Building Accountable Machine Learning Systems in Practice

By Rebecca N. White

WORKS REFERENCED

Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. 2016. "Machine Bias." ProPublica. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Association for Computing Machinery US Public Policy Council. 2017. "Statement on Algorithmic Transparency and Accountability." January 12. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

Ananny, Mike, and Kate Crawford. 2016. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." New Media & Society. http://journals.sagepub.com/doi/abs/10.1177/1461444816676645.

"Asilomar AI Principles." 2017. Future of Life Institute. https://futureoflife.org/ai-principles/.

Bamberger, Kenneth A., and Deirdre K. Mulligan. 2011. "Privacy on the Books and on the Ground." *Stanford Law Review* 63, No. 2 (January 2011): 247–315.

Bamman, David. "Interpretability in Human-Centered Data Science." https://cscw2016hcds.files.wordpress.com/2015/10/bamman_hcds.pdf.

Barabas, Chelsea, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency PMLR 81*: 62–76. http://proceedings.mlr.press/v81/barabas18a.html.

Barocas, Solon. n.d. "Understanding Inscrutability." http://www.law.nyu.edu/sites/default/files/upload_documents/Solon%20Barocas%20Slides_0.pdf.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." 104 *California Law Review* 617 (2016). http://dx.doi.org/10.15779/Z38BG31.

Barocas, Solon, and danah boyd. 2017. "Engaging the Ethics of Data Science in Practice." *Communications of the ACM* 60 (11): 23–25. https://cacm.acm.org/magazines/2017/11/222176-engaging-the-ethics-of-data-science-in-practice/fulltext.

Bogost, Ian. 2005. "Procedural Literacy: Problem Solving with Programming, Systems, and Play." *Journal of Media Literacy* 52 (1). http://bogost.com/writing/procedural_literacy/.

Bornstein, Aaron M. 2016. "Is Artificial Intelligence Permanently Inscrutable?" *Nautilus*. September 1, 2016. http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable.

boyd, danah. 2018. "Beyond the Rhetoric of Algorithmic Solutionism." *Data & Society: Points.* January 11, 2018. https://points.datasociety.net/beyond-the-rhetoric-of-algorithmic-solutionism-8e0f9cdada53.

boyd, danah and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication, & Society* 15 (5): 662–679.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–231.

Brennan-Marquez, Kiel. 2017. "Plausible Cause: Explanatory Standards in the Age of Powerful Machines." *Vanderbilt Law Review* 70 (2017). https://ssrn.com/abstract=2827733.

Brink, Henrik, and Joshua Bloom. n.d. "Overcoming the Barriers to Production-Ready Machine-Learning Workflows." https://cdn.oreillystatic.com/en/assets/1/event/105/Overcoming%20the%20Barriers%20to%20Production-Ready%20Machine-Learning%20Workflows%20Presentation%201.pdf.

Brynjolfsson, Erik, and McAfee, Andrew. 2017. "The Business of Artificial Intelligence." Harvard Business Review. July 2017. https://hbr.org/2017/07/the-business-of-artificial-intelligence.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency PMLR 81*: 77–91. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1). https://doi.org/10.1177/2053951715622512.

Business and Human Rights Resource Centre. n.d. "UN 'Protect, Respect and Remedy' Framework and Guiding Principles." Accessed April 21, 2018. https://www.business-humanrights.org/en/un-secretary-generals-special-representative-on-business-human-rights/un-protect-respect-and-remedy-framework-and-guiding-principles.

Calhoun, Craig. 2011. "Civil society and the public sphere." In Edwards, Michael, ed., *The Oxford Handbook of Civil Society*. New York: Oxford University Press: 311–323.

Calo, Ryan. 2017. "Artificial Intelligence Policy: A Primer and Roadmap." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3015350.

Carnegie Mellon University. 2018. "Carnegie Mellon Launches Undergraduate Degree in Artificial Intelligence." May 10, 2018. http://www.cmu.edu/news/stories/archives/2018/may/ai-undergraduate-degee.html.

Carr, Nicholas. 2014. *The Glass Cage: How Our Computers Are Changing Us*. New York: W. W. Norton.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. KDD '15. New York, NY, USA: ACM. https://doi.org/10.1145/2783258.2788613.

Ann Cavoukian. "Privacy by Design: The 7 Foundational Principles." https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf.

Cavoukian, Ann, and Michelle Chibba. 2018. "Start with Privacy by Design in All Big Data Applications." In S. Srinivasan, ed., *Guide to Big Data Applications*, Studies in Big Data 26, https://www.springer.com/cda/content/document/cda_downloaddocument/9783319538167-c1.pdf?SGWID=0-0-45-1608242-p180657462.

