Using Explainability to Help Children Understand Gender Bias in AI

Gaspar Isaac Melsión
KTH Royal Institute of Technology
Stockholm, Sweden
gimp@kth.se

Eva Vidal
UPC Universitat Politècnica de Catalunya
Barcelona, Spain
eva.vidal@upc.edu

Ilaria Torre
KTH Royal Institute of Technology
Stockholm, Sweden
ilariat@kth.se

Iolanda Leite
KTH Royal Institute of Technology
Stockholm, Sweden
iolanda@kth.se

ABSTRACT

Machine learning systems have become ubiquitous into our society. This has raised concerns about the potential discrimination that these systems might exert due to unconscious bias present in the data, for example regarding gender and race. Whilst this issue has been proposed as an essential subject to be included in the new AI curricula for schools, research has shown that it is a difficult topic to grasp by students. We propose an educational platform tailored to raise the awareness of gender bias in supervised learning, with the novelty of using Grad-CAM as an explainability technique that enables the classifier to visually explain its own predictions. Our study demonstrates that preadolescents (N=78, age 10-14) significantly improve their understanding of the concept of bias in terms of gender discrimination, increasing their ability to recognize biased predictions when they interact with the interpretable model, highlighting its suitability for educational programs.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → **Interactive learning environments**; • **Computing methodologies** → *Machine learning*; *Artificial intelligence*.

KEYWORDS

Explainable AI; Interpretability; Transparency; Gender Bias; Education; Children

ACM Reference Format:

Gaspar Isaac Melsión, Iliara Torre, Eva Vidal, and Iolanda Leite. 2021. Using Explainability to Help Children Understand Gender Bias in AI. In *Interaction Design and Children (IDC '21)*, June 24–30, 2021, Athens, Greece. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3459990.3460719>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IDC '21, June 24–30, 2021, Athens, Greece

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8452-0/21/06...\$15.00

<https://doi.org/10.1145/3459990.3460719>

1 INTRODUCTION

The rapid integration of machine learning into our society is raising concerns about its potential discrimination in terms of gender and race caused by the unconscious bias introduced in it [7]. This has led to the necessity to incorporate new sustainable practices to reduce its societal implications, in which the field of Explainability in AI [42] has emerged as an essential piece to bring closer the advances in machine learning to society, including schools [11].

Recent studies have demonstrated the discrimination exerted by intelligent systems in different scenarios. Examples include the Facebook ad delivery system targeting mainly women for super-market cashier jobs, and men for those in the lumber industry [2]; the COMPAS system used in the US criminal justice services to assess recidivism risking being biased against black population [5]; and commercial facial recognition systems having the highest error rate for gender classification of darker-skin females [8]. These examples show that machine learning is subject to implicit bias. We approach this issue from an educational standpoint, proposing an educational platform to teach about gender bias in early ages. We evaluate a novel way to increase children understanding of the effects of biased data into a classification problem, and consider that our findings may contribute to the existing efforts of developing an inclusive AI curricula for school-age children and adolescents [38].

While previous studies focused on teaching machine learning and other AI techniques to children [15, 21, 47], the issue of the implications of bias in these systems has only been addressed in very few cases [3]. However, in these studies it has been noted that children have difficulties in understanding concepts associated with bias, posing an important challenge to be handled when teaching AI in schools. For example, one way in which gender bias manifests itself in machine learning is when the system consistently classifies an image as containing a woman when there are other stereotypically “feminine” objects in the same image, such as ironing boards or kitchen utensils. During similar image classification activities, Druga [15] noticed that children did not associate the surroundings of the object as a main factor for the system to misclassify it. Hence, Druga suggested that the concept of bias should be addressed when children experiment with this kind of technologies; one way of doing this could be to show the connection between the training dataset and the final outcome of the model. We propose the use of interpretable models to tackle this challenge, given that research

Figure 1: Our bias visualization tool integrated into our online educational platform. On top, images from the COCO-BinaryGender dataset are used to train the model. The highlighted parts in red on the predicted image represent the model visual explanation from Grad-CAM.

has shown that these models facilitate understanding of the models and their predictions [25, 34]. Starting from the claim in [30] that explainability techniques can have a positive impact in the education of AI to the general public, we present the first study evaluating its effectiveness among preadolescents.

We propose an online educational platform (<https://biaix.now.sh>) built using our bias visualization tool that integrates interpretable machine learning techniques to provide an explanation of the model decisions. Our platform was tailored for preadolescents to improve their understanding of the concept of bias, and increase their ability to discern the impact of gender bias. To our knowledge, this is the first tool integrating a machine learning explainability technique in an educational platform. Furthermore, this work also reports the first investigation on which methods can increase the understanding of algorithmic bias among children concretely.

2 RELATED WORK

2.1 Teaching children computer-related concepts

Technology is a useful resource in education that may contribute to improve the learning environment in specific contexts. For instance, hands-on activities [23] and new tools such as block-based programming languages [31] have introduced children to more advanced, commonly used computational techniques, such as tools

for data science, robotics, and AI [12, 47]. Notwithstanding, there are fewer studies addressing the possibility to teach children how machine learning works [1, 4, 15, 21, 49]. Notably, Hitron et al. [21] showed that preadolescents understand these concepts better when they have had the opportunity to experiment with every step of the process; and Alonso [4] developed a tool for children to learn what Explainable AI is. Similarly to some of these studies, we propose a tool that allows children to participate in the three main steps of a machine learning classifier: labeling, training and evaluation; however our work is distinctive, in that it focuses on the issue of gender bias and leverages state of the art explainability techniques to increase preadolescents' understanding of the notion of bias.

