Document Title	**Practitioner Guidelines**
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.	D6.1
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) J. Otterbacher (CYENS/OUC)
Reviewers	S. Kleanthous (OUC)
Status (D: draft; RD: revised draft; F: final)	F
File Name	D6.1_Practitioner_Guidlines.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	10/1/22	E. Christoforou	e.christoforou@cyens.org.cy	Initial document
v2.0	31/3/22	M. Kasinidou	maria.kasinidou@ouc.ac.cy	Updated with draft guidelines
v3.0	20/6/22	M. Kasinidou	maria.kasinidou@ouc.ac.cy	Post-session draft
v4.0	23/6/22	E. Christoforou	e.christoforou@cyens.org.cy	Review version
v5.0	30/7/22	E. Christoforou	e.christoforou@cyens.org.cy	Final version

Abstract

This deliverable describes the development of practical guidelines aimed at helping machine learning practitioners mitigate social bias in manually-labelled datasets. We focus specifically on datasets collected and/or enhanced through crowdsourcing, and that are meant for supervised machine learning tasks. The guidelines represent our efforts to “translate” the state-of-the-art research findings and the experience gained in WP4 on the topic of crowdsourcing and social bias, into easy-to-use guidelines.

Through a set of consultation meetings with the practitioners we have assessed the practical contribution of these guidelines and we present here a summary of the conclusions extracted from these meetings. Particular emphasis is given to how these guidelines were received by the practitioners, which can serve as an indication as to how these guidelines could be received by a larger community of experts and stakeholders.

Keyword(s): Practitioners, Crowdsourcing Platform, Monetary Reward, Human Intelligence Task (HIT), Requester, Worker, Crowdsourced data, Bias, Demographics.

Contents

1. About this Deliverable

Work package 6 of the DESCANT project aims to generate impact from the project's research results, through the development of an easy-to-use set of guidelines for machine learning practitioners on how to *minimize the problem of social bias in manually-labelled datasets.* The guidelines comprise lessons learned from our research work within the DESCANT project, as well as state-of-the-art research, articles, and other published work in the field from experts, cited where appropriate.

The deliverable is structured as follows:

- Section 2 provides an introduction, which motivates the need for our work on mitigating social biases in machine learning datasets that have been labelled via crowdsourcing. It concludes with a brief overview of the paid, micro-task crowdsourcing framework as well as a presentation of the most common types of biases that have been discussed in the research to date.
- In Section 3, we present an initial research study carried out in DESCANT, aimed at exploring the knowledge - and also attitudes - of computer science and data science students surrounding ethical issues in machine learning (including data practices). The aim was to assess the extent to which these *future practitioners* have already been exposed to key concepts and principles surrounding AI ethics, when entering the workforce.
- Section 4 presents the guidelines that have been developed, based on our survey of the relevant literature as well as our own DESCANT research.
- The deliverable concludes with Section 5, which describes the key insights gained from the stakeholder engagements. (Please see D6.2 for the description of this process, as well as the notes and materials that were generated during our stakeholder engagements.)

2. Introduction

This section provides background information that motivates the need for our guidelines on the mitigation of social bias in crowdsourced training datasets for machine learning. In particular, we explain the link between the Democratization of AI and the wider availability for platforms and tools for the creation and use of AI components - including data - to the rise of ethical issues, including social bias.

2.1 The Democratization of AI

Recent years have witnessed a dramatic increase in the use of data-driven artificial intelligence[1] (AI) in products and services used across the public and private sectors. For instance, Forbes describes an "AI Revolution" that is driving the Fourth Industrial Revolution.[2] This revolution

[1] Note that we use the term *artificial intelligence* (AI) in the general sense put forward in the European Commission's White Paper on AI, "*a collection of technologies that combine data, algorithms, and computing power*." https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[2] https://www.forbes.com/sites/forbestechcouncil/2021/08/25/the-ai-revolution-is-happening-now/?sh=72b209f028c8

depends on rapid adoption of AI technologies by organizations across sectors; to this end, mechanisms to facilitate *participation and uptake* have become crucial.

In the earlier years of the Internet, access to tools, platforms, and data sources determined who would be able to participate in the Artificial Intelligence field. If a person (or organization) did not have the required infrastructures (e.g., machines with high processing power, large datasets for training), know-how in machine learning, and/or time and resources to invest in obtaining them – or be a member of a closed group with access to certain platforms and tools – the person or organization simply would not be able to build an AI application.

However, these days it is possible to find large datasets online for free, follow online tutorials about techniques required to train an AI model, and even "rent" the processing power required to train the model. It is even possible, if the goal of the model is commonly sought in the industry, to find a pre-trained model offered by a trusted, big technological company, and pay a very low fee per data point to get the necessary outputs – without any effort or knowledge regarding the process to train an algorithm. *Cognitive Computing* refers to the use of AI components (e.g., concepts, algorithms, datasets) on the large-scale, but also a strategy for deployment.[3] While industry giants have used cognitively inspired algorithms for years, they are now available to third parties as "AI as a Service." Microsoft describes "democratiz[ing] AI by packaging it into discrete components that are easy for developers to use in their own apps."

This increasing ease of access to AI applications/methods is lately referred to as "AI Democracy" (and the process by which it happens, the "Democratization of AI"), as this trend has started to shift the power from a small elite to the broader society. However, despite the use of this term, the other key elements of a political democracy – namely the protection of people's freedom, and the access to social benefits for everyone – are arguably not reflected in "AI Democracy," or at least not in the current AI market.[4] This is increasingly apparent in the efforts that try to make AI "more fair" – the applications more trustworthy, and the outputs more beneficial for a wider group of people. One crucial question must be asked, however – are these efforts also reaching the decentralized groups (or even individuals) who now have access to the techniques/tools required to build AI systems?

2.1 Bias in AI - The challenges of data bias

Ethical concerns about AI, and in particular, algorithmic bias, have taken center stage on the agenda for stakeholders in industry, government and researchers. There is widespread recognition that AI should behave in a manner that respects human values, and that developers and owners should mitigate the potential for their applications to bring about harm in the social world. There is also growing public concern, as cases of data-driven AIs and their *social biases* are frequently discussed in the press and on social media. Such cases concern a wide range of domains and applications, for example:

[3] Kelly III, J., & Hamm, S. (2013). *Smart Machines*. Columbia University Press.

