




# MindSet: A Bias-Detection Interface Using a Visual Human-in-the-Loop Workflow

Senthuran Kalanathan<sup>1</sup>, Alexander Kichutkin<sup>2</sup>(✉), Ziyao Shang<sup>1</sup>,  
András Strausz<sup>1</sup>, Francisco Javier Sanguino Bautiste<sup>3</sup>,  
and Mennatallah El-Assady<sup>1,4</sup> 

<sup>1</sup> Department of Computer Science, ETH Zürich, Switzerland  
{skalanan, zshang, strausza}@ethz.ch

<sup>2</sup> Department of Mathematics, ETH Zürich, Switzerland  
akichutkin@ethz.ch

<sup>3</sup> Department of Information Technologies and Electrical Engineering, ETH Zürich,  
Switzerland

<sup>4</sup> AI Center, ETH Zürich, Switzerland

**Abstract.** Handling data artifacts is a critical and unsolved challenge in deep learning. Disregarding such asymmetries may lead to biased and socially unfair predictions, prohibiting applications in high-stake scenarios. In the case of visual data, its inherently unstructured nature makes automated bias detection especially difficult. Thus, a promising remedy is to rely on human feedback. Hu et al. [14] introduced a three-stage theoretical study framework to use a human-in-the-loop approach for bias detection in visual datasets and ran a small-sample study. While showing encouraging results, no implementation is available to enable researchers and practitioners to study their image datasets. In this work, we present a dataset-agnostic implementation based on a highly flexible web app interface. With this implementation, we aim to bring this theoretical framework into practice by following a user-centric approach. We also extend the framework so that the workflow can be adjusted to the researcher's needs in terms of the granularity of detected anomalies.

**Keywords:** User Interfaces · Dataset Bias · Bias in Machine Learning

## 1 Introduction

Since the appearance of the CNN [20] and subsequently the transformer [34] architectures, deep learning yielded remarkable achievements in various computer vision tasks. We have seen improvements in all the different branches of visual pattern recognition, such as image segmentation [23], classification [19], or most recently in image generation [25]. Moreover, these techniques have long left academia and have been deployed in real-life scenarios, often involving such

---

S. Kalanathan, A. Kichutkin, Z. Shang and A. Strausz—Equal contribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

S. Nowaczyk et al. (Eds.): ECAI 2023 Workshops, CCIS 1948, pp. 93–105, 2024.

[https://doi.org/10.1007/978-3-031-50485-3\\_8](https://doi.org/10.1007/978-3-031-50485-3_8)

where an ethical and fair decision is indispensable [22]. Unfortunately, prior research has shown that many models fall short of this criteria [5, 8, 18].

The primary challenge in developing fair and trustworthy models comes from the lack of a precise quantitative formulation of bias [7]. In the context of visual data, which involves datasets composed of visual components and is visually interpreted, this challenge has gained even greater prominence. As a result, either proxy measures are used aiming to grasp parts of the contained biases [35] or human judgments are included. While the prior can provide specific, typically technical, fairness guarantees, it often falls short in ensuring a universally fair model [17]. On the other hand, humans can naturally detect visual biases, and their assessments can be later included in the Machine Learning pipeline. Even when personal judgments are influenced or led astray by prejudices, this can be counteracted by collecting a large sample of opinions. According to the *wisdom of the crowd* hypotheses [28], when sufficient and diverse opinions are gathered, the common understanding of a diverse crowd would lead to more reasonable and thorough judgments of possible biases. It is worth noting that the diversity of the sample is an integral part of this concept.

Biases could occur in various stages of the ML pipeline, including data input, training, and model applications [2, 22]. Deep learning techniques have been shown to rely heavily on the training data, and biases included in the training will be reflected by the models' prediction [32]. Since raw images are unintelligible for computers and learned representation may carry biases already, it is natural to involve human judgments at this stage and employ them to filter the training data. Hu et al. [14] propose a three-stage study technique to detect sample biases in visual datasets. In their evaluation, they show that the framework allows for finding both commonly known and dataset-specific yet unknown biases among images. However, the authors did not develop any implementation for their study but used static forms that were tailored for a single dataset and were not made available to the public.

