Document Title	**Notes and materials generated from consultations with stakeholders**
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.	D6.2
Work package No.	WP6
Work package title	Consultation with Stakeholders
Authors (Name and Partner Institution)	E. Christoforou (CYENS) M. Kasinidou (OUC)
Contributors **(Name and Partner Institution)**	P. Barlas (CYENS) K. Orphanou (CYENS - OUC) J. Otterbacher (CYENS)
Reviewers	S. Kleanthous (CYENS - OUC)
Status **(D: draft; RD: revised draft; F: final)**	F
File Name	D6.2_Material_from_Consultations.docx
Date	30 July 2022

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**
v1.0	1/5/22	M. Kasinidou	maria.kasinidou@ouc.ac.cy	Initial document
v2.0	31/5/22	E. Christoforou	e.christoforou@cyens.org.cy	Updated with session design
v3.0	16/6/22	E. Christoforou	e.christoforou@cyens.org.cy	Review version
v4.0	30/7/22	E. Christoforou	e.christoforou@cyens.org.cy	Final version

Abstract

This final deliverable of WP6 documents the process used for our stakeholder consultations, as well as the main findings that were generated. Specifically, it includes the materials generated for the consultations as well as the highlights of our discussion with the practitioners.

Keyword(s): Guidelines, Machine learning practitioners, Roundtable consultations, Stakeholder feedback

Contents

1. About this Deliverable

This deliverable provides a record of the stakeholder consultations that were conducted in the context of WP6 - *Consultation with Stakeholders*. In particular, we first provide a description of the stakeholders who were recruited and participated in our consultations. We then describe the agenda for the roundtable consultations, along with the material that was presented. Finally, we report on the conversations and discussions that took place during the consultations, and the feedback on the practitioner guidelines (presented in D6.1).

2. Description of the stakeholders

This section provides a description of the stakeholders that participated in our consultation meetings. As described in WP6, we aimed to engage with machine learning (ML) practitioners from the local industry, who are using crowdsourcing in their work either directly (as requesters using crowd platforms to create datasets) or indirectly (by using datasets created by others via crowdsourcing). Our aim was two-fold. Firstly, we wanted to share the key practical findings from DESCANT that resulted in our practitioner guidelines. Secondly, we aimed to understand the challenges faced by practitioners, and solicit their feedback regarding the set of guidelines.

Via the CYENS network, we reached out to ML practitioners in the local community, with an emphasis on those coming from small, start-up companies. This choice was made in line with our emphasis on the “Democratization of AI" and the growing inclusion of practitioners from a wide variety of educational and professional backgrounds in the creation and use of AI tools and practices. In other words, we avoided approaching those who work in larger and/or multi-national companies, who likely enjoy more support for their ML needs and who likely have incorporated more standarized methodologies and tools into their work practices.

2.1 Description of the Practitioners

Before the roundtable consultation session, the ML practitioners (participants) were asked to complete a brief questionnaire (see Annex 1), which collected demographic information and assessed their experience with crowdsourcing and/or crowdsourced datasets. Finally, it also aimed to gauge their beliefs concerning the prevalence of bias in ML algorithms and training datasets, as well as the origin of such biases. The demographic information of the participants as collected from the questionnaire is displayed in Table 1.

Participant No.	Gender	Age	Profession / Industry	Education level
1	M	35-44	Banking industry	PhD
2	M	25-34	CS Academic/Start-up	PhD
3	M	25-34	CS Academic	PhD candidate
4	M	35-44	Consulting industry	Msc
5	M	35-44	Algolysis (start-up company)	PhD
6	M	35-44	Algolysis (start-up company)	PhD
7	M	35-44	Ceranext	PhD

Table 1: Basic information on the practitioners participating in our consultations.

Based on their responses to the questionnaire, most of the participants have many years of experience with ML algorithms working on several projects and multiple application domains such as bioinformatics, banking, biometrics, cybersecurity, health, criminal - justice. However, only one of the participants used crowdsourcing to collect data directly. He reported using the collected data to build participatory services for trajectory similarity and crowdsourced occupancy/congestion count.

Almost all of the participants reported that they believe that the *input data* is the main reason for experiencing bias in the output of the developed algorithms which, as they all mentioned, is the most difficult type of bias to handle. Some of the actions they mentioned for minimizing unwanted bias in the developed algorithms is: a) to study and understand the input data, b) to develop a framework to identify, understand and address the bias, c) to build a system (or develop an algorithm) using also techniques to mitigate bias and algorithmic fairness and d) to use an inclusive way to collect input data with the minimal bias as is possible. Moreover, the participants believed that the impact that a biased output might have is an important factor to consider before deciding how to handle/minimize bias in the system.

