Kamar et al. (Kamar, Kapoor and Horvitz, 2015) propose a method for automatically recognizing and correcting task-specific biases using probabilistic graphical models. Ipeirotis et al., (Ipeirotis, Provost and Wang 2010) presented statistical methods for eliminating systematic bias. Notice that their approach has as an objective the elimination of worker bias which might not always be the objective of bias mitigation and it depends on the subjective nature of the task.

# 6. Conclusion

In this deliverable, we have presented a review of the literature most relevant to the objectives of the DESCANT project. In particular, after explaining the nature of crowdsourcing, we have defined the scope of the project, which focuses on the use of paid, micro-task crowdsourcing via commercial platforms. This review has help to identify the key parameters that are important when considering biases in the types of crowdwork, which are most interesting and relevant for researchers and practitioners who are using it to develop training datasets for AI systems.

# References

Abbas, Tahir, Vassilis-Javed Khan, Ujwal Gadiraju, Emilia Barakova, and Panos Markopoulos. 2020. “Crowd of Oz: a crowd-powered social robotics system for stress management.” *Sensors* (Multidisciplinary Digital Publishing Institute) 20 (2): 569.

Aipe, Alan, and Ujwal Gadiraju. 2018. “Similarhits: Revealing the role of task similarity in microtask crowdsourcing.” *In Proceedings of the 29th on Hypertext and Social Media.* 115-122.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.” *ProPublica.*

Barbosa, Natã M., and Monchu Chen. 2019. “Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning.” *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1-12.

Bodenhausen, Galen V. . 1993. “Emotions, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping.” *In Affect, cognition and stereotyping.* Academic Press. 13-37.

Boratto, Ludovico, Gianni Fenu, and Mirko Marras. 2019. “The effect of algorithmic bias on recommender systems for massive open online courses.” *In European Conference on Information Retrieval.* Springer. 457-472.

Breazeal, Cynthia, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. 2013. “Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment.” *Journal of Human-Robot Interaction* 2 (1): 82-111.

Buolamwini, Joy, and Timnit Gebru. 2018. “Gender shades: Intersectional accuracy disparities in commercial gender classification.” *In Conference on fairness, accountability and transparency.* PMLR. 77-91.

Cai, Carrie J., Shamsi T. Iqbal, and Jaime Teevan. 2016. “Chain reactions: The impact of order on microtask chains.” *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* 3143-3154.

Callison-Burch, Chris, and Mark Dredze. 2010. “Creating speech and language data with amazon’s mechanical turk.” *In Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon’s Mechanical Turk.* 1-12.

Chang, Joseph Chee, Saleema Amershi, and Ece Kamar. 2017. “Revolt: Collaborative crowdsourcing for labeling machine learning datasets.” *In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* 2334-2346.

Chen, Irene, Fredrik D. Johansson, and David Sontag. 2018. “Why is my classifier discriminatory?” *arXiv preprint arXiv:1805.12002.*

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. “A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions.” *Conference on Fairness, Accountability and Transparency.* PMLR. 134-148.

Dawid, Alexander Philip, and Allan M. Skene. 1979. “Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28 (1): 20-28.

Demartini, Gianluca. 2015. “Hybrid human--machine information systems: Challenges and opportunities.” *Computer Networks* 90: 5-13.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. “Imagenet: A large-scale hierarchical image database.” *In IEEE conference on computer vision and pattern recognition.* IEEE. 248-255.

Domingos, Pedro. 2012. “A few useful things to know about machine learning.” *Communications of the ACM* (ACM New York, NY, USA) 55 (10): 78-87.

Dong, Wei, and Wai-Tat Fu. 2010. “Cultural difference in image tagging.” *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 981-984.

Dong, Zhenhua, Chuan Shi, Shilad Sen, Loren Tervee, and John Riedl. 2012. “War versus inspirational in forrest gump: Cultural effects in tagging communities.” *In Proceedings of the International AAAI Conference on Web and Social Media.*

Dressel, Julia, and Hany Farid. 2018. “The accuracy, fairness, and limits of predicting recidivism.” *Science advances* (American Association for the Advancement of Science) 4 (1): eaao5580.