In this work, we face the challenge of bringing this theoretical framework into practice by developing an interactive web interface that implements the framework of Hu et al. [14]. By making this tool accessible to various users, we allow examinations, applications, and future extensions to the framework. In particular, we:

- Create a user-friendly online survey platform for detecting biases in image datasets
- Enable the examination of any image dataset and the fine-tuning of the study with respect to the dataset
- Extend the framework by Hu et al. with an interactive dashboard for study parameter selection

## 2 Related Work

### 2.1 Bias Discovery

Fabrizzi et al. [10] gives a framework for categorizing machine-centric bias detection methods for image data. Most notably, they argue that even in carefully

curated bias-aware datasets, disparities exist, making methods for bias exploration crucial. They cluster prior works as follows:

1. Reduction to tabular data: such methods convert visual data into a tabular form and use bias detection techniques designed for tabular datasets e.g. count/demographic parity [9] or causality [36]
2. Biased image representations: Bias detection methods in this category analyze distances and geometric relationships among images utilizing the lower dimensional representation to identify the presence of bias. This includes distance-based methods [15] and interventions [3]
3. Cross-dataset bias detection: Methods in this category aim to identify the distinct signature of each dataset by comparing various datasets, e.g. [32,33]
4. Other methods: They include a wide range of methods such as crowdsourcing frameworks to ad-hoc trained classification models. Examples are [24] and [31]

To the best of our knowledge, prior human-in-the-loop methods for bias identification all focused on tabular data. A line of work relied on the assumption that humans can evaluate small graphical causal models. Silva [41] offers a visual interface to detect biases based on the causal relationships between the features. D-BIAS [12] follows a similar methodology but also allows to alter the causal links and thus actively mitigate bias. Other works use individual or group-level fairness measures to detect asymmetries among features, that are then visualized in different ways. Examples are FairRankViz [40] for bias detection in graph mining or DiscriLens [37], which offers novel visualizations of group-level bias attributes. In conclusion, there does not yet exist a human-centric approach for bias detection in visual datasets.

## 2.2 *Wisdom of the Crowd* bias detection

We summarize the study procedure by Hu et al. [14] in more detail, as this serves as the basis of our interface. We only describe the main stages of the study here and defer any specifics or changes to Sect. 3. The study can be described by the following three stages:

1. *Question generation*: The study starts by asking the participants to enter question-answer pairs that describe a similarity among the set of images that are currently shown. Participants are encouraged to ask questions starting with *What*, *Where*, *When* or *How* and avoid questions describing common characteristics of objects. These questions are then merged to filter reformulations of the same concept.
2. *Answer collection*: The collected questions are shown to the users again but with a different sample of images. The user is then asked to enter an answer to the question if at least half of the images share the same answer; otherwise, the user should skip it. Afterward, similar answers are merged to avoid ambiguities from different spellings or synonyms.

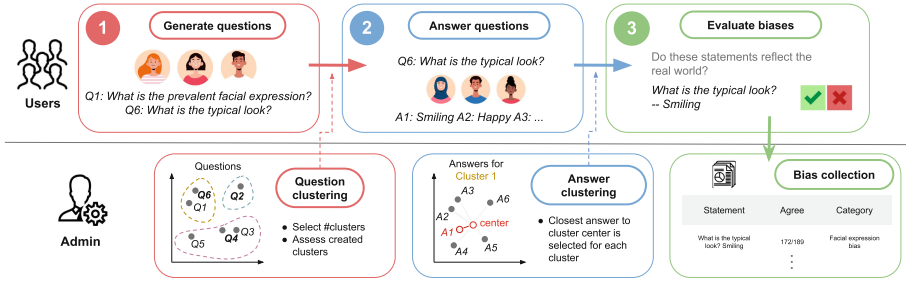


Fig. 1: Workflow illustrating the three step study framework and the corresponding admin tasks.