Despite our efforts to reach out to women practitioners, we must note that all of our participants were men. Participants, who volunteered their time for our study, also likely had a previous interest in the topics of data quality / data bias. While our group was not required to be representative of the local population of practitioners in order to solicit helpful feedback, we do have in mind to extend our future efforts in engaging a broader audience of stakeholders.

3. Session Agenda and Material

Two roundtable consultations were organized. In order to best accommodate our participants' schedules, but also, to keep each group relatively small, two groups were formed. Four of our team's researchers participated in both sessions; two researchers facilitated the sessions while two more participated as observers and were responsible for keeping notes during each session. Before each session, the participants were asked to complete a brief questionnaire (see Annex 1), which collected demographic information and assessed their experience with crowdsourcing and/or crowdsourced datasets. This questionnaire also helped in giving practitioners an idea of the session's content.

The agenda for each session was as follows:

- Welcome and introductions
- Introductory video to the DESCANT project
- Presentation of the RECANT demonstration tool
- Exercise with the demo (concerning the in-group v. outgroup effect)
- Discussion on the DESCANT Practitioner Guidelines
- Final comments

Annex 2 presents the slides used to facilitate the session. In the next Section, we discuss in depth the DESCANT Demonstration Tool as well as the exercises introduced to the practitioners during the session to help them familiarise with the tool.

4. Exploring Social Biases in an Crowdsourced Image Dataset: The DESCANT Demonstration Tool

In this section, we provide some exercises, using the DESCANT Demonstration Tool (D5.3), which are aimed at helping practitioners better understand the strong influence that a training dataset has on the performance of a machine-learned algorithm. The Tool is accessible online: https://recant.cyens.org.cy/

Image Dataset. The Chicago Face Database[1] (CFD) was developed at the University of Chicago by a team of social psychologists. It provides high-resolution, standardized photographs of male and female faces of varying ethnicity between the ages of 17-65. Extensive norming data are available for each individual model. These data include both physical attributes (e.g., face size) as well as subjective ratings by independent judges (e.g., attractiveness). The main CFD set consists of images of 597 unique individuals. They include self-identified Asian, Black, Latino, and White female and male models, recruited in the United States. All models are represented with neutral facial expressions. A subset of the models is also available with happy (open mouth), happy (closed mouth), angry, and fearful expressions. Norming data are available for all neutral expression images. Subjective rating norms are based on a U.S. rater sample.

Crowdsourcing. We asked crowdworkers on the Appen platform, who were based in the U.S., to analyze a subset of the attributes that were studied by the CFD researchers. In our HIT, the workers were first asked to identify their own demographic attributes, in the same manner that CFD models had been asked to do. Then, they were shown one CFD image. They were asked to indicate, via closed-form responses, the gender and race of the depicted person, as well as their perceived attractiveness and trustworthiness. For these last two attributes, workers indicated their responses from 1 (low) to 7 (high); however, we then converted these scores to responses of low, medium, and high, as to train a three-class prediction model.

[1] https://www.chicagofaces.org/

Machine learning models. In the spirit of using tools that promote "AI Democratization," we trained models using the lobe.ai tool. By using open-source machine learning architectures, the lobe.ai tool is able to automate Deep Learning classification tasks without the need to perform a rigorous manual model optimization process, ensuring that all models under comparison are trained using exactly the same training and model optimization procedures. Furthermore, the lobe.ai tool is able to achieve excellent performance at low computational costs.

Model Results. The tool provides a visualization of the results we obtained when applying our crowdsourced image dataset to four different computer vision / biometrics tasks: i) prediction of the depicted person's gender; ii) predicted race; iii) perceived level of attractiveness; iv) perceived level of trustworthiness.

The user can explore the predictions for 20 CFD images, when we train the models on eight different (sub)sets of crowdworker annotations: i) using all data, ii) using all data provided by workers identifying as men; iii) using the data from workers identifying as women; iv) based on the data from workers identifying as Black, v) Asian, vi) White, vii) Latino, and finally, viii) a random sample of the data.

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 disproportionate error rates across demographic groups,[2] and due to their nature, reduce 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[3]. 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)[4]. 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?

Exercise 2: Gender differences and perceived attractiveness

In a similar vein, there are results from evolutionary psychology suggesting that physical attractiveness is more important to men than women, and that women may be less judgemental than men in this respect.[5] For the task of predicting a person's attractiveness, do we find evidence that

[2] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

[3] Rule, N. O., & Sutherland, S. L. (2017). Social categorization from faces: Evidence from obvious and ambiguous groups. *Current Directions in Psychological Science*, *26*(3), 231-236.

[4] Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., & Nowak, M. A. (2012). Evolution of in-group favoritism. *Scientific reports*, *2*(1), 1-6.

