How Does the Crowd Impact the Model?

A Tool for Raising Awareness of Social Bias in Crowdsourced Training Data

Periklis Perikleous
CYENS Centre of Excellence
Cyprus
pericles@outlook.com

Andreas Kafkalias
CYENS Centre of Excellence
Cyprus
antreaskafkalias7@gmail.com

Zenonas Theodosiou
CYENS Centre of Excellence
Cyprus
z.theodosiou@cyens.org.cy

Pınar Barlas
CYENS Centre of Excellence
Cyprus
pin.barlas@gmail.com

Evgenia Christoforou
CYENS Centre of Excellence
Cyprus
e.christoforou@cyens.org.cy

Jahna Otterbacher
CYENS Centre of Excellence & Open University of Cyprus
Cyprus
j.otterbacher@cyens.org.cy

Gianluca Demartini
The University of Queensland
Australia
g.demartini@uq.edu.au

Andreas Lanitis
CYENS Centre of Excellence &
Cyprus University of Technology
Cyprus
a.lanitis@cyens.org.cy

ABSTRACT

It is increasingly easy for interested parties to play a role in the development of predictive algorithms, with a range of available tools and platforms for building datasets, as well as for training and evaluating machine learning (ML) models. For this reason, it is essential to create awareness among practitioners on the ethical challenges, such as the presence of *social bias* in training data. We present RECANT (*Raising Awareness of Social Bias in Crowdsourced Training Data*), a tool that allows users to explore the behaviors of four biometric models – predicting the gender and race, as well as the perceived attractiveness and trustworthiness, of the person depicted in an input image. These models have been trained on a crowdsourced dataset of passport-style people images, where crowd annotators described attributes of the images, and reported their own demographic characteristics. With RECANT, users can explore the correct and wrong predictions made by each model, when using different subsets of the data in training, based on annotator attributes. We present its features, along with sample exercises, as a hands-on tool for raising awareness of potential pitfalls in data practices surrounding ML.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Social and professional topics**;

KEYWORDS

algorithmic bias, biometrics, crowdsourcing, data bias, education

CIKM'22, October 17–22, 2022 , Atlanta, Georgia, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557178

ACM Reference Format:
Periklis Perikleous, Andreas Kafkalias, Zenonas Theodosiou, Pınar Barlas, Evgenia Christoforou, Jahna Otterbacher, Gianluca Demartini, and Andreas Lanitis. 2022. How Does the Crowd Impact the Model?: A Tool for Raising Awareness of Social Bias in Crowdsourced Training Data. In *Proceedings of the 31st ACM Int'l Conference on Information and Knowledge Management (CIKM '22), Oct. 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3511808.3557178

1 INTRODUCTION

There has been a rise in the popularity of Artificial Intelligence (AI)[1] systems, in part due to the wide availability of training materials (e.g., fast.ai, lobe.ai) that provide relevant skills to practitioners of diverse backgrounds. While this has many benefits, such as the application of AI across domains, it also comes with the risk of building systems that reinforce social bias and stereotypes caused by low quality and/or unbalanced data used as input to the trained models. As AI models are usually data-hungry, a popular way to collect labelled data at scale has been crowdsourcing. Online crowdsourcing platforms (e.g., Amazon MTurk) allow practitioners to tap into a large pool of human annotators available around the clock. This makes it easy to quickly collect large volumes of data annotations to train AI.

At the same time, the number of AI bias cases is growing rapidly over time. Popular examples include police resource allocation algorithms focusing on the race of the resident of a certain suburb [1], automatic lending decision systems declining applications from members of certain demographic groups [8], and arrest decisions based on the district of residence [4]. These critical, algorithmic mistakes are often due to the data (or lack of) used to train AI. If a class is under-represented in the training dataset, then the learned model may not be able to accurately classify it. Furthermore, it

[1] We use "AI" in the general sense – "a collection of technologies that combine data, algorithms [machine learning], and computing power." https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

might not even be aware of making mistakes as indicated by low confidence classification decisions.

To raise practitioners' awareness of the impact of the data used to train AI on model performance, in this paper, we describe a demo system designed to raise awareness of social bias in crowdsourced annotations and the AI trained with it. The key innovation of this system is presenting ML classification results in a way that the annotation sources used to train ML models is transparent to the end-user. This is a tool that can be used with/by anyone with basic data science knowledge, and can be a great "conversation starter" with students and practitioners alike.

