International Joint Conference on Neural Networks (IJCNN). IEEE, 1–7.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[7] Haydn Belfield. 2020. Activism by the AI community: Analysing recent achievements and future prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 15–21.

[8] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[9] Ethan Corey. 2020. New Data Suggests Risk Assessment Tools Have Little Impact on Pretrial Incarceration. https://theappeal.org/new-data-suggests-risk-assessment-tools-have-little-impact-on-pretrial-incarceration/

[10] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. *SSRN Electronic Journal* (11 2017). https://doi.org/10.2139/ssrn.3064761

[11] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[12] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior* 46, 2 (2019), 185–209.

[13] John Fox, David Glasspool, Dan Grecu, Sanjay Modgil, Matthew South, and Vivek Patkar. 2007. Argumentation-based inference and decision making–A medical perspective. *IEEE intelligent systems* 22, 6 (2007), 34–41.

[14] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[15] Ben Green. 2020. The false promise of risk assessments: epistemic reform and the limits of fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 594–606.

[16] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[17] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[18] Ben Green and Yiling Chen. 2020. Algorithm-in-the-loop decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13663–13664.

[19] Philipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann. 2020. Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law* (2020), 1–25.

[20] Denis J Hilton. 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning* 2, 4 (1996), 273–308.

[21] Nicholas C Hunt and Andrea M Scheetz. 2019. Using MTurk to distribute a survey or experiment: Methodological considerations. *Journal of Information Systems* 33, 1 (2019), 43–65.

[22] Hripsime A Kalaian and Stephen W Raudenbush. 1996. A multivariate mixed linear model for meta-analysis. *Psychological methods* 1, 3 (1996), 227.

[23] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[24] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[25] Jeff Larson, Julia Angwin, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. Retrieved Mar 1, 2021 from http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[26] P. Law, Sana Malik, Fan Du, and M. Sinha. 2020. The Impact of Presentation Style on Human-In-The-Loop Detection of Algorithmic Bias. In *Graphics Interface*.

[27] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[28] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[29] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[30] P Baumgartner MDeMichele, M Wenger, K Barrick, M Comfort, and S Misra. 2018. The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky.(2018).

[31] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[32] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[33] Carlos Montemayor, Jodi Halpern, and Abrol Fairweather. 2021. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *Ai & Society* (2021), 1–7.

[34] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617.

[35] 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. (02 2018).

[36] Northpointe. 2010. http://www.northpointeinc.com/files/technical_documents/Selected_Compas_Questions_Posed_by_Inquiring_Agencies.pdf

[37] Leila Ouchchy, Allen Coin, and Veljko Dubljević. 2020. AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY* 35, 4 (2020), 927–936.

[38] Wolter Pieters. 2011. Explanation and trust: what to tell the user in security and AI? *Ethics and information technology* 13, 1 (2011), 53–64.

[39] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[41] Maria Riveiro and Serge Thill. 2021. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence* 298 (2021), 103507. https://doi.org/10.1016/j.artint.2021.103507

[42] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.

[43] Philipp Schmidt and Felix Biessmann. 2020. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 431–449.

[44] Ed Yong. 2018. A popular algorithm is no better at predicting crimes than random people. *The Atlantic* 17 (2018), 2018.

[45] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.