[4] Clough & Otterbacher.

- **Automated decision support**, e.g., used in the judicial system[5], the financial sector[6], in educational assessment[7], for human resources decisions[8];
- **Information filtering and access**, e.g., Web search engines[9], social media[10], music recommendation engines[11];
- **Computer vision**, e.g., face and gender recognition[12], image tagging[13], photo cropping[14];
- **Natural language processing**, e.g., chatbots[15], search auto-complete[16], translation[17].

How did we arrive at the point of having so many AI applications behaving so badly? Drawing an analogy to the concept of *technical debt* in software engineering, Cristianini[18] describes the *ethical debt* of modern AI. He explains that the unintended problems of AI are a direct result of the technical shortcuts taken in developing the new paradigm of data-driven AI. Three key shortcuts are described: i) the emphasis on Big Data and machine learning (ML) models, which are based on correlation rather than causation; ii) the use of data captured in the wild rather than bespoke training datasets; iii) the use of proxies and implicit feedback to infer what users want.

Thus, the quality of training data, and the extent to which it captures and reflects the aspects to be modelled by an algorithmic process, are core issues surrounding the problem of AI bias. Data is obviously a crucial resource for training models via machine learning. Olteanu and colleagues, in a comprehensive survey paper, provide a framework for understanding the challenges of using large datasets captured via social media and other "Big Data" sources.[19] As expected, they cite data processing - or the way that data is cleaned, enriched and aggregated, as being a potential place for human bias to make its way into a dataset. One of the most common ways in which datasets are enriched with human judgements is crowdsourcing, the involvement of people who are typically not experts in data science or machine learning, in completing small, well-defined data augmentation tasks.

The DESCANT project has specifically considered the issue of social bias in crowdsourced datasets in an effort to provide guidance on how we can mitigate these issues. Specifically, we focus on

[5] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of Data and Analytics* (pp. 254-264). Auerbach Publications.

[6] https://www.technologyreview.com/2020/12/04/1013068/algorithms-create-a-poverty-trap-lawyers-fight-back/

[7] https://www.theguardian.com/education/2020/aug/20/england-exams-row-timeline-was-ofqual-warned-of-algorithm-bias

[8] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[9] https://www.theguardian.com/technology/2016/jun/09/three-black-teenagers-anger-as-google-image-search-shows-police-mugshots

[10] https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html

[11] https://www.thetimes.co.uk/article/spotifys-sexist-algorithm-prefers-to-recommend-male-musicians-rhdn3fqn5

[12] 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.

[13] https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[14] https://www.wired.com/story/twitters-photo-cropping-algorithm-favors-young-thin-females/

[15] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

[16] https://www.wired.com/story/google-autocomplete-vile-suggestions/

[17] https://theconversation.com/online-translators-are-sexist-heres-how-we-gave-them-a-little-gender-sensitivity-training-157846

[18] Cristianini, N. (2019). Shortcuts to artificial intelligence. In *Machines We Trust*. Cambridge, MA: MIT Press.

[19] Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13.

paid, microtask crowdsourcing via platforms such as Amazon's Mechanical Turk[20], Appen[21], or Clickworker[22]. Such platforms have also contributed to the Democratization of AI, by enabling groups and individuals to design and deploy a human intelligence task (or "HIT"), in an easy and affordable way. In other words, crowd platforms enable ML practitioners and researchers to get over the data bottleneck.

Figure 1 illustrates the role of micro-task crowdsourcing, in the context of machine learning, for the purpose of developing an intelligent AI system or application. As can be seen, the key stakeholders are the requester (i.e., the practitioner who wants to develop a dataset), the system or process (i.e., the machine learning task(s) to be trained), and the crowdworkers. Crowdworkers may be motivated to participate in HITs for a variety of reasons[23], from personal interest to financial gain; in the case of paid, micro-tasking, they are paid a pre-defined amount per completed HIT.