[5] Cunningham, M. R., Barbee, A. P., & Pike, C. L. (1990). What do women want? Facialmetric assessment of multiple motives in the perception of male facial physical attractiveness. *Journal of personality and social psychology*, *59*(1), 61.

using data collected from men, results in better prediction accuracy, as compared to that from women (or from any worker)?

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[6]. Recent research suggests that facial trustworthiness is more important to women as compared to men[7]. 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?

4. Consultation with Stakeholders

In this section, we present a general account of our interactions with the stakeholders and report on the highlights of our consultation with them, documenting the key themes that emerged from the discussions.

Feedback on the RECANT Demonstration tool

Practitioners in both sessions showed an interest in the demonstration tool and provided us with feedback on improving it. One of the stakeholders' comments regarding the RECANT demonstration tool was the necessity for including the ground truth and the confidence scores for the classification outcome. The practitioner mentioned that in his opinion it is very important to have the ground truth value in order to measure the performance of the models by measuring the confidence level of the model.

Another comment we received was that more detail regarding the annotation task could be given in the demo itself like the number of annotators, the demographics of the annotators i.e. race etc, the data collection process i.e "When you ask the workers, you ask them for the same image to give you the labels of the different categories."

Data quality and ground truth

Given the comments on the demonstration tool, the topic of dealing with datasets where the ground truth can be obtained, as opposed to datasets where only subjective judgements can be acquired, naturally emerged. In that respect, our participants (i.e., machine learning practitioners) who train their models on data collected from experts insisted that dealing with datasets where the ground truth can be obtained is the only kind of data that one can have to train algorithms since accountability and transparency towards the client of the AI algorithm is important for them.

Alongside the above topic, that of data quality emerged. Practitioners were concerned about the quality of data collected from the crowdsourcing task and whether the stakeholders will prefer crowdsourcing for collecting data to build a classification model especially when the ground truth

[6] Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *Journal of cognitive neuroscience*, *19*(9), 1508-1519.

[7] Duan, Y., Hsieh, T. S., Wang, R. R., & Wang, Z. (2020). Entrepreneurs' facial trustworthiness, gender, and crowdfunding success. *Journal of Corporate Finance*, *64*, 101693.

values are absent. Regarding the quality of data, one of the stakeholders mentioned that it is important to be strategic in terms of crowdsourcing, targeting the workers who give the best data. Another stakeholder stated that we should get data from a diverse group of workers and design the HIT such as to get the best quality data, i.e., change the questions based on the group of workers. The above discussions aided in the presentation of our practitioners guidelines, where in practice we received an affirmation regarding the value and timeliness of our guidelines.

Nature of the crowdsourcing task carried out for the needs of RECANT

As far as concerns the use of crowdsourcing for images annotation, there was a disagreement between the stakeholders. One of the stakeholders appeared to be categorically against these types of tasks, mentioning that it is easy to offend people and/or discriminate by asking crowdworkers to annotate the images of people. A few other stakeholders felt that if we can get large numbers of workers and control the quality, it is acceptable/feasible. However, another practitioner noted that, even if a large sample is more appropriate, someone will have to devote time to get more data and this is something that is difficult for both people who are building ML models and performing crowdsourcing research and might face a lot of restrictions (i.e., economic limitations).

General perception on crowdsourcing as a mean for data generation

One of the stakeholders mentioned that the cost of auditing the dataset is a key concern, which will influence whether or not he would use crowdsourcing. Another issue is that of the application domain. For example, if someone wants data that are representative to train the model then the data needs to be close to the ground truth. However, in two of the cases showcased in our demo (prediction of one's trustworthiness and attractiveness), crowdsourcing is not the best solution.

In the cases where the ground truth of the dataset is missing, a few stakeholders mentioned the use of automated methods such as distant supervision tasks or asking a small group of experts to provide the ground truth. The use of a diverse group of crowdworkers with different backgrounds/experiences was also an option, but rather than using crowdsourcing for discovering ground truth, one can use it to model the opinions of people from different perspectives. The majority of applications are mainly human-centric (recommendation, asking their opinion). Building and training localized models from data generated from different groups of crowdworkers would also be beneficial in many applications.

Two of the stakeholders who worked on the Coronasurveys project used that as an example. They stated that in that project, people (i.e., members of the crowd) were reporting information for others rather than for themselves and the focus was not to find out if they are biased. In general, it was noted that in many ML classification models, we do not know how the algorithm might be trained and if all the findings are correct. Hence, it was confirmed by the participants that it is important to create guidelines but not explicit rules.

Finally, all the stakeholders agree that transparency on the data collection and training of the ML model is very important, especially in industry where customers want to know how the model arrives as a given output. Transparency on building ML models can be achieved through transparent models or explainability in AI techniques but also, as was emphasized in the guidelines and in the discussion held with the practitioners, proper documentation of the crowdsourcing procedure is fundamental.