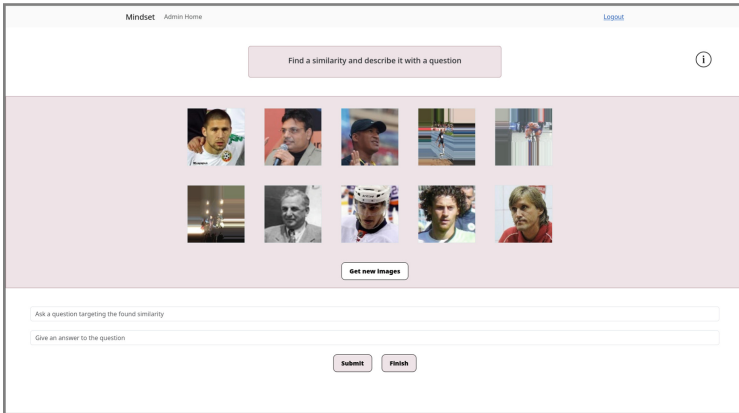


Fig. 2: Stage 1 with 10 images shown: Users enter a pair of question/answer that characterize the image set.

3. *Bias Judgement*: Lastly, questions and their corresponding answers are used to generate universal statements describing a possible bias. Users are then asked whether this statement is true in the real world or is a specific attribute of the dataset.

### 3 The MindSet Interface

To better support bias mitigation in visual datasets, we implement a user-friendly interface for the study framework of Hu et al. [14]. The implementation is publicly accessible<sup>1</sup>.

#### 3.1 Implementation Details

In the following, we describe our implementation as well as all extensions to the framework of Hu et al. [14]. Figure 1 shows an overview of the workflow.

<sup>1</sup> <http://a10-bias-assessment-with-human-feedback.course-xai-impl23.isginf.ch/>.

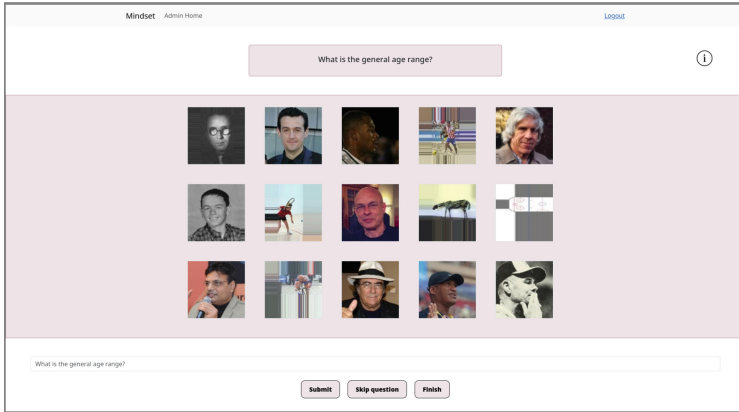


Fig. 3: Stage 2 with 15 images shown: According to the image set, users provide an answer to the given question.

**User Types.** We differentiate between two types of users, *participants* and *study admins*. To deal with spurious inputs, we ask users to first register with their email accounts. Registrations of participants can be verified by the participant itself through a code sent by email. Registration of study admins has to be accepted by the developers.

*Participants* can only access the current stage of the study and are notified when the study progresses to the next stage. Study Admins can choose to move their study from one stage to another and see overview statistics as well as detected biases. Admins are also responsible for setting the number of images (randomly sampled from the whole dataset) provided to each participant during each step and choosing parameters when proceeding with the study to the next stage.