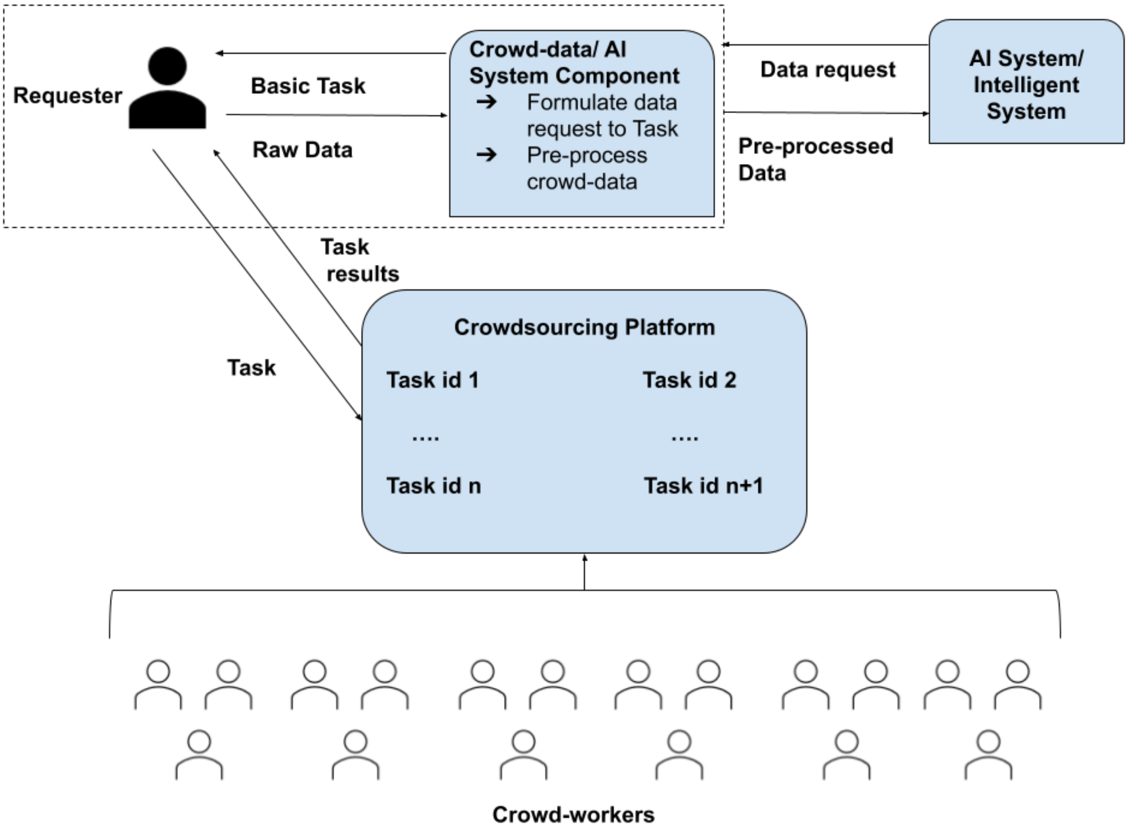


Figure 1. Crowdsourcing process and main actors when creating data for an AI system.

Although the Requester formulates the task and provides instructions to the Workers, there is the potential for the data to contain human biases or perpetuate social stereotypes. Below, we provide

[20] https://www.mturk.com/

[21] https://appen.com/solutions/crowd-management/

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

[23] Posch, L., Bleier, A., Lechner, C. M., Danner, D., Flöck, F., & Strohmaier, M. (2019). Measuring motivations of crowdworkers: The multidimensional crowdworker motivation scale. *ACM Transactions on Social Computing*, *2*(2), 1-34.

a short description of some of the most common or critical biases generated via paid, micro-task crowdsourcing, which have been discussed in the previous literature.

Demographic biases:

Crowdworkers may behave and annotate data differently, based on their own demographic characteristics. Depending on the nature of the task, the background of a Worker may influence how he or she approaches the task. For example, Dong and Fu[24] compared the manner in which people described the content of digital images, using word-tags. This represents a very common task at platforms such as MTurk, which has been used to create important image datasets such as ImageNet[25]. What Dong and Fu found was that workers' approach to tagging images differed along the lines of their ethnicity. In particular, in their study, American crowdworkers tagged the main object first, while Chinese workers provided tags about the overall properties of the image first.

Cognitive biases:

Some HITs may ask workers to make judgements that have a subjective element. In those cases, workers may create their own "subjective social reality."[26] There are several types of cognitive biases that might affect the quality of judgement provided in a HIT. For instance, priming biases are those that result from a worker having been influenced by the first presentation of some information (e.g., information found during completion of a first HIT, which influences what she does in later HITs, assuming that she is permitted to complete multiple tasks). Another common cognitive bias is implicit bias, or stereotyping. Particularly when approaching a HIT that relates to the social world, the Worker's unconscious biases might affect her responses. For instance, imagine a sentiment analysis task, in which the Worker is asked to indicate if a given sentence carries a positive, negative or neutral valence. If the text is written in a dialect (e.g., Black American English) rather than formal language (e.g., Standard American English), a Worker may be more likely to label the text as being negative, reflecting underlying social stereotypes[27]. Another example might be an image labelling task, where Workers are asked to provide labels or a caption. Gender and racial stereotypes are often reflect here (e.g., given an image depicting a man and a women, the Workers may be more likely to describe the man as being in a position of power, and/or to describe the physical attractiveness of the woman)[28]. Similarly, as will be discussed in the next section, major societal events can impact the workers' attitudes and annotations, which can create biases in the data.

Temporal variations:

As mentioned above, major societal events can impact workers' attitudes and annotations. For example, Christoforou and colleagues found that the same HIT run a few years later displayed

[24] Dong, W., & Fu, W. T. (2010, April). Cultural difference in image tagging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 981-984).

[25] https://www.image-net.org/

[26] Eickhoff, C. (2018). Cognitive Biases in Crowdsourcing. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18). Association for Computing Machinery, New York, NY, USA, 162–170.

[27] Shen, J. H., Fratamico, L., Rahwan, I., & Rush, A. M. (2018). Darling or babygirl? investigating stylistic bias in sentiment analysis. *Proc. of FATML*.

[28] Van Miltenburg, E. (2016). Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*.

differences that were indicative of the major societal events during the period.[29] Namely, shortly before the second HIT, the COVID-19 pandemic had started along with social unrest in the U.S. around racial injustice. As a result, research showed that the workers' responses contained more references to the depicted person's racial identity as well as their health.

Annotation sparsity:

Often, HITs are set up such that one worker can, or is expected to, annotate data multiple times. However, most workers only do a small number of annotations, which can result in missing or imbalanced data[30].

3. Initial investigation with future practitioners

As mentioned previously, before designing our practitioner guidelines and the subsequent round of consultation with stakeholders, we wanted to understand the perception that future practitioners (i.e., students in the fields adjacent to computing) hold about algorithmic (un)fairness and its sources. This has allowed us to create more focused guidelines and pay extra attention to the practitioners' views and the challenges they face, which can potentially be responsible for introducing a form of bias in the applications developed.

We conducted an online survey with student participants across regions to investigate how they perceive fairness, accountability, transparency, and ethics (FATE) in algorithmic decision-making.[31] The survey instrument first focused on assessing participants' definitions of algorithmic fairness. Next we used the scenario of a hypothetical recruitment system (i.e., decision support system) to explore their views around the following: i) what they believe to be the underlying causes of unfair/biased decisions, ii) ways to make such a system more transparent to its users, iii) in the event of an unfair outcome, who should be held accountable.

