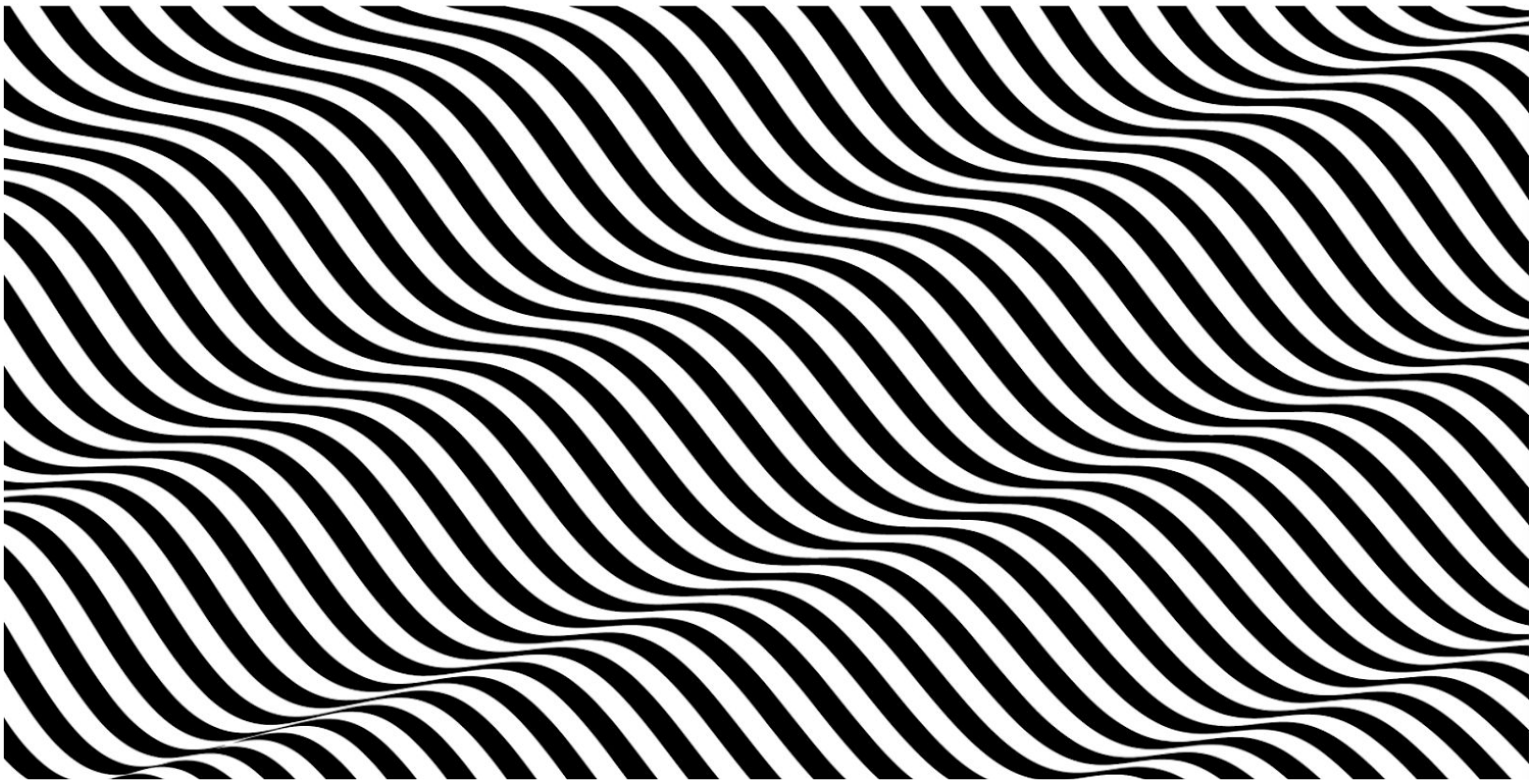
Shea Molloy
Designing Methods for Media.A.Sp20