The advances in education of the latest technologies related to AI and its ubiquity in society have prompted researchers and educators from different parts of the world to propose and define a specific curriculum to teach AI in schools [3, 9, 22, 38]. The contributions presented in this work are aligned with the guidelines proposed by Touretzky et al. [38], which suggest that preadolescents should understand that biased datasets can affect system performance. Specifically, it is important that they learn that systems need to be capable of generalizing the given inputs; they also need to appraise the issues of transparency and fairness as relevant aspects to account for when identifying the societal impacts of AI. We also take inspiration from the program described by Payne et al. [3, 33], who designed an activity to teach the importance of training data into

machine learning algorithms, and how a biased dataset can affect its functioning. We recognize the key role of available online tools¹ into bringing machine learning technologies closer to non-expert populations, and include some of them partially into our study design as done in [33]. Our paper extends this body of work by introducing state of the art explainability techniques into an online platform aimed to raise awareness of the issue of gender bias in machine learning to preadolescents.

Moreover, children's perception of robots and agents powered by AI algorithms has been shown to be strongly impacted by previous interactions with them [46]. For example, a study conducted by Druga et al. [16] holds that participants aged 6 to 10 consider these agents to be friendly, trustworthy, and smart, with more than 70% considering voice-assistants to be smarter than themselves and always telling the truth. The fact that young children nowadays have more frequent interactions with AI could explain this finding. This also supports other findings that a greater exposure with robots at this age increases their perceived intelligence [6]. Similarly, Ullman et al. [39] could see a change in participants' perception of trustworthiness and intelligence depending on the behavior of a robot, which when acting in a dishonest manner was perceived to be even more intelligent than humans.

In this sense, the wide exposure of the population to AI has motivated HCI researchers to investigate what skills are necessary for people to interact effectively and critically with this technology, and what design principles should be used to improve public understanding of it [30]. The direct correlation between the dataset in use and bias in machine learning models is also considered important for educational matters, due to the different relationships that exist depending on the learner's social and environmental context in regards to the data used [45]. Our work is aligned to this line of thought, whereby we try to foster AI literacy in preadolescents. We do this mainly by following the defined competencies of "Critically Interpreting Data" and "Ethics", through the design considerations of "Explainability", "Contextualizing Data" and "Critical Thinking" defined by Long and Magerko [30].

2.2 Bias in machine learning

Current advances in AI and machine learning are occurring at a fast pace, becoming ubiquitous among the population. Many of us now make use of this technology on a regular basis, even in social and political institutions including education, hiring and criminal justice [32]. However, this societal change brings with it emerging concerns about the potential gender and racial discrimination that these systems could be reinforcing into the society. This has led to researchers and institutions to study and produce frameworks for AI systems to be deployed in a sustainable manner, taking into account their potential and shortcomings from a sustainability perspective [40].

Unconscious bias in machine learning systems is commonly derived from the use of a biased dataset to train the models [50]. It is also believed that these kinds of algorithms amplify an existing bias [48], worsening their discrimination potential. For instance, two very popular datasets used in computer vision research, ImageNet [13] and Microsoft COCO [29], contain biases regarding

ethnicity: more than 50% of the images in ImageNet come from the USA and UK, whereas China and India contribute to only 3% of the corpus [50]. There are also studies demonstrating gender bias. For example, image recognition systems perform more poorly for female subjects [8], and gender stereotypes tend to be magnified whenever contextual cues are taken into account for gender prediction [48]. Moreover, several research studies evaluated the presence and amplification of bias, providing the required background to propose different methods to mitigate gender bias [36].

Additionally, the majority of these systems reduce the expression of gender to the traditional view of it being binary and physiological. This fails to capture the complexities of gender identities [19], and there are clear concerns of the social and ethical impacts that automated gender recognition has in society [24]. In our work, we propose a visualization tool to raise the awareness of the gender bias present in the most popular research datasets, with the objective of making explicit that the traditional view of gender, i.e., that based on the binary gender construct and sexist stereotypes, is present in the data collected to train machine learning models. We claim that explainability may be useful to accomplish this within the general public, even at young ages.

2.3 Interpretable machine learning

The field of Fairness, Accountability and Transparency (FAccT) has recently seen an increase in popularity as a multidisciplinary area to address the challenges of discrimination and understandability of automated decisions. As such, it has the potential to encourage further research into the dangers of inadvertently encoding bias into these algorithms. Furthermore, explainability in AI systems has become an essential aspect to understand how bias can make these systems produce certain outcomes. It can also help determine which attributes have influenced the AI's decision [44], giving people the opportunity to better comprehend the functioning of the system. In recent years, there has also been a high proliferation of interpretability techniques focused on different aspects of AI reasoning, such as the kind of problem faced, differences in explanation, and the type of model and input data [18].

As a result, many recent HCI studies have included an interpretable model, for example focusing on the perception of trust, fairness and intelligibility of intelligent systems that are able to present an explanation of their own predictions [14, 17, 25, 28, 34]. For instance, Ribeiro et al. [34] demonstrated that the system's own explanations enable users to discern the best of two classifiers, even when the test accuracy measure suggested the contrary. Similarly, in [25] participants showed an improvement in identifying the features that did not help the model generalize well, enabling them to provide valuable feedback to refine the algorithm. Several factors have been shown to influence fairness perceptions of algorithmic decisions from a human perspective. Lee et al. [26, 27] suggested that mathematically fair algorithms are not perceived as fair as those made through group discussion by the participants; Wang et al. [43] found that people rate algorithms to be more fair when the outcome is favorable to them. Moreover, different kinds of explanations shown with the corresponding algorithmic decisions influence people's perceived fairness in various ways [14]. These findings highlight the potential and impact of interpretable machine

¹e.g., AI for Oceans, ml5.js, ML for Kids, Teachable Machine, Quick Draw.

learning models into issues of fairness, trust and transparency in regards to the user, and warrant further examination. We contribute to this area of research by applying some of the aforementioned techniques to increase children’s understanding of bias in machine learning.

