FairTargetSim: An Interactive Simulator for Understanding and Explaining the Fairness Effects of Target Variable Definition

Dalia Gala[1] , **Milo Phillips-Brown**[2,3] , **Naman Goel**[1] , **Carinal Prunkl**[4] , **Laura Alvarez Jubete**[5] , **medb corcoran**[5] and **Ray Eitel-Porter**[5]

[1]University of Oxford
[2]University of Edinburgh
[3]Jain Family Institute
[4]Utrecht University
[5]Accenture

Abstract

Machine learning requires defining one's target variable for predictions or decisions, a process that can have profound implications on fairness: biases are often encoded in target variable definition itself, before any data collection or training. We present an interactive simulator, FairTargetSim (FTS), that illustrates how target variable definition impacts fairness. FTS is a valuable tool for algorithm developers, researchers, and non-technical stakeholders. FTS uses a case study of algorithmic hiring, using real-world data and user-defined target variables. FTS is open-source and available at: http://tinyurl.com/ftsinterface. The video accompanying this paper is here: http://tinyurl.com/ijcaifts.

1 Motivation

Machine learning requires translating real-world problems into numerical representations. Sometimes, the translation is straightforward—e.g. in predicting whether someone defaults on a loan. Other times, things are not so simple. When developing an algorithm to predict which job applicants will be good employees, for example, one must make precise the notion of a "good" employee. This is an ambiguous, subjective notion about which reasonable minds may disagree. How one translates this notion numerically—how one *defines the target variable*—can have profound implications for fairness [Passi and Barocas, 2019]. Defining "good" employee one way rather than another may result, e.g., in fewer applicants being hired from certain demographics. These issues arise in many domains. For a college admissions algorithm, one must determine who counts as a "good" student; for a search engine, one must determine what counts as a "good" search result; etc. How these notions are defined may likewise have weighty implications for fairness: which university applicants are admitted [Kizilcec and Lee, 2023]; which items appear at the top of search results [Phillips-Brown, manuscript]; etc. Target variable definition, then, is not a merely technical matter. Defining "good" employee, student, or search result is a value-laden process: it calls for close attention and transparency [Fazelpour and Danks, 2021].

But all too often, target variables are defined without transparency or attention to fairness. On one hand, technical developers may take target variable definition as a given, focusing instead on issues such as data quality, variance, accuracy of predictions, etc. On the other hand, stakeholders who are not a part of the technical process—like (hiring) managers in non-technical roles, or those working in upper management—either do not understand, or are simply unaware of, the implications of target variable definition in algorithmic settings. There is thus a pressing need for the fairness implication of target variable definition to be understood—and foregrounded—for stakeholders of all kinds.

To help meet this need, we developed an *interactive target variable simulator*, FairTargetSim (FTS): http://tinyurl.com/ftsinterface. FTS introduces its users to target variable definition, and reveals and explains its impact on fairness. FTS uses a case study: hiring algorithms. FTS invites the user to imagine that they are building a hiring algorithm, which mirrors a widely-used style of hiring algorithm based on psychometric tests. The user defines two target variables, using real-world psychometric test data from [Jaffe *et al.*, 2022]. With these two definitions, FTS builds two corresponding models and gives visualizations of how the models and training data differ in matters of fairness and overall performance. FTS's interactive and visualization elements bring these issues to life—offering a more compelling and memorable illustration than one can get by, for example, reading a text.

FTS's code is public and freely available. Its use is, then, not limited to hiring algorithms or to the dataset we use in our case-study: it can be extended to uses beyond education, and to different datasets and models.

2 FairTargetSim's audience

FTS is a valuable tool for a wide range of audiences. The first target audience is technical developers who often want to develop algorithms responsibly but have less understanding of non-algorithmic factors such as target variable definition. With FTS, they can better the behavior of their abstract algorithms under different target variable definitions. This technical audience may also have less control over non-algorithmic factors, and can use FTS to better advocate—to decision-makers with non-technical backgrounds—for responsible algorithmic development. This leads us to the second target audience: non-technical stakeholders: e.g. those who use algorithms for making decisions or those who are impacted by the decisions. When these stakeholders better understand the fairness implications of target variable definition, the way is paved for more responsible and accountable use of algorithms in the real world. The third target audience is educators. There is a pressing need for more responsible AI education and training in universities ([Grosz *et al.*, 2018], [Kopec *et al.*, forthcoming]), government, and the private sector [Eitel-Porter, 2021]. The ethical implications of technical issues can be challenging to explain to learners. FTS gives educators an accessible, hands-on way to illustrate them.