2 THE SCENARIO

RECANT focuses on biometrics tasks aimed at identifying a person based on their characteristics – in our case, their visual appearance in an input photo. Biometrics provides an ideal scenario for the exploration of social bias in crowdsourced image data, as its tasks typically involve supervised ML and thus, rely on large training datasets, particularly given the latest advances in deep learning models [12]. RECANT illustrates the impact of labelled data on four tasks – two aimed at predicting demographic attributes (gender, race) and two that infer characteristics related to perceived traits (attractiveness, trustworthiness).

Algorithmic inference of demographic attributes can obviously benefit a number of applications. For instance, such algorithms are often used to enhance user experience, through adaptive and personalized interfaces. Likewise, they can be used to address security and privacy concerns, automating identity verification and/or access control. However, biometrics has also faced growing controversy. In particular, the Gender Shades project, which audited a number of commercial gender recognition services, found that they consistently have higher error rates on images of Black individuals, especially Black women, as compared to other demographic groups [2]. Other researchers have categorically objected to the use of an algorithm to detect social identities, which are often complex [11]. Nonetheless, biometrics remains an enabling technology in applications such airport security and screening[2] or for the identification of the dead in a warzone or in disaster hit areas.[3]

In contrast to demographic inferences, inferences on a depicted person's physical attractiveness and trustworthiness might seem more esoteric. However, these are actually processes that people do automatically when encountering someone new. According to psychologists, judging others' attractiveness is related to evolutionary motivations (i.e., choosing a mate) [3], while judging someone's trustworthiness is related to the need to protect ourselves, and can happen solely based on a photograph [5]. Although not yet mainstream, there is a slow uptake of such algorithmic processes within the human resources domain, with researchers taking note of the emerging ethical concerns [7].

In summary, biometrics was selected as a good scenario for our awareness-raising tool, given its potential to be deployed into high-stakes applications such as the above, as well as the criticism surrounding its tendencies to exhibit disparate error rates across demographic groups. Furthermore, people do not always perform accurately on biometric tasks. For instance, they may struggle to correctly identify the social identities of people whose appearances are not "typical" for their social group(s) [10]. Likewise, human judgements of others are subject to in-group favoritism [6] (e.g., describing a person who belongs to one's demographic group in more favorable terms, as compared to out-group members). Thus, it is to be expected that our crowdsourced set of image annotations will exhibit such social biases. RECANT allows exploration of their impact on the predictions of the trained biometrics tasks.

3 THE RECANT TOOL

The RECANT tool allows users to explore the ways in which the characteristics of the training dataset affect the resulting machine learning models. By selecting an image and classification task (Gender, Age, Attractiveness, Trustworthiness), the tool presents the classification results from machine learning models trained using data annotated by different groups of annotators. The RECANT tool is available online at https://recant.cyens.org.cy/.

3.1 Data, Annotations, and ML Models

RECANT uses the Chicago Face Database (CFD) [9]. The CFD contains photographs of male and female faces of different races, between the ages of 17-65, which were taken in a standardized way. Additional data including both physical attributes (e.g., face size) as well as subjective ratings by independent judges (e.g., attractiveness) are available for each subject. Specifically, the dataset consists of 597 images of female and male human subjects recruited in the United States, who self-identified as Asian, Black, Latino, or White. Each image also has associated "norming data" (i.e., labels on all attributes collected from 50 CFD annotators, who were hired and trained by the dataset creators).

We presented crowdworkers on the Clickworker [4] platform with a CFD image, asking them to identify the person's gender (Male, Female, Other) and race (Asian, Black, Latino, White[5]), as well as to rate the attractiveness and trustworthiness of the depicted person. The latter two ratings were done on a scale from 1 (minimum – "not at all attractive/trustworthy") to 7 (maximum – "extremely attractive/trustworthy"). Crowdworkers were also asked a set of demographic questions about themselves, including their gender and race. Participants received a fair payment according to the average time of the task completion and the average wage in the United States (where the crowdworkers are located). Our research protocol received ethical approval from the Cyprus National Bioethics Committee prior to commencing the study.

For each of the 597 images, we asked four crowdworkers (two males, two females) to label/rate the gender, race, attractiveness and trustworthiness of the person depicted in the image. After we collected all the data, we excluded the responses arriving from spammers (i.e., crowdworkers who did not pass our control questions). Out of the 2.388 responses received, we were left with 2.370 at the end of this process. Notice that for each image we were left with data arriving from a maximum of four distinct crowdworkers.

[2] https://www.cntraveler.com/story/how-airports-are-using-biometrics-so-you-can-spend-less-time-waiting-in-lines

[3] https://www.biometricupdate.com/202203/clearview-ai-facial-recognition-being-used-to-deliver-news-of-russian-war-dead-to-families

[4] https://www.clickworker.com/

[5] We use these terms as they are used in the CFD metadata. The CFD treats gender and race as being discrete constructs.