Duan., Xiaoni, Chien-Ju Ho, and Ming Yin. 2020. “Does Exposure to Diverse Perspectives Mitigate Biases in Crowdwork? An Explorative Study.” *In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8 (1): 155-158.

Dumitrache, Anca. 2015. “Crowdsourcing disagreement for collecting semantic annotation.” *In European Semantic Web Conference.* Springer. 701-710.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. “Fairness through awareness.” *In Proceedings of the 3rd innovations in theoretical computer science conference.* 214-226.

Eickhoff, Carsten. 2018. “Cognitive biases in crowdsourcing.” *In Proceedings of the eleventh ACM international conference on web search and data mining.* 162-170.

Estellés-Arolas, Enrique , and Fernando González-Ladrón-de-Guevara. 2012. “Towards an integrated crowdsourcing definition.” *Journal of Information Science* 38 (2): 189-200.

Faltings, Boi, Radu Jurca, Pearl Pu, and Bao Duy Tran. 2014. “Incentives to counter bias in human computation.” *In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing.*

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. “Certifying and removing disparate impact.” *In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining.* 259-268.

Frénay, Benoît, and Michel Verleysen. 2013. “Classification in the presence of label noise: a survey.” *IEEE transactions on neural networks and learning systems* 25 (5): 845-869.

Friedman, Batya , and Helen Nissenbaum. 1996. “Bias in computer systems.” *ACM Transactions on Information Systems (TOIS)* (ACM New York, NY, USA) 14 (3): 330-347.

Gadiraju, Ujwal, Jie Yang, and Alessandro Bozzon. 2017. “Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing.” *In Proceedings of the 28th ACM Conference on Hypertext and Social Media.* 5-14.

Gao, Huiji, Geoffrey Barbier, and Rebecca Goolsby. 2011. “Harnessing the crowdsourcing power of social media for disaster relief.” *IEEE Intelligent Systems* 26 (3): 10-14.

Ghai, Bhavya, Q. Vera Liao, Yunfeng Zhang, and Klaus Mueller. 2020. “Measuring Social Biases of Crowd Workers using Counterfactual Queries.” *arXiv preprint arXiv:2004.02028.*

Hansen, Derek L., Patrick J. Schone, Douglas Corey, Matthew Reid, and Jake Gehring. 2013. “Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at familysearch indexing.” *In Proceedings of the 2013 conference on Computer supported cooperative work.* 649-660.

Hara,, Kotaro, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. “A data-driven analysis of workers' earnings on Amazon Mechanical Turk.” *In Proceedings of the 2018 CHI conference on human factors in computing systems.* 1-14.

Hilton, James L., and William Von Hippel. 1996. “Stereotypes.” *Annual review of psychology* 47 (1): 237-271.

Houser, Kimberly A. 2019. “Hou Can AI Solve the Diversity Problem in the Tech Industry: Mitigating Noise and Bias in Employment Decision-Making.” *Stan. Tech. L. Rev.* 22, 290.

Howe, Jeff. 2006. “The rise of crowdsourcing.” *Wired magazine* 14 (6): 1--4.

Hube, Christoph, Besnik Fetahu, and Ujwal Gadiraju. 2019. “Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments.” *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1-12.

Hutson, Jevan A., Jessie G. Taft, Solon Barocas, and Karen Levy. 2018. “Debiasing desire: Addressing bias \& discrimination on intimate platforms.” *Proceedings of the ACM on Human-Computer Interaction.* 1--18.

Ipeirotis, Panagiotis G, Foster Provost, and Jing Wang. 2010. “Quality management on amazon mechanical turk.” *In Proceedings of the ACM SIGKDD workshop on human computation.* 64-67.