Diversity of data in machine learning

Crossing research limitations between machine learning projects and sociological critique

Abstract | This proposal outlines the current landscape of ethics in machine learning work including the policy, research and technological work being executed to drive change. While inequitable outcomes from machine learners are universally acknowledged, they are not largely well understood and the fragmented research fails to speak specifically to the technological community. This project proposes in-depth research into the technological processes and outcomes of the last 30 years of artificial intelligence work to lay the groundwork to build new collaborative tools for iterative technological risk assessment. This proposal offers methods of engaging technologists in critical thinking about their work as well as connecting them to the communities affected by it.

Shea Molloy
Designing Methods for Media.A.Sp20

Table of Contents

Shea Molloy
Designing Methods for Media.A.Sp20

Machine Learning and Social Science

With the field as new as it is, there are many definitions and interpretations of machine learning. In Ethem Alpaydin's definition, "Machine Learning is one way to achieve artificial intelligence. By training on a data set, or by repeated trials using reinforced learning, we can have a computer program behaving so as to maximize performance criterion, which in a certain context appears intelligent."[1]. In identifying machine learning as a single facet of artificial intelligence and simplifying the steps to build such projects, Alpaydin suggests that machine learners themselves are neutral and logical. Apaydin insinuation is largely echoed by the technological community, and much of the critique of the effects of machine learners comes retroactively and from outside the community of those who build them. As the development of these projects continues in creativity and scale, the neutrality of these projects comes into questions and we must consider how to incorporate greater sociological critique into the projects themselves.

One of the friction points of applying ethics and sociological research into the field of machine learning so far has been the incompatibility of qualitative research into quantitative practice. Long histories of social science work provide a foundational understanding of the racial, gender, economic discriminations in societal systems that are difficult to condense into mathematical problems. Even the language of "bias" and "discrimination" mean different things to sociological researchers and statisticians. While educational programs in academia and professional industries have begun the work of bridging this understanding in the form of workshops, conferences, and frameworks of open dialogues [2] [3] [4] [5], there is still a need to incorporate ongoing critical thinking into workflows as they are being designed and built. One could argue that there could never be too much nuance or critical dialogue, as the field of technology grows so expansively. There is

[1] Ethem Apaydin, "Machine Learning: The New AI", page 161, 2020
[2] Annelie Berner, The Ethical Stack (beta), 2020
[3] Institute for the Future and Omidyar Network, Ethical OS: A Guide to anticipating the future impact of today's technology, 2018
[4] Daniels et. al "Racial Literacy in Tech", 2020
[5] Joy Buolamwini, The Algorithmic Justice League, 2020

currently a gap in connecting technologists to the outcomes of their work and the communities their work affects. As much of the process of building technology is iterative and agile, frameworks to assess work (and machine learners specifically) should also be iterative and agile, which is somewhat the inverse of traditional sociological and ethical critique.

The research field of ethical applications into technological practice expands into academic theoretical critique as well as critical practice. From theoretical criques, we learn it is universally understood that big data hides systemic biases both when data is recorded and when it is not. By accepting this premise before building with data, we can approach data projects more critically and, to scale, aim for "data with depth rather than big data" [6]. Much literature is dedicated to the acknowledgement of serious flaws in projects that have already been developed. From here, the critical practice methods in the current landscape range from tight, prescriptive frameworks to ongoing educational initiatives. Both of these approaches limit the ongoing critique as it ought to be incorporated into agile technological practice.

In order to best address concerns about equity in machine learners, sociological critiques will have to speak more of the language of technology and insert itself into the process of designing algorithms. Agile software development is characterized by the "division of tasks into short phases of work and frequent reassessment and adaptation of plans" [7] and often describes an entire modern method of building technology. Social scientists will have to condense and reframe the immensity of sociological critique into a form that is digestible by technologists, iterative in nature, and concise in assessment.

