From: Liz O'Sullivan
Chief Executive Officer
Parity Technologies, Inc.
liz@getparity.ai

To: National Institute of Standards and Technology
Attn: Information Technology Laboratory
100 Bureau Drive
Gaithersburg, Maryland 20899-2000
ai-bias@list.nist.gov

September 10, 2021

To whom it may concern:

Feedback on *Draft NIST Special Publication 1270: A Proposal for Identifying and Managing Bias in Artificial Intelligence* (June 2021, doi:10.6028/NIST.SP.1270-draft)

The team at Parity welcome the opportunity to comment on S.P. 1270. We are sympathetic to NIST's efforts toward reducing bias in AI and welcome any and all such efforts towards reducing bias. However, the ambitious goal of a framework wanting to "[manage and reduce] the impacts of harmful biases across contexts" [Lines 227-8] needs to be battle-tested against specific use cases and address industry-specific learnings, as "the associated biases that come with [the use of AI] create harm in context-specific ways" [Lines 256-7]. A broad approach in universalizing this pursuit to all contexts may in fact be less useful than similar guides and frameworks tailored to specific applications of AI. In particular, we want to share how working with the financial services industry has revealed the practical necessities and operational challenges for AI risk management.

There is much to celebrate in this framework. In particular, we celebrate the assertion that some datasets reflecting historical inequities will codify those inequities into the future, and should not be used. S.P. 1270 also acknowledges the structural biases and feedback loops that complicate effective bias management and remediation, although some key research seems to have been overlooked[1]. We also welcome the advice to seek

[1] (a) Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Long-Term Impacts of Fair Machine Learning. *Ergonomics in Design: The Quarterly of Human Factors Applications*, article 106480461988416, October 2019. doi: 10.1177/1064804619884160. (b) Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed Impact of Fair Machine Learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI '19)*, volume 7, pages 6196–6200, California, August 2019. doi: 10.24963/ijcai.2019/862. (c) Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. *Proceedings of the*

outside assistance from "interdisciplinary professionals from the law and social sciences", although this practice is in fact essential to the development of responsible AI[2], not just a "nice to have". This should come as more of a requirement from NIST than a suggestion.

In the rest of this letter, we would like to share our multiple years of experience building fairness-related technology for the enterprise marketplace, constituting lessons learned from bias management in multiple industries, in the hopes of improving the bias management framework proposed.

1. Industry-specific responsible AI regulations already exist.

We are surprised to see little mention of existing regulations in S.P. 1270 from industries like financial services and healthcare. The entirety of financial services is covered in just two citations on Line 317. Rather, we believe that these industries have many instructive lessons for overarching regulations and frameworks[3] such as the one proposed in S.P. 1270; a general purpose framework should at least be able to address the existing context-specific needs on file. S.P. 1270 already alludes to fairness in employment as enforced by the Equal Employment Opportunity Commission. However, we did not find any references to other similar regulations. For example, the healthcare industry has data privacy and security requirements outlined in the Health Insurance Portability and Accountability Act (1996). But the most extensive regulatory needs by far exist in the financial service industry[4]. Large investment banks need to comply with Basel standards such as the Basel Committee on Banking Supervision's Standard No. 239, "Principles for effective risk data aggregation and risk reporting", and must confront responsible AI challenges such as ontology drift[5]. Consumer financial institutions have many of the

---

Conference on Fairness, Accountability, and Transparency (FAT)*, pages 525–534, 2020. doi: 10.1145/3351095.3372878.

[2] (a) Donald Martin, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. Participatory problem formulation for fairer machine learning through community based system dynamics. *ICLR Workshop on Machine Learning in Real Life*, 2020. URL https://arxiv.org/abs/2005.07572. (b) Sina Fazelpour and Zachary C. Lipton. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 57–63, New York, NY, USA, February 2020. doi: 10.1145/3375627.3375828.

[3] (a) Jiahao Chen, Victor Storchan, and Eren Kurshan. Beyond fairness metrics: Roadblocks and challenges for ethical AI in practice. *ACM SIGKDD Workshop on Machine Learning in Finance (KDD-MLF)*, 11 August 2021. URL https://arxiv.org/abs/2108.06217. (b) Jiahao Chen and Victor Storchan. Seven challenges for harmonizing explainability requirements. *ACM SIGKDD Workshop on Machine Learning in Finance (KDD-MLF)*, 11 August 2021. URL https://arxiv.org/abs/2108.05390.