Feedback on the guidelines

For completeness, we provide here the feedback received from the practitioners on the guidelines as was documented in D6.1 (Section 5).

In general, we can classify our practitioners in two categories. The first consists of machine learning practitioners that train their models on data collected from experts and, according to them, can obtain in a large degree the ground truth of their datasets. The second category comprises practitioners that are part of the industry but also have links to research and collect data from crowdworkers.

The guidelines were discussed in the general spirit of the RECANT demonstration tool (documented in D5.3) that accompanied the session material and aided the interaction and "example-oriented" presentation of the guidelines. Additionally, we asked the practitioners to compare the feasibility and applicability of the guideline to their experiences with crowdsourced datasets. One practitioner noted that the guidelines for dataset generation via crowdsourcing follows a set of instructions that have the right "flow" between them and are in line with actions he uses when creating datasets. Both categories of practitioners noted that there is value in creating such guidelines, especially in recent days where transparency in AI is becoming a necessity. To this extent practitioners also asked for a copy of the guidelines for future reference.

Furthermore, one practitioner belonging to the second category noted that such a set of guidelines are of real value and in the near future, once the EU AI Act will be activated, there will be the need for auditing the use and creation of data. On this last point, another practitioner emphasised that our set of guidelines present a very valid point: "data must be documented thoroughly" in order to be replicated and re-used in the future. On this topic, a lively discussion took place, affirming our observation throughout the project that researchers and practitioners take little, if any, notice of the documentation that accompanies a dataset. In fact, the set of guidelines we have established for use of Datasets Generated via Crowdsourcing is partly not applicable if a detailed documentation of the datasets and the process of using crowdsourcing to obtain the datasets is not appropriately documented. Thus, it was made obvious during the second session with our practitioners that there is a real need for disseminating these set of guidelines and educating practitioners in the field.

All in all, practitioners in both sessions that we held agreed that the discussions that emerged during the sessions were very interesting and they mentioned that they were willing to participate in similar efforts in the future.

5. Conclusions

This deliverable includes all the material used during the consultation sessions together with the feedback received from the practitioners. We believe that guidelines such as the ones created in DESCANT will be of value and sought out by practitioners in the very near future, as once we have the EU AI Act, it is probable that there will be the need for auditing the use and creation of the datasets accompanying the AI algorithms. Hence, this deliverable will be a good starting point when presenting the created guidelines. It still remains to be seen how guidelines like the ones we have created can be used to educate a larger, more diverse and geographically sparse audience.

Annex 1 - Pre-consultation questionnaire

1. What is your gender?*
    - Male
    - Female
    - Other:
2. What is your race?*
    - White
    - Black
    - Asian
    - Latino/a
    - Other:
3. What is your age?*
    - 18-24
    - 25-34
    - 35-44
    - 45-54
    - 55-64
    - 65+
4. What is your profession?*
    - Student with studies relevant to AI
    - Student with studies relevant to Computer Science
    - Student
    - ML practitioner
    - Other
5. What is your education level?*
    - Elementary
    - High-school graduate
    - Bachelor or equivalent
    - Master or equivalent
    - Doctoral or equivalent
6. Do you have experience with machine learning (ML) algorithms?*
    - Yes
    - No
7. Can you elaborate briefly on your experience with ML algorithms (i.e., years of experience, field, etc.)

8. Would you say that you are working with critical applications in your line of work?*
    - Yes
    - No
9. If you replied "Yes" to the question above, could you elaborate on the critical applications you are working on:
    - Health
    - Justice/Criminal
    - Mobility
    - Other
10. Do you have experience with biometrics tasks?*
    - Yes
    - No
11. Can you elaborate briefly on your biometrics experience? (i.e., years of experience, specialization, etc.)
12. Do you have experience collecting data through crowdsourcing?*
    - Yes
    - No
13. Can you elaborate briefly on your crowdsourcing experience? (i.e., years of experience, platforms used, etc.)
14. Have you ever developed an algorithm (or ML model) that you suspected had biased output?*
    - Yes
    - Maybe
    - No
15. Have you ever developed an algorithm that you suspected might propagate undesirable social stereotypes?*
    - Yes
    - Maybe
    - No
16. What was the main reason for experiencing bias in the output of your designed algorithms?*
    - Input data
    - ML model
    - Developer's bias
    - Mismatch between algorithmic output and intended audience

- Other

17. In your professional opinion, from the reasons listed below for bias to emerge, which is the most difficult to handle/address?*
    - Input data
    - ML model
    - Developer's bias
    - Mismatch between algorithmic output and intended audience
    - Other
18. What actions - if any - did you take to minimize unwanted bias in your algorithms? Please elaborate briefly
19. If you were asked to design an algorithm that had minimal bias, what do you think is the most important thing to address first?

Annex 2 - Slides used during the sessions