Our results indicated that participants identify the use of sensitive attributes, like demographics, as the most probable cause of algorithmic unfairness. It was interesting to observe that few participants referred to the system or model, while fewer still mentioned the *training dataset*. This indicates a superficial understanding of how biases end up in an algorithmic system. The focus of the participants (i.e., young practitioners) to sensitive person attributes, without considering the rest of the dataset (such as seemingly non-sensitive variables that could act as a proxy for sensitive ones), or the way these attributes are handled/manipulated by the algorithm, can be blind siding, impeding them from getting to the bottom of the problem. Hence, this study is yet another confirmation of the importance of research and educational initiatives surrounding data biases. In particular, it is crucial to inform both current but also future practitioners concerning the ways in which bias can

[29]Christoforou, E., Barlas, P., & Otterbacher, J. (2021, May). It's About Time: A View of Crowdsourced Data Before and During the Pandemic. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

[30] Ipeirotis, Panagiotis G., Demographics of Mechanical Turk (March 2010). NYU Working Paper No. CEDER-10-01, Available at SSRN: https://ssrn.com/abstract=1585030

[31] Kleanthous, S., Kasinidou, M., Barlas, P., & Otterbacher, J. (2022). Perception of fairness in algorithmic decisions: Future developers' perspective. *Patterns*, *3*(1), 100380.

be introduced and propagated in the algorithm through training datasets. DESCANT, of course, focuses on the issue of using crowdsourcing to create such datasets.

Furthermore, the study participants most often discussed the importance of having “objective factors”, i.e., not including subjective judgements in the model, in order to have a more fair algorithm. Similarly, they remarked that the fairness judgements (e.g., whether or not the algorithmic screening process that determined who would be called for an in-person interview issued fair decisions) are context-dependent. In particular, such fairness judgements depend on factors such as the end-application and the application’s own context. In addition, when discussing the potential strategies for promoting transparency, very few opted for disclosing the training data – often, they chose to explain the algorithm or the output.

As a general remark, the study showed that current students require more, targeted education around algorithmic fairness, accountability, transparency, and ethics (FATE) issues. The current guideline aims to provide a compilation of best practices both for generating datasets through crowdsourcing and for using such datasets in an attempt to bridge part of the gap among the current perceptions of practitioners’ on FATE issues and the actual challenges faced during development and employment of their machine learned algorithms.

4. Practitioner Guidelines

This section presents an outline together with a description of best practices that we have gathered from our experience during the DESCANT project. This includes lessons learned while surveying the literature, as well as while crowdsourcing our own datasets and experimenting with the use of some of the collected datasets in training machine learning algorithms.

This practical guideline is composed of two parts; the guidelines for dataset generation via crowdsourcing and the guidelines for use of datasets generated via crowdsourcing. Practitioners are often obligated by the circumstances to use crowdsourced datasets that were created for purposes other than the one for which they intend to use them, thus *repurposing* them. In this event, knowing how to use these datasets without propagating or augmenting unwanted bias is as important as knowing how to create the dataset from scratch.

The practitioner guidelines are meant to encourage the deliberation, discovery, and conscious management of biases in datasets created via crowdsourcing on a case-by-case basis. The reader must use these guidelines with caution, identifying first of all the specific parameters involved since what is considered a (wanted or unwanted) bias in a dataset, is often context- and application-dependent. In Figure 1, we provide once more a view of the crowdsourcing process and the main actors when creating data for an machine learned (i.e., AI) system or process, as a way to compliment the guidelines given below.

4.1 Guidelines for Dataset Generation via Crowdsourcing

As we have mentioned during the development of the framework in D3.2, the developer of the AI application will act as the requester of the dataset, creating a task to assign to the crowdworkers according to the needs of the AI application. Thus, practitioners must be very attentive to the whole process of creating the dataset as presented in Figure 4.1. Below we highlight the main points a practitioner should take into consideration during the generation of the dataset via crowdsourcing.

I. *Task creation must reflect the goal of the final AI application.*

If the final dataset will be used by the same team that is generating the dataset, the crowdsourcing task must be focused narrowly on the goal of the application that the dataset will be used to develop. The broader the crowdsourcing task, the more chances for the generated data to diverge into unnecessary concepts and unwanted subjective judgements from the workers.

II. *Data used within the crowdsourcing task must be thoroughly vetted.*

Practitioners must pay extra attention when generating a dataset through crowdsourcing with the help of another dataset. For example, a task can ask crowdworkers to label photos of people from previously-published image dataset. In this case the practitioner must carefully consider the use of the image dataset within the created task according to the principles outlined in the next subsection (Subsection 4.2).

III. The dataset's needs should be taken into consideration during the selection of the crowdsourcing platform

Each crowdworking platform has a different pool of workers, representing different countries/regions. For example, Amazon Mechanical Turk and Appen have many participants from the U.S., Latin America, and India, while Clickworkers has a larger European crowd. Based on the goals of the task, one (or a combination of) platforms may be preferred to recruit a suitable crowd.

The nature of the task will also impact this decision. Some platforms, such as Clickworkers, are more dedicated to running surveys, while others like Microworkers have very specific frames for the development of tasks. As the task design can rely on what is available on the platform, and the design affects how the crowdworkers perceive and interact with the task, this is not a decision to be taken lightly.

Practitioners must take into account all the above considerations before taking a conscious decision when choosing the platform as it might affect the dataset generation.

IV. When applicable, given the AI application context, the practitioner must aim at gathering data from a diverse crowd.

In many cases, social diversity – in race, gender, age, culture, and other backgrounds/identities – will result in a better dataset. This is true not only for a more diverse team creating and a more diverse team using the dataset, but also a more diverse crowd recruited for the annotations. A diverse team is more likely to notice different (potential) sources of bias in the data, and annotations from a diverse crowd would reflect the different ways the same data may be interpreted by different communities. Applications developed and datasets created without such diverse teams and crowds may contain biases, such as a Western- and rich-household bias.[32]

[32] De Vries, T., Misra, I., Wang, C., & Van der Maaten, L. (2019). Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 52-59).

Because of the above, practitioners during the task creation and execution, must take into consideration a number of parameters such as the timing of the task (what time of the day it is in a certain part of the world), the payment to the crowdworkers (respecting local wage standards) and how cultural differences will affect each group of workers differently.

