Document Title	**Conceptual Framework Description**
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.	D3.2
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) A. Lanitis (CYENS) G. Demartini (UQ)
Reviewers	S. Kleanthous (OUC)
Status (D: draft; RD: revised draft; F: final)	F
File Name	D3.2_Framework.docx
Date	30 May 2020

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	20/3/20	E. Christoforou	e.christoforou@cyens.org.cy	initial document
1.2	1/4/20	E. Christoforou	e.christoforou@cyens.org.cy	addition of content
1.3	20/5/20	E. Christoforou	e.christoforou@cyens.org.cy	review version
1.4	30/5/20	E. Christoforou	e.christoforou@cyens.org.cy	final version

Abstract

This deliverable provides a general overview of the crowdsourcing framework and provides in a structured way the possible parameters in the crowdsourcing process responsible for the emergence of bias in the collected data. This framework sets the foundations for the work carried out in WP4.

Keyword(s):	Requester, Worker, Micro-task, Crowdsourced data, Bias, Stereotypes, Demographics

Contents

1. About this Deliverable

The goal of this deliverable is to set the bases of the framework used throughout WP4 and forms the cornerstone of this project. Through this document the main actors are identified, their role explicitly mentioned, and all the parameters affecting the creation of biases in the collected data associated with each actor presented. In particular, we identify three main actors/processes that create and propagate biases through the crowdsourcing process: (1) the task requester, (2) the crowdworkers and (3) the micro-task itself.

2. General Micro-task Crowdsourcing Setting

Deliverable D3.1 provided a general review of the concept of crowdsourcing and the different forms it can take. In this deliverable, we are interested in micro-task crowdsourcing (from this point onward we will refer to it simply as *crowdsourcing)*, which is the main process of interest in the DESCANT project. Our objective is to study the way biases and stereotypes are created and introduced into the crowdsourced data. Thus, in this section, we will first review the crowdsourcing setting and its various components.

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 the requester with workers, who engage in 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**.

DESCANT, and hence also this deliverable, considers that data creation/collection through crowdsourcing happens in order to create and/or enhance already existing data to be used by an intelligent system. Figure 1. presents an overview of the crowdsourcing process and its main actors when the purpose of crowdsourcing is the collection of data for an AI system. We will use Figure 1. when describing our framework as a conceptual guideline that will help us identify and present in a structured way, the sources of bias.

Through Figure 1, which presents a high-level perspective of an intelligent system in need of data collected through crowdsourcing, we can identify the different sources of bias. The intelligent system in need of data through a *process*, either operated manually by one or multiple humans or in an automated way, will formulate a task that will best serve the purpose of collecting data through crowdsourcing. Then, a requester (which can be operated by the same entity) will formally compose a crowdsourcing task fitting the specifications of the crowdsourcing platform and the needs of the data to be collected (i.e., only female workers above 18 years of age replying to the task).