**Study Workflow.** The interface for stages 1, 2, and 3 are depicted, respectively, in Fig. 2, Fig. 3, and Fig. 4, where participants are guided through the interface at the start and further aided with hints. We aim to create a neutral interface with as little text as possible to avoid influencing the participant. Participants receive a short introduction to every state and are guided through the interface before starting the study. A hint is also available in case the participant loses track. A difference from the original framework is that we do not provide any suggestions at any stage of the study in order to avoid influencing the participant.

**Administrator View.** We create a separate overview for study administrators where they can manage their currently ongoing study.

One of the main tasks of study administrators is to proceed with the study from one stage to the next one. To process Step 1 (question generation), the administrator needs to extract certain representative questions from

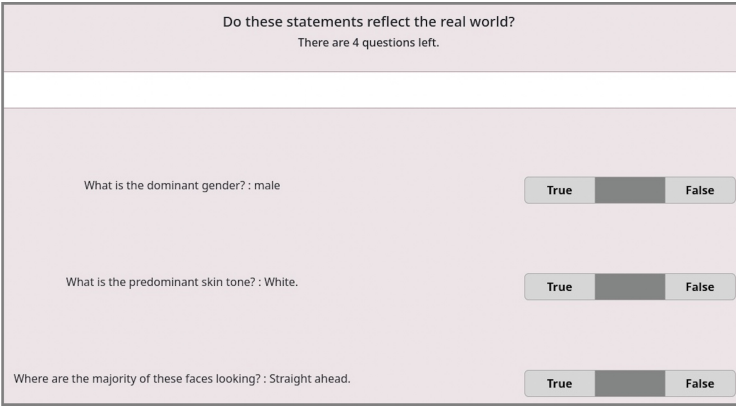


Fig. 4: Stage 3: users evaluate whether the biases are reflected in real life.

existing questions. The extraction is done via clustering. Each question is first embedded into a vector by an NLP model, for which we currently use the `all-MiniLM-L12-v2` pre-trained sentence transformer. Then, the embeddings are clustered using K-means clustering. For each resulting cluster, the question whose embedding is closest to the cluster centroid is chosen to represent that cluster. Thus, the administrator has to choose the number of questions they want to keep. This decision is aided by an interactive visualization of the elbow method [29] and the clustering for the currently chosen setting.

For the elbow method, the administrator specifies a range of centroid numbers. The visualization would be able to plot the Within Cluster Sum of Squares (WCSS) of all centroid numbers within that range. Generally, the elbow point of this graph would be a sound choice for the number of clusters. The administrator can hover over the data points on the visualization to see their actual WCSS values. Next, if the administrator clicks on one of the data points, a preview of the clustering results will be presented.

To create the preview, the embedding for each question is reduced to a 2D vector using Principal Component Analysis [11]. The cluster to which each point belongs would be encoded using the color (hue) of the points. Hovering on the points, the administrator would be able to see the actual question behind the point. The preview also contains a legend containing the questions chosen for

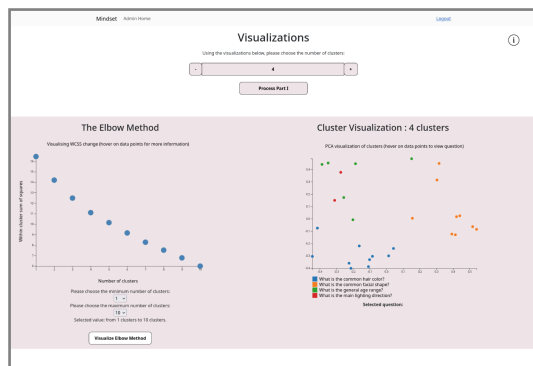


Fig. 5: Interface for processing inputs after Step 1, the admin is provided with visualizations of the WCSS distribution and the clusters.

each centroid. An example state of the dashboard for processing Step 1 is shown in Fig. 5

Once the study administrator decides to finish Stage 2, the answers are processed. For each question, all its answers are embedded. Then, the answer that is closest to the centroid is chosen. At this step, our implementation also deviates from [14], as they decide to construct statements from the question-answer pairs. Although such simple statements may be easier to understand at Stage 3, since statements must be generated with language models, they may be slightly imprecise or contain biases originating from the language model. To avoid this, we directly present the question-answer pairs to the participants.