V. Crowdsourcing tasks must be designed carefully to account for the inclusion of unwanted bias

When designing a crowdsourcing task, practitioners must take into consideration several factors that might influence the answers of the crowdworkers. One important factor is priming, asking the workers a series of questions before the main task activity could prime their answers and introduce (un)wanted bias. Whether this bias is desirable or not is task- and goal-dependent. For instance, it may be the case that the Requester wants to receive annotators of a certain type. Alternatively, the Requester may wish to make crowdworkers aware of a certain unwanted bias, in order to deter them from biased responses. This may require that the HIT presents some questions prior to the annotation section.[33]

Furthermore, introducing "gold standard" questions can also serve as a quick and simple way of filtering the responses based on accuracy and worker attention. Gold standard or control questions can be a series of questions of which the requester knows the answer and can be used to quickly filter out low quality workers or workers that lack the necessary attention to complete the task.[34] When gold standard questions are used, the requester must account for the cost-benefit of introducing them, since it can make the task longer and thus the requester can be required to pay more or even fatigue the worker before even having the chance to respond to the main task.

VI. Properly document the designed task and collected data

Developers must have mechanisms in place to properly document the designed task as well as the collected data (i.e., participation crowd, time that the data were collected, significant events taking place globally and locally, payment, etc.). This process is especially important, as it will allow the requester to easily re-use the collected data and aggregate them with future data collected through the same process.

Similarly, the HIT can be set up to record information about each crowdworker, including their own relation to a national/global event (whether they are aware of or affected by it), any personal experiences which make them more/less familiar with the subject of the HIT, and similar information in addition to the typical demographic questions included. For example, it has been found that the worker's political standpoint can bias their fact-checking task responses, or that their discriminatory beliefs/stereotypes may create biased data against social (e.g. race or gender)

[33] Christoforou, E., Barlas, P., & Otterbacher, J. (2021, May). It's About Time: A View of Crowdsourced Data Before and During the Pandemic. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

[34] Han, L., Maddalena, E., Checco, A., Sarasua, C., Gadiraju, U., Roitero, K., & Demartini, G. (2020, January). Crowd worker strategies in relevance judgment tasks. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 241-249).

groups.[35] The phrasing and order of questions related to these issues must be considered carefully, as stated in an earlier point.

VII. Appropriate payment must be made to the workers

Ideally, the payment to the workers must respect at least the minimum wage in the country in which the worker is based. Thus, the practitioner must have a good estimation of the average time it takes workers to complete the task. Advertising a task that rewards workers excessively more or excessively less for their work can result in a less qualified, biased or not diverse enough crowd (i.e., socio-economically or geographically) for the purposes of the task.

VIII. Ensure the (annotation) quality of the data[36]

After the crowdsourcing task, a check is made based on the “gold standard” questions introduced in the task to remove the spam or low quality responses. Some crowdworkers may annotate in ways that pass through quality checks of the platform and the initial “cleaning” of the data, but do not contain the information needed in the annotation. The data must be checked to ensure there is a consistent, high quality of annotations. If the dataset does not have good enough annotations for training, the data may need to be re-annotated through manual (e.g., crowdsourcing) or automated (e.g., through ML applications) methods. These methods however need to be monitored closely as well, similar to the concerns outlined in the previous sections, as they may result in more imbalance or unwanted biases.

In addition, some of the data may contain personal information about the people represented in the data or about the crowdworkers who performed the annotations. This includes sensitive attributes or confidential information about specific individuals, such as race, gender, language, religion, political affiliations, and more. This type of information must either be removed or encrypted to preserve the privacy of those involved.

IX. Keep a record of the raw dataset

Datasets are often built with a certain use in mind but often enough get repurposed. For this reason, maintaining access to the raw dataset can allow practitioners repurposing this data in the future to identify temporal variations and other biases inherent in the data that may go unnoticed during the creation of the dataset. With access to the “raw” data, future users of the dataset can more readily investigate and mitigate these “temporal” biases.

X. Accompany the dataset with a documentation

Similar to the contextual information about the HIT and workers mentioned earlier, other information about the data collection and preparation process would help in recognizing and managing the bias in the data. Therefore, datasets should be accompanied by documentation

[35] Otterbacher, J., Checco, A., Demartini, G., & Clough, P. (2018, June). 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* (pp. 933-936).

[36] From the cycat whitepapers on generating + using datasets

including the motivation, composition, collection process, cleaning/pre-processing, potential uses, distribution, and maintenance of the dataset, as described by Gebru and colleagues.

4.2 Guidelines for Use of Datasets Generated via Crowdsourcing

When practitioners use datasets generated via crowdsourcing to train a machine learning algorithm, they need to take certain action in order to discover and manage potential unwanted biases introduced due to the crowdsourcing process. In addition to that, practitioners should not neglect to follow best practices in general for data management and use during the training process of an ML algorithm.

I. Read the documentation accompanying the dataset (if applicable)

Practitioners using crowdsourcing data must familiarize themselves with the data by reading the available documentation. In case documentation is not available, practitioners must take the time to make sure that enough information can be retrieved about the dataset in order to be usable. Available documentation or gathered information must at a minimum give information about any codification/anonymization in the data, any synthetic/missing/incomplete data, and usage rights associated with the dataset. In the case of a crowd-annotated dataset, the documentation should also include information about the crowd that is not present in the dataset itself, such as the platform setup of the task.

II. Audit the dataset

The dataset might be imbalanced due the choice of data points or the crowd recruited. If the demographics of the crowd or characteristics of data points are not balanced, generating more data might be necessary to ensure that the data represents the diversity of the end users of the application, or the diversity of concepts required to train a working application. Automated tools or other auditing methods can uncover biases in the dataset, with respect to any feature.

However, it is important to remember that what makes bias unwanted or beneficial is context- and application-dependent; as such, upon discovering a type of bias in the data, practitioners may choose to mitigate or enhance the said bias, depending on how it would affect the developed ML application.

III. Clean the dataset if necessary

In the documentation of the dataset, the cleaning and other pre-processing methods should be laid out. If it is discovered that the processed data contains some unwanted biases, it may be valuable to see whether the “raw” data was also released in the dataset. This may allow for the unwanted biases and imbalances introduced in the annotation/cleaning process to be reversed and for a different cleaning method to be applied, more aligned with the end application’s goal. If the “raw” data is not available, the processed data may be cleaned with appropriate methods to overcome the issues discovered.