We emphasize that FTS illustrates not "only" the fairness implications decisions about target variable definition. It also illustrates, more generally, the ethical implications of decisions at the intersection of technical and non-technical aspects of algorithmic development. While it is well understood among theorists that such decisions are value-laden ([Friedman and Nissenbaum, 1996], [Johnson, forthcoming]), they often do not wear their ethical dimensions on their sleeves. FTS allows audiences of all kinds to see—through a concrete example—such decisions for what they are.

3 Related Work

A wealth of research has established the importance of understanding and addressing the fairness implications of target variable definition—in algorithmic systems generally ([Passi and Jackson, 2018], [Obermeyer *et al.*, 2019], [Martin Jr. *et al.*, 2020], [Levy *et al.*, 2021], [Barocas *et al.*, 2023]) and hiring algorithms specifically ([Bǎžgu and Cernea, 2019], [Raghavan *et al.*, 2020], [Tilmes, 2022]).

A number of systems have been developed for practitioners—and in some cases, non-technical stakeholders—to understand, identify, and address algorithmic bias. We list just some, and note that various of them, like FTS, have a visualization element: [Tramèr *et al.*, 2017], [Bellamy *et al.*, 2019], [Ribeiro *et al.*, 2018], [Cabrera *et al.*, 2019], [Microsoft and contributors, 2019], [Saleiro *et al.*, 2019], [Vincent and ManyOthers, 2019], [Ahn and Lin, 2020], [Wexler *et al.*, 2020], [Johnson *et al.*, 2023], [Liu *et al.*, 2023]. FTS is an important addition to these systems because it is, to our knowledge, the only one that addresses target variable definition.

Compared to previous demonstrations at IJCAI on related subjects (e.g. [Sokol and Flach, 2018; Juan *et al.*, 2021; Yu *et al.*, 2019; Miguel *et al.*, 2021; Henderson *et al.*, 2021; Baumann *et al.*, 2023]), our demonstration will focus on the problem of fairness implications of target variable definition.

4 Overview of FairTargetSim

FTS's interface works with most modern browsers; Firefox is advised. FTS has four pages that the user visits in order.

4.1 Key Concepts Explained

This page introduces target variable definition to a non-technical audience, explains how it impacts fairness, and gives an overview of the other pages of FTS.

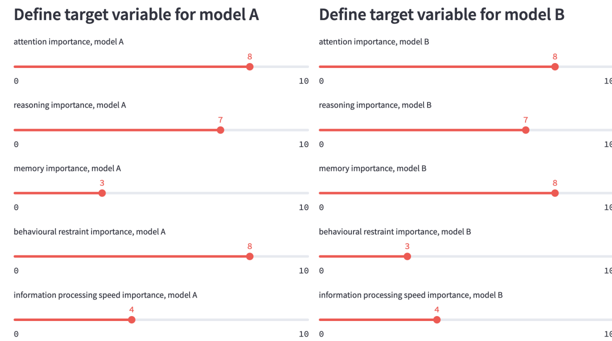

Figure 1: The user defines two target variables, using sliders representing the importance of traits of "good" employees.

4.2 User Defines Target Variables

This page has the user define two different target variables (Figure 1), which FTS uses to train two models, A and B.

In the real-world hiring algorithms that are based in cognitive tests, developers often define "good" employee by having an employer identify a group of current employees whom the employer deems "good" for a given role [Wilson *et al.*, 2021]. These employees then play cognitive-test games, and a model is trained to identify applicants that share cognitive traits with these employees.