3 BIAS VISUALIZATION TOOL

In this section we present our bias visualization tool that leverages an interpretable model offering visual explanations of its own predictions. We have integrated the Grad-CAM [35] explainability technique into an image captioning system, using the COCO dataset [29] for training, and built an online educational platform based on our tool to teach preadolescents about gender bias in machine learning. To the best of our knowledge, this is the first time an interpretability technique is used into an educational platform aimed at raising awareness of algorithmic bias and its societal implications among the population, specifically tailored for children.

Our bias visualization tool uses the information extracted from Grad-CAM [35] to create a visual representation of the relevant parts used by the model for a certain inference, with the same approach and system used in [20]. It is out of the scope of this study to develop a new model or dataset, and therefore the case study is limited to the binary genders “man” and “woman” used in the system. Although this poses an ethical dilemma of not including gender non-conforming identities, in this particular study this is considered suitable for showing the inherent gender bias and construct present in the COCO dataset.

The system is adapted to be presented as a classification problem, and the confidence intervals for each class – man and woman – are computed normalizing the obtained probabilities from a list of synonyms for each of the binary gender classes in the captioning system. A saliency map is then computed from the predicted class using Grad-CAM, which enables us to curate the visualization of the model explanation highlighting the parts on the image that have been more relevant for the inference. Fig. 1 contains a snapshot of our bias visualization tool, where the model’s explanation are the highlighted parts in red on the predicted image.

3.1 Dataset

We used the COCO dataset [29] for the development of the experiment. This dataset was chosen because, despite the fact that it has been shown to contain gender stereotypes [20, 48], it is used very often in machine learning research. Since the dataset does not contain a gender label, the process described by Zhao et al. [48] was adopted to label the images as man and woman. This process consists of adding the gender label when at least one of the captions mention any of these two words, and removing any image containing both. This subset of the dataset is referenced as *COCO-BinaryGender*. In order to provide evidence of the gender bias present in the dataset, here we report the results obtained in [48]. In fig. 2, it can be noticed that 86.6% of the objects are biased towards men by a small average value of 0.65, and 37.6% with a value higher than 0.7. Specifically, it can be seen that objects like snowboard or boat are biased towards men, whereas terms such as fork and handbag are biased towards women, and this bias is increased even after training. Therefore, for example an image including a person riding a snowboard will

Figure 2: Gender bias and amplification analysis of the COCO-BinaryGender dataset. It contains the bias score of the category man before and after training. Values near 1.0 indicate bias towards man, while values near 0.0 indicate bias towards woman. Red dots represent the categories, and the stars highlight some of those. Extracted from [48].

be more likely to be classified as man, and a person in the kitchen as woman.

3.2 Interpretable model

Our bias visualization tool is based on an interpretable model capable of giving an explanation of its own predictions. Given previous advances in natural language processing regarding gender bias [36], an image captioning system was used, adapted to finally provide the predicted gender word as a classifier interface for the purpose of the experiment. Specifically, we used the baseline model presented by Hendricks et al. [20] based on the *Show and Tell* architecture [41], which combines a Convolutional Neural Network (CNN) – InceptionV3 [37] – acting as image encoder with a Long-Short Term Memory (LSTM) as decoder to generate the sentences. It makes use of Grad-CAM [35] technique to make the model explainable. The model from [20] is pre-trained on the full COCO dataset, with an additional round of training on the actual COCO-BinaryGender subset for fine-tuning.

The Grad-CAM technique is specifically designed to be used on CNNs with image inputs providing a visual saliency mask of the relevant parts of the image for the inference task. This provides an intuitive interface to be used with children, and was compatible with the model and dataset in use. Grad-CAM exploits the gradient information on the last layer of the CNN to extract the importance of each neuron for the model output, and a more detailed description can be found in [35]. We apply Grad-CAM on the gender-specific word, i.e., woman or man, and not to a whole caption as suggested in [35], in order to obtain the saliency map corresponding specifically to the gender. Thereafter, the Grad-CAM mask obtained from this process is resized using a bilinear interpolation to the original image size.

Figure 3: Procedure of the experiment consisting of the interaction task with pre- and post-steps. The green boxes represent collected measures and the violet boxes are activities performed by the participants.

4 EXPERIMENT

4.1 Hypotheses

We formulated three hypotheses: one centered on the understanding of bias and its ethical implications (H_A); this is our main hypothesis, assessing the interactive activity given by our bias visualization tool. The second hypothesis focuses on generalising examples where bias may be present (H_B), aiming to identify children’s understanding of bias from an educational perspective. The third hypothesis addresses the relevance of training data into machine learning and the bias issue (H_C), to gain a better understanding of children’s ideas about how machine learning models work after completing the task.

H_A : The bias visualization tool will help preadolescents recognize that a machine learning system might perform biased predictions.

H_B : The bias visualization tool will help preadolescents identify biased models in unrelated settings.

H_C : The bias visualization tool will help preadolescents recognize training data as an essential part in machine learning, and consequently bias.

4.2 Experimental design

The study follows the procedure shown in fig. 3, which consists of the main interaction task accompanied with pre- and post-interaction steps. The bias visualization tool presented in section 3 is used with participants assigned to our treatment group; a modified version without the visualization of the model explanation is employed to create the control condition. Thus, in the *treatment* condition (2B in fig. 3), the visual explanation of the classifier is included together with the prediction, whereas in the *control* condition (2A in fig. 3) the system provides only the prediction results, that is the image without any highlighted parts.

From this configuration, two types of analysis can be performed, *between-* and *within-subject* comparisons, as depicted in fig. 3. The first type allows a direct examination of the effect of using the

interpretable model, derived from the comparison of the responses between the control and treatment groups. Secondly, the measures taken before and after the interactive part are used to inspect whether there is a difference in participants’ answers once they have completed the task, and compare their evolution depending on the two conditions.

Below, we first describe the activities that participants performed during the experiment, and then the related measures that we collected from these activities.