5. Knowledge Gained from Interaction with Practitioners

In this section we wish to highlight some of the most noteworthy observations and knowledge gained from presenting our practitioner guidelines to researchers in the field of machine learning and AI. A full report on our interaction with the industry practitioners and the material used is presented in D6.2. Here we wish to briefly note how the guidelines were received by the practitioners and their general comments on the effort of introducing such guidelines.

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.

6. Conclusions

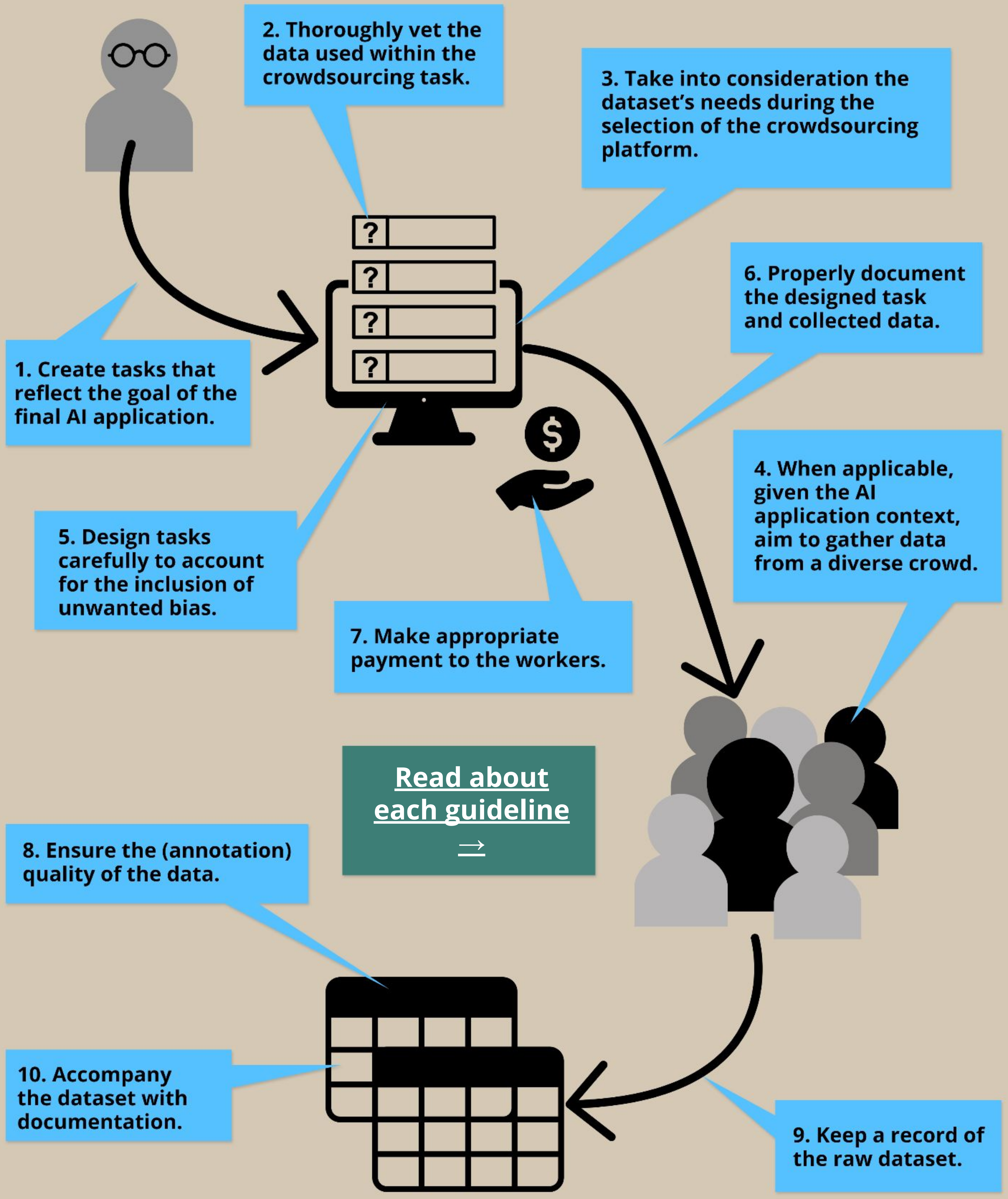
Apart from presenting the set of guidelines for generating and using datasets through crowdsourcing that we have created as part of the DESCANT project we have also identified a real need for distributing these guidelines to current and future practitioners. It is our intention to keep distributing and educating practitioners regarding our findings in this project as we believe they can be of real value for the creation of trustworthy AI and transparency in AI.

7. Annex

Infographics versions of the Practitioner Guidelines.

DESCANT: Detecting Stereotypes in Human Computational Tasks

Guidelines for Dataset Generation via Crowdsourcing



1. Create tasks that reflect the goal of the final AI application.

2. Thoroughly vet the data used within the crowdsourcing task.

3. Take into consideration the dataset's needs during the selection of the crowdsourcing platform.

4. When applicable, given the AI application context, aim to gather data from a diverse crowd.

5. Design tasks carefully to account for the inclusion of unwanted bias.

6. Properly document the designed task and collected data.

7. Make appropriate payment to the workers.

8. Ensure the (annotation) quality of the data.

9. Keep a record of the raw dataset.

10. Accompany the dataset with documentation.

Read about each guideline →

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.

Icons from Andrejs Kirma, SAM Designs, Nico Ilk, Fasobrun Jamil, Adrien Coquet, and SANB on NounProject.com

DESCANT: Detecting Stereotypes in Human Computational Tasks

Guidelines for Dataset Generation via Crowdsourcing

1) *Create tasks that strictly reflect the goal of the final AI application.*

Ideally, the practitioners who will be training an AI application on the dataset, are also the ones generating the dataset. When this is the case, the crowdsourcing task must be focused narrowly on the goal of the application to be developed. The broader and less focused the crowdsourcing task, the more chances for the generated data to diverge into unnecessary concepts and unwanted subjective judgements from the workers.

