## REFERENCES


[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016).

[2] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[3] Alexandra Chouldechova, Diana Benavides Prado, Oleksandr Fialko, Emily Putnam-Hornstein, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the First Conference on Fairness, Accountability, and Transparency*.

[4] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[5] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2016. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64, 3 (2016).

[6] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. (2017). CoRR arXiv:1702.08608.

[7] Mary T. Dzindolet, Linda G. Pierce, Hall P. Beck, and Lloyd A. Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44, 1 (2002), 79–94.

[8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[9] Raymond Fisman, Sheena S Iyengar, Emir Kamenica, and Itamar Simonson. 2006. Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics* 121, 2 (2006), 673–697.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. (2018). CoRR arXiv:1803.09010.

[11] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the Twenty-Fourth International World Wide Web Conference*.

[12] Kartik Hosanagar and Apoorv Saxena. 2017. The Democratization of Machine Learning: What It Means for Tech Innovation. Knowledge@Wharton, retrieved from http://knowledge.wharton.upenn.edu/article/democratization-ai-means-tech-innovation/.

[13] Matthew Kay, Shwetak N Patel, and Julie A Kientz. 2015. How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 347–356.

[14] Ryan Kennedy, Philip D. Waggoner, and Matthew Ward. 2018. Trust in Public Policy Algorithms. Working paper.

[15] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

[16] Zachary C. Lipton. 2016. The mythos of model interpretability. (2016). CoRR arXiv:1606.03490.

[17] Jennifer M. Logg, Julia Minson, and Don A. Moore. 2018. Algorithm Appreciation: People prefer algorithmic to human judgment. (2018). Harvard Business School NOM Unit Working Paper No. 17-086.

[18] Polina Marinova. 2017. How Dating Site eHarmony Uses Machine Learning to Help You Find Love. http://fortune.com/2017/02/14/eharmony-dating-machine-learning/.

[19] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Second Conference on Fairness, Accountability, and Transparency*.

[20] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. (2018). CoRR arXiv:1802.00682.

[21] David W Nickerson and Todd Rogers. 2014. Political campaigns and big data. *Journal of Economic Perspectives* 28, 2 (2014), 51–74.

[22] Dilek Önkal, Paul Goodwin, Mary Thomson, and Sinan Gönül. 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22 (2009), 390–409.

[23] Umberto Panniello, Michele Gorgoglione, and Alexander Tuzhilin. 2016. Research note–In CARSs we trust: How context-aware recommendations affect customers? Trust and other business performance measures of recommender systems. *Information Systems Research* 27, 1 (2016), 182–196.

[24] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. (2018). CoRR arXiv:1802.07810.

[25] 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 Conference on Knowledge Discovery and Data Mining*.

[26] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 141–148.

[27] Jennifer Wortman Vaughan and Hanna Wallach. 2017. The Inescapability of Uncertainty. In *CHI Workshop on Designing for Uncertainty in HCI: When Does Uncertainty Help?*

[28] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

[29] Peng Xia, Hua Jiang, Xiaodong Wang, Cindy X Chen, and Benyuan Liu. 2014. Predicting User Replying Behavior on a Large Online Dating Site.. In *Proceedings of the International Conference on Web and Social Media*.

[30] Michael Yeomans, Anuj K. Shah, Sendhil Mullainathan, and Jon Kleinberg. 2018. Making sense of recommendations. Working paper.

[31] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 223–227.

[32] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 307–317.