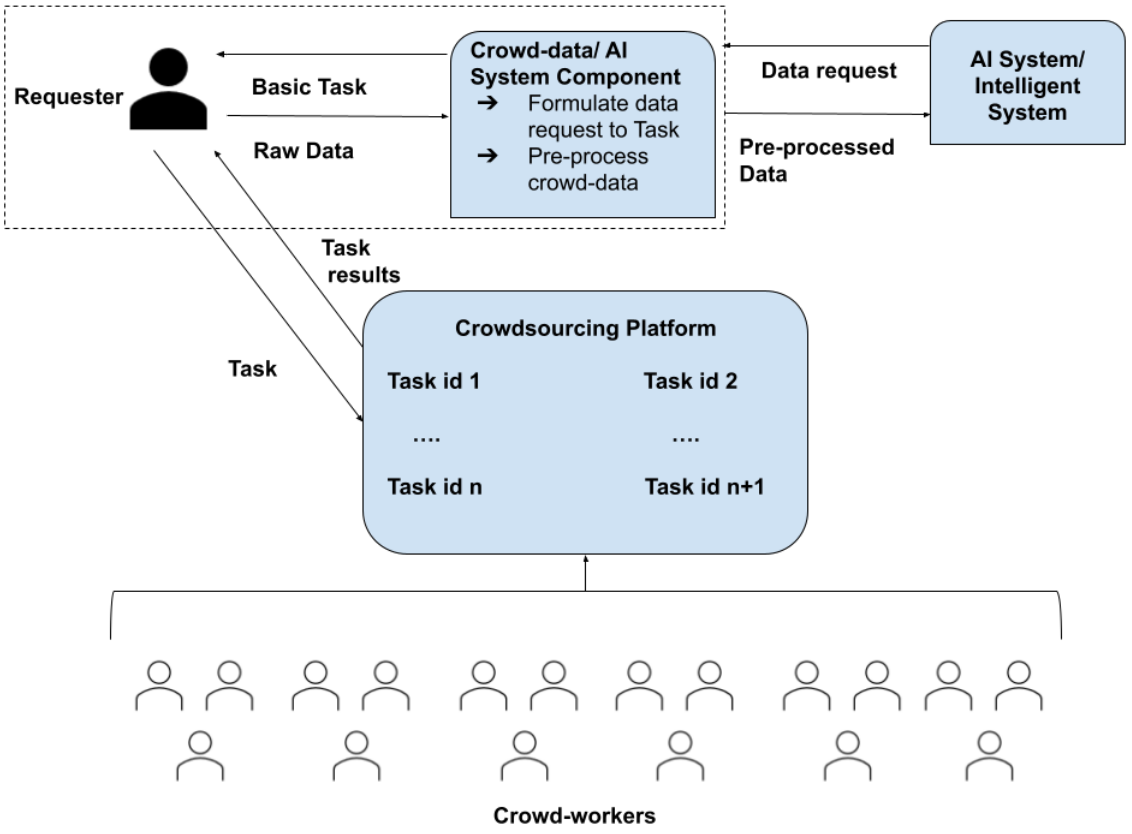

Figure 1. Crowdsourcing process and main actors when creating data for an AI system.

The requester will place the task in the platform, declaring the monetary rewards provided to each worker replying to the task. Interested and eligible workers will respond to the crowdsourcing task and the platform will forward back to the requester the collected replies. It is the responsibility of each requester and the intelligent system that is supporting the crowdsourcing process, to pre-process the crowdsourced data to fit the specific goal of the system.

Notice that in a crowdsourcing platform multiple tasks from multiple requesters co-exist at the same time. How many replies must a specific task receive and from how many different workers is an option set by the requester. Additionally, notice that in the crowdsourcing platform multiple requesters, tasks and crowdworkers co-exist and are possibly in competition with each other.

In the following sections, we analyse one by one the crowdsourcing components for creating data for intelligent systems and identify the main parameters in each component that can potentially introduce biases.

3. Requester

In this section we will identify the parameters that play a role in the propagation of biases and are under the control of the requester and the Crowd-data component (Figure 1, dotted frame).

Goal of data collection. One of the most common reasons for collecting data through crowdsourcing is for the creation of ***gold standard data*** or data expressing the ***ground truth.*** These data are used for training or testing machine learning algorithms (Snow, et al. 2008, Nowak and Rüger 2010). It has been shown that gold standard data created through the process of crowdsourcing can include cultural biases correlated with the crowdworkers demographics (Sen, et al. 2015).

Given the goal of the AI system and the requested set of data, many times it is recognized that collecting gold standard data is infeasible and the aim shifts to forcing the disagreements among workers to surface. This is another valuable goal of the data collection because it will indicate the possible issues with the collection of the data or indicate the existence of more clusters or categories than the requester was expecting.

Acquiring high quality data. Several techniques were proposed over the years for achieving high quality data whether those will serve as gold standard or not. These techniques focus on a combination of collecting redundant answers and applying some majority technique together with some auditing technique like asking workers to reply to gold standard (already known) questions to evaluate their accuracy/skills. Another approach for acquiring high quality data is to train the workers before introducing them to the actual task (Gadiraju, Fetahu and Kawase, Training workers for improving performance in crowdsourcing microtasks 2015). A lot of studies also propose the use of incentive schemes towards the crowdworkers (Eickhoff and de Vries, Increasing cheat robustness of crowdsourcing tasks 2013) but this can make the task susceptible to participation biases.