Crowdworkers were allowed to annotate more than one image, and in total, 388 crowdworkers participated. Among them, 52% identified themselves as male, 47% as female and 1% as other. Regarding their race, 69% identified themselves as White, 12% as Black, 9% as Asian, 6% as Latino and the remaining 4% as "none of the above."

Nine different models were trained on the same images for each task, with different (sub)sets of crowdworker annotations. The trained models include:

- A model trained using the norming data / annotations provided with the CFD ("CFD Annotators");
- A model trained using all the annotations for all images ("All Annotators");
- Two models trained using annotations that were created by crowdworkers who identified as men and women;
- Four models created by crowdworkers who identified as Black, Asian, White and Latino;
- A "random" model, which simulates the case where annotators generate labels without considering the image content.

Model training was performed using the publicly available lobe.ai tool. The lobe.ai tool allows the training and deployment of Deep Learning Networks without the need to perform a rigorous manual model optimization process, ensuring that all models under comparison are trained using the same training and model optimization procedures. However, although lobe.ai allows the export of trained models for use in the most popular deep learning libraries, it does not provide explicit details of the model architecture and/or the training algorithms used for training the models.

3.2 Walk-through

After clicking the "Get Started" button on the splash page of the RECANT website, users are presented with a sequence of choices to be made (see Figure 1). First, users are able to select an image from a subset of 20 CFD images, or to keep the default selection (Step 1). Then, users select a classification task among those available or keep the default Gender classification task (Step 2). Finally, users can execute the ML models for the selected classification task of the selected image (Step 3).

After a few seconds, users are presented with the results of the classification. They can observe the labels predicted by each of the models for the selected classification task in the table. Here, they can compare the classification decision for each model trained with annotations provided by certain human crowdworkers (i.e., Men, Women, Black, Asian, White, Latino) as well as a classifier trained with all available labels. Finally, the table also presents the random model's decision.

Users are able to continue their exploration by selecting different images and/or classification tasks. Once the user has explored three different CFD images *or* three different classification tasks, a button appears inviting the user to complete a post-experience questionnaire. The questionnaire might be configured as to assess the user experience with the tool and/or to report on the results collected through the awareness-raising exercises. In future user studies, we plan to explore how the user impression might correlated to the time spent using the demo and/or the "moves" made.

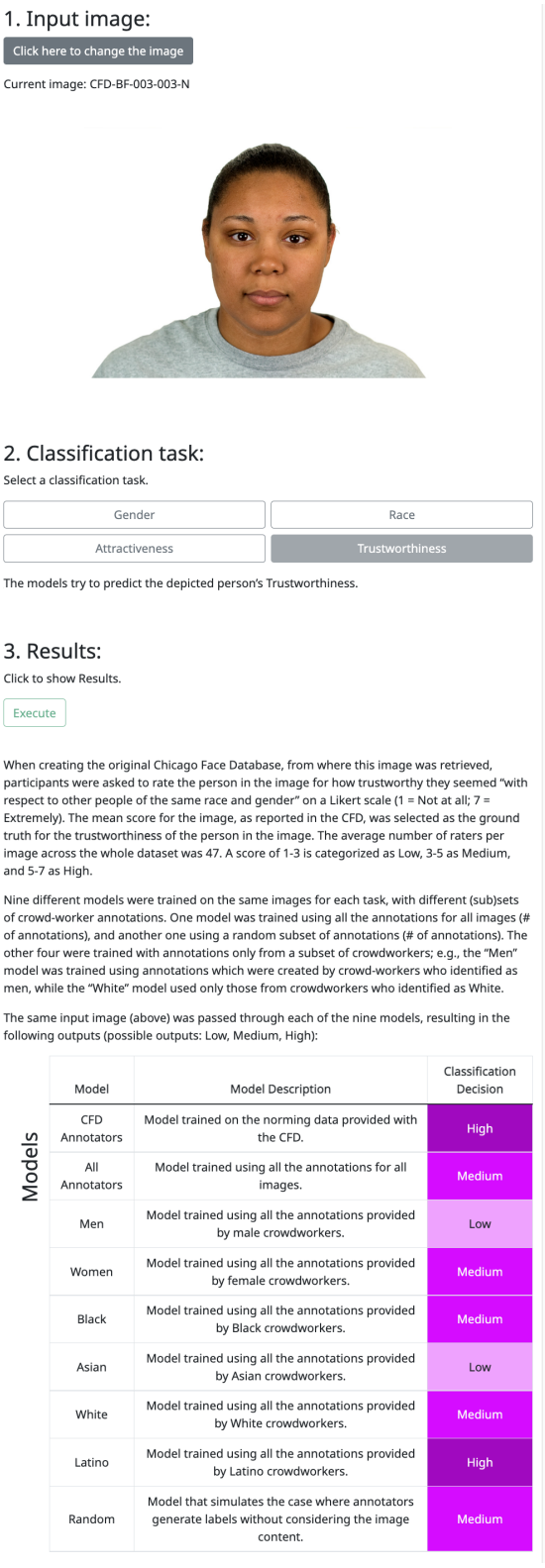

Figure 1: The user interface of the RECANT tool.

