A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education

Jay Quedado
Computing and Software Systems,
University of Washington Bothell
Bothell, Washington, USA
jayque@uw.edu

Annuska Zolyomi
Computing and Software Systems,
University of Washington Bothell
Bothell, Washington, USA
annuska@uw.edu

Afra Mashhadi
Computing and Software Systems,
University of Washington Bothell
Bothell, Washington, USA
mashhadi@uw.edu

ABSTRACT

As demonstrated by media attention and research, Artificial Intelligence systems are not adequately addressing issues of fairness and bias, and more education on these topics is needed in industry and higher education. Currently, computer science courses that cover AI fairness and bias focus on statistical analysis or, on the other hand, attempt to bring in philosophical perspectives that lack actionable takeaways for students. Based on long-standing pedagogical research demonstrating the importance of using tools and visualizations to reinforce student learning, this case study reports on the impacts of using publicly-available visualization tools used in HCI practice as a resource for students examining algorithmic fairness concepts. Through qualitative review and observations of four focus groups, we examined six open-source fairness tools that enable students to visualize, quantify and explore algorithmic biases. The findings of this study provide insights into the benefits, challenges, and opportunities of integrating fairness tools as part of machine learning education.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Social and professional topics** → *Computing literacy*.

KEYWORDS

Fairness, Ethics, Tools

ACM Reference Format:
Jay Quedado, Annuska Zolyomi, and Afra Mashhadi. 2022. A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3491101.3503568

1 INTRODUCTION

In the past decade, technologists have created, researched, and deployed machine learning (ML) applications and underlying artificial intelligence (AI) systems in increasingly diverse domains with direct societal impact on daily lives. Machine learning is no longer the invisible engine behind recommendations and spam filters; it is increasingly used for consequential decision-making in many scenarios, including filtering loan applicants, allocating policing to areas of crime, and informing bail and parole decisions. Researchers, technologists, and the media have expressed significant concerns about the potential of these data-driven methods to introduce and perpetuate discriminatory and unfair practices [4, 27, 35].

At the same time, academia has seen an unprecedented amount of interest in studying fairness in machine learning [15]. Conferences such as the ACM Conference on Fairness, Accountability, and Transparency publish nearly 100 papers on fairness and transparency every year. Many other top-tier machine learning conferences have added a track on societal implications and algorithmic biases to their main tracks. Other academic conferences have revised their peer-review processes to assess ethical implications regarding data practices and possible biases. Despite the volume and velocity of published work, computer science (CS) education falls short in educating *students* on quantifying and assessing algorithmic biases. Indeed to date, the majority of CS student's training about responsible AI comes from the notion of *the ethical responsibilities*, typically through a course dedicated to societal aspects of computing. For example, in 2019, educators proposed a new pedagogical approach, Embedded EthiCS [23], that combines two key tactics: (1) interspersing ethical discussions throughout the curriculum, and (2) engaging the help of faculty from philosophy backgrounds to co-teach relevant material. Although Embedded Ethics is an excellent example of promoting ethics integration into CS education, it does not meet some of the educational components required for CS students to understand responsible AI practices. Making fairness a central element of machine learning courses would empower students to think not only about what algorithm they *could* create but also whether they *should* create the technology and, if so, how to make decisions that lead to designing inclusive and equitable algorithms and systems.

Pedagogical studies have demonstrated that tools can help reinforce student learning [34, 36]. Interactive tools can provide opportunities for experiential learning through questioning, investigating, reflecting, and conceptualizing based on direct experiences. Recently, researchers have shown that interactive algorithm visualizations strongly support active learning [33]. In our research, we hypothesize that student learning about algorithm biases and technical competencies can be shaped by using *tools that support live interaction with data.* In this case study, we examined our hypothesis and explored how tools can: (1) create a deeper reflection of fairness topics for novice machine learning students, (2) create opportunities for active learning, and (3) foster opportunities for students to hold ethical debates through exploratory analysis. We

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3503568

conducted a survey and focus groups with a class of undergraduate CS students at a public university. The students reviewed six existing tools designed to assess and *visualize* algorithmic fairness. Our findings demonstrated that student learning was reinforced through the use of interactive tools and tools that allow users to examine custom high-dimensional datasets (such as images). Such tools helped bridged the gap between theoretical concepts of fairness and observable consequences of biased decision-making by enabling students to examine counterfactual points. Based on the study results, our work calls for ML and AI visualization tools to become more transparent, interactive, and adaptive to help students explore, learn, and debate fair choices in AI.