2) *Thoroughly vet the data used within the crowdsourcing task.*

Practitioners must pay extra attention when generating a dataset through crowdsourcing with the help of another dataset. For example, a task can ask crowdworkers to label photos of people from previously-published image dataset. In this case, the practitioner must carefully consider the use of the image dataset within the created task.

3) *Take into consideration the dataset's needs during the selection of the crowdsourcing platform.*

Each crowdworking platform has a different pool of workers, representing different countries/regions. For example, Amazon Mechanical Turk and Appen have many participants from the U.S., Latin America, and India, while Clickworkers has a larger European crowd. Based on the goals of the task, one (or a combination of) platforms may be preferred to recruit a suitable crowd.

The nature of the task will also impact this decision. Some platforms, such as Clickworkers, are more dedicated to running surveys, while others like Microworkers have very specific frames for the development of tasks. As the task design can rely on what is available on the platform, and the design affects how the crowdworkers perceive and interact with the task, this is not a decision to be taken lightly.

Practitioners must take into account all the above considerations before taking a conscious decision when choosing the platform as it might affect the dataset generation.

4) *When applicable, given the AI application context, aim to gather data from a diverse crowd.*

In most cases, social diversity – in race, gender, age, culture, and other backgrounds/identities – will result in a better dataset. This is true not only for a more diverse team creating and a more diverse team using the dataset, but also a more diverse crowd recruited for the annotations. A diverse team is more likely to notice different (potential) sources of bias in the data, and annotations from a diverse crowd would reflect the different ways the same data may be interpreted by different communities. Applications developed and datasets created without such diverse teams and crowds may contain biases, such as a Western- and rich-household bias.

Because of the above, practitioners during the task creation and execution, must take into consideration a number of parameters such as the timing of the task (what time of the day it is in a certain part of the world), the payment to the crowdworkers (respecting local wage standards) and how cultural differences will affect each group of workers differently.

5) *Design tasks carefully to account for the inclusion of unwanted bias.*

When designing a crowdsourcing task, practitioners must take into consideration several factors that might influence the answers of the crowdworkers. One important factor is priming. Asking the workers a series of questions before the main task activity could prime their answers and introduce (un)wanted bias. Whether this bias is desirable or not is task- and goal-dependent. For instance, it may be the case that the Requester wants to receive annotators of a certain type. Alternatively, the Requester may wish to make crowdworkers aware of a certain unwanted bias, in order to deter them from biased responses. This may require that the HIT presents some questions prior to the annotation section.

Furthermore, introducing "gold standard" questions can also serve as a quick and simple way of filtering the responses based on accuracy and worker attention. Gold standard or control questions can be a series of questions of which the requester knows the answer and can be used to quickly filter out low quality workers or workers that lack the necessary attention to complete the task. When gold standard questions are used, the requester must account for the cost-benefit of introducing them, since it can make the task longer and thus the requester can be required to pay more or even fatigue the worker before even having the chance to respond to the main task.

(cont'd)

6) *Properly document the designed task and collected data.*

Developers must have mechanisms in place to properly document the designed task as well as the collected data (i.e., participation crowd, time that the data were collected, significant events taking place globally and locally, payment, etc.). This process is especially important, as it will allow the requester to easily re-use the collected data and aggregate them with future data collected through the same process.

Similarly, the HIT can be set up to record information about each crowdworker, including their own relation to a national/global event (whether they are aware of or affected by it), any personal experiences which make them more/less familiar with the subject of the HIT, and similar information in addition to the typical demographic questions included. For example, it has been found that the worker's political standpoint can bias their fact-checking task responses, or that their discriminatory beliefs/stereotypes may create biased data against social (e.g. race or gender) groups. The phrasing and order of questions related to these issues must be considered carefully, as stated in an earlier point.

7) *Make appropriate payment to the workers.*

Ideally, the payment to the workers must respect at least the minimum wage in the country in which the worker is based. Thus, the practitioner must have a good estimation of the average time it takes workers to complete the task. Advertising a task that rewards workers excessively more or excessively less for their work can result in a less qualified, biased or not diverse enough crowd (i.e., socio-economically or geographically) for the purposes of the task.

8) *Understand the (annotation) quality of the data.*

After the crowdsourcing task, a check is made based on the "gold standard" questions introduced in the task to remove the spam or low quality responses. Some crowdworkers may annotate in ways that pass through quality checks of the platform and the initial "cleaning" of the data, but do not contain the information needed in the annotation. The data must be checked to ensure there is a consistent, high quality of annotations. If the dataset does not have good enough annotations for training, the data may need to be re-annotated through manual (e.g., crowdsourcing) or automated (e.g., through ML applications) methods. These methods however need to be monitored closely as well, similar to the concerns outlined in the previous sections, as they may result in more imbalance or unwanted biases.

In addition, some of the data may contain personal information about the people represented in the data or about the crowdworkers who performed the annotations. This includes sensitive attributes or confidential information about specific individuals, such as race, gender, language, religion, political affiliations, and more. This type of information must either be removed or encrypted to preserve the privacy of those involved.