[6] Kate Crawford, "The Hidden Biases in Big Data", Harvard Business Review, 2013
[7] Lexico, The Oxford Dictionary, 2020

Tools for Change

There are many such projects already speaking the language of technologists. Tools such as The Ethical Stack offer the integration of critical questions directly into the practice of building smart objects i.e. the "internet of things" (IoT) while utilizing mental models familiar to IoT technologists. Frameworks such as Ethical OS offer programs to critique projects at every level of their development in a managed and transparent way. Most interestingly, The model developed in Suraj Acharya's "Tackling Bias in Machine Learning" builds an adversarial network assessing fairness. His project examines features, targets, and sensitive attributes to predict sensitive labels using a neural network, which is then applied as a classifier on top of a predictive model. His project quantifies the fairness of a predictive model using a fairness metric, determines whether or not the model satisfies a certain threshold value for fairness, and if a model is not fair, de-biases it with an adversarial network. The model was trained on a recidivism risk Dataset from ProPublica, which helped him analyze the trade-off between accuracy and fairness. In this dataset, there was a 6% decrease in accuracy of the classification in exchange for for 200% increase in fairness [8].

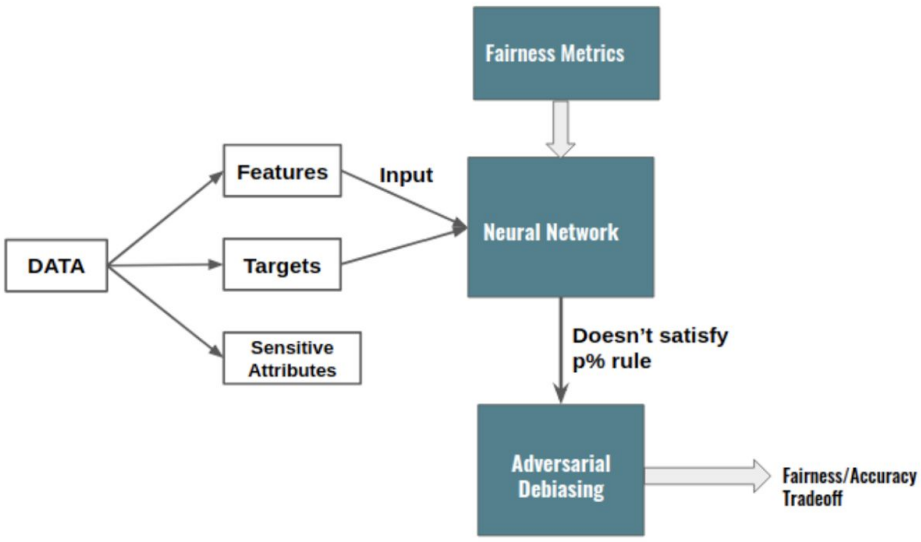


Fig. 1: The Fairness Model developed by Suraj Acharya

[8] Suraj Acharya, "Tackling Bias in Machine Learning", 2019

Acharya's approach is the most effective integration of qualitative and quantitative data because it directly incorporates both schools of thought. He translates the concerns about the predictive models' approaches into digestible metrics and constructs a machine learner that directly applies the knowledge of both analytical and critical thinking. It further puts this knowledge into ethical practice by applying values of transparency and accessibility to the research. Acharya establishes the need for ensuring equity in machine learning models and defines his metrics for addressing that need with the probability variables needed by machine learners. Archarya details his techniques for mitigating bias in the specific language used in the machine learning field, making it easier for technologists to apply those steps to their own work. The source of his training data is also well-known in sociological fields, and addresses the specific concern of racial discinimiation in artificial intelligence research. It is the epitome of the kind of work that needs to be done in ethical technological critique. While this one model is limited in its scale, the transparency in the data used and the rationale behind the model structure make this work easier to build on.