Chen, Daizhuo, Samuel P. Fraiberger, Robert Moakler, and Foster Provost. 2017. "Enhancing Transparency and Control When Drawing Data-Driven Inferences About Individuals." *Big Data* 5 (3): 197–212.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency PMLR 81*: 134–48. http://proceedings.mlr.press/v81/chouldechova18a.html.

"CITP/UCHV Workshop: A.I. & Ethics." n.d. University Center for Human Values. Accessed May 2, 2018. https://uchv.princeton.edu/events/citpuchv-workshop-ai-ethics.

Citron, Danielle Keats, and Benjamin Wittes. 2017. "The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity." *Fordam Law Review* 86: 401–423. http://fordhamlawreview.org/wp-content/uploads/2017/10/Citron_November_v86.pdf.

Citron, Danielle Keats, and Frank A. Pasquale. 2014. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89. https://ssrn.com/abstract=2376209.

Clark, Jack. 2018. Presentation at Transformers: Artificial Intelligence. https://www.washingtonpost.com/post-live-2018-transformers-artificial-intelligence/

Crawford, Kate. 2016. "Artificial Intelligence's White Guy Problem." *The New York Times*, June 25, 2016, sec. Opinion. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

Crawford, Kate and Ryan Calo. 2016. "There is a blind spot in AI research." *Nature*, October 13, 2016. http://www.nature.com/news/there-is-a-blind-spot-in-ai-research-1.20805.

Social Media Collective. "Critical Algorithm Studies: A Reading List." n.d. https://socialmediacollective.org/reading-lists/critical-algorithm-studies/.

De Mooy, Michelle. 2018. "The Ethics of Design: Unintended (But Foreseeable) Consequences." Center for Democracy and Technology. January 31, 2018. https://cdt.org/blog/the-ethics-of-design-unintended-but-foreseeable-consequences/.

Diakopoulous, Nicholas. 2016. "Accountability in Algorithmic Decision Making." *Communications of the ACM* 59 (2). https://cacm.acm.org/magazines/2016/2/197421-accountability-in-algorithmic-decision-making/fulltext.

Diakopoulos, Nicholas, and Sorelle Friedler. 2016. "How to Hold Algorithms Accountable." *MIT Technology Review.* November 17, 2016. https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/.

Diakopoulos, Nicholas, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, et al. n.d. "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML." FAT/ML. Accessed May 2, 2018. https://www.fatml.org/resources/principles-for-accountable-algorithms.

Domingos, Pedro. 2012. "A Few Useful Things to Know About Machine Learning." *Communications of the ACM* 55 (10): 78–87. https://doi.org/10.1145/2347736.2347755.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning." *ArXiv:1702.08608*. Last updated March 2, 2017. http://arxiv.org/abs/1702.08608.

Doshi-Velez, Finale, 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." *ArXiv:1711.01134*. Last updated November 21, 2017. http://arxiv.org/abs/1711.01134.

Dwork, Cynthia, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2017. "Decoupled Classifiers for Fair and Efficient Machine Learning." *ArXiv:1707.06613*. July 20, 2017. https://arxiv.org/abs/1707.06613.

Dwork, Cynthia, and Deirdre K. Mulligan. 2013. "It's Not Privacy, and It's Not Fair." *Stanford Law Review*. September 3, 2013. https://www.stanfordlawreview.org/online/privacy-and-big-data-its-not-privacy-and-its-not-fair/.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. "Fairness Through Awareness." *ArXiv:1104.3913*. Last updated November 29, 2011. http://arxiv.org/abs/1104.3913.

Elish, M. C., and danah boyd. 2018. "Situating Methods in the Magic of Big Data and AI." *Communication Monographs* 85 (1): 57–80. https://doi.org/10.1080/03637751.2017.1375130.

Eubanks, Virginia. 2017. *Automating Inequality*. New York: St. Martin's Press.

Executive Office of the President, National Science and Technology Council Committee on Technology. 2016. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. May 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

Executive Office of the President, National Science and Technology Council Committee on Technology. 2016. *Preparing for the Future of Artificial Intelligence.* October 2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

Ford, Martin. 2015. *Rise of the Robots: Technology and the Threat of a Jobless Future*. New York: Basic Books.

Friedler, Sorelle. 2018. "Benchmarks and Reproducibility in Fair ML." *Algorithmic Fairness* (blog). February 19, 2018. https://algorithmicfairness.wordpress.com/2018/02/19/benchmarks-and-reproducibility-in-fair-ml/.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. "On the (Im)Possibility of Fairness." *ArXiv:1609.07236*. September 23, 2016. http://arxiv.org/abs/1609.07236.