4.2.1 Activities. The activity in the pre-interaction step is meant to give participants a first general introduction to machine learning, specifically on the topic of image classification; here, the intention is to ensure that participants have a minimal understanding of the concepts required to perform the interaction step. It consists of playing with the Teachable Machine by Google [10], which is a tool tailored for children that allows people to train an image classifier using their own camera as input data.

Then, during the activity on gender bias in the interaction step, participants go through the three main steps of labeling, training and prediction to create a gender classifier from the images in the COCO-BinaryGender dataset using the interpretable model described in section 3.2. A balanced subset of the dataset containing 500 images of the two labeled genders in the dataset is used for the labeling part, where participants are asked to classify as many images as they wish with the man or woman labels. Whenever they wish to stop labeling the images, a button automatically completes the rest for them. Then, they can click on another button to simulate the training phase, after which the system will show its predictions. The prediction stage is divided in two parts. In the first part, a set of images with their prediction and confidence levels are shown to the participants. For the treatment group, these examples contain the explanation interface from our bias visualization tool (fig. 4a), whereas the control group observes the image without explanation. This is the only difference between conditions; everything else is constant, so that we can be confident that any difference in the

(a) First part of the prediction stage for the treatment group. The bias visualization tool includes the model’s explanation in the form of highlighted parts of the image. The control group would have the same without the explanation. Image shown is COCO ID 410019.

(b) Second part of the prediction stage including the *Bias Understanding* interactive measure. This is the same for both treatment and control groups. Image shown is COCO ID 14029.

Figure 4: Prediction stage of the gender bias activity during the interaction step in the experiment.

between-subject comparisons will be due to our condition manipulation. In the second part of the prediction, a different set of images is shown with their respective predictions, but without the interface from our bias visualization tool for all participants, regardless of condition. Instead, at this point the images are presented with their annotations to perform the interactive step, as seen in fig. 4b. The interactive step activity is described in section 4.2.2.

4.2.2 Measures. The list of measures can be seen in table 1. The questions on pre- and post-interaction are numbered as per order of appearance during the task, and they are all multiple-choice questions with one selectable answer. We divide the questions into four different groups depending on what they measure, as shown in table 1: (i) *Bias Understanding* contains the interactive measure during the main task assessing participants’ ability to grasp bias; (ii) *Bias Impacts* consists of Q_1 and Q_6 and is related to the impact of bias in society; (iii) *Unrelated Bias* comprises questions Q_3 , Q_4 and Q_5 and tackles the capacity of participants to apply their new knowledge of bias to unrelated settings; and (iv) *Machine Learning Concepts* includes questions Q_2 and Q_7 , related to general concepts on the functioning of machine learning and its relation to data.

The *Bias Understanding* measure in the interactive step is the same for every participant regardless of the condition group, and is repeated for a total of 13 images. From these images, 7 correspond to the man class (all classified correctly) and 6 to woman (3 misclassified). Participants are asked to select the part of the image they think was most relevant for the model to make the shown prediction. These parts are highlighted in different colors on the image as shown

in fig. 4b, and they correspond to the annotations provided in the COCO dataset, in addition to the background. The annotations have been limited to a maximum of three including the person in the image. That is, all images were presented with the following selectable parts: the person, one or two objects, and the background.

For this measure, the frequency of selecting the person annotation is compared to that of selecting any other annotation in the image, i.e., aggregating the background and the other objects (such as surfboard, oven, umbrella...). This measure is an evaluation of participants’ understanding of the effects of biased datasets, which might cause the model to make predictions based on wrong evidence. This is relevant for the analysis because an unbiased model is expected to base its gender prediction on the person in the image, rather than other elements. These other elements will likely be directly related to a gender stereotype, therefore enforcing gender inequalities in classification systems.

The *Unrelated Bias* measures were taken from the program “*An Ethics of Artificial Intelligence Curriculum for Middle School Students*” [3, 33]; to the best of our knowledge, ours is the first study to incorporate them in an experiment setting for children. We added a variation on the images accompanying these questions to show different classification examples on pre- and post-interaction steps. These questions simplify the evaluation of bias understanding using the examples in classification of dogs/cats and fish/reptiles.

4.3 Procedure

The experiment was run online, but advertised only by the teachers at a primary school in Stockholm to their students in Years 5 and 6

Table 1: List of measures during the experiment. The type of measure, answer options and at which step of the experiment takes place is included. The multiple-choice answers contain a shortened version of the actual text in the task.

Measure	Data	Answer options	Step
<i>Bias Understanding</i>			
Q_{INT} : What parts of the image do you think the Machine Learning system used to make this prediction? (select one)	Frequency of selection	Selectable parts: person, objects or background	Interaction
<i>Bias Impacts</i>			
Q_1 : Do you think all Machine Learning systems work equally well for everybody?	5-point Likert scale	Never–Always	Pre/Post
Q_6 : How did the Machine Learning system work?		Very bad–Very good	Post
<i>Unrelated Bias</i>			
Q_3 : A Machine Learning system has been trained with the following images. How do you expect it to work?	One correct answer	Adapted from [33, Supervised ML Quiz]	Pre/Post
Q_4 : Which set of images would you use to create the best Machine Learning system for both cats/reptiles and dogs/fish?			
Q_5 : A Machine Learning system has been trained with the following images. How do you expect it to classify the image on the right?			
<i>Machine Learning Concepts</i>			
Q_2 : Why do you think a Machine Learning system might work worse for certain people?	Conceptual answer	a) Depends on training data b) Depends on algorithm c) Same for everyone	Pre/Post
Q_7 : How could you improve the system?		a) Not possible b) Show more images c) Remove background	Post

(age group of 10–14 years old). The experiment was conducted in English, the school’s primary language. Before starting the experiment, each participant was asked to answer a demographic questionnaire recording their age, gender, nationality, self-assessment of experience in computer science, and English level. They were then directed to watch an introductory video² on machine learning for children; they then performed the activities and related questions described in sections 4.2.1 and 4.2.2. Finally, they were thanked and debriefed at the end of the study.

