NIST A.I. Reference Library
Bibliography: Bias in Artificial Intelligence

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2019). The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*.

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020, January). Roles for computing in social change. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252-260.

Aggarwal, A., Shaikh, S., Hans, S., Haldar, S., Ananthanarayanan, R., & Saha, D. (2021). Testing framework for black-box AI models. *arXiv preprint.* Retrieved from https://arxiv.org/pdf/2102.06166.pdf

Ahmed, N., & Wahed, M. (2020). The De-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*. Retrieved from https://arxiv.org/ftp/arxiv/papers/2010/2010.15581.pdf.

AI Now Institute. Algorithmic Accountability Policy Toolkit. (2018). Retrieved from: https://ainowinstitute.org/aap-toolkit.html.

Aitken, M., Toreini, E., Charmichael, P., Coopamootoo, K., Elliott, K., & van Moorsel, A. (2020, January). Establishing a social licence for Financial Technology: Reflections on the role of the private sector in pursuing ethical data practices. *Big Data & Society.* doi:10.1177/2053951720908892

Ajunwa, I. (2016). Hiring by Algorithm. *SSRN Electronic Journal*. doi:10.2139/ssrn.2746078

Ajunwa, I. (2020, Forthcoming). *The Paradox of Automation as Anti-Bias Intervention*, 41 Cardozo, L. Rev.

Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019, January). Uncovering and mitigating algorithmic bias through learned latent structure. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society,* 289-295.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Andrade, N. N. G. & Kontschieder, V. (2021). AI impact assessment: A policy prototyping experiment. *Open Loop*. Retrieved from: https://openloop.org/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM, 61*(6), 54-61.

Baig, E. C. (2018, December 2). Who's going to review your college applications—a committee or a computer? *USA Today.* Retrieved from https://www.usatoday.com/story/tech/2018/12/02/college-admissions-when-ai-robots-decide/2147201002/

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev., 104,* 671.

Bary, E. (2018, October 29). How artificial intelligence could replace credit scores and reshape how we get loans. *Market Watch.* Retrieved from https://www.marketwatch.com/story/ai-based-credit-scores-will-soon-give-one-billion-people-access-to-banking-services-2018-10-09

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943.* Retrieved from https://arxiv.org/abs/1810.01943

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6,* 587-604.

Benjamin, R. (2019a). Assessing risk, automating racism. *Science, 366*(6464), 421-422.

Benjamin, R. (2019b). *Race after technology: Abolitionist tools for the new jim code.* John Wiley & Sons.

Bietti, E. (2020, January). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 210-219.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Journal of Machine Learning Research*, 1-11.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's Reducing a Human Being to a Percentage'. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. doi:10.1145/3173574.3173951

Bird, S., Kenthapadi, K., Kiciman, E., & Mitchell, M. (2019, January). Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining,* 834-835.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint arXiv:2005.14050.* Retrieved from https://arxiv.org/pdf/2005.14050.pdf

Bogen, M. (2019, May 6). All the ways hiring algorithms can introduce bias. *Harvard Business Review*. Retrieved from https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias

Bogen, M., & Rieke, A. (2018). Help wanted: an examination of hiring algorithms, equity. and bias. *Technical report, Upturn.* Retrieved from https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems,* 4349-4357.

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121.*

Boyarskaya, M., Olteanu, A., & Crawford, K. (2020). Overcoming failures of imagination in AI infused system development and deployment. *arXiv preprint.* Retrieved from https://arxiv.org/pdf/2011.13416.pdf

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society, 15*(5), 662-679.

Boza, P., & Evgeniou, T. (2021). Implementing Ai Principles: Frameworks, Processes, and Tools. *INSEAD Working Paper No. 2021/04/DSC/TOM*. Retrieved from http://dx.doi.org/10.2139/ssrn.3783124

Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does predictive policing lead to biased arrests? Results from a randomized controlled trial. *Statistics and Public Policy, 5*(1), 1-6.

Broussard, Meredith (2018). *Artificial Unintelligence: How Computers Misunderstand the World,* The MIT Press.

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. Big Data & Society, 8(1). DOI: 10.1177/2053951720983865

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A. and Vaithianathan, R. (2019, May). Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 41*, 1-12. doi:10.1145/3290605.3300271

Brown, L., Richardson, M., Shetty, R., Crawford, A., & Hoagland, T. (2020, October). Challenging the use of algorithm-driven decision making in benefits determinations: Affecting people with disabilities. *Center for Democracy & Technology.* Retrieved from https://cdt.org/wp-content/uploads/2020/10/2020-10-21-Challenging-the-Use-of-Algorithm-driven-Decision-making-in-Benefits-Determinations-Affecting-People-with-Disabilities.pdf

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Maharaj, T. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213.* Retrieved from https://arxiv.org/pdf/2004.07213.pdf

Brunet, M. E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the origins of bias in word embeddings. *arXiv preprint arXiv:1810.03611.*

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). *Notes from the AI frontier: Modeling the impact of AI on the world economy.* McKinsey Global Institute.

Buolamwini, J. (2018, June 25). Letter to Mr. Jeffrey P. Bezos *"Re: Audit of Amazon Rekognition Uncovers Gender and Skin-Type Disparities."*

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77-91.

Burget, M., Bardone, E., & Pedaste, M. (2016). Definitions and conceptual dimensions of responsible research and innovation: A literature review. *Science and Engineering Ethics, 23*(1), 1–19. doi:10.1007/s11948-016-9782-1.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). doi: 10.1177/2053951715622512.

Burt, A., & Hall, P. (2020, May 18). What to do when AI fails. *O'Reilly*. Retrieved from www.oreilly.com/radar/what-to-do-when-ai-fails/

Burt, A., Leong, B., Shirrell, S., & Wang, X. G. (2018). Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models. *Future of Privacy Forum*.