FTS's models are similar to these real-world systems in two key ways. First, like those systems, FTS uses support vector machine models to identify people who share cognitive traits with those who are identified as "good" employees. Second, FTS's models are trained on data of real peoples' cognitive tests; the data we use is from Jaffe et al.'s (2022) battery 26, which has eleven tests that we grouped into five traits: memory, information processing speed, reasoning, attention, and behavioural restraint.[1]

FTS's models differ from the real-world systems in one key way: how the target variable is defined. With FTS, the user explicitly defines, using sliders depicted in Figure 1, how important the five cognitive traits are to what makes for a "good employee." The user does this twice, creating two different target variables. Then FTS calculates the weighted average of tests scores, given the slider weightings, and assigns class label "0" to those in the bottom 85th percentile. From the top 15% subset, we randomly sample 100 "good" employees to whom we assign the class label "1" with weights ranging from 0.99 for the highest scoring candidate to 0.01 for the lowest scoring candidate, using linear distribution with the following equation for those in between:

$$f(x) = \frac{0.98}{1-n}x + \frac{0.01 - 0.99n}{1-n}$$

We assign a class label "0" to those not selected, thus intro-

[1]Our five categories are based on the following tests: *Memory* (forward memory span, reverse memory span, verbal list learning, delayed verbal list learning); *Information Processing Speed* (digit symbol coding, trail making part A, trail-making part B); *Reasoning* (arithmetic reasoning, grammatical reasoning); *Attention* (divided visual attention); and *Behavioral Restraint* (go/no-go).

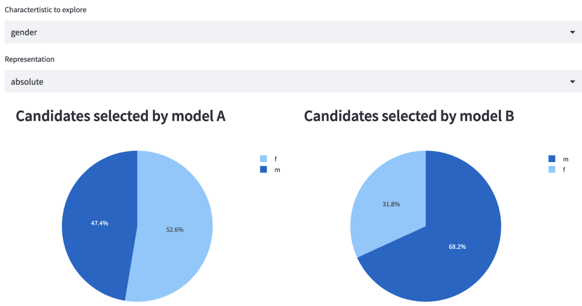

Figure 2: Charts display how the percentage of selected male and female applicants differs between models A and B.

Candidate ID	Ranking, model A	Ranking, model B	Predicted label A	Predicted label B
12	0.148166	0.987394	0	1
9	0.936614	0.978990	1	1
11	0.912242	0.960691	1	1
15	0.795276	0.953701	1	1
21	0.016506	0.952070	0	1

Figure 3: A table illustrates how individual applicants are evaluated differently by the two models.

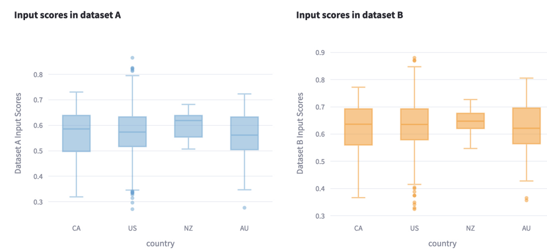

Figure 4: Bar graphs show how choice of features of importance affects the model input scores achieved for different candidates depending on the demographic group—in this case, country of origin. For example, for model A, the median score for American candidates is approximately 0.57, while for model B, it is 0.63.

ducing randomness. FTS then generates two labeled datasets and corresponding models, each with different target variable definitions.

FTS works with user-defined target variables because, first, we do not have access to real-world target variables, and, second, the lessons FTS offers are brought to life for the user when she can see how her very own choices in target variable definition can have implications for fairness. As we explain further in Section 4.4, having user-defined target variables is not a fundamental constraint on the idea of FTS; FTS can be extended to use real-world labels when they are available.

4.3 Visualize Effects of Target Variable Definition

This page contains visualizations that illustrate how the user's two target variable definitions impact issues of fairness and overall model performance. The visualizations are categorized in to *Demographic* and *Non-demographic* sections, and further divided into categories that (i) show features of the models and (ii) features of the training data.

In the *Demographic* section, charts as in Figure 2 show how models A and B differ in, e.g., the proportions of selected applicants across demographic groups (gender, education level, age, and nationality—these are the demographic groups that the Jaffe *et al.* dataset has information on). Other charts show how the models differ across groups with respect to "fairness metrics" ([Angwin *et al.*, 2016], [Corbett-Davies and Goel, 2018]), such as true and false positive rates and positive and negative predictive value.