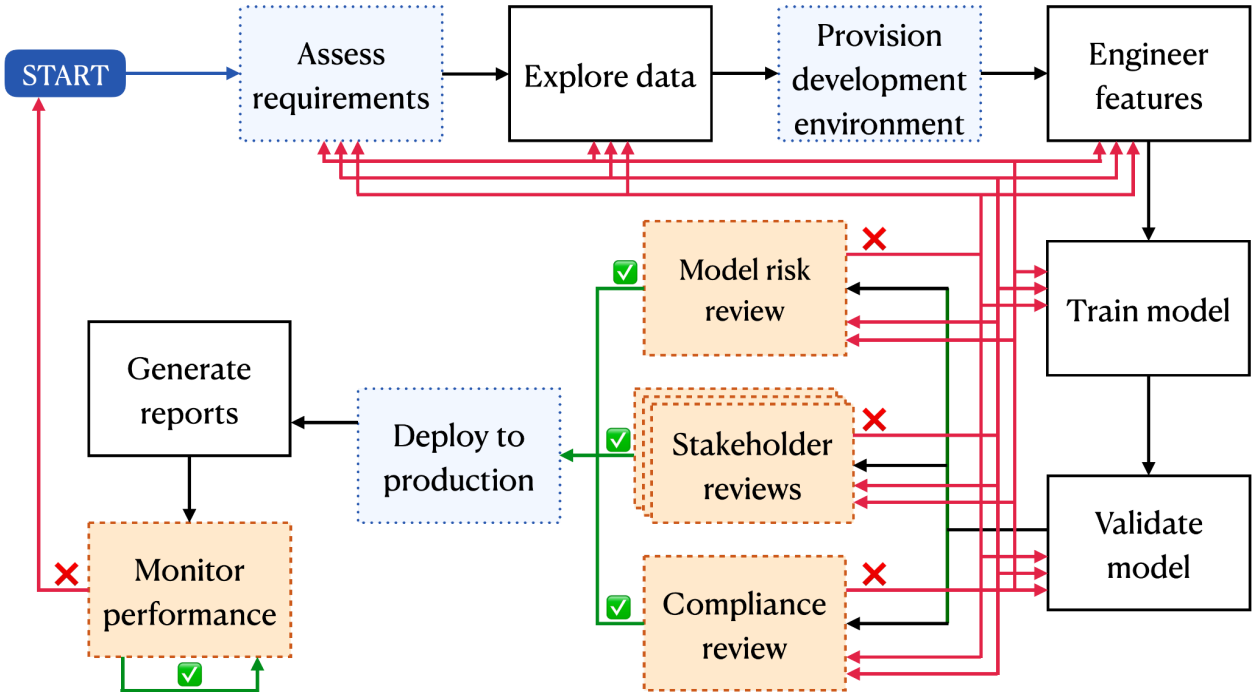
[4] Marc Labonte. Who Regulates Whom? An Overview of the U.S. Financial Regulatory Framework. *Congressional Research Service*, 17 August 2017.

[5] Jiahao Chen. Ontology drift is a challenge for explainable data governance, 2021. URL https://arxiv.org/abs/2108.05401.

longest standing regulatory requirements related to AI bias[6], such as the consumers' right to explanation, rights to appeal incorrect information, and nondiscrimination in credit decisions, which are codified in regulations such as the Fair Housing Act (1968), Fair Credit Reporting Act (1970), Equal Credit Opportunity Act (1974), and protections against unfair, deceptive, or abusive acts and practices (UDAAP) as part of regulations such as the Dodd-Frank Wall Street Reform and Consumer Protection Act (2010).

Of critical relevance for S.R. 1270 is the *Supervisory Guidance on Model Risk Management*, which is an existing model risk management regulation required for AI/ML models in the consumer finance industry, and is enforced by the Office of the Comptroller of Currency (OCC 2011-12), Federal Reserve System (SR 11-7), and the Federal Deposit Insurance Corporation (FIL-22-1017). This regulation already touches many of the key points of S.P. 1270 such as model monitoring [Line 654-655], suitability for purpose [Lines 554-555; 583-584; 601-602]. Furthermore, SR 11-7 lays out a risk management structure that is significantly more complex than that of Figure 1 [Line 415], and instead describes what is known as the three lines of defense (3LOD) for model risk management, representing the different organization stakeholders of business teams/model development teams (first line), model risk management (second line), and audit (third line). The resulting needs for model risk management lead to a significantly more complex risk management process that is shown in the figure below[7]. Crucially, there are multiple stakeholders that necessitate multiple rounds of interaction to resolve conflicting goals., which is absent from Figure 1 of S.P. 1270. A model that fails review by any one line of defence must be re-engineered, resulting in many months of total time to development. A practical model risk management framework must therefore be conscious of the overhead introduced into the model development and deployment process.