Building More

To build on that procedure in building scalable tools for technologists, I propose a mixed methodology approach to building an interactive assessment tool for machine learners including historical research, community surveys, and quantitative risk assessment. The first phase of this project would involve a deep dive into the historical case studies, resulting in a report on the machine learning projects to date that have produced inequitable results to be learned from. It would include surveys, first-person interviews, archival research of regulatory efforts, and machine learning component assessments. Setting the landscape of inequitable machine learning projects to date would lay the foundation for building tools to address these issues in addition to making machine-learning specific terminology more accessible to social science researchers, and sociological methods of critique more applicable to technologists. This research would result in an open database of machine learning projects and their outcomes, to be maintained by researchers from all backgrounds.

The next phase of this project would include an interactive digital tool to help identify risk assessment areas for "attunement" [9]). This tool would provide input areas for technologists regarding the individual components of their machine learning projects including their models, features, and training and testing data. From these components, the interactive tool would assess risk areas based on multiple frameworks detailing ethical concerns in technological projects [10] [11]. The risk assessment areas called out by the model developed would link technologists to an "action tract", presented as a decision tree with steps to take that would further connect them to communities affected and questions raised by their project.

The effect of highlighting problematic areas in technologists' research in a guiding, rather than prescriptive, way would allow for iterative assessment as projects develop and rewarding

[9] Sareeta Amrute, "of techno-ethics and techno-affects", page 123, 2019
[10] Catherine D'Ignazio and Lauren F. Klein, "Our Values and Our Metrics for Holding Ourselves Accountable", Data Feminism, 2020
[11] Nicol Turner Lee and Paul Resnick, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms", 2019

feedback as projects progress. A tool like this would be best distributed through the communities technologists are already using to answer technological questions, such as Github and Stackoverflow, as well as by communities that provide active support for ethical projects such as the Mozilla Foundation. Making this tool available through Github would also have the benefit of adding to a transparent process and building confidence in the approach. Both the initial research report, the tool construction and the community collaborators would be made available through these distribution methods, lending to the hybridization of qualitative and quantitative research methods.