The differences are stark: different target variable definitions often result in major differences in the demographics of selected applicants and in fairness metrics (see e.g. Figure 2). Visualizations in the *Demographics* section also show how target variable definition affects models' training data: e.g. how positive and negative labels are distributed across demographic groups.

In the *Non-demographic* section, visualizations show how the models and training data differ in ways other than fairness: e.g. how the models rank particular applicants (Figure 3), overall model confusion matrices, and accuracy metrics.

4.4 Further uses of FairTargetSim

This page gives recommendations for using FTS not just for providing explanations and educating stakeholders, but also for directly impacting practices in hiring and other domains.

As noted, FTS's code is available publicly; an organization can extend FTS to use with their own data, models, and target variables. And, as also noted, in real-world target variable definition, employers do not directly identify cognitive characteristics of "good" employees; they identify certain current employees as "good." We give guidance on how to this in a way that can promote fairness. For example, (i) consult various managers on whom they judge "good;" these judgments can be weighted in different ways—just as FTS weights the cognitive tests in different ways—resulting in different target variables. Or, (ii) use various performance metrics to evaluate current employees (e.g., number of years to promotion, length of tenure at a company, or role-specific metrics, such as number of sales with a sales role); these metrics can, again, be weighted in different ways, resulting in different target variables. We also explain how to weight different judgements and metrics in other domains: e.g. in a college admissions algorithm or a search algorithm.

5 Future work

FTS opens up various avenues for future work, of which we will highlight a few. One, as noted in Section 4.4, is to apply FTS to real-world hiring settings. Another, facilitated by the fact that FTS is flexible and openly available, is to invite the community to add more features to the simulator by, for example using different kinds of data sets, models, or visualizations. Likewise, FTS could be extended to cases beyond algorithmic hiring, such as college admissions or search engines. Finally, FTS affords opportunities for human-centered research. For example, user-studies could be run—with both technical and non-technical stakeholders—to test how FTS affects how they think about, develop, and use algorithms for hiring and beyond.

References

[Ahn and Lin, 2020] Yongsu Ahn and Yu-Ru Lin. Fairsight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1086–1095, 2020.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Matthu, and Lauren Kirchner. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 23 2016. *ProPublica*.

[Bãžgu and Cernea, 2019] Drago Bãžgu and Mihail-Valentin Cernea. Algorithmic bias in current hiring practices: an ethical examination. *Proceedings of the International Management Conference*, 13(1):1068–1073, 2019.

[Barocas *et al.*, 2023] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[Baumann *et al.*, 2023] Joachim Baumann, Alessandro Castelnovo, Andrea Cosentini, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. Bias on demand: investigating bias with a synthetic data generator. In *32nd International Joint Conference on Artificial Intelligence (IJCAI), Macao, SAR, 19-25 August 2023*, pages 7110–7114. International Joint Conferences on Artificial Intelligence Organization, 2023.

[Bellamy *et al.*, 2019] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019.

[Cabrera *et al.*, 2019] Angel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, October 2019.

[Corbett-Davies and Goel, 2018] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

[Eitel-Porter, 2021] Ray Eitel-Porter. Beyond the promise: implementing ethical AI. *AI and Ethics*, 1:73–80, 2021.

[Fazelpour and Danks, 2021] Sina Fazelpour and David Danks. Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8), 2021.

[Friedman and Nissenbaum, 1996] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems*, 3(14):330–347, 1996.

[Grosz *et al.*, 2018] Barbara J. Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. Embedded ethics: Integrating ethics broadly across computer science education, 2018.

[Henderson *et al.*, 2021] Jette Henderson, Shubham Sharma, Alan Gee, Valeri Alexiev, Steve Draper, Carlos Marin, Yessel Hinojosa, Christine Draper, Michael Perng, Luis Aguirre, et al. Certifai: a toolkit for building trust in ai systems. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 5249–5251, 2021.

[Jaffe *et al.*, 2022] Paul I. Jaffe, Aaron Kaluszka, Nicole F. Ng, and Robert J. Schafer. A massive dataset of the neurocognitive performance test, a web-based cognitive assessment. *Scientific Data*, 9(1), 2022.