During the final stage, the administrator can see a table overview of the biases detected in the dataset. This overview contains a list view containing the text of each bias statement, the number of users that agree with the statement, and the ratio of users agreeing, where the administrator can label biases, filter biases based on the labels, and save labels into the database. The users are also able to download the table as a .csv file.

**Demo Study.** To showcase the workflow, we used the CelebFaces Attributes Dataset (CelebA), a comprehensive collection comprising more than 200,000 celebrity images [1]. Additionally, we have pre-filled the database with dummy text data for demonstration purposes, allowing for a comprehensive illustration of the system’s functionality.

### 3.2 Use Cases

The following section will outline two possible use cases for the proposed workflow and showcase the flexibility of our application.

**Human-Assisted Compilation of Bias-Aware Datasets.** In an ideal scenario, researchers would examine their dataset for biases that might amplify societal stereotypes before training models. To that end, Wang et al. [38] proposed the measure of dataset leakage that describes how much information an image is leaking about a protected attribute when this attribute is obscured. For instance, assume that we protect for gender and consider a set of images containing adult males and females. Then, after obscuring the people in the visual data, dataset leakage will be measured on the extent to which it is still possible to infer the gender attribute from the remaining objects on the images.

Training a model on data with high dataset leakage exhibits the risk of further increasing these biases. The authors only provided a measure but not a solution for improving the dataset. To this end, we can leverage human feedback through our proposed workflow and interface. Our application allows researchers to pinpoint the objects leaking societal stereotypes to subsequent models. The first two steps localize strong signals in the data, while the third step judges whether these signals might be responsible for, in this case, gender biases. We allow the researcher to tag these findings for further processing.

An example workflow, in this case, would be that: at first, the researcher measures the data leakage of his dataset according to Wang et al. [38] revealing that it leaks societal biases about gender, which is a protected attribute here. Then, utilizing our crowdsourcing application, the researcher localizes the objects within images that leak information about the protected attribute. For instance, human feedback might suggest that an overly large portion of images containing females contain cooking utensils. The human crowd decides that this does not reflect the real world. Thus, this signals the researcher that the dataset amplifies old societal gender stereotypes. The researcher then can tag this finding on the summary page as such, that cooking utensils in his images strongly correlate with females being depicted. Assuming that he has a certain level of control over the data collection process, he can adjust the dataset to mitigate this phenomenon. Afterward, the process may be repeated over several iterations until a satisfactory upper limit on the data leakage is reached. Hence, our workflow can be leveraged together with a quantitative measure of dataset bias to build an iterative loop that results in a bias-aware dataset.

**Bias Reducing Training of Generative Models.** The idea of generative models is not new. They have been around as early as the 1960s with the introduction of the ELIZA chatbot [39]. However, with recent advances, the topic has attracted a lot of new attention outside the scientific community. Frameworks such as Stable Diffusion (Rombach et al. [25]) can produce images that are difficult to be recognized as synthetically generated at first sight. However, similar to any other machine learning models, they tend to mirror biases reflected in their training data. To this end, we can utilize our workflow to examine the outputs of generative models more closely. We can consider a set of outputs of a deep generative model as a synthetically produced dataset. This dataset can then be analyzed in our study interface similarly to any other dataset. The first two steps point towards anomalies in the outputs, while the third step gives human feedback on whether these anomalies reflect the real world. As in Sect. 3.2, the researcher can use the tagging feature of the interface to categorize found anomalies. There are frameworks allowing the researcher to steer its generative model towards certain attributes. One such instance for GANs [13] are Style-Based Generators as introduced by Karras et al. [16]. They allow the researcher to steer the generative models to address these biases. To this end, our application can be incorporated into a training loop for generative models aiming to minimize exhibited biases. First, the models produce outputs. These outputs are considered the input dataset for the next step, which is crowdsourced for bias discovery using our application. The discovered biases can be directly addressed by guiding the outputs of the generative models. These steps can then be repeated until a satisfactory performance is reached.