Ipeirotis, Panagiotis G. 2010. “Analyzing the amazon mechanical turk marketplace.” *XRDS: Crossroads, The ACM Magazine for Students* 17 (2): 16-21.

Joel, Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. “Who are the crowdworkers? Shifting demographics in Mechanical Turk.” *In CHI'10 extended abstracts on Human factors in computing systems.* 2863-2872.

Kadija, Ferryman, and Mikaela Pitcan. 2018. “Fairness in precision medicine.” *Data & Society* 1.

Kairam, Sanjay, and Jeffrey Heer. 2016. “Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks.” *In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* 1637-1648.

Kaiyu, Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. “Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy.” *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 547-558.

Kallus, Nathan, and Angela Zhou. 2018. “Residual unfairness in fair machine learning from prejudiced data.” *In International Conference on Machine Learning.* PMLR. 2439-2448.

Kamar, Ece. 2016. “Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.” *In IJCAI.* 4070-4073.

Kamar, Ece, Ashish Kapoor, and Eric Horvitz. 2015. “Identifying and accounting for task-dependent bias in crowdsourcing.” *In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing.*

Khosla, Aditya, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. 2012. “Undoing the damage of dataset bias.” *In European Conference on Computer Vision* (Springer, Berlin, Heidelberg) 158-171.

Kittur, Aniket, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. “The future of crowd work.” *In Proceedings of the 2013 conference on Computer supported cooperative work.* 1301-1318.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human decisions and machine predictions.” *The quarterly journal of economics* 133(1), 237-293.

Klinger, Ulrike , and Jakob Svensson. 2018. “The end of media logics? On algorithms and agency.” Edited by England SAGE Publications Sage UK: London. *New Media & Society* 20 (12): 4653-4670.

Kovashka, Adriana, Olga Russakovsky, Li Fei-Fei, and Kristen Grauman. 2016. “Crowdsourcing in computer vision.” *arXiv preprint arXiv:1611.02145.*

Krasanakis, Emmanouil , Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. “Adaptive sensitive reweighting to mitigate bias in fairness-aware classification.” *Proceedings of the 2018 World Wide Web Conference.* 853--862.

Kumar, Shamanth, Fred Morstatter, and Huan Liu. 2014. *Twitter data analytics.* New York: Springer.

Lasecki, Walter S., Christopher Homan, and Jeffrey P Bigham. 2015. “Architecting real-time crowd-powered systems.” *Human Computation* 1 (1).

Laws, Florian, Christian Scheible, and Hinrich Schutze. 2011. “Active learning with amazon mechanical turk.” *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing.* 1546--1556.

Leung, Weiwen, Zheng Zhang, Daviti Jibuti, Jinhao Zhao, Maximilian Klein, Casey Pierce, Lionel Robert, and Haiyi Zhu. 2020. “Race, Gender and Beauty: The Effect of Information Provision on Online Hiring Biases.” *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1-11.

Liu, Anqi, Lev Reyzin, and Brian Ziebart. 2015. “Shift-pessimistic active learning using robust bias-aware prediction.” *In Proceedings of the AAAI Conference on Artificial Intelligence.*

McGarty, Craig Ed, Vincent Y. Yzerbyt, and Russell Ed Spears. 2002. “Stereotypes as explanations: The formation of meaningful beliefs about social groups.” *Cambridge University Press.*

Newell, Edward, and Derek Ruths. 2016. “How one microtask affects another.” *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* 3155-3166.

Nicholls, Michael ER, Owen Churches, and Tobias Loetscher. 2018. “Perception of an ambiguous figure is affected by own-age social biases.” *Scientific reports* 8 (1): 1-5.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. “Social data: Biases, methodological pitfalls, and ethical boundaries.” (Frontiers in Big Data) 2: 13.

Otterbacher, Jahna, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. “Investigating user perception of gender bias in image search: the role of sexism.” *In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* 933-936.

Pan, Sinno Jialin, and Qiang Yang. 2009. “A survey on transfer learning.” *IEEE Transactions on knowledge and data engineering* 22 (10): 1345-1359.