Caliskan, A., & Lewis, M. (2020, July 16). Social biases in word embeddings and their relation to human cognition. Retrieved from https://doi.org/10.31234/osf.io/d84kg

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017, April 14). Semantics derived automatically from language corpora contain human-like biases, *Science, 356*(6334), 183-186.

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of science, 22*(3), 595-612.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). AI now 2017 report. *AI Now Institute at New York University*.

Carr, S. (2020). 'AI gone mental': engagement and ethics in data-driven technology for mental health. *Journal of Mental Health, 29*(2), 125-130. doi: 10.1080/09638237.2020.1714011

Centre for Data Ethics and Innovation. (2020, November). Review into bias in algorithmic decision-making. *CDEI.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf

Cevolini, A., & Esposito, E. (2020). From pool to profile: Social consequences of algorithmic prediction in insurance. *Big Data & Society, 7*(2). Doi: 10.1177/2053951720939228

Chaney, A. J. B., Stewart, B. M., & Engelhardt, B. E. (2018). How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. *Proceedings of the 12th ACM Conference on Recommender Systems - RecSys '18.* doi:10.1145/3240323.3240370

Chen, V. & Hooker, J. N. (2021, January). Fairness through optimization. *arxiv preprint.* Retrieved from https://arxiv.org/pdf/2102.00311.pdf

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020, November). Generative pretraining from pixels. *International Conference on Machine Learning,* 1691-1703.

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially responsible AI algorithms: Issues, purposes, and challenges. arXiv preprint. Retrieved from https://arxiv.org/pdf/2101.02032.pdf

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Carnegie Mellon University*. arXiv:1610.07524

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810.*

Chouldechova, A., Benavides-Prado, D., Fialko, O. & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in PMLR 81,* 134-148.

Christian, B. (2020) *The Alignment Problem*. W. W. Norton Company

Cihon, P. (2019, April). Standards for AI governance: International standards to enable global coordination in AI research & development. *Future of Humanity Institute, University of Oxford*. Retrieved from https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Commission Nationale de l'Informatique et des Libertés, European Data Protection Supervisor, Garante per la protezione dei dati personali. (2018, October 23). Declaration on Ethics and Data Protection in Artificial Intelligence. *40th International Conference of Data Protection & Privacy Commissioners, Brussels.* Retrieved from https://www.huntonprivacyblog.com/wp-content/uploads/sites/28/2018/10/ICDPPC-40th_AI-Declaration_ADOPTED.pdf

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023.*

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797-806.

Costanza-Chock, S. (2018, July 18). Design justice, A.I., and escape from the matrix of domination. *Journal of Design and Science, MIT Media Lab*. Retrieved from https://jods.mitpress.mit.edu/pub/costanza-chock/release/4

Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. (2020, July). Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. *Proceedings of the 21st ACM Conference on Economics and Computation,* 679-681.

Cowgill, B. (2018). Bias and productivity in humans and algorithms: Theory and evidence from resume screening. *Columbia Business School, Columbia University, 29*.

Crawford, K. (2016, June 25). Artificial Intelligence's White Guy Problem. *The New York Times*. Retrieved from https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*(7625), 311-313.

Crawford, K., & Joler, V. (2018). Anatomy of an AI System-The Amazon Echo as an anatomical map of human labor, data and planetary resources. *AI Now Institute and Share Lab, 7*. Retrieved from https://anatomyof.ai/.

Criado Perez, Caroline (2019). *Invisible Women: Data Bias in a World Designed for Men*, Abrams Press

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. *MIT Press*. Retrieved from https://data-feminism.mitpress.mit.edu/

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data, 5*(2), 120-134.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Hormozdiari, F. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395.*

Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. *IJCAI*, 4691-4697.

Dastin, J. (2018, October 9). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Demolino, D., Gebre, B., & Whitehouse, M. (2018). Artificial intelligence unleashed: How agencies can use AI to automate & augment operations to improve performance. *Accenture Federal Services.*

Deshpande, K. V., Pan, S., & Foulds, J. R. Mitigating Demographic Bias in AI-based Resume Filtering. *University of Maryland, Baltimore County.* Retrieved from http://jfoulds.informationsystems.umbc.edu/papers/2020/Deshpande%20(2020)%20-%20Mitigating%20Demographic%20Bias%20in%20AI-based%20Resume%20Filtering%20(FairUMAP).pdf

Diakopoulos, N. (2016, May 23). We need to know the algorithms the government uses to make important decisions about us. *The Conversation*. Retrieved from https://theconversation.com/we-need-to-know-the-algorithms-the-government-uses-to-make-important-decisions-about-us-57869

Diaz, F. (2016, June). Worst practices for designing production information access systems. *ACM SIGIR Forum, 50*(1), 2-11. Retrieved from https://dl.acm.org/doi/pdf/10.1145/2964797.2964799

Dickerson, S., Haggerty, P., Hall, P., Kannan, A. R., Kulkarni, R., Prochaska, K., Schmidt, N., & Wiwczaroski, M. (2020). Machine learning: Considerations for expanding access to credit fairly and transparently. *BLDS, LLC., Discover Financial Services Inc., & H2O.ai.* Retrieved from http://info.h2o.ai/rs/644-PKX-778/images/Machine%20Learning%20-%20Considerations%20for%20Fairly%20and%20Transparently%20Expanding%20Access%20to%20Credit.pdf

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *SSRN Electronic Journal.* doi:10.2139/ssrn.2466040

Dogruel, L., Facciorusso, D., & Stark, B. (2020). 'I'm still the master of the machine.'Internet users' awareness of algorithmic decision-making and their perception of its effect on their autonomy. *Information, Communication & Society*, 1-22. Doi:10.1080/1369118X.2020.1863999

Dormehl, L. (2014, November 19). Algorithms are great and all, but they can also ruin lives. *Wired*. Retrieved from https://www.wired.com/2014/11/algorithms-great-can-also-ruin-lives/

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., & Busch, C. (2020). Demographic Bias in Biometrics: A Survey on an Emerging Challenge. *IEEE Transactions on Technology and Society, 1*(2), 89–103. Retrieved from https://doi.org/10.1109/TTS.2020.2992344

Dwork, C., & Ilvento, C. (2018). Fairness under composition. *arXiv preprint arXiv:1806.06122.*

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-226.