## 4 Discussion

*MindSet* could be used in various situations, such as examining machine-generated images or detecting biases in common visual datasets, and has a large



potential for adaptations and extensions.

However, the framework is not complete yet and is subject to some constraints, which will be discussed in this section.

#### 4.1 User Selection

The MindSet framework does not specify rules on how to choose study participants. This task is left completely to the study admin. However, participant selection is crucial. Depending on the individual participants and possible incentives or rewards, there is a risk of selection bias being introduced into the framework by the participants. This would defy its intended purpose. By design, there is a risk of self-selection and under-coverage. Failing to choose participants from a broad and diverse background can lead to the user selection mechanism failing to capture a sufficient representation of the population [4]. In that case, detected biases would only represent a one-sided perspective from a particular population group and would likely fail to capture the majority of existing biases in the specific dataset. However, it is not the emphasis of the *MindSet* framework to direct the user selection for a survey. It assumes that the study admin is familiar with guidelines on selecting participants such that the survey can be used to infer usable insights. Such guidelines can be found in plenty of literature in medical [21] or business domains [30].

Another point to consider is the effect of incentives or rewards. For instance, there could be a monetary incentive to include as many *Question-Answer pairs* as possible in Step 1. This could potentially lead to participants submitting *Question-Answer pairs* which are not suitable to the subset of images they are presented with and lead to an accumulation of redundant information. Again, *MindSet* does not aim to provide a specific study setup but rather focuses on enabling practitioners to perform bias detection in image datasets through crowd-sourcing. There exists a rich literature on how incentives and rewards can be used to optimize survey setups such as in Singer and Ye [27].

#### 4.2 Measuring Bias

Biases in image-based machine learning workflows are usually measured in variations of the following two ways. First, given an existing and available protected attribute (e.g. race) for a set of images, we can measure the performance of subsequent machine learning applications conditioned on the protected attribute (e.g. Facial recognition accuracy for different ethnicities [6]). However, this approach is more output-focused rather than working with the dataset at hand. To that end, another popular approach is to look at label distributions within the dataset [26].

If the distribution is skewed towards specific labels, it implies a higher occurrence than for other labels. The performance of subsequent applications could then depend on label occurrences. However, this approach has its limitations. First, it assumes that the dataset is well-annotated, which, besides larger benchmark datasets, is not necessarily the case. Also, a class label does not capture

the intra-class variability within the images. For instance, a subset of 100 images labeled as containing a human does not tell us anything about the distribution of important attributes such as age, gender, and race. The dataset might suffer from label bias [32], and subsequent real-life applications might take the subset of 100 images as the ground truth for how humans are defined, posing the risk of discriminatory decisions. As mentioned in the first section, automated systems have difficulties detecting intra-class biases towards often specific attributes, while humans have an innate understanding of images and their details.

The *MindSet* framework leverages human nature to detect biases in images where algorithms would fail. However, we need to define our own measure of bias. No general all-encompassing measure is available, but we could exploit heuristics to construct a proxy variable for how biased a dataset is. The output of the interface is a table containing the number of participants agreeing with the statements aggregated in Step 3. If there is a strong agreement towards a particular statement, this implies that the dataset seems to capture a real-world property very well. A low agreement suggests that the dataset depicts properties that do not occur in reality. If there are many statements with which the participants agree, it implies that the dataset seems to capture overall real-life properties very well. However, if there are many statements with low agreement numbers (e.g. disagreeing with the statement), it suggests that the dataset fails at representing real-life properties. Overall, this provides the admin with a quantitative and qualitative assessment of possible biases in his image data. First, by observing the number of statements with high and/or low agreements, the admin gets a sense of how many and how strongly the dataset captures real-life features or fails to do so. The individual statements themselves give a qualitative pointer to the admin about which features, in particular, are well or badly captured by the dataset.