2 INTEGRATION OF TOOLS AND ETHICS IN CS EDUCATION

2.1 Tools in CS Education

When teaching AI, educators need to teach students about the construction and use of statistical models. Typically, educators focus on understanding and using algorithms through text-heavy materials and tools rarely supplemented with algorithmic visualizations (AVs). However, educators have found that interactive AVs—those that support control of animation and manipulation of visual representations—are more effective and enjoyable for learners [7]. Amershi et al. [7] discuss the value of interactive AVs as including increasing student motivation and focus, decreasing student stress, and improving long-term learning. They identify five pedagogical goals in the design of interactive AVs: increase understanding of the target domain, support individual differences, motivate the focus student attention, promote active engagement, and support various learning activities. Naps et al. [33] also outlines eleven good practices of integrating AV into CS pedagogy, including support for custom datasets and complementing visualizations with explanations. Similar to those goals, we believe that fairness tools can help students explore dimensions of ML fairness through visualization, communicate biases of data and algorithms more effectively and allow students to think critically and debate ethical decisions.

2.2 Ethics in CS Education

A common criticism of ethics training in CS is that ethics has traditionally been treated as an after-thought worthy of a module or guest lecture at the conclusion of regular coursework, as opposed to being more seamlessly woven into the course material at every stage [18, 32]. Others such as [11, 23] advocate for a more "modular" approach that would focus on developing an *inquiry framework* to prompt students to ask critical ethical questions, for example, "how do you know the *data* is ethically available for its intended use". Stepping away from ethical data challenges and focusing on the *ethical algorithms*, there are fewer studies and pedagogical frameworks to follow. In [26], the authors highlight the ongoing need to develop new tools and methods for cultivating students' ethical sensibilities. They suggest engaging more directly with projects of certain ('banned') technologies, such as facial recognition systems, to promote debate. They argue that in this way, data science ethics education has an opportunity to learn from its precursors in engineering and computer science, and thus, develop a more expansive terrain of ethical engagement and debate. Finally, in terms of examining tools as a means for quantifying fairness, related work is limited to a recent study by Lee and Singh [30]. They studied fairness toolkits to explore the needs of data science *practitioners*. Their study did not encompass novice programmers or student populations. They identified unmet needs regarding tool functionality, user-friendliness, contextualization, and customization. In contrast, our work explores fairness tools in the context of education and curriculum development with an emphasis on integrating responsible AI principles into machine learning education.

3 METHODOLOGY

3.1 Definition of Fairness

The vast majority of work to date on fairness in machine learning has focused on the task of batch classification. Most of the literature on fair classification focuses on statistical definitions of fairness. This set of definitions specifies a small number of protected groups (e.g., demographic) and then asks for (approximate) parity of some statistical measure across all of these groups. Popular measures include statistical parity [21] (also known as demographic parity, independence, and group fairness), conditional statistical parity [16], false positive and false negative rates (also sometimes known as equalized odds) [24], and predictive parity[14].

3.2 Scope

Recently, various open source *fairness toolkits* have emerged to make the fairness methods more widely accessible. Although not explicitly mentioned by toolkit creators, it is assumed that many of these toolkits were designed for model developers in commercial settings and for researchers in their work to improve fairness testing and bias mitigation. In our survey of tools, we established the following inclusion criteria:

- The tool must be open-source and freely available to the public. We excluded tools with a restricted user base, such as FairFlow [22], which is only available to Facebook registered developers, since our surveyed tools need to be accessible to educators and students.
- The tool must have been designed to focus on fairness assessment. We excluded tools and packages used by the research community for *mitigation only*. We included this criterion to align the focus of the tool with common AI/ML curriculum, which currently focuses on fairness assessment as opposed to other considerations such as bias mitigation.
- The tool must have a visualization component that is integral to its function and allows users to visualize fairness metrics. We excluded fairness packages that did not support embedded visualizations, including scikit-fairness [40], Themis-ml [9], and DeepInspect [13].