Game Changers: Artificial Intelligence Part III, Artificial Intelligence and Public Policy, Hearing Before the Subcommittee on Information Technology. 2018. Statement of Ben Buchanan Postdoctoral Fellow, Belfer Center Cybersecurity Project Harvard University. https://oversight.house.gov/wp-content/uploads/2018/04/Buchanan-Harvard-Statement-AI-III-4-18.pdf.

Game Changers: Artificial Intelligence Part III, Artificial Intelligence and Public Policy, Hearing Before the Subcommittee on Information Technology. 2018. Statement of Jack Clark, Strategy and Communications Director OpenAI. https://oversight.house.gov/wp-content/uploads/2018/04/Clark-OpenAI-Statement-AI-III-4-18.pdf.

Gleicher, Michael. 2016. "A Framework For Considering Comprehensibility In Modeling." *Big Data* 4 (2): 75–88. http://online.liebertpub.com/doi/abs/10.1089/big.2016.0007.

Horvitz, Eric. 2017. "Reflections On the meaningful understanding of the logic of automated decision-making." Presented at Berkeley Center for Law & Technology Privacy Law Forum: Silicon Valley. March 24, 2017. https://www.law.berkeley.edu/wp-content/uploads/2017/03/BCLT_Eric_Horvitz_March_2017.pdf.

Howell, Chuck. "A framework for addressing fairness in consequential machine learning." https://fatconference.org/static/tutorials/howell-framework18.pdf.

Ito, Joi. 2018. "AI Engineers Must Open Their Designs to Democratic Control." American Civil Liberties Union. April 2, 2018. https://www.aclu.org/issues/privacy-technology/surveillance-technologies/ai-engineers-must-open-their-designs-democratic.

Jackman, Molly, and Lauri Kanerva. Evolving the IRB: Building Robust Review for Industry Research,
72 Wash. & Lee L. Rev. Online 442 (2016), http://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/8.

Jones, Meg Leta. 2017. "The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood." Social Studies of Science 47 (2):216–39. https://doi.org/10.1177/0306312717699716.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores."

Knobel, Cory, and Geoffrey C. Bowker. 2011. ''Values in Design.'' Communications of the ACM 54 (7): 26-28.

Kroll, Joshua, Joanna Huey, Solon Barocas, Edward Felten, Joel Reidenberg, David Robinson, and Harlan Yu. 2017. "Accountable Algorithms." University of Pennsylvania Law Review 165 (3):633. http://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.

Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." ArXiv:1606.03490 [Cs, Stat], June. http://arxiv.org/abs/1606.03490.

Levendowski, Amanda. gnovicon 2018.

MacCarthy, Mark. STANDARDS OF FAIRNESS FOR DISPARATE IMPACT ASSESSMENT OF BIG DATA ALGORITHMS. *Cumberland Law Review Vol. 48:1* (forthcoming).

McKinnon, Alex. 2018. "AIs Have Replaced Aliens as Our Greatest World-Destroying Fear." *Quartz*. February 8. https://qz.com/1201846/ais-have-replaced-aliens-as-our-greatest-world-destroying-fear/.

Menlo Report. http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, Luciano Floridi. "The ethics of algorithms: Mapping the debate." *Big Data & Society*, Vol 3, Issue 2, December 2016 at http://journals.sagepub.com/doi/pdf/10.1177/2053951716679679.

Molnar, Christoph. n.d. *Interpretable Machine Learning*. Accessed May 2, 2018. https://christophm.github.io/interpretable-ml-book/.

Naranyanan, Arvind. "Tutorial: 21 fairness definitions and their politics." https://www.youtube.com/watch?v=jIXIuYdnyyk

Ng, Andrew. n.d. "Machine Learning Yearning." Accessed May 2, 2018. http://www.mlyearning.org.

Nissenbaum, Helen. 1996. "Accountability in a Computerized Society." *Science and Engineering Ethics* 2 (1): 25–42. https://doi.org/10.1007/BF02639315.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. n.d. "Ethically Aligned Design, Version 2 (EADv2)." Accessed May 2, 2018. https://ethicsinaction.ieee.org/.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

O'Neil, Cathy. 2016. "Auditing Algorithms." *Mathbabe* (blog). July 22, 2016. https://mathbabe.org/2016/07/22/auditing-algorithms/.

"Partnership on Artificial Intelligence to Benefit People and Society." n.d. Accessed May 2, 2018. https://www.partnershiponai.org/.

Phillips, Richard, Kyu Hyun Chang, Sorelle A. Friedler. 2018. "Interpretable Active Learning." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency PMLR 81*: 49–61. http://proceedings.mlr.press/v81/phillips18a.html.