This heuristic attempts to measure bias by examining how strongly a human crowd agrees with statements describing a dataset. Nonetheless, the robustness of this measure likely correlates with the selection of survey participants. The response to statements might differ from group to group, and selection biases are possible.

### 4.3 Conclusion and Future Work

The *MindSet* interface is a practical implementation of the crowdsourcing framework proposed and validated by Hu et al. [14]. It makes additions to the original framework to increase usability. The demo case using the CelebA data-set [1] in Sect. 3 serves as a proof-of-concept for the interface.

However, it was not yet validated in a real-life environment. It would be interesting to see how it holds up when interacting with real human participants as part of a bias detection workflow and if it works with arbitrary image datasets as well. Overall, the validity of the interface is based on the case study done by Hu et al. [14], which tested the theoretical framework in a real-life environment. The interface enables the survey to be conducted in a scalable and user-

friendly manner. A bias measurement is provided using a heuristic, giving the admin a quantitative and qualitative response to the dataset used. The biggest uncertainty remains the selection of users. As a workflow depending largely on human feedback, the selection of participants can influence the results of the workflow drastically. However, *MindSet* is a platform for enabling surveys, while the ultimate responsibility regarding study setup lies with the practitioner. It is important that the practitioner acknowledges the possibility of selection bias and acts in a responsible way. In terms of future use cases, a possible adaptation of our interface would be converting it into a deductive interaction workflow for evaluating synthetic image generation pipelines. Our interface could be used to compare real-world and synthetic data and probe whether the data generation repeats, amplifies, or mitigates real-life biases.

Lastly, *MindSet* could be extended to enable the direct refinement of the dataset. Given the detected biases, it would be convenient if the user could fine-tune the dataset based on the study results, preferably through an interactive workflow that contains data refinement methods.

## References

1. Large-scale celebfaces attributes (celeba) dataset. <https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>
2. Baer, T.: Understand, manage, and prevent algorithmic bias: a guide for business users and data scientists. Apress, New York, NY (2019)
3. Balakrishnan, G., Xiong, Y., Xia, W., Perona, P.: Towards causal benchmarking of bias in face analysis algorithms. In: European Conference on Computer Vision (2020)
4. Bethlehem, J.: Selection bias in web surveys. *Int. Statist. Rev.* **78**(2), 161–188 (2010). <https://doi.org/10.1111/j.1751-5823.2010.00112.x>
5. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. PMLR (2018)
6. Cavazos, J.G., Phillips, P.J., Castillo, C.D., O’Toole, A.J.: Accuracy comparison across face recognition algorithms: where are we on measuring race bias? *IEEE Trans. Biometrics, Behav. Identity Sci.* **3**(1), 101–111 (2021). <https://doi.org/10.1109/TBIOM.2020.3027269>
7. Corbett-Davies, S., Gaebler, J., Nilforoshan, H., Shroff, R., Goel, S.: The measure and mismeasure of fairness. *J. Mach. Learn. Res* (2023)
8. De-Arteaga, M., et al.: Bias in bios: a case study of semantic representation bias in a high-stakes setting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (2019)
9. Dulhanty, C., Wong, A.: Auditing ImageNet: towards a model-driven framework for annotating demographic attributes of large-scale image datasets. ArXiv (2019)
10. Fabbrizzi, S., Papadopoulos, S., Ntoutsis, E., Kompatsiaris, Y.: A survey on bias in visual datasets. *Comput. Vis. Image Underst.* **223** (2021)
11. F.R.S., K.P.: LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh Dublin Philos. Mag. J. Sci.* **2**(11) (1901)
12. Ghai, B., Mueller, K.: D-bias: a causality-based human-in-the-loop system for tackling algorithmic bias. *IEEE Trans. Vis. Comput. Graph.* (2022)

13. Goodfellow, I., et al.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc. (2014)
14. Hu, X., et al.: Crowdsourcing detection of sampling biases in image datasets. In: *Proceedings of The Web Conference 2020. WWW '20*, Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3366423.3380063>
15. Kärkkäinen, K., Joo, J.: FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019)
16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019)
17. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent Trade-Offs in the fair determination of risk scores. *Conf. Innov. Theoret. Comput. Sci.* **67**, 23 (2017). <https://doi.org/10.4230/LIPICs.ITCS.2017.43>
18. Koenecke, A., et al.: Racial disparities in automated speech recognition. *Proc. Natl. Acad. Sci.* **117**(14) (2020)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**(6) (2017)
20. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
21. Martínez-Mesa, J., González-Chica, D.A., Duquia, R.P., Bonamigo, R.R., Bastos, J.L.: Sampling: how to select participants in my research study? *An. Bras. Dermatol.* **91**, 326–330 (2016)
22. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**(6) (2021). <https://doi.org/10.1145/3457607>
23. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(7) (2021)
24. Model, I., Shamir, L.: Comparison of data set bias in object recognition benchmarks. *IEEE Access* **3** (2015)
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022). <https://doi.org/10.1109/cvpr52688.2022.01042>
26. Rudd, E., Günther, M., Boulton, T.: Moon: a mixed objective optimization network for the recognition of facial attributes, 9909 (2016). [https://doi.org/10.1007/978-3-319-46454-1\\_2](https://doi.org/10.1007/978-3-319-46454-1_2)
27. Singer, E., Ye, C.: The use and effects of incentives in surveys. *Ann. Am. Acad. Polit. Soc. Sci.* **645**(1), 112–141 (2013). <https://doi.org/10.1177/0002716212458082>
28. Surowiecki, J.: *The Wisdom of Crowds*. Anchor (2005)
29. Syakur, M., Khotimah, B., Rochman, E., Satoto, B.D.: Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In: *IOP Conference Series: Materials Science and Engineering*, vol. 336. IOP Publishing (2018)
30. Taherdoost, H.: Sampling methods in research methodology; how to choose a sampling technique for research. *How to choose a sampling technique for research* (2016)

31. Thomas, C., Kovashka, A.: Predicting the politics of an image using webly supervised data. In: *Advances in Neural Information Processing Systems* 32 (2019)
32. Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T.: A deeper look at dataset bias (2017). [https://doi.org/10.1007/978-3-319-58347-1\\_2](https://doi.org/10.1007/978-3-319-58347-1_2)
33. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: *CVPR 2011* (2011)
34. Vaswani, A., et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017)
35. Verma, S., Rubin, J.: Fairness definitions explained. In: *Proceedings of the International Workshop on Software Fairness* (2018)
36. Wachinger, C., Rieckmann, A., Pölsterl, S.: Detect and correct bias in multi-site neuroimaging datasets. *Med. Image Anal.* 67 (2020)
37. Wang, Q., Xu, Z., Chen, Z., Wang, Y., Liu, S., Qu, H.: Visual analysis of discrimination in machine learning. *IEEE Trans. Vis. Comput. Graph.* **27**, 1470–1480 (2020)
38. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced datasets are not enough: estimating and mitigating gender bias in deep image representations. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019)
39. Weizenbaum, J.: Eliza-a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**(1) (1966)
40. Xie, T., Ma, Y., Kang, J., Tong, H., Maciejewski, R.: FairRankVis: a visual analytics framework for exploring algorithmic fairness in graph mining models. *IEEE Trans. Vis. Comput. Graph.* (2022)
41. Yan, J.N., Gu, Z., Lin, H., Rzeszutarski, J.M.: Silva: interactively assessing machine learning fairness using causality. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20, Association for Computing Machinery, New York, NY, USA (2020)