Elish, M. C., Barocas, S., Plasek, A., and Ferryman, K. (2016, July 7). The social & economic implications of artificial intelligence technologies in the near-term. *AI Now 2016 Symposium, New York.* Retrieved from https://ainowinstitute.org/AI_Now_2016_Primers.pdf

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M. F. (2020). Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. *Administrative Conference of the United States*. Retrieved from https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf

EPIC. (n.d.). Algorithms in the criminal justice system: Pre-trial risk assessment tools. *Electronic Privacy Information Center.* https://epic.org/algorithmic-transparency/crim-justice/

Esterle, L., Guckert, M., Anh Han, T., and Lewis, P. R. (2018, December). Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine*. doi: 10.1109/MTS.2018.2876107

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Commission (2018, December 18). Draft Ethics guidelines for trustworthy AI. *European Commission.* Retrieved from https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai

European Commission, Directorate-General for Communications Networks, Content and Technology (2017, March). Attitudes towards the impact of digitisation and automation on daily life. *European Commission, Special Eurobarometer, 460.*

European Commission. (2020, July 17). Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment. *European Commission Robotics and Artificial Intelligence Unit A.1*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Evans, M., & Wilde Mathews, A. (October 26, 2019). New York regulator probes UnitedHealth algorithm for racial bias. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/new-york-regulator-probes-unitedhealth-algorithm-for-racial-bias-11572087601

Farrand, T., Mireshghallah, F., Singh, S., & Trask, A. (2020). Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. *arXiv preprint arXiv:2009.06389.* Retrieved from https://arxiv.org/pdf/2009.06389.pdf

Fast, E., & Horvitz, E. (2017, February). Long-term trends in the public perception of artificial intelligence. *Thirty-First AAAI Conference on Artificial Intelligence.*

Fazelpour, S. & Lipton, Z. C. (2020). Algorithmic fairness from a non-ideal perspective. *2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES'20)*, New York, NY, USA. Doi: 10.1145/3375627.3375828

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2020). Bias and Discrimination in AI: a cross-disciplinary perspective. *arXiv preprint arXiv:2008.07309*. Retrieved from https://arxiv.org/pdf/2008.07309.pdf

Fischer, M. (2020, February 20). Machine learning can't fix algorithmic bias. But humans can. *Quartz at Work.* Retrieved from https://qz.com/work/1805067/machine-learning-wont-fix-algorithmic-bias/?fbclid=IwAR3vSqRrGLgJPT_zxdXVmJlIBPhGe41DzU0nAKVPFXB0iwYdoK2FzCJOB40

Fish, B., & Stark, L. (2020). Reflexive design for fairness and other human values in formal models. *arXiv preprint arXiv:2010.05084.* Retrieved from https://arxiv.org/pdf/2010.05084.pdf

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019, January). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329-338.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS), 14*(3), 330-347.

Fry, H. (2018). *Hello world: being human in the age of algorithms.* WW Norton & Company.

Fussell, S. (2019, October 9). How an attempt at correcting bias in tech goes wrong. *The Atlantic*. Retrieved from https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/

Future of Privacy Forum. (2017, December 11). Unfairness by algorithm: Distilling the harms of automated decision-making. *Future of Privacy Forum*. Retrieved from https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/

Galhotra, S., Brun, Y., & Meliou, A. (2017, August). Fairness testing: testing software for discrimination. *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 498-510. Retrieved from https://people.cs.umass.edu/~brun/pubs/pubs/Galhotra17fse.pdf

Garvie, C., & Frankle, J. (2016, April 7). Facial-recognition software might have a racial bias problem. *The Atlantic.* Retrieved from https://www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., ... & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *npj Digital Medicine, 4*(1), 1-8. DOI:10.1038/s41746-021-00385-9

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumeé III, H. and Crawford, K. (2018). Datasheets for Datasets. *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning: Stockholm, Sweden.* Retrieved from: https://arxiv.org/abs/1803.09010.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine, 178*(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Goel, K., Rajani, N., Vig, J., Tan, S., Wu, J., Zheng, S., ... Ré, C. (2021). Robustness gym: Unifying the NLP evaluation landscape. *arxiv preprint.* Retrieved from https://arxiv.org/pdf/2101.04840.pdf

Goel, S., Shroff, R., Skeem, J., & Slobogin, C. (2019, January). The accuracy, equity, and jurisprudence of criminal risk assessment. *Social Science Research Network.* Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3306723

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862.*

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine, 38*(3), 50-57.

Green, B., & Viljoen, S. (2020). Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*)*.

Gupta, A., Royer, A., Heath, V., Wright, C., Lanteigne, C., Cohen, A., ... & Akif, M. (2020, October). The state of AI ethics report (October 2020). *arXiv preprint arXiv:2011.02787.*

Hajian, S., Bonchi, F., & Castillo, C. (2016, August). Algorithmic bias: From discrimination discovery to fairness-aware data mining. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2125-2126.

Hall, P., Gill, N., & Cox, B. (2020). *Responsible machine learning: Actionable strategies for mitigating risks and driving adoption.* Sebastopol, CA: O'Reilly Media Inc.