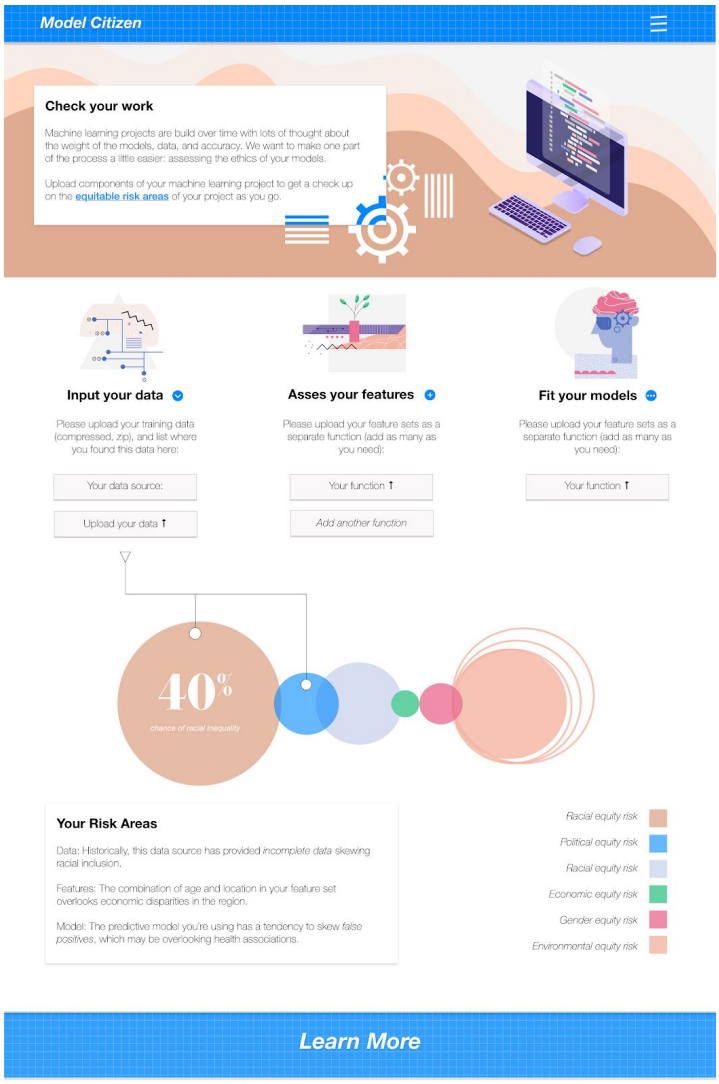


Fig. 2: Model Citizen Prototype Sketch

In Practice

The initial research phase of this project would take about 1 month of dedicated resources to lay a solid foundation for future steps forward. From there, the model and feature construction could take from 3 to 6 months to develop in tandem with the input user interface. Another 3 months would be needed to build out the decision tree and recommendation output, likely using D3. js.

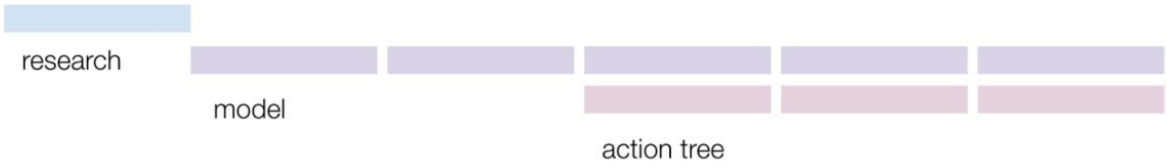


Fig. 3: A project timeline, with proposed overlap

My background in technology, media studies, and design and my research focus areas on gamification and interaction completion, lends well to the need to make this tool functional and accessible. Additional work in content usage and accessible technology curriculums and interaction design also address the needs of this project to span multiple fields and levels of understanding. I believe I can build this work to expand upon the ethical practices taught in the Parson's Data Visualization program, making them more available to researchers and technologists in other programs. All of my professional and personal work has been built to incorporate the values of transparency and collaboration that this research will aim to inspire other technologists to build upon.

Additional ethical concerns are more than welcome to be addressed by outside researchers and communities. In the interest of continued usefulness of this work, I ask that machine learning practitioners bring any potentially overfit recommendations and feedback on the usefulness of the action tracts to the community as a whole through the Git repo. The engagement metrics on the tool would also provide insight into the pain points of the

technologists using this tool as well as their follow up actions. The relevancy of this work will be maintained with continuous research in the field to avoid neglecting future conversations and affected communities by opening this research in real-time. Others working in the field have built upon the practice of intersectional research by adopting opening draft processes and model construction[12] to educate as well as invite feedback. Building a practice of iterative, accessible, intersectional and transparent technology will be crucial for building a more equitable internet, and will need to invite as many questions as it provides solutions in order to fully disrupt current methods.

[12] Erica Kochi, "How to Prevent Discriminatory Outcomes in Machine Learning", 2018

Bibliography

Acharya, Suraj. “Tackling Bias in Machine Learning.” Community. Medium, March 18, 2019. https://blog.insightdatascience.com/tackling-discrimination-in-machine-learning-5c95fde95e95.

Alpaydin, Ethem. *Machine Learning: The New AI*. MIT Press Essential Knowledge Series. Cambridge, MA: MIT Press, 2016. https://ebookcentral-proquest-com.libproxy.newschool.edu/lib/newschool/detail.action?docID=4714219.

Amrute, Sareeta. “Of Techno-Ethics and Techno-Affects.” *Feminist Review*, no. 123 (2019): 56–73. https://doi.org/10.1177/0141778919879744.