Peesapati, Tejaswi S. , Hao-Chuan Wang, and Dan Cosley. 2010. “Intercultural human-photo encounters: how cultural similarity affects perceiving and tagging photographs.” *In Proceedings of the 3rd international conference on Intercultural collaboration.* 203-206.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. “Mitigating bias in algorithmic hiring: Evaluating claims and practices.” *In Proceedings of the 2020 conference on fairness, accountability, and transparency.* 469-481.

Raykar, Vikas C., Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. “Learning from crowds.” *Journal of Machine Learning Research* 11 (4).

Raykar, Vikas , Shipeng Yu, Linda Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. “Supervised learning from multiple experts: whom to trust when everyone lies a bit.” *In Proceedings of the 26th Annual international conference on machine learning.* 889-896.

Richardson, Rashida, Jason M Schultz, and Kate Crawford. 2019. “Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice.” *NYUL Rev. Online* (HeinOnline) 94: 15.

Roitero, Kevin, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. “Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background.” *In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 439-448.

Russakovsky, Olga, Li-Jia Li, and Li Fei-Fei. 2015. “Best of both worlds: human-machine collaboration for object annotation.” *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2121-2131.

Sánchez-Monedero, Javier, Lina Dencik, and Lilian Edwards. 2020. “What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems.” *In Proceedings of the 2020 conference on fairness, accountability, and transparency.* 458-468.

Sen, Shilad, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao Wang, and Brent Hecht. 2015. “Turkers, Scholars," Arafat" and" Peace" Cultural Communities and Algorithmic Gold Standards.” *In Proceedings of the 18th acm conference on computer supported cooperative work & social computing.* 826-838.

Sen, Shilad, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. 2006. “Tagging, communities, vocabulary, evolution.” *In Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work.* 181-190.

Snow, Rion, Brendan O’connor, Dan Jurafsky, and Andrew Y. Ng. 2008. “Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks.” *In Proceedings of the 2008 conference on empirical methods in natural language processing.* 254-263.

Speicher, Till, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 2018. “Potential for discrimination in online targeted advertising.” *In Conference on Fairness, Accountability and Transparency.* PMLR. 5-19.

Stol , Klaas-Jan, and Brian Fitzgerald. 2014. “Two's company, three's a crowd: a case study of crowdsourcing software development.” *In Proceedings of the 36th International Conference on Software Engineering.* 187-198.

Sun, Kaiyuan, Jenny Chen, and Cécile Viboud. 2020. “Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study.” *The Lancet Digital Health* 2 (4): e201-e208.

Tommasi, Tatiana, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. 2017. “A deeper look at dataset bias.” *In Domain adaptation in computer vision applications* (Springer, Cham) 37-55.

Van Den Eede, Yoni. 2011. “In between us: On the transparency and opacity of technological mediation.” *Foundations of Science* 16, no. 2-3: 139-159.

Vaughan, Jennifer Wortman. 2017. “Making better use of the crowd: How crowdsourcing can advance machine learning research.” *The Journal of Machine Learning Research* (JMLR. org) 18 (1): 7026--7071.

Verbeek, Peter-Paul. 2005. *What things do: Philosophical reflections on technology, agency, and design* . Penn State Press.

Wauthier, Fabian L., and Michael Jordan. 2011. “Bayesian bias mitigation for crowdsourcing.” *Advances in neural information processing systems* 1800-180.

West, Mark, Rebecca Kraut, and Han Ei Chew. 2019. *I'd blush if I could: closing gender divides in digital skills through education.* UNESCO.

White, Ryen. 2013. “Beliefs and biases in web search.” *In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.* 3-12.

Willson, Michele. 2017. “Algorithms (and the) everyday.” *Information, Communication & Society* 20(1), 137-150.

Zafar, Muhammad Bilal , Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. “Fairness constraints: Mechanisms for fair classification.” *In Artificial Intelligence and Statistics.* PMLR. 962-970.