Hao, K. (2021, February 5). This is how we lost control of our faces. *MIT Technology Review.*

Hao, K. (2019, April 15). Congress wants to protect you from biased algorithms, deepfakes, and other bad AI. *MIT Technology Review.* Retrieved from https://www.technologyreview.com/s/613310/congress-wants-to-protect-you-from-biased-algorithms-deepfakes-and-other-bad-ai/

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. arXiv preprint arXiv:1610.02413. Retrieved from https://arxiv.org/pdf/1610.02413.pdf

Hardt, M. (2014, September 26). How big data is unfair: Understanding unintended sources of unfairness in data driven decision making. *Medium.* Retrieved from https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Harel, M., Mannor, S., El-Yaniv, R., & Crammer, K. (2014, January). Concept drift detection through resampling. *International Conference on Machine Learning*, 1009-1017.

Harlan, E., & Schnuck, O. (2021, February). Objective or biased: On the questionable use of Artificial Intelligence for job applications. *Bayerischer Rundfunk.* Retrieved from https://web.br.de/interaktiv/ki-bewerbung/en/

Harwell, D. (2019, December 19). Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use. *The Washington Post.* Retrieved from https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/

Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in machine learning What is it good for?. *arXiv preprint arXiv:2004.00686.* Retrieved from: https://arxiv.org/pdf/2004.00686.pdf

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing System,* 1-16.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058.* Retrieved from https://arxiv.org/pdf/2010.03058.pdf.

Howard, J. J., Rabbitt, L. R., & Sirotin, Y. B. (2020). Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. *PloS one, 15*(8). Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237855.

Hurley, D. (2018, January 2). Can an algorithm tell when kids are in danger? *The New York Times Magazine.* Retrieved from https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html

Hurlin, C., Perignon, C., & Saurin, S. (2021). The Fairness of Credit Scoring Models. *Available at SSRN 3785882*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3785882

Hutchinson, B., Pittl, K. J., & Mitchell, M. (2019). Interpreting Social Respect: A Normative Lens for ML Models. *arXiv preprint arXiv:1908.07336.*

Ipsos, M. O. R. I. (2017). Public views of machine learning. *Royal Society*.

Jacobs, A. Z., Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020, January). The meaning and measurement of bias: Lessons from natural language processing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 706-706.

Jacobs, A. Z., & Wallach, H. (2019). Measurement and fairness. *arXiv preprint arXiv:1912.05511.* Retrieved from https://arxiv.org/pdf/1912.05511.pdf

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2020). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *arXiv preprint arXiv:2010.07487.* Retrieved from https://arxiv.org/pdf/2010.07487.pdf.

Jaume-Palasí, L., & Spielkamp, M. (2017). Ethics and algorithmic processes for decision making and decision support. *AlgorithmWatch, 2*, 1-18. Working paper.

Johndrow, J. E., & Lum, K. (2017). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*.

Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems*, 325-333.

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2017). Simple rules for complex decisions. *SSRN*. Retrieved from http://dx.doi.org/10.2139/ssrn.2919024

Kamiran, F. & Calders, T. (2012, December 3). Data processing techniques for classification without discrimination. *Knowledge and Information Systems, 33*, 1-33. doi:10.1007/s10115-011-0463-8

Kamiran, F., Karim, A., Verwer, S., & Goudriaan, H. (2012). Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset. *2012 IEEE 12th International Conference on Data Mining Workshops.* doi:10.1109/icdmw.2012.117

Kamiran, F., Mansha, S., Karim, A., & Zhang, X. (2018). Exploiting reject option in classification for social discrimination control. *Information Sciences, 425,* 18–33. doi:10.1016/j.ins.2017.09.064

Karkkainen, K., & Joo, J. (2021). FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1548-1558.* Retrieved from https://openaccess.thecvf.com/content/WACV2021/papers/Karkkainen_FairFace_Face_Attribute_Dataset_for_Balanced_Race_Gender_and_Age_WACV_2021_paper.pdf

Katwala, A. (2020, August 15). An algorithm determined UK students' grades. Chaos ensued. *Wired UK.* Retrieved from https://www.wired.com/story/an-algorithm-determined-uk-students-grades-chaos-ensued/

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design.* Oxford University Press.

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society, 7*(1), 2053951720915939.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big data & society, 1*(1), 2053951714528481.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14-29.

Klein, A. (2020, July 10). Reducing bias in AI-based financial services. *Brookings Institution.* Retrieved from https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2020). Algorithms as discrimination detectors. *Proceedings of the National Academy of Science*s. Retrieved from https://www.pnas.org/content/pnas/117/48/30096.full.pdf

Kleinberg, J., & Mullainathan, S. (2019, May). Simplicity creates Inequity: Implications for fairness, stereotypes, and interpretability. *National Bureau of Economic Research, NBER Working Paper Series.* Retrieved from http://www.nber.org/papers/w25854

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics.* doi:10.1093/qje/qjx032

Knight Foundation & Gallup. (2020, March). Techlash? America's growing concern with major technology companies. *Knight Foundation.* Retrieved from https://knightfoundation.org/wp-content/uploads/2020/03/Gallup-Knight-Report-Techlash-Americas-Growing-Concern-with-Major-Tech-Companies-Final.pdf

Knowles, B., & Richards, J. T. (2021). The sanction of authority: Promoting public trust in AI. *arXiv preprint.* Retrieved from https://arxiv.org/ftp/arxiv/papers/2102/2102.04221.pdf

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684-7689.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684-7689.

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., ... & Lomas, E. (2021). Towards algorithm auditing: A survey on managing legal, ethical and technological risks of AI, ML and associated algorithms.

Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180084. doi:10.1098/rsta.2018.0084

Kroll, J.A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*.

Kusner, M., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. *Conference on Neural Information Processing Systems*.

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15.* doi:10.1145/2783258.2788620

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science, 65*(7), 2966-2981.

Langford, C. (2017, May 8). Houston schools must face teacher evaluation lawsuit. *Courthouse News.* Retrieved from https://www.courthousenews.com/houston-schools-must-face-teacher-evaluation-lawsuit/

Latzer, M., & Festic, N. (2019). A guideline for understanding and measuring algorithmic governance in everyday life. *Internet Policy Review, 8*(2).