The experiment was built as a web application, accessible at <https://biaix.now.sh>, to conform with the ongoing COVID-19 recommendations. The model in the interaction step was pre-trained to improve the online experience, i.e., to avoid long delays, as well as the computed predictions with their explanation. The students participated in the experiment individually at the school or at home using a computer equipped with a camera, without specific supervision. All tasks were guided and unidirectional, that is, the participants were not allowed to go to previous steps, to assure the students proceeded in the task in the most comparable way. Participants were randomly allocated to the treatment and control conditions. The experiment took an average of 16 minutes.

4.4 Participants

In total, 84 participants completed the task. Of these, 8 took less than 5 minutes to complete the study, which is significantly less than the total duration of the whole task in our pilot experiments, and therefore this data was discarded from the analysis. In table 2

Table 2: Number of participants divided by condition.

	Total	Control	Treatment
Started	156	80	76
Finished	84	44	40
Duration >5min	76	38	38

Table 3: Contingency table by condition of *Bias Understanding* measure. It states the aggregated frequency of selecting person annotation.

Condition	Person	Other
Control	316	178
Treatment	272	222

we report the number of participants in our experiment grouped by condition. As can be noted, the final number of analyzed participants was 38 per condition.

The students were between 10 and 14 years old ($M = 12.25$, $SD = 0.82$). Regarding gender, 46 were females, 28 male, 1 other and 1 chose to not disclose it, and this proportion was balanced for each condition. The majority of participants (72) came from Sweden, and only one of them reported to have a low level of English. In terms of self-reported experience in computer science-related topics, 90% chose the options “none” or “little”.

²Code.org. AI: What is Machine Learning? Accessible at <https://youtu.be/OeU5m6vRyCk> (January, 2021)

Figure 5: Participants' answers to *Bias Understanding* measure on the interaction step, counting the selection of the person annotation. (a) Frequency of selecting the person in the image compared to other regions/objects aggregated for all the images, by condition. (b) Frequency of selecting the person for each image individually by condition, and in magenta is represented the percentage change between conditions in absolute value.

5 RESULTS

We first analyzed the results from the *Bias Understanding* and *Bias Impacts* measures (see table 1), which are associated to hypothesis H_A – evaluating the effects of the explainable model towards participants' understanding of bias and its implications in terms of gender. In fig. 5, we show the aggregated answers to question Q_{INT} , from the *Bias Understanding* measure. These data correspond to the main evaluation of hypothesis H_A , since they assess people's interaction with our proposed bias visualization tool. These results are studied from two different points of view: the overall frequency of selecting a certain part of the image (fig. 5a), and the same frequency broken down by individual image (fig. 5b).

From the aggregated results, we can see that the control group selected the person annotation more frequently than participants in the treatment condition (64% and 55% of the time, respectively). In addition, looking at each image separately, it is observed that participants in the treatment condition selected the person annotation less frequently in every image except one – where the two annotations were selected an equal number of times. A chi-square test of independence on the frequency of selecting the person annotation (table 3) confirmed that these differences are statistically significant ($\chi^2(1, N = 988) = 8.1, p = .004$). These results support hypothesis H_A , that using an interpretable model (as part of our bias visualization tool) has a positive effect on participants' ability to recognize the impact of unconscious bias introduced into supervised learning systems, being more capable to identify the reason behind its biased predictions – in this case, the presence of objects associated with certain gender stereotypes.

Secondly, the results from the *Bias Impacts* measure (Q_1 and Q_6) can be seen in fig. 6. In this case, the difference between conditions

Table 4: Contingency table by condition of Q_6 from *Bias Impacts* measures. It states the aggregated count of negative, neutral and positive answers.

Condition	Negative	Neutral	Positive
Control	3	20	15
Treatment	8	19	11

Table 5: Contingency table by condition of *Unrelated Bias* measures. It states the aggregated number of correct and incorrect answers.

Condition	Correct	Incorrect
Control	53	61
Treatment	60	54

are not statistically significant, as specified by a t-test of independence ($t(75) = 0.45, p = .655$) between the means of the answers in Q_1 after the interaction; and a chi-square test of independence comparing the aggregated answers as reported in table 4 from Q_6 , $\chi^2(2, N = 76) = 2.9, p = .233$. Despite this, some interesting trends can be observed from the *Bias Impacts* measure: in terms of Q_1 (Do you think all Machine Learning systems work equally well for everybody?), overall the answers moved from positive toward neutral after the interaction step (fig. 6a); and regarding Q_6 (How did the ML system work?), in fig. 6b we can see an increase in the number of participants answering a *Negative* option (*Bad* or *Very bad*) when our bias visualization tool is used, with 21.1% of participants in the treatment group selecting one of those answers compared to 7.9% in the control group. These results suggest that, after interacting with the bias visualization tool, participants seem to get some understanding of the impact of data bias in the model predictions.

Hypothesis H_B , centered on the capacity of participants to identify biased models on unrelated settings to those of the interactive activity, is associated with the *Unrelated Bias* measure (Q_3 , Q_4 and Q_5) reported in fig. 7. A chi-squared test of independence on the frequency of selecting a correct answer in the treatment and control condition was not statistically significant ($\chi^2(1, N = 28) = 0.63, p = .427$, table 5).

From fig. 7a we can still observe a trend for participants in the control condition to perform more poorly after the interaction (median number of correct answers = 2 pre-interaction and = 1 post-interaction). On the other hand, participants in the treatment condition performed equally well pre- and post-interaction (median number of correct answers = 1.5), with more people overall selecting all 3 correct answers. Fig. 7b further shows the difference between the number of correct responses after the interaction and before the interaction: the control group shows a larger frequency of negative values compared to the treatment group (31.6% and 21.1%, respectively). These results suggest that participants using our bias visualization tool with the interpretable model outperformed those

(a) Q_1 . Distribution of the answers on the pre- and post-interaction for each group.

(b) Q_6 . Frequency by condition.