Our inclusion criteria guided us in curating a set of fairness toolkits to examine within the context of machine learning education; note that our intention was not conduct a holistic review of all exiting tools.

3.3 Method

Prior research on ML fairness involved focus groups, interviews, and surveys of data science practitioners [30], and we followed a

similar methodological approach for our study focusing on higher education. Focus groups allow research participants to share experiences and co-construct ideas in unfamiliar knowledge areas. We conducted research with undergraduate CS students at a public university who were enrolled in a course on *fairness in machine learning*. The prerequisites for the course included familiarity with programming and a basic understanding of statistics. At the onset of the course, students self-reported via a survey that their existing knowledge of machine learning was, on a scale of 1 to 5, an average of 1.5 (ranging from 1 to 3). The class introduced machine learning concepts in classification and regression problems, combined with evaluating biases of the training data.

Over the course of four focus groups with a total of 20 students, students discussed their initial notions of ML fairness concepts, fairness tools, and perspectives on ML fairness based on their explorations of visualization tools. At the time of the focus groups, all students had a well-developed understanding of statistical biases and were familiar with the fairness criteria described earlier in this paper and defined in the course textbook [10]. The students were novice users of ML systems and fairness evaluations. We conducted an initial focus group for exploratory analysis of the student's familiar with ML fairness. During that focus group, students were given guidance for conducting a tool exploration, which they completed over a two week time frame. The students were instructed to work in groups to find three fairness tools that met the inclusion criteria described earlier. Any tools that were outside of this scope were deemed not relevant and thus were not discussed in the later stage. The groups each selected three tools, completed a sample demo of each tool, and constructed a tutorial. Upon completion of their tool explorations, we conducted a second session to collect information about their learning and concerns based on their explorations. We concluded data collection with a survey asking the students the following open-ended questions:

- What aspects of each tool do you believe contributed to your learning?
- What aspects of each tool did you find challenging?
- What are the opportunities to improve the tools to integrate it into the ML curriculum?

The research team conducted a thematic analysis of the focus group discussions and survey results to identify meaningful themes.

4 SURVEYED TOOLS

In this section, we review a set of existing tools that could support the education of CS students learning about responsible AI, such as within introductory machine learning classes. In our survey of each tool, we describe platform requirements and the supported fairness criteria. To provide a stepping stone for fairness training in CS education, we have created a repository with an example notebook of each tool [37].

4.1 Aequitas

Aequitas is a bias auditing toolkit developed by the Center for Data Science and Public Policy at the University of Chicago through their Data Science for Social Good program (DSSG). Aequitas tools support three steps: indication of which sensitive groups are analyzed and which are set as the reference group, analysis of bias

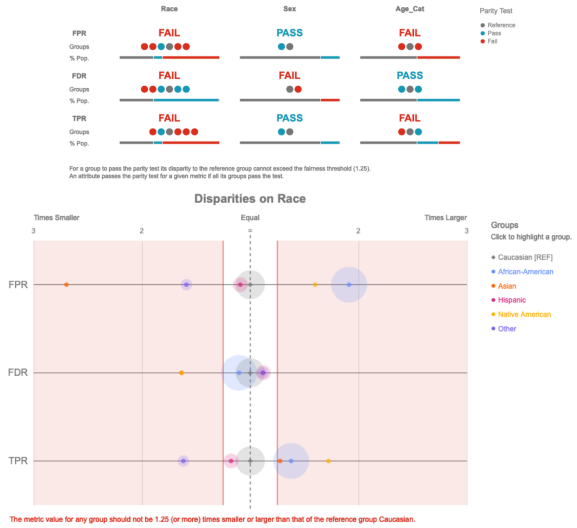

Figure 1: Aequitas visualization of the fairness criteria.

based on desired metrics (false negative, false discovery, etc.), and determination of a final pass/fail fairness indicator. Overall the tools are meant to generate reports highlights disparities in predictive classification models. Aequitas offers visualizations of group parity examples shown in Figure 1 (left), including graphics for parity tests and the magnitude of disparities between races using a simple traffic light fail/pass visualization. Figure 1 (right) shows the magnitude of racial disparities and their impacts on the fairness metrics. Aequitas offers tutorials for visualizing different disparities against user-designed attributes.

Visualizations can be implemented through an online web application, a Python library, and a command-line tool. The Aequitas Bias Report web application uses two datasets for their demo, the COMPAS Recidivism Risk Assessment dataset and the US Adult Income dataset [20], although users have the option to upload their own dataset. To use a custom dataset, the dataset must adhere to a very strict format and contain a "score" column indicating the correctness of a prediction model and a "label_value" column indicating the actual value of the predicted label. In terms of the clarity of the tool and tutorial, the Aequitas website offers helpful descriptions of bias metrics and includes a *fairness tree* to help users decide which metrics to evaluate and thresholds for fairness. This tool, however, lacks the interactivity of some other dashboards such as Fairlearn and What-If-Tool, as we will describe later.

4.2 AI Fairness 360

As a comprehensive toolkit, IBM's AI Fairness 360 [39], or AIF360, contains multiple tools and algorithms to identify and mitigate bias and can be used for the analysis of different metrics based on various use cases. These fairness metrics include Group Fairness using the *DatasetMetric* class, Individual Fairness using the *SampleDistortionMetric* class, or both Individual and Group Fairness using the *ClassificationMetric* class. AIF360 also provides tools to measure fairness at different stages in the machine learning

process. The *DatasetMetric* class can be used to provide fairness metrics on the training data, while the *ClassificationMetric* class should be used to provide fairness metrics on the models themselves. In addition to fairness assessment, it also offers pre-processing, in-processing, and post-processing mechanisms for mitigating biases. This toolkit is ideal for allocation and risk-assessment problems with defined protected attributes. AIF360 also is an open-source library with API documentation provided. User support includes tutorial guides, videos, demos, and a notebook repository of working examples of the tool. The main website also provides resources for understanding applications of fairness and a glossary of introductory terms to fairness in ML.

4.3 Dalex

Dalex [8] (moDel Agnositc Language for Exploration and eXplanation) is an explainability tool created by ML2DataLab at Warsaw University of Technology and University of Warsaw. This project started in R [12] and recently has been expanded to python. The Dalex Python package implements a main *Explainer* class to provide an abstract layer between the model API's and explainability and fairness methods. The *fairness_check* method compares the most common statistical fairness measures (Table 1) and provides a detailed textual description of the group fairness analysis. It also contains the result attribute and plot methods, which provide various visualizations depending on the type parameter. Figure 2 provides two sample visualizations of the fairness methods. Both aim to draw a comparison between multiple models. Dalex comes with a full fairness tutorial on COMPAS recidivism [19] and has a credit dataset based in Germany [20] integrated within the package. Additionally, it provides tutorials for introductory explain-ability in AI to inform the appropriate methods available to explore a model.

4.4 Fairlearn

Fairlearn's post-processing algorithms take an already-trained model and transform its predictions so that they satisfy the constraints implied by the selected fairness metric (e.g., demographic parity) while maximizing model performance (e.g., accuracy rate). For example, given a model that predicts the probability of defaulting on a loan, a post-processing algorithm will try to find a threshold above which an applicant should get a loan. Fairlearn's reduction algorithms treat any standard classification or regression algorithm as a black box, and iteratively (a) re-weights the data points and (b) retrains the model after each re-weighting so that the model will satisfy the constraints implied by the selected fairness metric while maximizing model performance. Unlike the post-processing approach, the reduction approach requires retraining of the model, which could make it a more time-consuming approach. Currently the mitigation techniques that are included are [5, 24] for classification models and [6] for regression. Fairlearn comes with three datasets integrated with the tool making it easily accessible for demonstration purposes; these are the UCI Adult and marketing dataset[20], and the Boston housing dataset [1].

4.5 Responsibly

Responsibly [25] is another fairness and auditing tool specifically developed for both researchers and learners and includes functionality specialized for natural language processing (NLP) analysis. It is compatible with data science and machine learning tools of the trade in Python, such as Numpy, Pandas, and especially scikit-learn. Like the other tools we examined, its primary focus is on auditing biases and its secondary aim is enabling mitigation. The tool considers fairness metrics for evaluating independence, separation, and sufficiency. It also analyzes various thresholds in post-processing methods. This tool is closely aligned with the content of the *Fairness in Machine Learning* book by Barocas et al. [10] that is used across various courses in ethics and fairness in multiple universities. [1] It showcases fairness analysis using various datasets to demonstrate these features, including visual plot capabilities, as shown in Figure 3 (bottom) showing decile scores by race. Additionally, it demonstrates Biases in Word Embedding with various datasets provided by Google, Facebook, and Stanford. Other available datasets include the Adult and the German Credit Dataset from UCI [20], the FICO Dataset from TransUnion [2], and COMPASS dataset by ProPublica [3].

4.6 What-If-Tool

The What-If Tool [41] from Google is built into the open source TensorBoard [38] web application and allows users to analyze a machine learning model performance and fairness. With the What-If Tool, users can test algorithmic fairness constraints, visualize inference results. One of the niche features of this tool is that it allows users to *edit* a data point to see how a model performs. This is useful for identifying counterfactual points in datasets. It also incorporates some inseparability into the model by allowing users to visualize Partial Dependence Plots (PDP). What-If Tool offers other interactivity with the performance of a classifier model (Confusion matrix) and the ability to adjust threshold values in an intuitive GUI. As the tool is incorporated into TensorFlow, it provides the most visual experience of the toolkits listed here and enables easy integration of deep neural network models such as those for Face Recognition.

5 QUALITATIVE REVIEW

In this section, we present the results of our surveys and focus groups. Three common themes emerged from our analysis: *transparency*, dataset *integration*, and *interactivity*. We discuss opportunities to promote ethical debates and teach students about representation harms towards improving the design of fairness tools that can be integrated into machine learning education.

5.1 Transparency

A key limitation of the ML fairness tools was a lack of transparency and explainablity. Students reported that the tools that offered mitigation techniques were the ones most lacking in terms of insight into how the mitigation techniques worked. For instance, the students found that AIF360's pre-processing mitigation techniques

[1] Berkeley CS 294: Fairness in machine learning; Cornell INFO 4270: Ethics and policy in data science; Princeton COS 597E: Fairness in machine learning; University of Washington CSS 444 Biases in Machine Learning

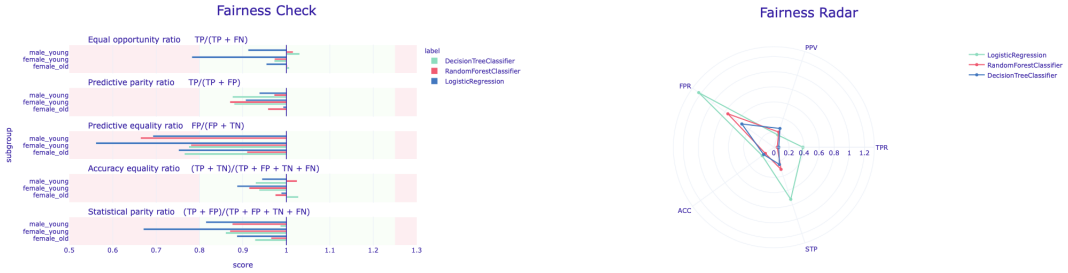

Figure 2: Dalex fairness comparison of multiple models.

Fairness Criteria	Aequitas	Dalex	FairLearn	Fairness 360	Responsibly	WIT
Balance for the negative/positive class [28]	✓	✓	✓		✓	✓
Conditional statistical parity [16]				✓	✓	
Equal opportunity [24]	✓	✓	✓	✓		✓
Equalized odds [24]		✓	✓	✓	✓	
Equalized correlations [42]		✓		✓		
Mitigation		✓	✓	✓	✓	
Predictive Parity [14]	✓	✓		✓	✓	✓
Statistical Parity [21]	✓	✓		✓	✓	✓
Interactivity	✓		✓			✓
Embedded Datasets	✓	✓	✓	✓	✓	✓

Table 1: Supported Fairness Criteria and other affordances of the surveyed tools.

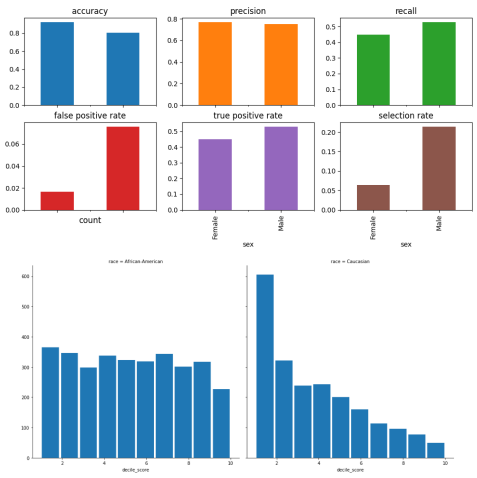

Figure 3: Example of Fairlearn Metrics visualization (top) and Responsibly graphs (bottom) presenting decile scores both for COMPAS dataset.

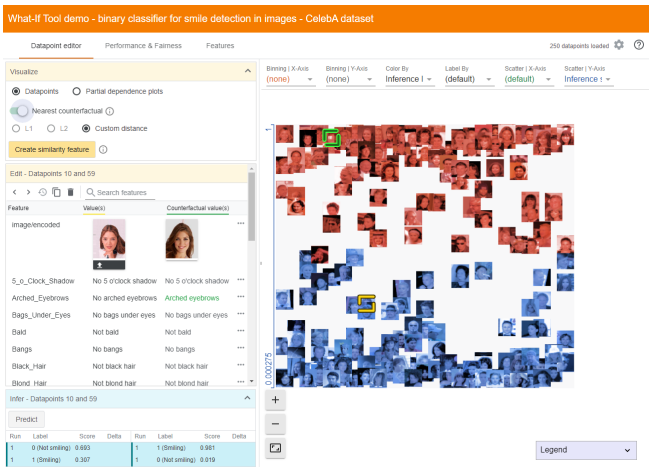

Figure 4: What-If-Tool Smile Detection Demo with some counterfactual points identified.

were not transparent about how the input data was reweighed and distributed to create a fairer model. In cases where tools provided tutorials (e.g., Fairlearn tutorial on their Grid Search), students reported that although there was a greater degree of transparency about the mitigation techniques, the tutorials were hard to understand. The tutorials' content was largely inaccessible to the undergraduate students and appeared to be targeted towards highly-trained academic researchers.

In contrast, students found that What-if-tool supports a greater level of transparency through its interactive design in allowing data points to be directly modified and the impact of this modification to be observed immediately through visualization. The What-If tool code was directly accessible within the Python Notebook, which allowed students to further explore how the data was analyzed

in the tool. There was some overlap between transparency and visualization capabilities, as students expressed that *"better visualizations are more clear to understand"*. Some students felt that the documentation was expansive, such as the case with AI Fairness 360. On the other hand, the demos and tutorials lacked transparency in explanation, with a student stating that: *"the tutorials did not necessarily make me feel like the program knew what it was doing, I was left to trust the code without too much explanation."*

Regarding integrating these tools into CS education, the use of tutorials and demos could be a sensible method for introducing fairness concepts and ideas. However, they may lack transparency in how the tools analyze data and generated visualizations. Indeed, these tools often contained additional technical documentation, which could be beneficial supplemental material for instructors and students. For students unfamiliar with data science topics or lacking strong statistical backgrounds, it may be challenging to comprehend technical documentation that was tailored for people with more expertise. Based on these research insights, we recommended that educators teach students how to read technical documentation and reference relevant documentation when using these tools in a classroom setting.

5.2 Flexibility with Custom Datasets

Another important theme that emerged from our focus groups was that student learning was limited by the integrated datasets of the tools and the inability to import custom datasets. Indeed the majority of the fairness literature tools and tutorials rely on a handful of common datasets sourced from UCI. Our focus groups highlighted that the integration of high-dimensional data (e.g., images) into tools was a high-priority need for students. Our findings resonate with those reported in [29] that, by allowing learners to specify their own input datasets, they engage more actively in the visualization process. It is worth mentioning that out of all the tools that we surveyed, only the What-If-Tool provided the ability to import high-dimensional data into the tool. Allowing for high-dimensional custom data enabled students to better apply and improve technical competencies. Furthermore, it allowed learners to freely explore the tool and discover how the algorithm executes on a range of data. On the other hand, some of the tools, such as Aequitas, were reported to be too limiting due to the strict data integration requirements. A student expressed that *"it has really strict pre-processing requirements in order to use their data and I couldn't get it to work with other data as easily."*

5.3 Visualizations and Interactivity

The final theme from our focus groups was that student learning was impacted by the tools' visualization styles and presentation of fairness criteria. Our focus group reflected on their positive experience in the simple and effective visualization that Aequitas provided in presenting fairness. Our focus group found the least effective visualization to AI Fairness 360. The What-if-tool received a mixture of feedback where our students liked the visualization but found a learning curve in getting familiar with many options provided by the tool in both. One student expressed that the What-if tool *"had nice visual elements, and seeing it really helps me wrap my head around the results and what they actually mean"*. Tools that presented options of interactivity were overall favored in the study.

6 CONCLUDING REMARKS

This paper presented a case study on the value of using interactive algorithmic visualization tools to teach Computer Science students about responsible AI. We reported our qualitative analysis of student perception of these tools and their impacts on student learning. Below, we propose a road map for how CS educators could integrate responsible AI into CS curriculum.

No One Right Tool: This research led us to conclude that we do not recommend one specific tool for CS curriculum. Instead, educators should select a set of tools that offer a range of capabilities. Each tool we surveyed was distinct in its functionality, usability for students, and capacity for instruction. While this paper highlights specific features around transparency and interaction usability, the needs of various curricula may require otherwise. Fortunately, the flexibility of the tools for importing custom datasets and models presented numerous opportunities for these tools to be embedded throughout the syllabi of courses. Opportunities ranged from visualizing data during data exploration to modeling assessments to mitigation. We recommend using this survey as a basis for evaluating and prioritizing tools and features when developing fairness and CS curriculum.

Striking the Right Balance: We believe that to foster students' interest in responsible AI and ML fairness, it is crucial to use datasets and models that are representative of fairness challenges of current ML models embedded in their daily lives. To advance students' understanding of fairness in ML, we recommend that students explore practical datasets and apply their emerging skills in data science and statistics. Indeed, most of the tools that we surveyed focused on offering statistical analysis of fairness on datasets that present *allocative harms* (e.g., mortgage qualifications and college admission). However, less attention has been paid to familiarizing students with *representation harms* [17], where systems reinforce the subordination of some groups along the lines of identity. We believe that it is highly desirable to balance the statistical examples of allocative harms and the use of interactive tools to explore representation harms (e.g., face recognition examples in What-If-Tool).

Promoting Group Conversation: Although in our study we did not measure the impact of group work, we believe tools such as those surveyed here present a meaningful opportunity for classroom discourse about ethical choices in machine learning and AI. These choices lead to technical systems that are embedded into our daily lives and have decision-making power on ethically sensitive topics, such as recidivism and hiring practices. As Zook et al. [43] argued, a crucial component of responsible big-data research is developing the capacity for students to engage in open-ended ethical debates. Ethical awareness and analysis of how computing systems are designed with inherent bias present opportunities within a CS curriculum to educate the next generation of computer scientists to recognize and confront biases in their work and research. As digital social scientist and ethicist Annette Markham [31] writes, "we can make [data ethics] an easier topic to broach by addressing ethics as being about choices we make at critical junctures; choices that will invariably have impact." We believe that there is

a significant opportunity for future research, for example, at the intersection of CHI and the Special Interest Group on Computer Science (SIGCSE), to more deeply understand, define, and measure the role of tool-oriented group work and its capacity to allow for ethical debates.

In summary, our case study and proposed roadmap contribute (1) a direction for enhancing authentic and practical learning of CS students in higher education and (2) actionable feedback on ML and AI tools used in education and HCI practice.