Leavy, S., O'Sullivan, B., & Siapera, E. (2020). Data, Power and Bias in Artificial Intelligence. *arXiv preprint arXiv:2008.07341.* Retrieved from https://arxiv.org/pdf/2008.07341.pdf

Ledford, H. (2019, October 31). Millions affected by racial bias in health-care algorithm. *Nature, 374*, 608-609. Retrieved from https://www.nature.com/articles/d41586-019-03228-6

Lee, E. W. J., & Yee, A. Z. H. (2020). Toward data sense-making in digital health communication research: Why theory matters in the age of big data. *Frontiers in Communication, 5*(11), 1-10.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 205395171875668. doi:10.1177/2053951718756684

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., ... & Legg, S. (2017). AI safety gridworlds. *arXiv preprint*. Retrieved from https://arxiv.org/pdf/1711.09883.pdf

Letzter, R. (2016, April 21). Amazon just showed us that 'unbiased' algorithms can be inadvertently racist. *Business Insider.* Retrieved from https://www.businessinsider.com/how-algorithms-can-be-racist-2016-4

Lewis, A. (2021). Reframing opportunity and fairness in the AI diversity pipeline. Retrieved from http://ceur-ws.org/Vol-2812/RDAI-2021_paper_7.pdf

Li, T., Khashabi, D., Khot, T., Sabharwal, A., & Srikumar, V. (2020). UNQOVERing Stereotyping biases via underspecified questions. *arXiv preprint.* Retrieved from: https://arxiv.org/pdf/2010.02428.pdf

Lin, Y. T., Hung, T. W., & Huang, L. T. L. (2020). Engineering equity: How AI can help reduce the harm of implicit bias. *Philosophy & Technology*, 1-26.

Liptak, A. (2017, May 1). Sent to prison by a software program's secret algorithms. *The New York Times*. Retrieved from https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html

Ma, X., Sap, M., Rashkin, H., & Choi, Y. (2020). PowerTransformer: Unsupervised controllable revision for biased language correction. *arXiv preprint arXiv:2010.13816.* Retrieved from https://arxiv.org/pdf/2010.13816.pdf.

Maddox, T. M., Rumsfeld, J. S., & Payne, P. R. (2018). Questions for artificial intelligence in health care. *JAMA, 321*(1), 31-32.

Mahtani, K., Spencer, E. A., Brassey, J., ... (2018). Catalogue of bias: observer bias. *BMJ Evidence-Based Medicine, 23*, 23-24.

Malik, M. M. (2020). A hierarchy of limitations in machine learning. *arXiv preprint*. Retrieved from https://arxiv.org/pdf/2002.05193.pdf

Manyika, J., Silberg, J., & Preston, B. (2019, October 25). What do we do about the biases in AI? *Harvard Business Review.* Retrieved from https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Matsakis, L. (May 29, 2020). Walmart employees are out to show its anti-theft AI doesn't work. *Wired.* Retrieved from https://www.wired.com/story/walmart-shoplifting-artificial-intelligence-everseen/

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561.*

McGraw, G., Figueroa, H., Shepardson, V., & Bonett, R. (2020). An architectural risk analysis of machine learning systems: Toward more secure machine learning. *Berryville Institute of Machine Learning,* Clarke County, VA. Retrieved from https://berryvilleiml.com/docs/ara.pdf

McStay, A. (2020). Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society.* https://doi.org/10.1177/2053951720904386

Mehrabi, N., Gowda, T., Morstatter, F., Peng, N., & Galstyan, A. (2020, July). Man is to person as woman is to location: Measuring gender bias in named entity recognition. *Proceedings of the 31st ACM Conference on Hypertext and Social Media,* 231-232. Retrieved from https://arxiv.org/pdf/1910.10872.pdf.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635.*

Metevier, B., Giguere, S., Brockman, S., Kobren, A., Brun, Y., Brunskill, E., & Thomas, P. S. (2019). Offline Contextual Bandits with High Probability Fairness Guarantees. *Advances in Neural Information Processing Systems,* 14893-14904. Retrieved from https://people.cs.umass.edu/~brun/pubs/pubs/Metevier19neurips.pdf

Miller, A. P. (2018, July 26). Want less-biased decisions? Use algorithms. *Harvard Business Review*. Retrieved from https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

Mishra, S., Clark, J., & Perrault, C. R. (2020). Measurement in AI policy: Opportunities and challenges. *arXiv preprint arXiv:2009.09071.* Retrieved from: https://arxiv.org/pdf/2009.09071.pdf

Misra, I., Zitnick, C. L., Mitchell, M., & Girshick, R. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2016.320

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229.

Mitchell, Melanie (2019). *Artificial Intelligence: A Guide for Thinking Human*. Farrar, Straus and Giroux.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), doi:2053951716679679.

Morgan, D. (2019, March 12). Are you there Google? It's me, a woman. *Medium*. Retrieved from https://medium.com/s/story/are-you-there-google-its-me-a-woman-3f4f527badb1

Morrison, K. (2020). Reducing discrimination in learning algorithms for social good in sociotechnical systems. *arXiv preprint*. Retrieved from https://arxiv.org/pdf/2011.13988.pdf

Morse, A. & Pence, K. (2020). Technological innovation and discrimination in household finance. Finance and Economics Discussion Series 2020-018. *Washington: Board of Governors of the Federal Reserve System.* Retrieved from https://doi.org/10.17016/FEDS.2020.018.

Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction, 3(CSCW)*, 1-36. Retrieved from https://arxiv.org/abs/1909.11869

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. *arXiv e-prints, arXiv-2010*. Retrieved from https://arxiv.org/pdf/2010.00133.pdf

National Security Commission on Artificial Intelligence. (2021). *NSCAI final report*. Retrieved from https://www.nscai.gov/2021-final-report/

Newman, J. C. (2020, May 5). Decision points in AI governance: Three case studies explore efforts to operationalize AI principles. *Center for Long-Term Cybersecurity, UC Berkeley*. Retrieved from https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism.* NYU Press.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Kompatsiaris, I. (2020). Bias in Data-driven AI Systems--An Introductory Survey. *arXiv preprint arXiv:2001.09762.*

O'Neil, Cathy (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,* Broadway Books.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. doi:10.1126/science.aax2342

Olszewska, J. I. (2019). D7-R4: software development life-cycle for intelligent vision systems. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2*, 435-441. SciTePress. Retrieved from https://doi.org/10.5220/0008354804350441.

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2*, 13.

Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, 1-17.

Paganini, M. (2020). Prune responsibly. *arXiv preprint.* Retrieved from https://arxiv.org/pdf/2009.09936.pdf

Palmiter Bajorek, J. (2019, May 10). Voice recognition still has significant race and gender biases. *Harvard Business Review.* Retrieved from https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health, 9*(2). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875681/

Pandey, A., & Caliskan, A. (2020). Iterative Effect-Size Bias in Ridehailing: Measuring Social Bias in Dynamic Pricing of 100 Million Rides. *arXiv preprint arXiv:2006.04599.*

Passi, S., & Barocas, S. (2019, January). Problem formulation and fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency,* 39-48.

Paul, W., Hadzic, A., Joshi, N., & Burlina, P. (2020). RENATA: Representation and training alteration for bias mitigation. *arxiv preprint.* Retrieved from https://arxiv.org/pdf/2012.06387.pdf

Pedersen, T., Johansen, C., & Johansen, J. (2020). Studying the Transfer of Biases from Programmers to Programs. *arXiv preprint arXiv:2005.08231*.

Peña, A., Serna, I., Morales, A., & Fierrez, J. (2020). Bias in Multimodal AI: Testbed for Fair Automatic Recruitment. *ArXiv preprint.* Retrieved from https://arxiv.org/pdf/2004.07173.pdf

Polack, P. (2019). AI discourse in policing criticisms of algorithms. *Evental Aesthetics, 8*, 57-92. Retrieved from https://eventalaesthetics.net/wp-content/uploads/2019/11/EAV8_2019_Polack_Policing_Algorithms_57_92.pdf

Polack, P. (2020). Beyond algorithmic reformism: Forward engineering the designs of algorithmic systems. *Big Data & Society, 7*(1), 205395172091306. doi:10.1177/2053951720913064

Prunkl, C. E., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence, 3*(2), 104-110. DOI:10.1038/s42256-021-00298-y

Raghavan, M., Barocas, S., Kleinberg, J.M, and Levy, K (2019) Mitigating bias in algorithmic hiring: evaluating claims and practices. *arXiv preprint.* Retrieved from https://arxiv.org/pdf/1906.09208.pdf

Raji, I. D., & Buolamwini, J. (2019, January). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429-435.

Raji, I. D., & Yang, J. (2019). Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles (ABOUT ML). *arXiv preprint arXiv:1912.06166.*

Ramey, C. (2020, September 20). Algorithm helps New York decide who gets free before trial: Point system used by judges awards defendants better chance at release based on data including convictions—and if they are reachable by phone. *The Wall Street Journal Online.*

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2020). Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *arXiv preprint arXiv:2010.07938*. Retrieved from https://arxiv.org/pdf/2010.07938.pdf.

Redden, J. (2018, November 1). The harm that data do. *Scientific American.* Retrieved from https://www.scientificamerican.com/article/the-harm-that-data-do/

Reike, A., Bogen, M., & Robinson, D. G. (2018) Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods. *Upturn and Omidyar Network*.

Reinhart, RJ. (2018, March 6). Most Americans already using artificial intelligence products. *Gallup*. Retrieved from https://news.gallup.com/poll/228497/americans-already-using-artificial-intelligence-products.aspx

Reisman, D., Schultz, J., Crawford, K., Whittaker, M. (2018). Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability. *AI Now Institute*. Retrieved from: https://ainowinstitute.org/aiareport2018.html.

Richardson, R. (2019, December 4). Confronting black boxes: A shadow report of the New York City automated decision system task force. *AI Now Institute.* Retrieved from https://ainowinstitute.org/ads-shadowreport-2019.html

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online,* Forthcoming.

Ricks, B., & Surman, M. (2020, December). Creating trustworthy AI: A Mozilla white paper on challenges and opportunities in the AI era. *Mozilla.* Retrieved from https://foundation.mozilla.org/en/insights/trustworthy-ai-whitepaper/executive-summary/

Rieland, R. (2018, March 5). Artificial intelligence is now used to predict crime. But is it biased? *Smithsonian Magazine*. Retrieved from https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/

Romanov, A., De-Arteaga, M., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., ... & Kalai, A. T. (2019). What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. *arXiv preprint arXiv:1904.05233.*

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review, 29*(5), 582-638.

Rossi, F., Sekaran, A., Spohrer, J., and Caruthers, R. (2019). Every ethics for artificial intelligence. *IBM Design Program Office.* Retrieved from https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

Rossi, F. (2019). AI Ethics for enterprise AI [PowerPoint slides]. Retrieved from https://economics.harvard.edu/files/economics/files/rossi-francesca_4-22-19_ai-ethics-for-enterprise-ai_ec3118-hbs.pdf

Roy, M. (2017, May 30). Is a chief AI officer needed to drive an artificial intelligence strategy?. *TechTarget*. Retrieved from https://searchcio.techtarget.com/feature/Is-a-chief-AI-officer-needed-to-drive-an-artificial-intelligence-strategy

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2019). Aequitas: A Bias and Fairness Audit Toolkit. *arXiv:1811.05577 [Cs]*. Retrieved from http://arxiv.org/abs/1811.05577

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., ... & Datta, D. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*.