Figure 6: Participants' answers to questions in *Bias Impacts* measure.

in the control group. This supports the notion of hypothesis H_B since being able to answer these questions correctly shows the potential in terms of transparency that an interpretable model may provide to grasp and generalize bias concepts.

The *ML Concepts* measure (Q_2 and Q_7) presented in fig. 8 provide a more conceptual insight into the participants' ability to understand the cause of bias. They are associated to hypothesis H_C , assessing whether participants recognize training data to be important in machine learning biased decisions. Regarding Q_2 , it can be observed in fig. 8a that all participants' frequency selecting the *Same for everyone* option increased after the interaction. This is accentuated for the treatment group, where this answer increased by more than double than in the control group (233.3% and 100%, respectively). A possible reason for this increment might be that visualizing the explanations from the model induced the idea of the system working equally for all images, where even correct predictions were supported by the other regions in the image than the person itself. This suggests that the explanation increased the transparency perception of the system, and participants

(a) Distribution of the number of correct answers on pre- and post-interaction for each condition.

(b) Frequency of the difference in number of correct answers between pre- and post-interaction. A positive value on the x-axis indicates that participants selected more correct answers after the interaction.

Figure 7: Participants' answers to *Unrelated Bias* measures group (Q_3 , Q_4 and Q_5).

could see that it was not capable of discerning the subject in the images.

Regarding Q_7 , fig. 8b shows that 94.7% of all participants considered it possible to improve the system; specifically, 58.8% of participants would show more images to the classifier and the rest considered removing the background from the images. Both answers are correct in this case because either option would improve the system to be less biased in general towards the gender stereotypes, although the option of removing the background from the image is more directly related to the idea of biased data. The explainable model did not show to have an impact on this results, since there is no important difference between conditions; however it is clear that in general after performing the task, the participants perceived the training data to be key in improving the system, partially supporting H_C .

Figure 8: Frequency of participants' answers to ML Concepts measure. (a) Q_2 Why do you think a ML system might work worse for certain people?; (b) Q_7 How could you improve the system?.

6 DISCUSSION

Our study provides evidence that using an interpretable model has a positive impact on preadolescents' understanding of bias in machine learning, thus supporting our working hypotheses. The results demonstrate that our bias visualization tool has a notable effect in understanding the impact of gender bias in supervised learning, as proposed in hypothesis H_A , where we envisaged that preadolescents would better recognize biased predictions when interacting with an explainable model. In addition, although our sample has been shown not to be statistically significant, there are promising insights into the analyzed data supporting the ideas behind hypotheses H_B and H_C , on the participants being able to generalize the bias concepts to unrelated settings and identify training data as a key aspect in it. We observe that participants using our bias visualization tool were better at recognizing bias in unrelated examples, and overall the task provided the knowledge to identify the training dataset as an essential part in the functioning of machine learning systems.

The results obtained from the *Bias Understanding* measure that confirm H_A resonate with the ideas introduced by Ribeiro et al. [34] and Kulesza et al. [25], where it was found that adult participants who were shown an explainable model improved on recognizing the features that did not help the model generalize well. Our work contributes to expanding the use of interpretability to increase public understanding of bias impacts in machine learning. Moreover, the confirmation of H_A demonstrates the benefits of using an explainable model to comply with the "AI for K-12" curriculum's guidelines by Touretzky et al. [38], where it is denoted the importance of children in the targeted age group to recognize the relevance of transparency and understand the effects of bias regarding the societal impacts of AI. At the same time, our study contributes to the validation of interpretability to be considered more recurrently in the design of educational tools to increase the AI literacy among the non-expert population, as suggested by Long and Magerko [30].

Hypothesis H_B was partially confirmed, in that using an explainable model in the main activity tended to increase participants'

ability to understand the idea of bias. To the best of our knowledge, no prior study evaluated the questions presented by Payne [33] on unrelated bias (extending the understanding of bias to other concepts, such as a cat and dog image classifier, as described in section 4.2.2). With our *Unrelated Bias* measure, we found evidence that preadolescents were better able to extend their knowledge of bias to new datasets when the explainable model was used, hence providing valuable insights for future studies employing these questions. This is a strong indicator of participants' better understanding of bias, because it suggests that they were capable of transferring this newly acquired knowledge to different situations. Additionally, in relation to H_C , participants were able to provide solutions to improve the system by identifying training data as an essential part in machine learning systems after the experiment, as observed in the answers to the *ML Concepts* questions. Our results are thus similar to the findings in [21], where the uncovering of all steps in supervised learning indicated a better grasp of concepts such as data labeling and sample size.

On the other hand, the responses to the *Bias Impacts* measure contrast with the figures in [16], where 70% of participants (6–10 years old) found intelligent assistants to always tell the truth. Participants in our study held a more critical view towards the machine learning system after performing the task – this was accentuated when they used our bias visualization tool. Specifically, they were not convinced by the idea of the system working equally well for everyone after seeing the biased results, and interacting with an explainable model through our bias visualization tool lowered their rating of the system's performance. This gives insight into how explainability could be used to raise awareness among the general public of the intelligence attribution to AI-powered systems and their limitations.

Finally, our work provides evidence of the positive effects of using an interpretable model to teach about the impacts of biased data in machine learning. Although the type of data and the specific instances shown of it, i.e., images from COCO dataset in our case, might affect to each student's notion of bias depending on their relationship to the data itself, in previous similar educational studies, general concepts related to bias in image classification were difficult for children to learn about. For example, in [15] students did not conceive the surroundings of an image as important factors in the AI misclassification; our results suggest that an explainability module would be helpful in this regard. Accordingly, our work can be relevant for those developing new AI curricula for schools; to our knowledge, this is the first controlled study demonstrating the effectiveness of using an explainable model in teaching bias and its implications in machine learning. As future ideators and users of AI technology, it is important that children are given the instruments to recognize all instances of possible bias, including unintentional one. Our visualization tool contributes to this goal.

6.1 Limitations and Future work

A limitation of our study is that the online nature of the experiment did not allow us to collect data about participants' impressions of the task. This would have potentially given us more insight into the reasoning behind their choices, and a more sound understanding

of what the participants thought about the different compounds of algorithmic bias, e.g., if it originated from the data or the model. Furthermore, given that we had two participants who chose one of the non-binary gender options, it would have been relevant to see their perception regarding a binary gender classifier and their opinion of the images conforming the COCO dataset. As an inherent limitation for any system and study that uses binary gender classification, there is the risk of exclusion of the gender non-conforming population.

Future work could explore whether repeated exposure (e.g., more than one learning session) to the tool would make the effects stronger. It would also be interesting to examine other visualization strategies or non-visual cues, e.g., natural language explanations combined with visuals; and expand our tool to work for systems other than an image classifier, for other biases, and to allow users to input their own images to understand bias on a smaller scale. While the online platform was originally primarily created to collect data for the current study, we plan to adapt it and release it as a free learning resource. We will modify it to be a self-contained activity that is also valuable for non-binary populations. Our aim is to enable interested children to learn about machine learning at home, as well as to allow teachers to use the tool at school.

7 CONCLUSION

We presented a novel bias visualization tool leveraging state of the art explainability techniques in machine learning to improve the understanding of bias among children. Following indications from ongoing research on learner-centered AI and on new AI curricula for schools, our work investigated the effectiveness of using an interpretable model to increase the ability of preadolescents to discern biased predictions from an image classifier. Concretely, we focused on gender discrimination that is perpetuated through machine learning models when the content present in popular training datasets used in academia is highly biased towards the historical stereotypes of women and men.

To the best of our knowledge, this is the first work combining the fields of educational machine learning for children and the latest advances of Explainability in AI, building on previous efforts to make explainable machine learning useful for the general public. The study we conducted at a local school through an online learning platform has shown that preadolescents using our bias visualization tool significantly improved their ability to identify gender-biased predictions. This contributes to the validation of Explainable AI for educational purposes in order to raise awareness on the importance of transparency and the societal implications of bias in machine learning models.

ACKNOWLEDGMENTS

We thank Elizabeth J. Carter for the insightful feedback on the study design, and are tremendously grateful to Sofia Partsinevelou for facilitating the study at IES Södertälje. This work was partially funded by a grant from the Swedish Research Council (reg. number 2017-05189), the Swedish Foundation for Strategic Research (SSF FFL18-0199), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) by the Knut and Alice Wallenberg Foundation.

SELECTION AND PARTICIPATION OF CHILDREN

The study was run in collaboration with the school IES Södertälje, and the teachers were the only ones advertising the activity to their students of 10–14 years old. The study was optional and anonymous for the students, and they conducted it in class or at home depending on COVID-19 regulations. It included an explanatory introduction, and a final debrief of the motivation and objectives of the study. The images used were pre-selected to contain appropriate content for children. The data excluded from the analysis corresponds only to the participants who did not finish the online task, or did it quicker than a predefined threshold. The study was approved by the teachers at the school, as well as by the university.

REFERENCES

- [1] Adam Agassi, Hadas Erel, Iddo Yehoshua Wald, and Oren Zuckerman. 2019. Scratch Nodes ML: A Playful System for Children to Create Gesture Recognition Classifiers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*. ACM Press, New York, New York, USA, 1–6. <https://doi.org/10.1145/3290607.3312894>
- [2] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 199 (Nov. 2019), 30 pages. <https://doi.org/10.1145/3359301>
- [3] Safinah Ali, Blakeley H Payne, Randi Williams, Hae Won Park, and Cynthia Breazeal. 2019. Constructionism, Ethics, and Creativity: Developing Primary and Middle School Artificial Intelligence Education. In *International Workshop on Education in Artificial Intelligence K-12 (EDUAI)*. 1–4.
- [4] Jose M. Alonso. 2020. Teaching Explainable Artificial Intelligence to High School Students. *International Journal of Computational Intelligence Systems* 13, 1 (Jul 2020), 974–987. <https://doi.org/10.2991/ijcis.d.200715.003>
- [5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [6] Debra Bernstein and Kevin Crowley. 2008. Searching for Signs of Intelligent Life: An Investigation of Young Children's Beliefs About Robot Intelligence. *Journal of the Learning Sciences* 17, 2 (April 2008), 225–247. <https://doi.org/10.1080/10580400801986116>
- [7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf>
- [8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. <http://proceedings.mlr.press/v81/buolamwini18a.html>
- [9] Harald Burgsteiner, Martin Kandhofer, and Gerald Steinbauer. 2016. IRobot: Teaching the Basics of Artificial Intelligence in High Schools. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (Mar. 2016). <https://ojs.aaai.org/index.php/AAAI/article/view/9864>
- [10] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable Machine: Approachable Web-Based Tool for Exploring Machine Learning Classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, 1–8. <https://doi.org/10.1145/3334480.3382839>
- [11] Cristina Conati, Kaska Porayska-Pomsta, and Manolis Mavrikis. 2018. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. *CoRR abs/1807.00154* (2018). [arXiv:1807.00154](http://arxiv.org/abs/1807.00154)
- [12] Sayamindu Dasgupta and Benjamin Mako Hill. 2017. Scratch community blocks: Supporting Children as data scientists. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, Vol. 2017-May. ACM Press, New York, New York, USA, 3620–3631. <https://doi.org/10.1145/3025453.3025847> [arXiv:1702.00112](http://arxiv.org/abs/1702.00112)

- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [14] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, 275–285. <https://doi.org/10.1145/3301275.3302310>
- [15] Stefania Druga. 2018. *Growing up with AI Cognimates: from coding to teaching machines*. Master's thesis. Massachusetts Institute of Technology. <https://www.media.mit.edu/posts/kids-teach-ai-a-little-humanity-with-cognimates/>
- [16] Stefania Druga, Cynthia Breazeal, Randi Williams, and Mitchel Resnick. 2017. "Hey Google is it ok if I eat you?" Initial explorations in child-agent interaction. In *IDC 2017 - Proceedings of the 2017 ACM Conference on Interaction Design and Children*. ACM Press, New York, New York, USA, 595–600. <https://doi.org/10.1145/3078072.3084330>
- [17] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International Journal of Human Computer Studies* 58, 6 (June 2003), 697–718. [https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7)
- [18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (Aug. 2018), 1–42. [arXiv:1802.01933](https://doi.org/10.1145/3236009)
- [19] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender reductionism? The social implications of Automatic Gender Recognition systems. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. ACM Press, New York, New York, USA, 1–13. <https://doi.org/10.1145/3173574.3173582>
- [20] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- [21] Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. 2019. Can Children Understand Machine Learning Concepts?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–11. <https://doi.org/10.1145/3290605.3300645>
- [22] Meng Jing. 2018. China looks to school kids to win the global AI race. <https://www.scmp.com/tech/china-tech/article/2144396/china-looks-school-kids-win-global-ai-race>
- [23] Rose Johnson, Venus Shum, Yvonne Rogers, and Nicolai Marquardt. 2016. Make or shake: An empirical study of the value of making in Learning about computing technology. In *Proceedings of IDC 2016 - The 15th International Conference on Interaction Design and Children*. ACM Press, New York, New York, USA, 440–451. <https://doi.org/10.1145/2930674.2930691>
- [24] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–22. <https://doi.org/10.1145/3274357>
- [25] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, Vol. 2015-Janua. ACM Press, New York, New York, USA, 126–137. <https://doi.org/10.1145/2678025.2701399>
- [26] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. ACM Press, New York, New York, USA, 1035–1048. <https://doi.org/10.1145/2998181.2998230>
- [27] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management that Allocates Donations to Non-Profit Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, Vol. 2017-May. ACM Press, New York, New York, USA, 3365–3376. <https://doi.org/10.1145/3025453.3025884>
- [28] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*. ACM Press, New York, New York, USA, 2119. <https://doi.org/10.1145/1518701.1519023>
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.
- [30] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3313831.3376727>
- [31] John Maloney, Mitchel Resnick, Natalie Rusk, Brian Silverman, and Evelyn Eastmond. 2010. The Scratch Programming Language and Environment. *ACM Transactions on Computing Education* 10, 4 (Nov. 2010), 1–15. <https://doi.org/10.1145/1868358.1868363>
- [32] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.
- [33] Blakeley H Payne. 2019. *An Ethics of Artificial Intelligence Curriculum for Middle School Students*. <https://docs.google.com/document/d/1e9wx9oBg7CR0s5O7YnYHVmX7H7pnTfoDxDxNdrSGkp60/edit>
- [34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 13-17-Aug. ACM Press, New York, New York, USA, 1135–1144. <https://doi.org/10.1145/2939672.2939778> arXiv:1602.04938
- [35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* (Oct. 2019). <https://doi.org/10.1007/s11263-019-01228-7>
- [36] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mei ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1630–1640. <https://doi.org/10.18653/v1/P19-1159> arXiv:1906.08976
- [37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2016-Decem. IEEE, 2818–2826. <https://doi.org/10.1109/CVPR.2016.308> arXiv:1512.00567
- [38] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for K-12: What Should Every Child Know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), 9795–9799. <https://doi.org/10.1609/aaai.v33i01.33019795>
- [39] Daniel Ullman, Iolanda Leite, Jonathan Phillips, Julia Kim-Cohen, and Brian Scasellati. 2014. Smart Human, Smarter Robot: How Cheating Affects Perceptions of Social Agency. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 36. 2996–3001. <https://escholarship.org/uc/item/2jh800n1>
- [40] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fusco Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications* 11, 1 (Jan. 2020), 233. <https://doi.org/10.1038/s41467-019-14108-y> arXiv:1905.00501
- [41] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (April 2017), 652–663. <https://doi.org/10.1109/TPAMI.2016.2587640> arXiv:1609.06647
- [42] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, 1–15. <https://doi.org/10.1145/3290605.3300831>
- [43] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–14. <https://doi.org/10.1145/3313831.3376813>
- [44] Maranke Wieringa. 2020. What to account for when accounting for algorithms. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona, Spain, 1–18. <https://doi.org/10.1145/3351095.3372833>
- [45] Michelle Hoda Wilkerson and Joseph L. Polman. 2020. Situating Data Science: Exploring How Relationships to Data Shape Learning. *Journal of the Learning Sciences* 29, 1 (Jan 2020), 1–10. <https://doi.org/10.1080/10508406.2019.1705664>
- [46] Randi Williams, Hae Won Park, and Cynthia Breazeal. 2019. A is for Artificial Intelligence: The Impact of Artificial Intelligence Activities on Young Children's Perceptions of Robots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland UK) (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3290605.3300677>
- [47] Randi Williams, Hae Won Park, Lauren Oh, and Cynthia Breazeal. 2019. PopBots: Designing an Artificial Intelligence Curriculum for Early Childhood Education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9729–9736. <https://doi.org/10.1609/aaai.v33i01.33019729>
- [48] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (July 2017), 2979–2989. <https://doi.org/10.18653/v1/D17-1323> arXiv:1707.09457

- [49] Abigail Zimmermann-Niefield, Makenna Turner, Bridget Murphy, Shaun K. Kane, and R. Benjamin Shapiro. 2019. Youth Learning Machine Learning through Building Models of Athletic Moves. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children (IDC '19)*. Association for Computing Machinery, 121–132. <https://doi.org/10.1145/3311927.3323139>
- [50] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist — it’s time to make it fair. *Nature* 559, 7714 (July 2018), 324–326. <https://doi.org/10.1038/d41586-018-05707-8>