9) *Keep a record of the raw dataset.*

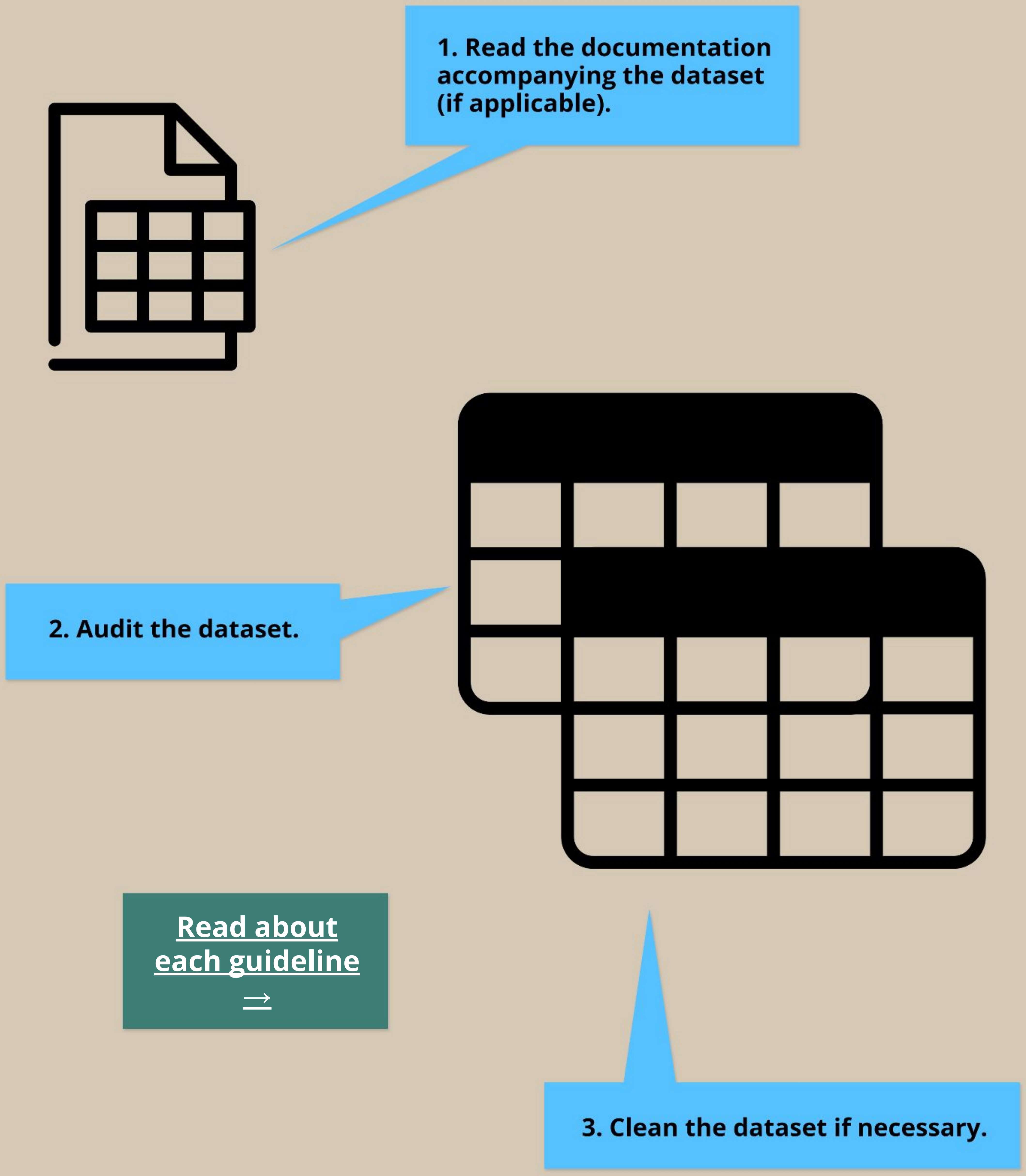
Datasets are often built with a certain use in mind but often enough get repurposed. For this reason, maintaining access to the raw dataset can allow practitioners repurposing this data in the future to identify temporal variations and other biases inherent in the data that may go unnoticed during the creation of the dataset. With access to the "raw" data, future users of the dataset can more readily investigate and mitigate these "temporal" biases.

10) *Accompany the dataset with documentation.*

Similar to the contextual information about the HIT and workers mentioned earlier, other information about the data collection and preparation process would help in recognizing and managing the bias in the data. Therefore, datasets should be accompanied by documentation including the motivation, composition, collection process, cleaning/pre-processing, potential uses, distribution, and maintenance of the dataset, as described by Gebru and colleagues.

DESCANT: Detecting Stereotypes in Human Computational Tasks

Guidelines for Use of Datasets Generated via Crowdsourcing



1. Read the documentation accompanying the dataset (if applicable).

2. Audit the dataset.

3. Clean the dataset if necessary.

Read about each guideline →

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.

Icons from Andrea Kinta and Iconparty on NounProject.com

DESCANT: Detecting Stereotypes in Human Computational Tasks

Guidelines for Use of Datasets Generated via Crowdsourcing

When practitioners use datasets generated via crowdsourcing to train a machine learning algorithm, they need to take certain actions in order to discover and manage potential unwanted biases introduced due to the crowdsourcing process. In addition to that, practitioners should not neglect to follow best practices in general for data management and use during the training process of an ML algorithm.

1) Read the documentation accompanying the dataset (if applicable).

Practitioners using crowdsourced data must familiarize themselves with the data by reading the available documentation. In case documentation is not available, practitioners must take the time to make sure that enough information can be retrieved about the dataset in order to be usable. Available documentation or gathered information must at a minimum give information about any codification/anonymization in the data, any synthetic/missing/incomplete data, and usage rights associated with the dataset. In the case of a crowd-annotated dataset, the documentation should also include information about the crowd that is not present in the dataset itself, such as the platform setup of the task.

2) Audit the dataset.

The dataset might be imbalanced due the choice of data points or the crowd recruited. If the demographics of the crowd or characteristics of data points are not balanced, generating more data might be necessary to ensure that the data represents the diversity of the end users of the application, or the diversity of concepts required to train a working application. Automated tools or other auditing methods can uncover biases in the dataset, with respect to any feature.

However, it is important to remember that what makes bias unwanted or beneficial is context- and application-dependent; as such, upon discovering a type of bias in the data, practitioners may choose to mitigate or enhance the said bias, depending on how it would affect the developed ML application.

3) Clean the dataset if necessary.

In the documentation of the dataset, the cleaning and other pre-processing methods should be laid out. If it is discovered that the processed data contains some unwanted biases, it may be valuable to see whether the “raw” data was also released in the dataset. This may allow for the unwanted biases and imbalances introduced in the annotation/cleaning process to be reversed and for a different cleaning method to be applied, more aligned with the end application’s goal. If the “raw” data is not available, the processed data may be cleaned with appropriate methods to overcome the issues discovered.