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020, January). What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 458-468.

Saria, S., & Subbaswamy, A. (2019). Tutorial: safe and reliable machine learning. *arXiv preprint arXiv:1904.07204.*

Scheffler, S., Smith, A. D., and Canetti, R. (2019, March 7). Artificial intelligence must know when to ask for human help. *The Conversation.* Retrieved from https://theconversation.com/artificial-intelligence-must-know-when-to-ask-for-human-help-112207

Schick, T., Udupa, S., & Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *arXiv preprint.* Retrieved from https://arxiv.org/pdf/2103.00453.pdf

Schwab, K. (2019, September 18). The real reason Google Assistant launched with a female voice: biased data. *Fast Company.* Retrieved from https://www.fastcompany.com/90404860/the-real-reason-there-are-so-many-female-voice-assistants-biased-data

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., & Lockhart, J. W. (2020). Diagnosing Gender Bias in Image Recognition Systems. *Socius*. https://doi.org/10.1177/2378023120967171

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68.

Serna, I., Peña, A., Morales, A., & Fierrez, J. (2020). InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics. *arXiv preprint arXiv:2004.06592.*

Seyfert, R. (2021): Algorithms as regulatory objects. *Information, Communication & Society.* DOI: 10.1080/1369118X.2021.1874035

Shah, D., Schwartz, H. A., & Hovy, D. (2019). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *arXiv preprint arXiv:1912.11078*. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.468v2.pdf

Shane, Jannelle (2019). *You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place,* Voracious Books.

Sharma, K. (February 9, 2018). Can we keep our biases from creeping into AI? *Harvard Business Review.* Retrieved from https://hbr.org/2018/02/can-we-keep-our-biases-from-creeping-into-ai

Shneiderman, B. (2016). Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences, 113*(48), 13538–13540. doi:10.1073/pnas.1618211113

Silva, S., & Kenney, M. (2019). Algorithms, platforms, and ethnic bias. *Communications of the ACM, 62*(11), 37-39.

Simonite, T. (2020, October 26). How an algorithm blocked kidney transplants to Black patients. *Wired.* Retrieved from https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/

Singh, R., Vatsa, M., & Ratha, N. (2020). Trustworthy AI. *arXiv preprint arXiv:2011.02272.* Retrieved from https://arxiv.org/pdf/2011.02272.pdf

Singh, M., & Ramamurthy, K. N. (2019). Understanding racial bias in health using the Medical Expenditure Panel Survey data. arXiv preprint arXiv:1911.01509. Retrieved from https://drive.google.com/file/d/1sUDjpqUQptSdaDUOgHWS4NzJPpQegTg1/view

Sixta, T., Jacque Jr., J. C. S., Buch-Cardona, P., Vazquez, E., & Escalera, S. (2020). FairFace Challenge at ECCV 2020: Analyzing bias in face recognition. *arXiv preprint.* Retrieved from https://arxiv.org/pdf/2009.07838.pdf

Smith., A., and Anderson, M. (2017, October). Automation in Everyday Life. *Pew Research Center.*

Smith., A., and Nolan, H. (2018, November). Public Attitudes Toward Computer Algorithms. *Pew Research Center.*

Snow, T. (2021). From satisficing to artificing: The evolution of administrative decision-making in the age of the algorithm. Data & Policy, 3. DOI: 10.1017/dap.2020.25

Specia, M. (2019, May 22). Siri and Alexa reinforce gender bias, U.N. finds. *The New York Times.* Retrieved from https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html

Srivastava, B., & Rossi, F. (2018). Towards composable bias rating of AI services. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18.* doi:10.1145/3278721.3278744

Stahl, B., Obach, M., Yaghmaei, E., Ikonen, V., Chatfield, K., & Brem, A. (2017). The Responsible Research and Innovation (RRI) Maturity Model: Linking Theory and Practice. *Sustainability, 9*(6), 1036. doi:10.3390/su9061036

Steed, R. & Caliskan, A. (2021). Image representations learned with unsupervised pre-training contain human-like biases. *Conference on Fairness, Accountability, and Transparency (FAccT '21) March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA*. DOI: 10.1145/3442188.3445932

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., ... & Leyton-Brown, K. (2016). Artificial Intelligence and Life in 2030. One hundred year study on artificial intelligence: Report of the 2015-2016 Study Panel. *Stanford University.* Retrieved from http://ai100.stanford.edu/2016-report.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002.*

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., & Kalai, A. T. (2019, January). What are the biases in my word embedding?. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society,* 305-311.

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint.* Retrieved from: https://arxiv.org/pdf/2102.02503.pdf

Tangermann, V. (April 26, 2019). Amazon used an AI to automatically fire low-productivity workers. *Futurism*. Retrieved from https://futurism.com/amazon-ai-fire-workers

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing,* 53-59.

The Economist. (2020, June 11). Humans will add to AI's limitations. *The Economist.* Retrieved from https://www.economist.com/technology-quarterly/2020/06/11/humans-will-add-to-ais-limitations

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020, January). The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 272-283.

Torralba, A., & Efros, A. A. (2011, June). Unbiased look at dataset bias. *CVPR 2011*, 1521-1528. IEEE.

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society.* Retrieved from https://pdfs.semanticscholar.org/e2d0/046cc76cf9a515acb29002a4d0b4e9776cde.pdf.

Tutt, A. (2017). An FDA for algorithms. *Administrative Law Review, 69*, 83.

United States House committee on Science, Space and Technology: Hearing on Artificial Intelligence: Societal and Ethical Implications, US House of representatives, 116th Cong. (2019) (Full Testimony).