Relying on techniques like the ones described above with the accompanying algorithmic process can in fact improve the quality of the data but can lead to the "dehumanization" of the crowd which can come into conflict with the goal of maintaining a balanced demographic crowd (Gadiraju, Demartini, et al. 2015; Kittur, Nickerson, et al., 2013).

Interaction method with workers. According to the goal of the AI system, the requester can select one of the following ways of interacting with the workers to gather the data for a single task: (1) single-shot interaction, (2) iterative, multiple interactions (Barbosa and Chen 2019, Chang, Amershi and Kamar 2017), (3) chat-based (Mavridis, et al. 2019).

Crowd selection. The crowd selected or allowed to complete a crowdsourcing task can have a large impact on the propagation of biases (Barbosa and Chen 2019). Given the nature of the task, different pre-selected features of the workers, such as gender, age, country, language, skills and experience can be selected according to some algorithm (Barbosa and Chen 2019). Additionally, as part of the crowd selection method the requester must identify whether workers that completed the same or similar task in the past can participate in the present task. Hube et al., (Hube, Fetahu and Gadiraju 2019) point out that experienced workers might also fail to distance themselves from their own opinion 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.

The number of workers to be recruited for the collection of the necessary data for a single task is another parameter that needs to be considered. Usually, workers are being selected based on the available resources and final goal of the data collection following some optimization method (Barbosa and Chen 2019). It is essential though that a uniform set of workers according to the required demographics is collected. It is also important to have in mind that 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).

Crowd selection can be controlled not only by the input parameters of the crowdsourcing platform but also from a questionnaire given to the crowdworkers, allowing the requester to identify other factor that might encourage the creation of biases like the cultural background of the workers in subjective tasks (Dong, et al. 2012).

Payment selection. Many studies focused on the payment of the crowdworkers as an optimization problem. For example, maximizing the quality of the collected data while having a budget restriction. In (Barbosa and Chen 2019) an optimization is proposed for minimizing the pay gap in order to respect the minimum wage in the workers' country. Proposing tasks with certain monetary rewards that do not satisfy workers in certain countries can lead to the collection of workers from certain countries and certain skills (Hara, et al. 2018), hence, encouraging the creation of biased data.

Platform selection. Different platforms attract crowds from different countries. Amazon Mechanical Turk and Appen have more participants from the US, Latin America and India, while the Clickworkers platform has a larger European crowd and Microworker a larger crowd from India. Moreover, some platforms have settings more dedicated to surveys (i.e., Clickworkers) and others provide very specific frames for the development of tasks (i.e., Microworkers). The selection of crowdsourcing platform dictates in a way the task design, the available crowd demographics and the perception of crowdwork (Kittur, Nickerson, et al., The future of crowd work 2013)

Time considerations. An additional consideration when selecting a platform and a crowd from a specific country is the time that the task should be executed. In countries where workers can have several hours of difference, the time that the task will be launched is crucial. For example, launching a task for US workers in the afternoon for the East Coast, will reduce the chances of receiving workers from the West Coast, because a lot of workers might be unavailable at that time.

4. Micro-task (task)

Subjective v. Objective tasks. Subjective tasks are particularly sensitive to biases (Otterbacher, Checco, et al., 2018; Nguyen, et al. 2014). For example, content moderation tasks recruiting a much larger percentage of males can introduce biases to the final intelligent system, which might

show content that is offensive to females. On the other hand, objective tasks are less relevant to the goal of the project, since by their nature they have a generally acceptable reply, for example, transcribing a receipt. It is clear that in DESCANT we are interested in subjective tasks.