4 AWARENESS-RAISING EXERCISES

As an awareness-raising tool, RECANT can be used with students and practitioners, in the context of both individual or group settings (in a classroom or training). Here, we provide sample exercises, to illustrate the types of issues that can be explored using the tool.

4.1 Exercise 1: In-group v. out-group effects

Computer vision algorithms for predicting the gender and/or race of a person depicted in an image have been under scrutiny in recent years. For instance, they still suffer from disparate error rates across demographic groups, and due to their nature, have the effect of reducing complex human identifies down into discrete categories.

In fact, there is some evidence that people are often inaccurate on the "task" of inferring others' identities in everyday life. Furthermore, there is much literature in social psychology suggesting that we may be more favorable towards others and/or understand others better when they are members of our in-group (based on social constructs such as gender and race). Given these challenges, and previous findings, we likely anticipate that who annotates image data will have an impact on the algorithms' performance.

In the CFD models for predicting gender and race, do we find evidence that the models trained on data collected from a depicted person's demographic in-group, have better prediction accuracy?

4.2 Exercise 2: Gender differences and perceived attractiveness

There are results from evolutionary psychology suggesting that physical attractiveness is more important to men than women, and that women may be less judgmental than men in this respect [3]. For the task of predicting attractiveness, do we find evidence that using data collected from men results in better prediction accuracy, as compared to that from women (or from any worker)?

4.3 Exercise 3: Gender differences and perceived trustworthiness

When encountering someone unfamiliar, we subconsciously make decisions as to how trustworthy we believe they are, based solely on their appearance; even a picture of a face is enough to trigger these judgments [5]. Recent research suggests that facial trustworthiness is more important to women as compared to men. Based on the performance of our CFD trustworthiness models, can we draw any conclusions regarding the differences in the image data provided by women versus men workers?

5 CONCLUSION & FUTURE WORK

With the continuing "democratization" of AI and the increasing diversity of AI practitioners, it becomes crucial to raise awareness of the challenges surrounding the data practices that underlie these technologies. We presented RECANT, a tool that can be used with anyone having basic data science skills to explore the impact of crowdsourced image annotations on four biometric tasks. In the near future, we plan to explore users' experience with our tool as well as the addition of more interactive features.

The current tool represents a first attempt at illustrating, in a concrete and accessible way, the importance of training ML models on high-quality, balanced training datasets. In future iterations of the tool, we aim to make it more interactive as well as to integrate it into our educational curriculum for data science and ML students. With broader deployment of such awareness-raising tools, we might positively impact the next generation of ML practitioners, and change the culture of the data work surrounding ML.

6 ACKNOWLEDGMENTS

This project has received funding from the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT), the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 739578 (RISE), and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

REFERENCES

[1] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E Gilbert. 2021. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law* (2021), 1–17.

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

[3] Michael R Cunningham, Anita P Barbee, and Carolyn L Pike. 1990. What do women want? Facial metric assessment of multiple motives in the perception of male facial physical attractiveness. *Journal of personality and social psychology* 59, 1 (1990), 61.

[4] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. 2019. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 170–179.

[5] Andrew D Engell, James V Haxby, and Alexander Todorov. 2007. Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *Journal of cognitive neuroscience* 19, 9 (2007), 1508–1519.

[6] Feng Fu, Corina E Tarnita, Nicholas A Christakis, Long Wang, David G Rand, and Martin A Nowak. 2012. Evolution of in-group favoritism. *Scientific reports* 2, 1 (2012), 1–6.

[7] Anna Lena Hunkenschroer and Christoph Luetge. 2022. Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics* (2022), 1–31.

[8] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 467–471.

[9] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47, 4 (2015), 1122–1135.

[10] Nicholas O Rule and Shelbie L Sutherland. 2017. Social categorization from faces: Evidence from obvious and ambiguous groups. *Current Directions in Psychological Science* 26, 3 (2017), 231–236.

[11] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[12] Kalaivani Sundararajan and Damon L Woodard. 2018. Deep learning for biometrics: A survey. *ACM Computing Surveys (CSUR)* 51, 3 (2018), 1–34.