Van de Poel, I. (2015). An Ethical Framework for Evaluating Experimental Technology. *Science and Engineering Ethics, 22*(3), 667–686. doi:10.1007/s11948-015-9724-3

Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data, 5*(3), 246-255. Retrieved from https://www.liebertpub.com/doi/abs/10.1089/big.2016.0051?journalCode=big

Veale, M., Binns, R., & Edwards, L. (2018). Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180083.

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* doi:10.1145/3173574.3174014

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1-7. IEEE.

Vogel, M. (June 14, 2020). COVID-19 could bring bias in AI to pandemic level crisis. *Thrive Global.* Retrieved from https://thriveglobal.com/stories/covid-19-could-bring-bias-in-ai-to-pandemic-level-crisis/

Von Schomberg, R. (2011). Prospects for Technology Assessment in a Framework of Responsible Research and Innovation. *SSRN Electronic Journal*. doi:10.2139/ssrn.2439112

Wachter-Boettcher, S. (2017). *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech.* WW Norton & Company.

Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?. *BMJ global health, 3*(4), e000798.

Weber, M., Yurochkin, M., Botros, S., & Markov, V. (2020). Black loans matter: Distributionally robust fairness for fighting subgroup discrimination. *arXiv preprint.* Retrieved from https://arxiv.org/pdf/2012.01193.pdf

West, D. (2018a, May 21). Brookings survey finds worries over AI impact on jobs and personal privacy, concern U.S. will fall behind China. *Brookings Institute.* Retrieved from https://www.brookings.edu/blog/techtank/2018/05/21/brookings-survey-finds-worries-over-ai-impact-on-jobs-and-personal-privacy-concern-u-s-will-fall-behind-china//

West, D. (2018b, August 29). Brookings survey finds divided views on artificial intelligence for warfare, but support rises if adversaries are developing it. *Brookings Institute.* Retrieved from https://www.brookings.edu/blog/techtank/2018/08/29/brookings-survey-finds-divided-views-on-artificial-intelligence-for-warfare-but-support-rises-if-adversaries-are-developing-it/

West, D. M., & Allen, J. R. (2020). *Turning Point: Policymaking in the Era of Artificial Intelligence.* Brookings Institution Press.

West, D., & Allen, J. (2018, April 24). How artificial intelligence is transforming the world. *Brookings Institute*. Retrieved from https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/

Wexler, R. (June 13, 2017).When a computer program keeps you in jail. *The New York Times.* Retrieved from https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

Whang, S. E., Tae, K. H., Roh, Y., & Heo, G. (2021). Responsible AI challenges in end-to-end machine learning. *arxiv preprint.* Retrieved from https://arxiv.org/pdf/2101.05967.pdf

Wheeler, T., Verveer, P., & Kimmelman, G. (2020, August). New digital realities; New oversight solutions in the U.S.: The case for a digital platform agency and a new approach to regulatory oversight. *Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy.* Retrieved from https://shorensteincenter.org/wp-content/uploads/2020/08/New-Digital-Realities_August-2020.pdf

Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., ... & West, S. M. (2019, November). Disability, bias, and AI. *AI Now Institute.* Retrieved from https://ainowinstitute.org/disabilitybiasai-2019.pdf?fbclid=IwAR31dX3o_nkVf-cirQ9P-yJqRRkT1vcKU3MgcEAeWVwUgA0Ue1c-60Zd9OE

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy, 8,* 78-115.

Wilson, M. (2017, August 22). This breakthrough tool detects racism and sexism in software. *Fast Company*. Retrieved from https://www.fastcompany.com/90137322/is-your-software-secretly-racist-this-new-tool-can-tell

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A Qualitative Exploration of Perceptions of Algorithmic Fairness. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* doi:10.1145/3173574.3174230

Wykstra, S. (2020). Government's use of algorithm serves up false fraud charges. *Undark.* Retrieved from https://undark.org/2020/06/01/michigan-unemployment-fraud-algorithm/

Xiang, A., & Raji, I. D. (2019). On the Legal Compatibility of Fairness Definitions. *arXiv preprint arXiv:1912.00761.*

Yang, K., & Stoyanovich, J. (2017, June). Measuring fairness in ranked outputs. *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 1-6.

Yates, K. (2020, January 11). Why do we gender AI? Voice tech firms move to be more inclusive. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2020/jan/11/why-do-we-gender-ai-voice-tech-firms-move-to-be-more-inclusive

Yeung, D., Khan, I., Kalra, N., & Osoba, O.A. (2021). Identifying systemic bias in the acquisition of machine learning decision aids for law enforcement applications. *RAND Corporation, Santa Monica, CA.* Retrieved from: https://www.rand.org/pubs/perspectives/PEA862-1.html.

Yuan, M., Kumar, V., Ahmad, M. A., & Teredesai, A. (2021). Assessing Fairness in Classification Parity of Machine Learning Models in Healthcare. *arXiv preprint.* Retrieved from https://arxiv.org/pdf/2102.03717.pdf

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values, 41*(1), 118-132.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017, November). Fa* ir: A fair top-k ranking algorithm. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1569-1578.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). Learning fair representations. *International Conference on Machine Learning*, 325-333.

Zhang, B. and Dafoe, A. (2019, January 9). Artificial Intelligence: American attitudes and Trends. *SSRN*. Doi:10.2139/ssrn.3312874

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335-340.

Zhang, X., Pérez-Stable, E. J., Bourne, P. E., Peprah, E., Duru, O. K., Breen, N., ... & Denny, J. (2017). Big data science: opportunities and challenges to address minority health and health disparities in the 21st century. *Ethnicity & Disease, 27*(2), 95.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zimmerman, A., Di Rosa, E., and Kim, H. (2020, January 9). Technology can't fix algorithmic injustice. *Boston Review.* Retrieved from https://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148.*

Zou, J. and Schiebiner, L. (2018, July 18). AI can be sexist and racists—it's time to make it fair. *Nature.* Retrieved from https://www.nature.com/articles/d41586-018-05707-8