Annelie Berner. “The Ethical Stack (Beta).” Interactive tool. The Ethical Stack. Accessed March 30, 2020. https://ethicalstack.virteuproject.eu/stack.html.

Barton, Genie, Nicol Turner Lee, and Paul Resnick. “Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms,” May 22, 2019. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

Betsy Anne Williams, Catherine F. Brooks, and Yotam Shmargad. “How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications.” *Penn State University Press*, Journal of Information Policy, 8 (2018): 78-115 (38 pages). https://doi.org/10.5325/jinfopoli.8.2018.0078 https://www.jstor.org/stable/10.5325/jinfopoli.8.2018.0078.

Bowman, Dr. Greg, Dr. John Chodera, and Dr. Vince Voelz. “Folding @ Home.” Interactive community and project. Accessed March 15, 2020. http://www.foldingathome.org.

Buolamwini, Joy. “The Algorithmic Justice League.” Community. Algorithmic Justice League United, 2020. https://www.ajlunited.org/.

Crawford, Kate. “The Hidden Biases in Big Data.” Though leadership. Harvard Business Review, April 1, 2013. https://hbr.org/2013/04/the-hidden-biases-in-big-data.

Daniels, Dr. Jessie, Mutale Nkonde, and Dr. Darakashan Mir. “Racial Literacy in Tech.” Operations. Data & Society. Accessed April 14, 2020. https://racialliteracy.tech/.

Dietz, Florian. “Why Your AI Might Be Racist and What to Do about It.” Online Community. Towards Data Science, November 9, 2019. https://towardsdatascience.com/why-your-ai-might-be-racist-and-what-to-do-about-it-c081288f600a.

D’Ignazio, Catherine, and Lauren F. Klein. *Our Values and Our Metrics for Holding Ourselves Accountable*. Vol. Data Feminism. Ideas Series. MIT Press, 2020. https://data-feminism.mitpress.mit.edu/pub/3hxh4l8o/release/1.

Elish, M.C, and Tim Hwang. “AI Pattern Language.” *Data & Society*, INTELLIGENCE & AUTONOMY INITIATIVE, n.d. https://www.datasociety.net/pubs/ia/AI_Pattern_Language.pdf.

Gitelman, Lisa, and Rita Raley. *Dataveillance and Counterveillance*. “Raw Data” Is an Oxymoron. Cambridge, MA: MIT Press, 2013.

Haven, Janet. “Data Society.” Research. Data & Society, 2014. https://datasociety.net/.

Institute for the Future and Omidyar Network. “Ethical OS: A Guide to Anticipating the Future Impact of Today’s Technology.” Organization. https://ethicalos.org/, 2018. https://ethicalos.org/.

Kochi, Erica. “How to Prevent Discriminatory Outcomes in Machine Learning.” Community. Medium, April 22, 2018. https://medium.com/@ericakochi/how-to-prevent-discriminatory-outcomes-in-machine-learning-3380ffb4f8b3.

Mackenzie, Adrian. *Machine Learners: Archeology of Data Practice*. The MIT Press, 2017.

Mateescu, Alexandra, and Madeleine Clare Elish. “AI in Context.” Community and research. Data Society, January 30, 2019. https://datasociety.net/library/ai-in-context/.

Milner, Yeshimabeit, and Lucas Mason-Brown. “Data 4 Black Lives.” Community. Data for Black Lives, 2020. http://d4bl.org/.

Noble, Safiya. “Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible.” *InVisible Culture*, no. 19 (October 29, 2013). http://ivc.lib.rochester.edu/google-search-hyper-visibility-as-a-means-of-rendering-black-women-and-girls-invisible/.

Noble, Safiya Umoja, and Brendesha M. Tynes. *The Intersectional Internet*. Digital Formations. United States, Beaverton: Ringgold Inc. Accessed April 15, 2020. https://login.libproxy.newschool.edu/login?url=https://search-proquest-com.libproxy.newschool.edu/docview/1787995263?accountid=12261.