[Johnson *et al.*, 2023] Brittany Johnson, Jesse Bartola, Rico Angell, Sam Witty, Stephen Giguere, and Yuriy Brun. Fairkit, fairkit, on the wall, who's the fairest of them all? supporting fairness-related decision-making. *EURO Journal on Decision Processes*, 11:100031, 2023.

[Johnson, forthcoming] Gabbrielle M. Johnson. Are algorithms value-free? feminist theoretical virtues in machine learning. *Journal Moral Philosophy*, pages 1–35, forthcoming.

[Juan *et al.*, 2021] Yi-Ning Juan, Yi-Shyuan Chiang, Shang-Chuan Liu, Ming-Feng Tsai, and Chuan-Ju Wang. Hive: Hierarchical information visualization for explainability. In *IJCAI*, pages 4988–4991, 2021.

[Kizilcec and Lee, 2023] René F. Kizilcec and Hansol Lee. Algorithmic fairness in education. In Wayne Holmes and Kaśka Porayska-Pomsta, editors, *The Ethics of Artificial Intelligence in Education*. Routledge, 2023.

[Kopec *et al.*, forthcoming] Matthew Kopec, Meica Magnani, Vance Ricks, Roben Torosyan, John Basl, Nicholas Miklaucic, Felix Muzny, Ronald Sandler, Christo Wilson, Adam Wisniewski-Jensen, Cora Lundgren, Kevin Mills, and Mark Wells. The effectiveness of embedded values analysis modules in computer science education: An empirical study. https://arxiv.org/abs/2208.05453, forthcoming. forthcoming in *Nature Machine Intelligence*.

[Levy *et al.*, 2021] Karen Levy, Kyla E. Chasalow, and Sarah Riley. Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science*, 17(1):309–334, October 2021.

[Liu *et al.*, 2023] Jessica Liu, Huaming Chen, Jun Shen, and Kim-Kwang Raymond Choo. Faircompass: Operationalising fairness in machine learning, 2023.

[Martin Jr. *et al.*, 2020] Martin Martin Jr., Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. Participatory problem formulation for fairer machine learning through community based system dynamics, 2020.

[Microsoft and contributors, 2019] Microsoft and contributors. Fairlearn. https://fairlearn.github.io/, 2019.

[Miguel *et al.*, 2021] Beatriz San Miguel, Aisha Naseer, and Hiroya Inakoshi. Putting accountability of ai systems into practice. In *Proceedings of the Twenty-Ninth International*

Conference on International Joint Conferences on Artificial Intelligence, pages 5276–5278, 2021.

[Obermeyer *et al.*, 2019] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[Passi and Barocas, 2019] Samir Passi and Solon Barocas. Problem formulation and fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 39–48, 2019.

[Passi and Jackson, 2018] Samir Passi and Steven J. Jackson. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018.

[Phillips-Brown, manuscript] Milo Phillips-Brown. Algorithmic neutrality. https://arxiv.org/abs/2303.05103, manuscript.

[Raghavan *et al.*, 2020] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, January 2020.

[Ribeiro *et al.*, 2018] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[Saleiro *et al.*, 2019] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit, 2019.

[Sokol and Flach, 2018] Kacper Sokol and Peter A Flach. Glass-box: Explaining ai decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In *IJCAI*, pages 5868–5870, 2018.

[Tilmes, 2022] Nicolas Tilmes. Disability, fairness, and algorithmic bias in ai recruitment. *Ethics and Information Technology*, 21(24), 2022.

[Tramèr *et al.*, 2017] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416, 2017.

[Vincent and ManyOthers, 2019] Matthijs Vincent and ManyOthers. Ccikit-fairness. https://github.com/koaning/scikit-fairness, 2019.

[Wexler *et al.*, 2020] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020.

[Wilson *et al.*, 2021] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 666–677, 2021.

[Yu *et al.*, 2019] Han Yu, Yang Liu, Xiguang Wei, Chuyu Zheng, Tianjian Chen, Qiang Yang, and Xiong Peng. Fair and explainable dynamic engagement of crowd workers. In *IJCAI*, pages 6575–6577, 2019.