REFERENCES

[1] 2021. https://www.kaggle.com/c/boston-housing
[2] 2021. https://docs.responsibly.ai/notebooks/demo-fico-analysis.html
[3] 2021. https://github.com/propublica/compas-analysis/
[4] 2021. Coded Bias.
[5] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.
[6] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.
[7] Saleema Amershi, Giuseppe Carenini, Cristina Conati, Alan K Mackworth, and David Poole. 2008. Pedagogy and usability in interactive algorithm visualizations: Designing and evaluating CIspace. *Interacting with Computers* 20, 1 (2008), 64–96.
[8] Hubert Baniecki, Wojciech Kretowicz, Piotr Piatyszek, Jakub Wisniewski, and Przemyslaw Biecek. 2020. dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. *arXiv preprint arXiv:2012.14406* (2020).
[9] Niels Bantilan. 2017. scikit-ml. https://themis-ml.readthedocs.io/en/latest/
[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.
[11] Benjamin S Baumer, Randi L Garcia, Albert Y Kim, Katherine M Kinnaird, and Miles Q Ott. 2020. Integrating data science ethics into an undergraduate major. *arXiv preprint arXiv:2001.07649* (2020).
[12] Przemysław Biecek. 2018. DALEX: explainers for complex predictive models in R. *The Journal of Machine Learning Research* 19, 1 (2018), 3245–3249.
[13] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. 2019. DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks.. In *IJCAI*. 4658–4664.
[14] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[15] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.
[16] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.
[17] Kate Crawford. 2017. "The Trouble with Bias".
[18] Michael Davis. 2006. Integrating ethics into technical courses: Micro-insertion. *Science and Engineering Ethics* 12, 4 (2006), 717–730.
[19] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc* 7, 7.4 (2016), 1.
[20] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[22] Facebook. 2020. FairFlow. https://www.facebook.com/FairFlowTech/
[23] Barbara J Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. 2019. Embedded EthiCS: integrating ethics across CS education. *Commun. ACM* 62, 8 (2019), 54–61.
[24] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).
[25] Shlomi Hod. 2018–. Responsibly: Toolkit for Auditing and Mitigating Bias and Fairness of Machine Learning Systems. http://docs.responsibly.ai/
[26] Anna Lauren Hoffmann and Katherine Alejandra Cross. 2021. Teaching Data Ethics: Foundations and Possibilities from Engineering and Computer Science Ethics Education. (2021).
[27] Michael Kearns and Aaron Roth. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.
[28] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[29] Ari Korhonen and Lauri Malmi. 2002. Matrix: concept animation and algorithm simulation system. In *Proceedings of the Working Conference on Advanced Visual Interfaces*. 109–114.
[30] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM New York, NY, USA, 1–13.
[31] A Markham. 2016. OKCupid data release fiasco: It's time to rethink ethics education. *Points: Data & Society* (2016).
[32] Jacob Metcalf, Kate Crawford, and Emily F Keller. 2015. Pedagogical approaches to data ethics. *Council for Big Data, Ethics, and Society* (2015).
[33] Thomas L Naps, Guido Rößling, Vicki Almstrum, Wanda Dann, Rudolf Fleischer, Chris Hundhausen, Ari Korhonen, Lauri Malmi, Myles McNally, Susan Rodger, et al. 2002. Exploring the role of visualization and engagement in computer science education. In *Working group reports from ITiCSE on Innovation and technology in computer science education*. 131–152.
[34] Željko Obrenović. 2012. Rethinking HCI education: teaching interactive computing concepts based on the experiential learning paradigm. *interactions* 19, 3 (2012), 66–70.
[35] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
[36] Kris Powers, Paul Gross, Steve Cooper, Myles McNally, Kenneth J Goldman, Viera Proulx, and Martin Carlisle. 2006. Tools for teaching introductory programming: what works?. In *Proceedings of the 37th SIGCSE technical symposium on Computer science education*. 560–561.
[37] Jay Quedado. 2021. Fairness Visualization Tools in ML Education. https://molochxte.github.io/ML-Fairness-Tools-in-Education/
[38] Google Brain Team. 2015. Tensorflow. https://themis-ml.readthedocs.io/en/latest/
[39] K Varshney. 2018. Introducing AI fairness 360. *IBM Research blog, September* 19 (2018).
[40] Matthijs Vincent. 2019. scikit-fairness. https://github.com/koaning/scikit-fairness
[41] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.
[42] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. In *Conference on Learning Theory*. PMLR, 1920–1953.
[43] Matthew Zook, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A Koenig, Jacob Metcalf, et al. 2017. Ten simple rules for responsible big data research.