[6] Jiahao Chen. Fair lending needs explainable models for responsible recommendation. In *FATREC Workshop on Responsible Recommendation*, 6 October 2018. URL https://arxiv.org/abs/1809.04684.
[7] Eren Kurshan, Hongda Shen, and Jiahao Chen. Towards self-regulating AI: Challenges and opportunities of AI model governance in financial services. In Proceedings of the 1st International Conference on AI in Finance, 15 October 2020. doi: 10.1145/3383455.3422564. URL https://arxiv.org/abs/2010.04827.



2. Demographic labels are often incomplete or absent.

S.P. 1270 does not cover the data quality issues in demographic labels that complicate the detection of bias in practice. In order to measure bias, we first need to know what demographic labels of race, gender, age, etc. (a.k.a. government monitoring information) to attach to each person in the data set. However, such labels are often incomplete or missing, which complicates the actual bias measurement[8] and lead to the persistence of discrimination in practice[9]. In practice, such GMI labels are imputed from publicly available census data using methods like Bayesian Improved Surname Geocoding[10], which introduce their own biases and ecological inference errors. These errors are so large that the wrong sign of discriminatory bias is measured: a model may be measured to be biased

[8] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2017. URL http://arxiv.org/abs/1707.00075.
[9] Marsha Courchane, David Nebhut, and David Nickerson. Lessons Learned: Statistical Techniques and Fair Lending. *Journal of Housing Research*, 11(2):277–295, 2000.
[10] (a) Kevin Fiscella and Allen M Fremont. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41(4p1):1482–1500, 2006. (b) Marc N Elliott, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole Lurie. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83, April 2009. doi: 10.1007/s10742-009-0047-1.

in favor of a disadvantaged minority when the ground truth is the exact opposite[11]. Such findings invalidate the standard approaches to measuring and mitigating biases that exist in the academic literature; instead, a careful quantification of the uncertainty in the bias measurement is necessary to avoid erroneous conclusions[12]. Such controversies are not merely academic, but have in fact led to disputes over the legal authority of regulatory agencies[13], with enormous financial consequences over the legality of assessing hundreds of millions of dollars in regulatory penalties[14].

We expect that situations with missing demographic information will be the norm, not the exception, and strongly urge that uncertainty quantification of bias metrics form an integral component of practical and relevant AI risk management frameworks.

3. Model and data documentation, monitoring, and testing must be integral parts of AI bias risk management

S.R. 1270 currently does not mention documentation requirements for model[15] and data[16], which already exist in financial regulations like SR 11-7 and BCBS 239. Such documentation is a crucial part of regulatory oversight in financial services, and brings benefits to every use of AI in industry beyond finance. In general, excellent documentation practices are crucial to translate design choices into plain English, where business stakeholders, legal experts, social scientists, and the algorithms' consumers can

---

[11] (a) Yan Zhang. Assessing fair lending risks using race/ethnicity proxies. *Management Science*, 64(1):178–197, January 2016. doi:10.1287/mnsc.2016.2579. (b) Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 30 January 2019. doi: 10.1145/3287560.3287594.

[12] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. *Management Science*, May 2021. doi: 10.1287/mnsc.2020.3850.

[13] Kevin M McDonald. Who's policing the financial cop on the beat? A call for judicial review of the Consumer FInancial Protection Bureau's non-legislative rules. *Review of Banking & Financial Law*, 35(1):224–271, 2016. URL http://ssrn.com/abstract=2786093.

[14] (a) Annie Nova. Congress eases rules against racial discrimination in the auto loan market. In *CNBC News*, 9 May 2018. URL https://www.cnbc.com/2018/05/09/congress-eases-rules-against-racial-discrimination-in-the-auto-loan-market.html. (b) Talia B. Gillis. False Dreams of Algorithmic Fairness: The Case of Credit Pricing. *SSRN Electronic Journal*, 2020. doi: 10.2139/ssrn.3571266.

[15] https://arxiv.org/abs/1810.03993 is a good start for model documentation, but specific industries or applications will have additional nuances. Citation: Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 29 January 2019. doi: 10.1145/3287560.3287596. URL http://dx.doi.org/10.1145/3287560.3287596.

[16] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2020. URL https://arxiv.org/abs/1803.09010.

understand clearly what limitations and intents are associated with the model and the data[17]. Good governance of AI requires such levels of explainability and transparency above and beyond the purely technical feature-based explanations like LIME or SHAP. The best way to mitigate bias in AI is to have a multistakeholder approach to its design, application, and ongoing governance. This will never be possible unless we can adopt shared documentation and translation frameworks to facilitate these conversations, and share knowledge between people of varied backgrounds.

We agree wholeheartedly that "selecting models based solely on accuracy is not necessarily the best approach for bias reduction" [Lines 511-512]. Notions of accuracy and bias reduction, even when they can be quantified, encompass several metrics[18] which can codify the operational needs of multiple stakeholders, which are not necessarily in alignment[19]. As a result, choosing which metrics to prioritize is not trivial, but rather one of the most important operational decisions behind the design and deployment of AI systems in practice.

Research also shows that proper data management is essential to remediating bias in practice[20]. Causally-informed methods such as counterfactual fairness [Lines 655-656] cannot work in practice without full knowledge of the underlying data generating process

---

[17] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality. *Proceedings of the ACM Conference on Human-Computer Interaction*, April 2021. DOI:https://doi.org/10.1145/3449081

[18] There are 50 performance metrics within the sklearn.metrics module of the well-used scikit-learn package (v0.24.2; April 2021): 22 metrics for classification, 11 for regression, 3 for multilabel ranking, 13 for clustering, and 1 for biclustering. Citations (a) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. (b) Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[19] (a) Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, June 2017. doi: 10.1089/big.2016.0047. (b) Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. FACT: A Diagnostic for Group Fairness Trade-offs. *Proceedings of Machine Learning Research*, 119:5264–5274, 2020. URL http://proceedings.mlr.press/v119/kim20a.html.

[20] (a) Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions. In *Proceedings of the 23nd International Conference on Extending Database Technology*, November 2019. URL http://arxiv.org/abs/1911.12587. (b) Julia Stoyanovich, Bill Howe, and H. V. Jagadish. Responsible data management. *Proceedings of the VLDB Endowment*, 13(12):3474–3488, 2020. doi: 10.14778/3415478.3415570.

and how the data is processed by the AI system[21]. Data and model documentation therefore constitute an essential component of AI model risk management, by requiring developers to understand the structure of the data being used and how the data are transformed within an AI system.

We also caution against adopting a standard workflow to the model development workflow, and advocate instead for contextual awareness in the selection of design criteria. For example, the industry standard practice of 80/20 split for train and test is often woefully inadequate to prevent bias in domain transfer at the moment of transition from lab to production. Data leakage issues can occur in many ways, and proper documentation of such potential risks is critical to preventing artificially inflated metrics and claims of performance that don't hold up in practice.

The pernicious notion that "accuracy" is the only currently worthwhile metric in AI is unfortunately persistent in marketing materials, notably by snake-oil vendors making unreasonable claims of product efficacy, as we saw with the Clearview CEO Han Ton That's outrageous claim that his facial recognition technology boasts "100% accuracy"[22]. In our experience creating models for paying clients, proper selection of performance and bias metrics, and the testing and monitoring of these metrics in production, are key systems design criteria that must be properly documented.

In fact, the need for guidance on testing is so crucial that we argue this concept warrants its own section in the AI lifecycle mapping, as does the question of "problem formulation",

---

[21] (a) Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2nd edition, 2009. (b) Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding Discrimination through Causal Reasoning. *Advances in Neural Information Processing Systems*, 30:656–666, 2017. URL http://arxiv.org/abs/1706.02744 (c) Junzhe Zhang and Elias Bareinboim. Fairness in decision-making the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, volume 32, pages 2037–2045, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11564. (d) Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Section 4, pages 582–593, January 2020. doi:10.1145/3351095.3372851 (e) Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. Fairness under feature exemptions: counterfactual and observational measures. *IEEE Transactions on Information Theory*, 2021. doi: 10.1109/TIT.2021.3103206.

[22](a) Donie O'Sullivan, Richa Naik, and John General. This man says he's stockpiling billions of our photos. CNN Business, 10 February 2020. URL https://www.cnn.com/2020/02/10/tech/clearview-ai-ceo-hoan-ton-that/index.html. (b) Caroline Haskins, Ryan Mac, and Logan McDonald. The aclu slammed a facial recognition company that scrapes photos from instagram and facebook. *BuzzFeed News*, 10 February 2020. URL https://www.buzzfeednews.com/article/carolinehaskins1/clearview-ai-facial-recognition-accurate-aclu-absurd (c) American Civil Liberties Union. ACLU sues Clearview AI, 28 May 2020. URL https://www.aclu.org/press-releases/aclu-sues-clearview-ai.

mentioned a few times in the text, but a massive source of bias that is difficult to quantify. We are glad to see specific mention of attempts to predict human qualities "that are only partially observable or capturable by data". However, by now there are more specifics that can be called out for applications of AI that are inherently biased towards the opinions of the modeller. To name a few: that cameras can infer internal qualities about humans, and that enhanced surveillance will yield greater safety for society. There is a mountain of research to oppose either of these claims' validity, and we would welcome more detail and specific attention to this problem.

When the goal of testing is to ensure a model's generalizability to the real world, special attention must be paid to the edge cases and undersampled minorities to ensure that their scenarios are represented in the test cases. While it may seem counterintuitive, we find that specially curated datasets for and testing and mitigation[23] should represent more than simply the unaltered distribution of scenarios one might find in the majority. When creating this partially curated data set, another kind of bias surfaces. This is the question of "what to test for": the decision of whose outcomes are of sufficient importance to warrant inclusion. White developers may inadvertently "forget" to include Black test subjects as we saw in Gender Shades[24], or an overly broad application of the label "Asian" may ensure that Pacific Islanders, Southeast Asians, and mainland Chinese individuals are lumped together even though their behaviors may deviate significantly from each other[25].

4. Bias mitigation techniques must be documented and tested for practical effect

S.R. 1270 does not currently discuss the need for validating the efficacy of bias mitigation techniques, which we have seen are fragile in practice and without careful monitoring, may give rise to unfounded hopes of having remediated unfairness. In our years of experience building fairness-related technology for the enterprise marketplace, the prevailing question that lingers, which is currently unanswered by vendors and the

[23]OpenAI found that fine-tuning iteratively on a curated dataset helped mitigate bias in GPT-3. Citation: (a) Irene Solaiman and Christy Dennison. Process for adapting language models to society (PALMS) with values-targeted datasets, 18 June 2021. URL https://arxiv.org/abs/2106.10328.
[24] Reference 24 of S.P. 1270
[25] (a) Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. *Proceedings of Machine Learning Research* 80, (2018), 2564–2572. URL http://proceedings.mlr.press/v80/kearns18a.html (b) Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax Pareto fairness: a multi-objective perspective. *Proceedings of Machine Learning Research* 119, (2020), 6755–6764. URL http://proceedings.mlr.press/v119/martinez20a.html (c) Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. 2021. Fairness Through Robustness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pp. 466–477. DOI:https://doi.org/10.1145/3442188.3445910 (d) Mark Weber, Mikhail Yurochkin, Sherif Botros, and Vanio Markov. 2020. Black Loans Matter: Distributionally Robust Fairness for Fighting Subgroup Discrimination. *NeurIPS Workshop on Fair AI in Finance*. (2020). URL http://arxiv.org/abs/2012.01193

government, is one of *what to do* when bias is found, rather than "how to find bias". In fact, the bias mitigations strategies popularized in toolkits such as IBM Fairness 360, Aequitas, Microsoft FairLearn, AWS Sagemaker Clarify often do not work in practice due to overfitting[26] or because they are geared only towards binary classification problems. These are serious limitations, and any purported claims of having "fixed" bias must be carefully validated in practice. Current research is not sufficiently advanced to provide guidance that will work for every application of AI in industry. Therefore, until the science of bias mitigation is sufficiently advanced, we have to rely on the best practices within specific vertical applications that have emerged for finding and mitigating disparities among protected classes. The "what to do" is the proverbial "holy grail" of bias mitigation, and something we hope to ultimately bring to life in our platform.

Thank you for your time and attention,

The Parity Team

[26] (a) Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pages 329–338, New York, New York, USA, 2019. doi: 10.1145/3287560.3287589. (b) Ashrya Agrawal, Florian Pfisterer, Bernd Bischl, Francois Buet-Golfouse, Srijan Sood, Jiahao Chen, Sameena Shah, and Sebastian Vollmer. Debiasing classifiers: is reality at variance with expectation?, 2021. URL https://arxiv.org/abs/2011.02407.