Some tasks can present themselves as objective but might include optical illusions (Kamar, Kapoor and Horvitz 2015). These types of tasks might still introduce biases that are not relevant to the subjectivity of the task and how the task is perceived by workers belonging to certain demographic groups. This is an example where the nature of the task introduces unwanted biases relevant to a worker's visual perception.

Task categories. Following the classification of Gadiraju et al., (Gadiraju, Kawase and Dietze, 2014) we consider the following task categories:

- Information Finding: Metadata finding
- Verification and Validation: Content Verification, Content Validation, Spam Detection, Data matching
- Interpretation and Analysis: Classification, Categorization, Media Transcription, Ranking, Data Selection, Sentiment Analysis, Content Moderation, Quality Assessment
- Content Creation: Media Transcription, Data Enhancement, Translation, Tagging
- Surveys: Feedback/Opinions, Demographics
- Content Access: Testing, Promoting

Notice that some tasks will create data from no initial "seed of data" (i.e., surveys, information finding) while other tasks require an initial set of data to exist (i.e., content creation). In the latter case, it is possible that biases will be created due to the very nature of the original dataset, i.e., tagging images from exotic destinations v. tagging natural disasters images. The original data can be a text, or some media artefact like images, audio, video, animation etc.

Difficult v. easy tasks. Considering the difficulty of the task is not a sufficient parameter to understand the task-dependent worker biases, but it is still required to model the explicit task characteristics (Kamar, Kapoor and Horvitz 2015).

Structure and clarity of the task. A crowdsourcing task can be composed of: (1) questions with an open answer, (2) closed form questions, choosing among a set of feasible answers, (3) questions requesting workers to provide a rank among a set of objects or a score, (4) questions combining open and closed form questions. Open-form questions and questions providing a rank make the aggregation of the data easier; on the other hand, closed-form questions allow for disagreements to surface, making bias detection easier (Kairam and Heer 2016).

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. The intertask effect (Newell and Ruths 2016) must also be taken into account when designing the task and the design must promote the focus of the workers.

5. Workers

Worker demographic. Workers in crowdsourcing platforms provide a diverse crowd from different nationalities, ethnicities, countries of residence, ages and gender. Workers also have different religions, cultures, education and economic status.

Worker reliability. According to whether the workers can be trusted to provide a correct answer to a task we refer to workers as reliable or unreliable. Judging the reliability of a worker usually happens through gold standard questions (Gadiraju, Demartini, et al. 2015). Unreliable workers can have different behaviours according to what incentivizes them to reply in the first place and might some time provide correct answers. We will follow the definitions in (Gadiraju, Demartini, et al. 2015) to characterize unreliable workers: Ineligible workers, Fast deceivers, Smart deceivers, Rule breakers, Gold standard preys.

A number of works study ways for incentivizing the unreliable workers to provide correct answers through rewards, for example. It has been shown that workers of different ethnicities (e.g., Americans v. Indians) have different perceptions on the incentives for non-monetary participation (Jiang, Wagner and Nardi 2015). Americans' value emotional benefits, while Indians' value self-improvement. This difference in the incentives of participation from different ethnicities can be another factor affecting bias.

Experienced crowdworkers use a number of tools that affect the way they select and process a crowdsourcing task, as well as how they recognize and approach attention and gold standard questions. Han et al., (Han, et al. 2020) provide quantitative and qualitative results examining the impact on the quality of responses.

Workers' preferences to task. Given various characteristics of the task like the completion time, the difficulty of the task, the reward, the topic and the number of other tasks currently available, a worker chooses the next task to complete. For example, easy tasks are often more attractive to workers (Gadiraju, Kawase and Dietze, 2014). Thus, the possibility of attracting a more diverse crowd compared to difficult tasks is higher. This is a factor that can affect the bias creation and propagation through crowdsourcing.

Workers, especially the ones spending a large amount of their time on crowdsourcing platforms are known to use tools to "capture" desirable tasks and increase their income. These methods can affect the quality of the produced work and the execution time (Gadiraju, Checco, et al. 2017, Han, et al. 2020) which in turn can influence the way worker bias can be accounted for. Notice that workers using forums as a tool to select a crowdsourcing task might suffer from participation bias in selecting a task.

Worker biases. Worker biases arise as the workers' personal opinion influences the way a worker completes a subjective task (Hube, Fetahu and Gadiraju 2019). For instance, workers holding sexist beliefs are less likely to report gender biases in image search results (Otterbacher, Checco, et al., 2018). Workers might exhibit gender and race bias when it comes to hiring people (Leung, et al. 2020). Also workers can have a biased behavior when fact checking tasks relevant to politics (Roitero, et al. 2020) measured. In general, strong opinions tend to produce biased annotations (Hube, Fetahu and Gadiraju 2019).

As we mentioned above, systemic biases may also arise from objective tasks that include an optical illusion for examples and is correlated with the way a worker processes visual or other senses stimulus (Kamar, Kapoor and Horvitz 2015). In this project we are interested in subjective tasks as they are the ones that will most probably create the training data that AI algorithms that pass judgements or form opinions will use. Cognitive biases (Eickhoff, Cognitive biases in crowdsourcing 2018) (ambiguity effect, anchoring, bandwagon and decoy effect) are a source of noise during the curation, annotation and generally the evaluation of the crowdsourced dataset.

Biases arising from a worker's association with the subjective task and due to the workers' beliefs (White 2013). Demographic characteristics such as a workers' origin can influence the task replies. (Dong and Fu 2010) especially when there is a connection among the person's origin and the cultural origin of the image (Peesapati, Wang and Cosley 2010). Different cultural communities can produce different gold standard datasets. (Sen, et al. 2015). Furthermore, the worker's own age can introduce age-bias in the data in face relevant tasks (Nicholls, Churches and Loetscher 2018) .

6. Summary

Extrapolating from our review of the literature (D3.1), this current deliverable presents a conceptual framework, which shall guide the crowd experiments to be carried out in WP4. In particular, we have described the particular parameters that will be of interest to use in our experiments, thus further refining the DESCANT Conceptual Framework.

References

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.

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.

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 Terveen, 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.*

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

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

Eickhoff, Carsten, and Arjen P. de Vries. 2013. “Increasing cheat robustness of crowdsourcing tasks.” *Information retrieval* 16 (2): 121-137.

Gadiraju, Ujwal, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. “Modus operandi of crowd workers: The invisible role of microtask work environments.” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies.* 1-29.

Gadiraju, Ujwal, Besnik Fetahu, and Ricardo Kawase. 2015. “Training workers for improving performance in crowdsourcing microtasks.” *In Design for Teaching and Learning in a Networked World.* Springer. 100-114.

Gadiraju, Ujwal, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. “Human beyond the machine: Challenges and opportunities of microtask crowdsourcing.” *IEEE Intelligent Systems* 30 (4): 81-85.

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.

Gadiraju, Ujwal, Ricardo Kawase, and Stefan Dietze. 2014. “A taxonomy of microtasks on the web.” *In Proceedings of the 25th ACM conference on Hypertext and social media.* 218-223.

Han, Lei, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. 2020. “Crowd worker strategies in relevance

judgment tasks." *In Proceedings of the 13th International Conference on Web Search and Data Mining.* 241-249.

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.

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.

Jiang, Ling, Christian Wagner, and Bonnie Nardi. 2015. "Not just in it for the money: a qualitative investigation of workers' perceived benefits of micro-task crowdsourcing." *In 2015 48th Hawaii International Conference on System Sciences.* IEEE. 773-782.

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.

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.*

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.

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.

Mavridis, Panagiotis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2019. "Chatterbox: Conversational interfaces for microtask crowdsourcing." *In Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization.* 243-251.

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.

Nguyen, Dong, Dolf Trieschnigg, Seza A. Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. "Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment." *In Proceedings of*

COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 1950-1961.

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.

Nowak, Stefanie, and Stefan Rüger. 2010. “How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation.” *In Proceedings of the international conference on Multimedia information retrieval.* 557-566.

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.

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.

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.

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.

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.

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.