Pontin, Justin. 2018. "The Limits of Artificial Intelligence and Deep Learning." *Wired*. February 2, 2018. https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning.

"Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML."
 n.d. Accessed January 31, 2018. https://www.fatml.org/resources/principles-for-accountable-algorithms.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. AI Now. https://ainowinstitute.org/aiareport2018.pdf.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *ArXiv:1602.04938*. Last updated August 9, 2016. http://arxiv.org/abs/1602.04938.

Roach, John. 2018. "Aiming to Fill Skill Gaps in AI, Microsoft Makes Training Courses Available to the Public." *The AI Blog* (blog). April 2, 2018. https://blogs.microsoft.com/ai/microsoft-professional-program-ai/.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." Paper presented to "Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA. http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

Selbst, Andrew D., and Julia Powles. 2017. "Meaningful Information and the Right to Explanation."
International Data Privacy Law 7 (4): 233–242,https://doi.org/10.1093/idpl/ipx022

Selbst, Andrew D., and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines." *Fordham Law Review* (forthcoming). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3126971.

Simonite, Tom. 2018. "Some Startups Use Fake Data to Train AI." *Wired*. April 25, 2018. https://www.wired.com/story/some-startups-use-fake-data-to-train-ai/.

Singer, Natasha. 2018. "Tech's Ethical 'Dark Side': Harvard, Stanford and Others Want to Address It." *The New York Times*, February 12, 2018, sec. Business Day. https://www.nytimes.com/2018/02/12/business/computer-science-ethics-courses.html.

Smith, Simson Garfinkel, Jeanna Matthews, Stuart S. Shapiro, Jonathan M. 2017. "Toward Algorithmic Transparency and Accountability." *Communications of the ACM* 60 (9): 5. https://dl.acm.org/citation.cfm?id=3125780.

Solove, Daniel J. 2006. "A Taxonomy of Privacy." *University of Pennsylvania Law Review* 154 (3): 477–560. https://www.law.upenn.edu/journals/lawreview/articles/volume154/issue3/Solove154U.Pa.L.Rev.477(2006).pdf.

Stark, Jennifer A., and Nicholas Diakopoulos. 2017. "Using Baselines for Algorithm Audits." Presented at European Data and Computational Journalism Conference, 2017. http://www.nickdiakopoulos.com/wp-content/uploads/2011/07/EDCJC_Extended_Abstract_ImageBaslines-FINAL.pdf.

"Structuring Machine Learning Projects." n.d. Coursera. Accessed May 2, 2018. https://www.coursera.org/learn/machine-learning-projects/.

Teamcore Group. n.d. University of Southern California. Accessed April 21, 2018. http://teamcore.usc.edu/.

United Nations. 2011. *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*. New York and Geneva: United Nations. http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.

United Nations. n.d. "The UN 'Protect, Respect and Remedy' Framework for Business and Human Rights." https://www.business-humanrights.org/sites/default/files/reports-and-materials/Ruggie-protect-respect-remedy-framework.pdf.

Rieke, Aaron, Miranda Bogen, and David G. Robinson. 2018. "Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods." An Upturn and Omidyar Network Report. https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods.

"Value Sensitive Design." n.d. http://vsdesign.org/

Wattenberg, Martin, Fernanda Viégas, and Moritz Hardt. n.d. "Attack Discrimination with Smarter Machine Learning." https://research.google.com/bigpicture/attacking-discrimination-in-ml/.

Weinberger, David. 2018. "Optimization over Explanation." *Berkman Klein Center Collection* (blog). January 28, 2018. https://medium.com/berkman-klein-center/optimization-over-explanation-41ecb135763d.

Wing, Jeanette M. 2016. "Computational Thinking, 10 Years Later." Microsoft Research. March 23, 2016. https://www.microsoft.com/en-us/research/blog/computational-thinking-10-years-later/.

Wright, Helen. 2018. "Data, Algorithms, and Fairness Panel." *CCC Blog*. January 11, 2018. http://www.cccblog.org/2018/01/11/data-algorithms-and-fairness-panel/.

Zarsky, Tal. 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): 118–32. https://doi.org/10.1177/0162243915605575.

Zarsky, Tal. 2013. "Transparent Predictions." *University of Illinois Law Review* 2013 (4). https://papers.ssrn.com/abstract=2324240.

Ziewitz, Malte. 2016. "Governing Algorithms Myth, Mess, and Methods." *Science, Technology, & Human Values* 41(1):1-16. http://journals.sagepub.com/doi/abs/10.1177/0162243915608948.

Zinkevich, Martin. n.d. "Rules of Machine Learning: Best Practices for ML Engineering."
http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf