From: Doshi-Velez, Finale <finale@seas.harvard.edu>
Sent: Thursday, July 01, 2021 9:50 PM
To: Pearson, Nikita
Cc: Decker, Debra A.
Subject: [EXTERNAL MESSAGE] RIN 3064-ZA24 - Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning [FR Doc. 2021–06607 Filed 3–30–21; 8:45 am]
Attachments: pastedImagebase640.png; Biblio Isaac Lage.docx; Biblio Finale Doshi-Velez.docx; Biblio Sarah Rathnam.docx; Biblio Weiwei Pan.docx; Cover Letter.pdf; 2021-07-01 DtAK Response to the Agencies on the RFI.pdf; 2021-07-01 DtAK Response to the Agencies on the RFI.docx; 2021-05-06 DNP Overview.pptx

Dear Director Pearson,

Thank you for the opportunity to submit comments to the Request for Information ('RFI') on Financial Institutions' Use of Artificial Intelligence, including Machine Learning (RIN 3064-ZA24) signed by Assistant Executive Secretary Sheesley on behalf of the Federal Deposit Insurance Corporation ('Corporation') as part of the collective agencies to the RFI.

Since its establishment in 2011,[1] the Office of Minority and Women Inclusion ('OMWI') of the Corporation and more recently with (a) the FDIC Diversity, Equity, and Inclusion Strategic Plan (2021-2023) and (b) the yearly data reports to Congress as part of the No Fear Act;[2] seems well positioned to support the **Mission-Driven Bank Fund's support of MDIs and CDFIs**. It seems like a **great opportunity to create data donation framework** for individual-level anonymized financial data donations for research to ensure accountability while measuring and monitoring systemic issues.

More broadly, DtAK commends the work of all the Agencies in proactively pursuing a diversity of viewpoints. We believe this multi-stakeholder process towards comprehensive AI regulation, which brings together key stakeholders – including academia – serves as a strong foundation for OMWI and the FDIC more broadly to lead Agencies to carry out comprehensive efforts to oversee the financial sector realize the potential of artificial intelligence while identifying and managing risks.

Specifically, to RFI RIN 3064-ZA24, we suggest that the current regulatory framework under review could benefit from a more practical definition of explainability, while the FDIC could use recent research to better define standards for the continuous monitoring of AI. We need a way of having an AI "Check Engine" light.

The work herein does not reflect the official or unofficial viewpoints of Harvard University or its Harvard John A. Paulson School of Engineering and Applied Sciences ('SEAS') and are submitted as part of a personal effort to support regulatory leadership with insights from our current research relating to accountability in AI for healthcare.

Respectfully submitted,

Finale Doshi-Velez

Finale Doshi-Velez (she/her/hers)

Gordon MacKay Full Professor of Engineering and Applied Sciences

Harvard's Data to Actionable Knowledge Lab
Cambridge, MA 02138
O: +1 (617) 495-3188 attn. Ms. Annalee S. Mendez
E-mail: finale@seas.harvard.edu Web: finale.seas.harvard.edu

[1] Pursuant to the Dodd-Frank Wall Street Reform and Consumer Protection Act, Section 342
[2] An example of the Corporation's own efforts to ensure accountability for anti-discrimination

Harvard John A. Paulson
School of Engineering
and Applied Sciences

SEAS, 29 Oxford Street, Cambridge, MA 02138
O: 617 495 3188
Finale@seas.harvard.edu
Finale.seas.harvard.edu

Finale Doshi-Velez
Gordon MacKay Full Professor of Engineering
and Applied Sciences

July 1st, 2021

To Whom It May Concern:

The Data to Actionable Knowledge ("DtAK") Lab appreciates the opportunity to provide feedback on the Agencies' request for information ("RFI") concerning the Financial Institutions' Use of Artificial Intelligence, including Machine Learning.

DtAK commends the work of the Agencies in proactively pursuing a diversity of viewpoints. We believe this multi-stakeholder process towards comprehensive A.i. regulation, which brings together key stakeholders – including academia – serves as a strong foundation for the Agencies to carry out their efforts to oversee the financial sector realize Artificial Intelligence's potential while identifying and managing risks.

Specifically, we suggest that the current regulatory framework under review could benefit from a more practical definition of explainability, while the FDIC could use recent research to better define standards for the continuous monitoring of AI. We need a way of having an AI "Check Engine" light.

The work herein does not reflect the official or unofficial viewpoints of Harvard University or its Harvard John A. Paulson School of Engineering and Applied Sciences ('SEAS') and are submitted as part of a personal effort to support regulatory leadership with insights from our current research relating to accountability in AI and healthcare.

Respectfully submitted,

[redacted]

Finale Doshi-Velez (she/her/hers)
Gordon MacKay Full Professor of Engineering and Applied Sciences

Harvard John A. Paulson
School of Engineering
and Applied Sciences

Harvard's Data to Actionable Knowledge lab
Cambridge, MA 02138
O: +1 (617) 495-3188 attn. Ms. Annalee S. Mendez
E-mail: finale@seas.harvard.edu Web: finale.seas.harvard.edu

Bibliography: Professor Finale Doshi-Velez

Amir, O., Doshi-Velez, F., & Sarne, D. (9 de July de 2018). Agent strategy summarization. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1203-1207. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6255044771133571112

Amir, O., Doshi-Velez, F., & Sarne, D. (1 de September de 2019). Summarizing agent strategies. *Autonomous Agents and Multi-Agent Systems, 33*(5), 628-644. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17814175098339224659

Andrew Slavin Ross, F. D.-V. (2021). *Benchmarks, Algorithms, and Metrics for Hierarchical Disentanglement.* arXiv. Fonte: https://arxiv.org/abs/2102.05185

Antorán, J., Yao, J., Pan, W., Hernández-Lobato, J. M., & Doshi-Velez, F. (17 de July de 2020). Amortised Variational Inference for Hierarchical Mixture Models. *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*, 1-11. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=486508101802156030

Ayanian, N. (1 de May de 2013). AI's 10 to Watch. *IEEE Intelligent Systems, 28*(3), 86-96. Fonte: https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/abs/10.1109/MIS.2013.57

Bertino, E., Doshi-Velez, F., Gini, M., Lopresti, D., & Parkes, D. (2020). *Artificial Intelligence & Cooperation.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=7783960080507753854

Blum, J. E., Kreienkamp, A. B., Doshi-Velez, F., & Radhakrishnan, M. L. (6 de March de 2014). Feature-based calculation of the electrostatic component of the free energy of protein binding. *Abstracts of Papers of the American Chemical Society, 247*. Fonte: https://scholar.google.com/scholar?hl=en&as_sdt=0,22&cluster=1788206292747178076

Buller, M., Cuddihy, P., Davis, E., Doherty, P., Doshi-Velez, F., Erdem, E., . . . Urken, A. B. (31 de October de 2011). Reports of the AAAI 2011 spring symposia. *Ai Magazine, 32*(2), 119-127. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13881558188560418801

Coker, B., Pan, W., & Doshi-Velez, F. (2021). *Wide Mean-Field Variational Bayesian Neural Networks Ignore the Data.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13049947611571259225

Coker, B., Pradier, M. F., & Doshi-Velez, F. (12 de December de 2019). *Towards Expressive Priors for Bayesian Neural Networks: Poisson Process Radial Basis Function Networks.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2119796613779441132

Coker, B., Pradier, M. F., & Doshi-Velez, F. (27 de August de 2020). PoRB-Nets: Poisson Process Radial Basis Function Networks. *Conference on Uncertainty in Artificial Intelligence, 124*, 1338-1347. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8376392909947316183

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (23 de May de 2016). Learning and policy search in stochastic dynamical systems with bayesian neural networks. *arXiv preprint*

arXiv:1605.07127, 1-14. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14568373457880481181

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (November de 2017). *Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4564853263001192019

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2017). *Uncertainty decomposition in bayesian neural networks with latent variables.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=18035330188583278265

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., & Udluft, S. (3 de July de 2018). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. *Proceedings of the 35th International Conference on Machine Learning, 80*, 1184-1193. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13563599882871713230

Doshi, F., & Gael, J. V. (s.d.). *Discovering Software Bugs with Bayesian NonParametric Models.* Poster. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13854909949350740587

Doshi-Velez, F. (2005). *6.867 Final Project: Applying Online Cotraining to a Vision-Based Hand Detection System.* Course Paper, Massachusetts Institute of Technology. Fonte: https://scholar.google.com/scholar?cluster=748817442671281472 4&hl=en&oi=scholarr

Doshi-Velez, F. (2008). *Efficient Inference in the Indian Buffet Process.* Fonte: https://scholar.google.com/scholar?cluster=10817723646322009682&hl=en&oi=scholarr

Doshi-Velez, F. (2009). *The Indian buffet process: Scalable inference and extensions.* University of Cambridge. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9215175749629133787

Doshi-Velez, F. (7 de December de 2009). The infinite partially observable Markov decision process. *Proceedings of the 22nd International Conference on Neural Information Processing Systems, 22*, 477-485. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10768110427383167189

Doshi-Velez, F. (5 de July de 2010). Nonparametric Bayesian Approaches for Reinforcement Learning in Partially Observable Domains. *Fifteenth AAAI/SIGART Doctoral Consortium*, 1976-1977. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5963548893230433483

Doshi-Velez, F. (2 de December de 2011). *To Infinity and Beyond… or… Bayesian Nonparametric Approaches for Reinforcement Learning in Partially Observable Domains.* Presentation Slides, Massachusetts Institute of Technology. Fonte: http://people.csail.mit.edu/finale/presentations/finale_defense_slides.pdf

Doshi-Velez, F. (2012). *Bayesian nonparametric approaches for reinforcement learning in partially observable domains.* Massachusetts Institute of Technology. MIT Press. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=848110038430606651

Doshi-Velez, F. (1 de November de 2017). From Electronic Health Records to Treatment Recommendations for Depression. *Neuropsychopharmacology, 42*, S100-S101. Fonte: https://scholar.google.com/scholar?cluster=11551551496929767776&hl=en&oi=scholarr

Doshi-Velez, F. (s.d.). *Model Learning for Dialog Management.* Abstract. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17711957998703251700

Doshi-Velez, F., & Ghahramani, Z. (14 de June de 2009). Accelerated sampling for the Indian buffet process. *Proceedings of the 26th annual international conference on machine learning*, 273-280. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15310891466322889089

Doshi-Velez, F., & Ghahramani, Z. (18 de June de 2009). Correlated non-parametric latent feature models. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 143–150. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=7094049166815964911

Doshi-Velez, F., & Ghahramani, Z. (July de 2011). A comparison of human and agent reinforcement learning in partially observable domains. *Proceedings of the Annual Meeting of the Cognitive Science Society, 33*(33), 2703-2708. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6703778567975104250

Doshi-Velez, F., & Kim, B. (March de 2017). *A roadmap for a rigorous science of interpretability.* arXiv. Fonte: https://arxiv.org/pdf/1702.08608.pdf

Doshi-Velez, F., & Kim, B. (28 de February de 2017). *Towards A Rigorous Science of Interpretable Machine Learning.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8789025022351052485

Doshi-Velez, F., & Kim, B. (2018). Considerations for evaluation and generalization in interpretable machine learning. Em H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. v. Gerven, *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 3-17). Springer Nature. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=12095494514140397496

Doshi-Velez, F., & Konidaris, G. (8 de December de 2012). Transfer learning by discovering latent task parametrizations. *NIPS 2012 Workshop on Bayesian Nonparametric Models for Reliable Planning and Decision-Making under Uncertainty*, 1-8. Fonte: https://scholar.google.com/scholar?start=10&hl=en&as_sdt=0,22&cluster=13886646994308848456

Doshi-Velez, F., & Konidaris, G. (9 de July de 2016). Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 1432–1440. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3270924689080010306

Doshi-Velez, F., & Palakodety, R. (2006). *Group 1: 6.375 Final Project Runahead Processor.* Course Paper, Massachusetts Institute of Technology. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5565255244945064225

Doshi-Velez, F., & Perlis, R. H. (12 de November de 2019). Evaluating machine learning articles. *Journal of the American Medical Association, 322*(18), 1777-1779. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15452153815990615586

Doshi-Velez, F., & Roy, N. (10 de March de 2007). Efficient model learning for dialog management. *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, 65-72. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5759348188893787326

Doshi-Velez, F., & Roy, N. (22 de July de 2007). Learning User Models with Limited Reinforcement: An Adaptive Human-Robot Interaction System. *Symposium on Language and Robotics*, 1-10. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=399550348607873317

Doshi-Velez, F., & Roy, N. (1 de December de 2008). Spoken language interaction with model uncertainty: an adaptive human–robot interaction system. *Connection Science, 20*(4), 299-318. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17586990007894260810

Doshi-Velez, F., & Roy, N. (12 de May de 2008). The permutable POMDP: fast solutions to POMDPs for preference elicitation. *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, 1*, 493–500. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13547905812581203405

Doshi-Velez, F., & Roy, N. (16 de December de 2011). An Analysis of Activity Changes in MS Patients: A Case Study in the Use of Bayesian Nonparametrics. *Workshop: Bayesian Nonparametrics, Hope or Hype?*, 1-2. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3722893354574273931

Doshi-Velez, F., & Williamson, S. A. (1 de September de 2017). Restricted Indian buffet processes. *Statistics and Computing, 27*(5), 1205-1223. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14673419880409111956

Doshi-Velez, F., Avillach, P., Palmer, N., Bousvaros, A., Ge, Y., Fox, K., . . . Kohane, I. (1 de October de 2015). Prevalence of inflammatory bowel disease among patients with autism spectrum disorders. *Inflammatory bowel diseases, 21*(10), 2281–2288. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9787709197227402824

Doshi-Velez, F., Brunskill, E., Shkolnik, A., Kollar, T., Rohanimanesh, K., Tedrake, R., & Roy, N. (October de 2007). Collision detection in legged locomotion using supervised learning. *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 317-322. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4414783219246469999

Doshi-Velez, F., Budish, R., & Kortz, M. (2017). *The Role of Explanation in Algorithmic Trust.* Harvard, Berkman Klein Center for Internet & Society. Artificial Intelligence and Interpretability Working Group. Fonte: https://bitlab.cas.msu.edu/trustworthy-algorithms/whitepapers/Finale%20Doshi-Velez.pdf

Doshi-Velez, F., Ge, Y., & Kohane, I. (January de 2014). Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics, 133*(1), e54-e63. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8952823131498776530

Doshi-Velez, F., Knowles, D. A., Mohamed, S., & Ghahramani, Z. (7 de December de 2009). Large scale nonparametric Bayesian inference: Data parallelisation in the Indian buffet process. *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 1294-1302. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8022303325169102009

Doshi-Velez, F., Kortz, M., Budish, R., Chris Bavitz, S. G., O'Brien, D., Scott, K., . . . Wood, A. (3 de November de 2017). *Accountability of AI under the law: The role of explanation.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13535939933778439444

Doshi-Velez, F., Li, W., Battat, Y., Charrow, B., Curthis, D., Park, J.-g., . . . Teller, S. (1 de July de 2012). Improving safety and operational efficiency in residential care settings with WiFi-based localization. *Journal of the American Medical Directors Association, 13*(6), 558-563. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16620197873028630726

Doshi-Velez, F., Miller, K., Gael, J. V., & Teh, Y. W. (15 de April de 2009). Variational inference for the Indian buffet process. *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, 5*, 137-144. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=12982039394924101433

Doshi-Velez, F., Pfau, D., Wood, F., & Roy, N. (1 de October de 2013). Bayesian nonparametric methods for partially-observable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(2), 394-407. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9789702462404251311

Doshi-Velez, F., Pineau, J., & Roy, N. (5 de July de 2008). Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs. *Proceedings of the 25th international conference on Machine learning*, 256-263. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4585349008724983456

Doshi-Velez, F., Wallace, B., & Adams, R. (25 de January de 2015). Graph-sparse lda: a topic model with structured sparsity. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2575–2581. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16621623251555081538

Doshi-Velez, F., Wingate, D., Roy, N., & Tenenbaum, J. (6 de December de 2010). Nonparametric Bayesian policy priors for reinforcement learning. *Proceedings of the 23rd International Conference on Neural Information Processing Systems, 1*, 532–540. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1163119569248222189

Doshi-Velez, F., Wingate, D., Tenenbaum, J. B., & Roy, N. (28 de June de 2011). Infinite dynamic Bayesian networks. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 913–920. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6524581654986168933

Downs, M., Chu, J. L., Yacoby, Y., Doshi-Velez, F., & Pan, W. (12 de July de 2020). CRUDS: Counterfactual Recourse Using Disentangled Subspaces. *ICML Workshop on Human Interpretability in Machine Learning*, 1-23. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14822408888462715234

Du, J., Futoma, J., & Doshi-Velez, F. (2020). *Model-based reinforcement learning for semi-markov decision processes with neural odes.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6882030783154485592

Du, J., Ross, A. S., Shavit, Y., & Doshi-Velez, F. (13 de December de 2019). Controlled Direct Effect Priors for Bayesian Neural Networks. *4th workshop on Bayesian Deep Learning*, 1-8. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17307132700561100484

Elibol, H. M., Nguyen, V., Linderman, S., Johnson, M., Hashmi, A., & Doshi-Velez, F. (1 de January de 2016). Cross-corpora unsupervised learning of trajectories in autism spectrum disorders. *The Journal of Machine Learning Research, 17*(1), 4597–4634. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16607715928057211631

Engelhardt, B. E., Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M., Li, K., & Doshi-Velez, F. (5 de January de 2021). *The importance of modeling patient state in reinforcement learning for precision medicine.* AI4Health. Fonte: https://ai4healthschool.org/wp-content/uploads/2021/01/AI4Health2020-bee.pdf

Fan, A., Doshi-Velez, F., & Miratrix, L. (2017). *Prior matters: simple and general methods for evaluating and improving topic quality in topic modeling.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=12889977940981975342

Fan, A., Doshi-Velez, F., & Miratrix, L. (2017). *Promoting domain-specific terms in topic models with informative priors.* arXiv. Fonte: https://www.semanticscholar.org/paper/Promoting-Domain-Specific-Terms-in-Topic-Models-Fan-Doshi-Velez/7f992c8ea80b7ee9640d67be92f377ee11cd01a1

Fan, A., Doshi-Velez, F., & Miratrix, L. (June de 2019). Assessing topic model relevance: Evaluation and informative priors. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 12*(3), 210-222. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4628884776712761559

Futoma, J., Hughes, M. C., & Doshi-Velez, F. (2 de December de 2018). Prediction-constrained POMDPs. *32nd Conference on Neural Information Processing Systems (NIPS)*, 1-8. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8362747442878248547

Futoma, J., Hughes, M. C., & Doshi-Velez, F. (August de 2020). Popcorn: Partially observed prediction constrained reinforcement learning. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 108*, 3578-3588. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2544924681479461357

Futoma, J., Masood, M. A., & Doshi-Velez, F. (30 de May de 2020). Identifying distinct, effective treatments for acute hypotension with soda-rl: Safely optimized diverse accurate reinforcement learning. *AMIA Summits on Translational Science Proceedings, 2020*, 181–190. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15014924363061995529

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (September de 2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital*

Health, 2(9), e489-e492. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17128416078609966243

Gafford, J., Doshi-Velez, F., Wood, R., & Walsh, C. (1 de September de 2016). Machine learning approaches to environmental disturbance rejection in multi-axis optoelectronic force sensors. *Sensors and Actuators A: Physical, 248*, 78-87. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1497978661435603804

Geramifard, A., Doshi-Velez, F., Redding, J., Roy, N., & How, J. P. (28 de June de 2011). Online discovery of feature dependencies. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 881-888. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5757284957182508738

Geramifard, A., Doshi-Velez, F., Redding, J., Roy, N., & How, J. P. (28 de June de 2011). *Online Discovery of Feature Dependencies Technical Report.* International Conference on Machine Learning. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15066633202434576157

Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., & Szolovits, P. (24 de August de 2014). Unfolding physiological state: Mortality modelling in intensive care units. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 75-84. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5798843096311556054

Ghassemi, M., Wu, M., Hughes, M. C., Szolovits, P., & Doshi-Velez, F. (26 de July de 2017). Predicting intervention onset in the ICU with switching state space models. *AMIA Summits on Translational Science Proceedings, 2017*, 82-91. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9563364752254266911

Ghosh, S., & Doshi-Velez, F. (2017). *Model selection in Bayesian neural networks via horseshoe priors.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9439867866709144171

Ghosh, S., Yao, J., & Doshi-Velez, F. (10 de July de 2018). Structured variational learning of Bayesian neural networks with horseshoe priors. *Proceedings of the 35th International Conference on Machine Learning, 80*, 1744-1753. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9128418694635359827

Ghosh, S., Yao, J., & Doshi-Velez, F. (2019). Model selection in Bayesian neural networks via horseshoe priors. *Journal of Machine Learning Research, 20*(182), 1-46. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13995759821761819294

Glueck, M., Naeini, M. P., Doshi-Velez, F., Chevalier, F., Khan, A., Wigdor, D., & Brudno, M. (January de 2017). PhenoLines: phenotype comparison visualizations for disease subtyping via topic models. *IEEE transactions on visualization and computer graphics, 24*(1), 371-381. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=7425015062402412575

Gottesman, O., & Doshi-Velez, F. (14 de July de 2018). Regularizing tensor decomposition methods by optimizing pseudo-data. *Workshop on Modern Trends in Nonconvex Optimization for Machine Learning. The 35 th International Conference on Machine Learning*, 1-5. Fonte:

https://finale.seas.harvard.edu/files/finale/files/2018regularizing_tensor_decomposition_models_by_optimizing_pseudodata.pdf

Gottesman, O., Futoma, J., Liu, Y., Parbhoo, S., Celi, L., Brunskill, E., & Doshi-Velez, F. (21 de November de 2020). Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. *Proceedings of the 37th International Conference on Machine Learning, 119*, 3658-3667. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3979059661142155029

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (7 de January de 2019). Guidelines for reinforcement learning in healthcare. *Nature medicine, 25*, 16-18. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=970534608763260270

Gottesman, O., Johansson, F., Meier, J., Dent, J., Lee, D., Srinivasan, S., . . . Doshi-Velez, F. (2018). *Evaluating reinforcement learning algorithms in observational health settings.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=7539099852161268532

Gottesman, O., Liu, Y., Sussex, S., Brunskill, E., & Doshi-Velez, F. (24 de May de 2019). Combining parametric and nonparametric models for off-policy evaluation. *International Conference on Machine Learning, 97*, 2366-2375. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5066391292071299163

Gottesman, O., Pan, W., & Doshi-Velez, F. (31 de March de 2018). Weighted tensor decomposition for learning latent variables with partial data. *International Conference on Artificial Intelligence and Statistics, 84*, 1664-1672. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11879085412951557959

Gottesman, O., Pan, W., & Doshi-Velez, F. (2019). *A general method for regularizing tensor decomposition methods via pseudo-data.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3178221463209035553

Guénais, T., Vamvourellis, D., Yacoby, Y., Doshi-Velez, F., & Pan, W. (2020). *BaCOUn: Bayesian Classifers with Out-of-Distribution Uncertainty.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=263723785578133163

Hughes, M. C., Elibol, H. M., McCoy, T., Perlis, R., & Doshi-Velez, F. (2016). *Supervised topic models for clinical interpretability.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16929944290679312128

Hughes, M. C., Hope, G., Weiner, L., Jr, T. H., Perlis, R. H., Sudderth, E. B., & Doshi-Velez, F. (31 de March de 2018). Semi-Supervised Prediction-Constrained Topic Models. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, 84*, 1067-1076. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17308418540210396368

Hughes, M. C., Hope, G., Weiner, L., Jr, T. H., Perlis, R. H., Sudderth, E. B., & Doshi-Velez, F. (2018). *Supplement: Semi-Supervised Prediction-Constrained Topic Models.* Supplementary Material to AISTATS accepted paper. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9358222559701059936

Hughes, M. C., Hope, G., Weiner, L., McCoy, T. H., Perlis, R. H., Sudderth, E. B., & Doshi-Velez, F. (2017). *Prediction-constrained topic models for antidepressant recommendation.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17015972370524450502

Hughes, M. C., Pradier, M. F., Ross, A. S., McCoy, T. H., Perlis, R. H., & Doshi-Velez, F. (20 de May de 2020). Assessment of a Prediction Model for Antidepressant Treatment Stability Using Supervised Topic Models. *Journal of the American Medical Association: Network Open, 3*(5), e205308-e205308. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17743588105347231287

Hughes, M. C., Pradier, M. F., Ross, A. S., McCoy, T. H., Perlis, R. H., & Doshi-Velez, F. (20 de March de 2020). Generating interpretable predictions about antidepressant treatment stability using supervised topic models. *medRxiv*, 1-27. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17618571076697370683

Hughes, M. C., Weiner, L., Hope, G., Jr, T. H., Perlis, R. H., Sudderth, E. B., & Doshi-Velez, F. (2017). *Prediction-constrained training for semi-supervised mixture and topic models.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=12318386511392406479

Jacobs, M., He, J., Pradier, M. F., Lam, B., Ahn, A. C., McCoy, T. H., . . . Gajos, K. Z. (6 de May de 2021). Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-14. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14830851749986601989

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (4 de February de 2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry, 11*(1), 1-9. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9602941028157969885

Jenna Wiens, S. S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., . . . Goldenberg, A. (September de 2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine, 25*(9), 1337-1340. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1921488170773999899

Jin, L., Doshi-Velez, F., Miller, T., Schuler, W., & Schwartz, L. (October de 2018). Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2721–2731. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8591212071315725016

Jin, L., Doshi-Velez, F., Miller, T., Schuler, W., & Schwartz, L. (1 de April de 2018). Unsupervised grammar induction with depth-bounded PCFG. *Transactions of the Association for Computational Linguistics, 6*, 211-224. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2125035178814160197

Jin, L., Doshi-Velez, F., Miller, T., Schwartz, L., & Schuler, W. (July de 2019). Unsupervised learning of PCFGs with normalizing flow. *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, 2442–2452. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9683284076824164052

Jin, L., Schwartz, L., Doshi-Velez, F., Miller, T., & Schuler, W. (5 de March de 2021). Depth-Bounded Statistical PCFG Induction as a Model of Human Grammar Acquisition. *Computational Linguistics, 47*(1), 181-216. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14543796888475076008

Joseph, J., Doshi-Velez, F., & Roy, N. (14 de May de 2012). A Bayesian nonparametric approach to modeling battery health. *2012 IEEE International Conference on Robotics and Automation*, 1876-1882. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11238423758068122176

Joseph, J., Doshi-Velez, F., Huang, A. S., & Roy, N. (1 de November de 2011). A Bayesian nonparametric approach to modeling motion patterns. *Autonomous Robots, 31*(4), 383-400. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11402996606679976479

Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., & Doshi-Velez, F. (2019). Explainable reinforcement learning via reward decomposition. *Proceedings at the International Joint Conference on Artificial Intelligence*, 1-7. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4762608020035398667

Killian, T., Konidaris, G., & Doshi-Velez, F. (2016). *Transfer learning across patient variations with hidden parameter markov decision processes.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3754116923727715072

Killian, T., Konidaris, G., & Doshi-Velez, F. (4 de December de 2017). Robust and efficient transfer learning with hidden parameter markov decision processes. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6251–6262. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6997382761578063659

Kim, B., & Doshi-Velez, F. (12 de April de 2021). Machine Learning Techniques for Accountability. *AI Magazine, 42*(1), 47-52. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17975239440196363863

Kim, B., Shah, J. A., & Doshi-Velez, F. (7 de December de 2015). Mind the gap: A generative approach to interpretable feature selection and extraction. *Proceedings of the 28th International Conference on Neural Information Processing Systems, 2*, 2260–2268. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5774689255374329461

Konidaris, G., & Doshi-Velez, F. (14 de September de 2014). Hidden parameter Markov decision processes: an emerging paradigm for modeling families of related tasks. *2014 AAAI Fall Symposium Series*, 46-48. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11889108073922919202

Krakovna, V., & Doshi-Velez, F. (2016). *Increasing the interpretability of recurrent neural networks using hidden Markov models.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13497099697736017612

Kreienkamp, A. B., Radhakrishnan, M. L., & Doshi-Velez, F. (7 de April de 2013). Estimating protein-protein electrostatic binding energetics: A feature-based approach. *Abstracts of Papers of the American Chemical Society, 245*. Fonte: https://scholar.google.com/scholar?cluster=7795246451078739358&hl=en&oi=scholarr

Kunes, R., Ren, J., & Doshi-Velez, F. (8 de December de 2019). Prediction Focused Topic Models Via Vocab Filtering. *Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing*, 1-12. Fonte: https://finale.seas.harvard.edu/files/finale/files/prediction_focused_topic_models_via_vocab_filtering.pdf

Lage, I., & Doshi-Velez, F. (2020). *Learning Interpretable Concept-Based Models with Human Feedback.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=644390616868598142

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., & Doshi-Velez, F. (28 de October de 2019). Human Evaluation of Models Built for Interpretability. *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, 7*(1), 59-67. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=18047780529940000065

Lage, I., Lifschitz, D., Doshi-Velez, F., & Amir, O. (2019). *Exploring computational user models for agent policy summarization.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4185786205509736655

Lage, I., Lifschitz, D., Doshi-Velez, F., & Amir, O. (8 de May de 2019). Toward robust policy summarization. *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2081-2083. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15646573855984201979

Lage, I., Ross, A. S., Kim, B., Gershman, S. J., & Doshi-Velez, F. (3 de December de 2018). Human-in-the-Loop Interpretability Prior. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 10180–10189. Fonte: https://scholar.google.com/scholar?start=10&hl=en&as_sdt=0,22&cluster=6925363852455924380

Lee, D., Srinivasan, S., & Doshi-Velez, F. (2019). *Truly batch apprenticeship learning with deep successor features.* arXiv. Fonte: https://arxiv.org/abs/1903.10077

Lingren, T., Chen, P., Bochenek, J., Doshi-Velez, F., Manning-Courtney, P., Bickel, J., . . . Savova, G. (29 de July de 2016). Electronic health record based algorithm to identify patients with autism spectrum disorder. *PLOS One, 11*(7), 1-16. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5434412872445599430

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A., Doshi-Velez, F., & Brunskill, E. (3 de December de 2018). Representation balancing mdps for off-policy policy evaluation. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2649–2658. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11325221239507449364

Lu, M., Shahn, Z., Sow, D., Doshi-Velez, F., & Lehman, L.-w. H. (2020). *Is Deep Reinforcement Learning Ready for Practical Applications in Healthcare? A Sensitivity Analysis of Duel-DDQN for Sepsis Treatment.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17007732522618687755

Masood, A., Pan, W., & Doshi-Velez, F. (2016). *An empirical comparison of sampling quality metrics: A case study for bayesian nonnegative matrix factorization.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1863580417329901184

Masood, M. A., & Doshi-Velez, F. (2016). *Rapid Posterior Exploration in Bayesian Non-negative Matrix Factorization.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8507157983734043039

Masood, M. A., & Doshi-Velez, F. (2018). *Diversity-Inducing Policy Gradient: Using MMD to find a set of policies that are diverse in terms of state-visitation.* Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8455214358698039867

Masood, M. A., & Doshi-Velez, F. (May de 2019). A Particle-Based Variational Approach to Bayesian Non-negative Matrix Factorization. *Journal of Machine Learning Research, 20*, 1-56. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2027101452273906656

Masood, M. A., & Doshi-Velez, F. (2019). *Diversity-inducing policy gradient: Using maximum mean discrepancy to find a set of diverse policies.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5089068101117259482

Masood, M. A., & Doshi-Velez, F. (s.d.). *Leveraging Geometry for Fast Mixing Bayesian Non-negative Matrix Factorization.* Harvard University. Fonte: https://finale.seas.harvard.edu/files/finale/files/2016robust_posterior_exploration_in_nmf.pdf

McMahon, A. W., Cooper, W. O., Brown, J. S., Carleton, B., Doshi-Velez, F., Kohane, I., . . . Califf, R. M. (1 de February de 2020). Big Data in the Assessment of Pediatric Medication Safety. *Pediatrics, 145*(2), e20190562. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13929182941958718072

Nair, Y., & Doshi-Velez, F. (2020). *PAC Bounds for Imitation and Model-based Batch Learning of Contextual Markov Decision Processes.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14910821833495810953

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2 de February de 2018). *How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=12029304911284070268

Ou, H. C., Wang, K., Doshi-Velez, F., & Tambe, M. (10 de May de 2020). Active Screening on Recurrent Diseases Contact Networks with Uncertainty: a Reinforcement Learning Approach. *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, 54-65. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17092379576894892280

Pan, W., & Doshi-Velez, F. (3 de April de 2016). *A characterization of the non-uniqueness of nonnegative matrix factorizations.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3099303622001653484

Parbhoo, S., Bogojeska, J., Roth, V., & Doshi-Velez, F. (5 de December de 2016). Combining Kernel and Model Based Reinforcement Learning for HIV Therapy Selection. *30th Conference on Neural Information Processing Systems*, 1-5. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6107298781484660643

Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V., & Doshi-Velez, F. (26 de July de 2017). Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings, 2017*, 239–248. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=657106867245270507

Parbhoo, S., Gottesman, O., & Doshi-Velez, F. (12 de December de 2020). Shaping Control Variates for Off-Policy Evaluation. *Offline Reinforcement Learning Workshop at Neural Information Processing Systems*, 1-8. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10136749941790753557

Parbhoo, S., Gottesman, O., Ross, A. S., Komorowski, M., Faisal, A., Bon, I., . . . Doshi-Velez, F. (12 de November de 2018). Improving counterfactual reasoning with kernelised dynamic mixing models. *PLOS One, 13*(11), e0205839. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15395330833371325776

Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Lehman, L.-w. H., . . . Doshi-Velez, F. (5 de December de 2018). Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. *AMIA Annual Symposium Proceedings, 2018*, 887–896. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3048790530901914801

Pradier, M. F., Hughes, M. C., & Doshi-Velez, F. (16 de October de 2019). Challenges in computing and optimizing upper bounds of marginal likelihood based on chi-square divergences. *JMLR: Symposium on Advances in Approximate Bayesian Inference*, 1-11. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5194056267031323011

Pradier, M. F., Hughes, M. C., McCoy, T. H., Barroilhet, S. A., Doshi-Velez, F., & Perlis, R. H. (January de 2021). Predicting change in diagnosis from major depression to bipolar disorder after antidepressant initiation. *Neuropsychopharmacology, 46*(2), 455-461. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14097824044734513131

Pradier, M. F., Jr, T. H., Hughes, M., Perlis, R. H., & Doshi-Velez, F. (6 de February de 2020). Predicting treatment dropout after antidepressant initiation. *Translational psychiatry, 10*(1), 1-8. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2188652166281340500

Pradier, M. F., Pan, W., Yao, J., Ghosh, S., & Doshi-Velez, F. (November de 2018). *Latent projection bnns: Avoiding weight-space pathologies by learning latent representations of neural network weights.* arXiv. Fonte: https://deepai.org/publication/latent-projection-bnns-avoiding-weight-space-pathologies-by-learning-latent-representations-of-neural-network-weights

Pradier, M. F., Pan, W., Yao, J., Ghosh, S., & Doshi-Velez, F. (7 de December de 2018). Projected BNNs: Avoiding weight-space pathologies by projecting neural network weights. *Third workshop on Bayesian Deep Learning*, 1-7.

Pradier, M. F., Pan, W., Yao, J., Ghosh, S., & Doshi-Velez, F. (2019). *Projected BNNs: Avoiding weight-space pathologies by learning latent representations of neural network weights.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=988884354241906835

Pradier, M. F., Pan, W., Yau, M., Singh, R., & Doshi-Velez, F. (7 de December de 2018). Hierarchical Stick-breaking Feature Paintbox. *3rd Bayesian Nonparametrics Workshop at NeurIPS*, 1-7. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15999838249257899565

Pradier, M. F., Zazo, J., Parbhoo, S., Perlis, R. H., Zazzi, M., & Doshi-Velez, F. (2021). *Preferential Mixture-of-Experts: Interpretable Models that Rely on Human Expertise As Much As Possible.* arXiv. Fonte: https://arxiv.org/abs/2101.05360

Prasad, N., Engelhardt, B. E., & Doshi-Velez, F. (2019). *Defining Admissible Rewards for High Confidence Policy Evaluation.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16234252057896164723

Prasad, N., Engelhardt, B., & Doshi-Velez, F. (2 de April de 2020). Defining admissible rewards for high-confidence policy evaluation in batch reinforcement learning. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 1-9. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=18316110452828974208

Raghu, A., Gottesman, O., Liu, Y., Komorowski, M., Faisal, A., Doshi-Velez, F., & Brunskill, E. (2018). *Behaviour policy estimation in off-policy policy evaluation: Calibration matters.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2553179708177407291

Ren, J., Kunes, R., & Doshi-Velez, F. (2019). *Prediction Focused Topic Models for Electronic Health Records.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=18217668241970505236

Ren, J., Kunes, R., & Doshi-Velez, F. (2019). *Prediction Focused Topic Models via Vocab Selection.* arXiv. Fonte: https://arxiv.org/abs/1910.05495v1

Ren, J., Kunes, R., & Doshi-Velez, F. (3 de June de 2020). Prediction Focused Topic Models via Feature Selection. *International Conference on Artificial Intelligence and Statistics, 108*, 4420-4429. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11979491999342264807

Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (19 de August de 2017). Right for the right reasons: Training differentiable models by constraining their explanations. *Proceedings of the 26th International Joint Conference on Artificial IntelligenceAugust 2017*, 2662–2670. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1999112949296082528

Ross, A. S., Pan, W., & Doshi-Velez, F. (2018). *Learning qualitatively diverse and interpretable rules for classification.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17466786373513063009

Ross, A. S., Pan, W., Celi, L. A., & Doshi-Velez, F. (3 de April de 2020). Ensembles of Locally Independent Prediction Models. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(04), 5527-5536. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14521572664817244396

Ross, A., & Doshi-Velez, F. (25 de April de 2018). Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1), 1660-1669. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10549843532884126759

Ross, A., Chen, N., Hang, E. Z., Glassman, E. L., & Doshi-Velez, F. (6 de May de 2021). Evaluating the Interpretability of Generative Models by Interactive Reconstruction. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1-15. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=815272469989940504

Ross, A., Lage, I., & Doshi-Velez, F. (December de 2017). The neural lasso: Local linear sparsity for interpretable explanations. *Workshop on Transparent and Interpretable Machine Learning in Safety Critical Environments, 31st Conference on Neural Information Processing Systems*, 1-5. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15073022973103931668

Shain, C., Bryce, W., Jin, L., Krakovna, V., Doshi-Velez, F., Miller, T., . . . Schwartz, L. (1 de December de 2016). Memory-bounded left-corner unsupervised grammar induction on child-directed input. *26th International Conference on Computational Linguistics*, 964-975. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3880418564438617660

Shain, C., Bryce, W., Jin, L., Krakovna, V., Doshi-Velez, F., Miller, T., . . . Schwartz, L. (s.d.). *Modeling syntax acquisition via cognitively-constrained unsupervised grammar induction.* Fonte: https://www.asc.ohio-state.edu/shain.3/pdf/cuny17uhhmmposter.pdf

Simons, M. G., Futoma, J. D., Gao, M., Corey, K., Sendak, M., Whalen, K. B., . . . Setji, T. (1 de June de 2019). 1185-P: Predictive Model for Hyperglycemic Events after High Dose Corticosteroid Administration. *Diabetes, 68*(Supplement 1). Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8875672999909086415

Singh, H., Joshi, S., Doshi-Velez, F., & Lakkaraju, H. (2021). *Learning Under Adversarial and Interventional Shifts.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9693492366965853958

Singh, R., Ling, J., & Doshi-Velez, F. (2017). Structured variational autoencoders for the beta-bernoulli process. *NIPS 2017 Workshop on Advances in Approximate Bayesian Inference*, 1-9. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2129901359852149828

Sonali Parbhoo, M. W.-V. (18 de September de 2020). Transfer Learning from Well-Curated to Less-Resourced Populations with HIV. *Proceedings of the 5th Machine Learning for Healthcare Conference, 126*, 589-609. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4619745818227670593

Srinivasan, S., & Doshi-Velez, F. (30 de May de 2020). Interpretable batch irl to extract clinician goals in icu hypotension management. *AMIA Joint Summits on Translational Science Proceedings, 2020*,

636–645. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11182009381876685964

Sussex, S., Gottesman, O., Liu, Y., Murphy, S., Brunskill, E., & Doshi-Velez, F. (15 de July de 2018). Stitched Trajectories for Off-Policy Learning. *International Conference on Machine Learning (ICML) Workshop on CausalML*, 1-6. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1246966712446552987

Tan, S., Doshi-Velez, F., Quiroz, J., & Glassman, E. (2017). *Clustering LaTeX Solutions to Machine Learning Assignments for Rapid Assessment.* Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9037807814473174850

Thakur, S., Lorsung, C., Yacoby, Y., Doshi-Velez, F., & Pan, W. (2020). *Learned Uncertainty-Aware (LUNA) Bases for Bayesian Regression using Multi-Headed Auxiliary Networks.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11848906772340612660

Tran, D., Kim, M., & Doshi-Velez, F. (2016). *Spectral M-estimation with Application to Hidden Markov Models: Supplementary Material.* Supplementary Material to JMLR accepted paper. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6851306845752047214

Tran, D., Kim, M., & Doshi-Velez, F. (2 de May de 2016). Spectral m-estimation with applications to hidden markov models. *Artificial Intelligence and Statistics, 51*, 1421-1430. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16606738410332986692

Tran, D., Ranganath, R., & Blei, D. M. (20 de November de 2015). *The variational Gaussian process.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13476964332561990649

Vaughan, D., Pan, W., Yacoby, Y., Seidler, E. A., Leung, A. Q., Doshi-Velez, F., & Sakkas, D. (1 de September de 2019). The application of machine learning methods to evaluate predictors of live birth in programmed thaw cycles. *Fertility and Sterility, 112*(3), e273. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16601754021436254072

Wang, K., Shat, S., Chen, H., Perrault, A., Doshi-Velez, F., & Tambe, M. (2021). *Learning MDPs from Features: Predict-Then-Optimize for Sequential Decision Problems by Reinforcement Learning.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10238009203271525593

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2015). *Bayesian or's of and's for interpretable classification with application to context aware recommender.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=7372300651005179578

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2015). *Or's of and's for interpretable classification, with application to context-aware recommender systems.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1841876631120950361

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (1 de January de 2017). A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research, 18*(1), 2357-2393. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17888042361257282144

Wu, M., Ghassemi, M., Feng, M., Celi, L. A., Szolovits, P., & Doshi-Velez, F. (May de 2017). Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association, 24*(1), 488-495. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13394270005196874867

Wu, M., Hughes, M., Parbhoo, S., & Doshi-Velez, F. (2 de February de 2018). Beyond sparsity: Tree-based regularization of deep models for interpretability. *The Thirty-Second AAAI Conference on Artificial Intelligence*, 1670-1678. Fonte: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/16285/15867

Wu, M., Hughes, M., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (25 de April de 2018). Beyond sparsity: Tree regularization of deep models for interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1), 1670-1678. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15277305634094866318

Wu, M., Parbhoo, S., Hughes, M. C., Roth, V., & Doshi-Velez, F. (2019). *Optimizing for interpretability in deep neural networks with tree regularization.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3379156586223781565

Wu, M., Parbhoo, S., Hughes, M., Kindle, R., Celi, L., Zazzi, M., . . . Doshi-Velez, F. (13 de August de 2019). *Regional tree regularization for interpretability in black box models.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10080078237025095815

Wu, M., Parbhoo, S., Hughes, M., Kindle, R., Celi, L., Zazzi, M., . . . Doshi-Velez, F. (3 de April de 2020). Regional tree regularization for interpretability in deep neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(04), 6413-6421. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6751398667967176999

Xia, X., Protopapas, P., & Doshi-Velez, F. (30 de June de 2016). Cost-Sensitive Batch Mode Active Learning: Designing Astronomical Observation by Optimizing Telescope Time and Telescope Choice. *Proceedings of the 2016 SIAM International Conference on Data Mining*, 477-485. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2899968381574441653

Yacoby, Y., Pan, W., & Doshi-Velez, F. (14 de June de 2019). Mitigating Model Non-Identifiability in BNN with Latent Variables. *ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning*. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10193865716492445484

Yacoby, Y., Pan, W., & Doshi-Velez, F. (3 de February de 2020). Characterizing and Avoiding Problematic Global Optima of Variational Autoencoders. *Symposium on Advances in Approximate Bayesian Inference, 118*, 1-17. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11829077304199688529

Yacoby, Y., Pan, W., & Doshi-Velez, F. (2020). *Failure Modes of Variational Autoencoders and Their Effects on Downstream Tasks.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10882037403625080405

Yacoby, Y., Pan, W., & Doshi-Velez, F. (2021). *Learning deep bayesian latent variable regression models that generalize: When non-identifiability is a problem.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2945272416496962114

Yang, W., Lorch, L., Graule, M. A., Lakkaraju, H., & Doshi-Velez, F. (December de 2020). Incorporating Interpretable Output Constraints in Bayesian Neural Networks. *Advances in Neural Information Processing Systems, 33*, 1-17. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=18422416496972996615

Yang, W., Lorch, L., Graule, M. A., Srinivasan, S., Suresh, A., Yao, J., . . . Doshi-Velez, F. (2019). *Output-constrained Bayesian neural networks.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2484614356474647950

Yao, J., Brunskill, E., Pan, W., Murphy, S., & Doshi-Velez, F. (2020). *Power-Constrained Bandits.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9242881172335977666

Yao, J., Killian, T., Konidaris, G., & Doshi-Velez, F. (2018). Direct policy transfer via hidden parameter markov decision processes. *Lifelong Learning: A Reinforcement Learning Approach Workshop, FAIM, 2018*, 1-7. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=12342080706571652113

Yao, J., Pan, W., Ghosh, S., & Doshi-Velez, F. (2019). *Quality of uncertainty quantification for Bayesian neural network inference.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6516073606226527796

Yi, K., & Doshi-Velez, F. (2017). *Roll-back hamiltonian monte carlo.* arXiv. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16133475510257504120

Zhang, K. W., Gottesman, O., & Doshi-Velez, F. (12 de December de 2020). A Bayesian Approach to Learning Bandit Structure in Markov Decision Processes. *The Challenges of Real World Reinforcement Learning Workshop at NeurIPS*, 1-12. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3923767981002832648

Zhang, K., Wang, Y., Du, J., Chu, B., Celi, L. A., Kindle, R., & Doshi-Velez, F. (11 de December de 2021). Identifying Decision Points for Safe and Interpretable Reinforcement Learning in Hypotension Treatment. *NeurIPS Workshop on Machine Learning for Health, 1*, 1-9. Fonte: https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5631931464143254765

Bibliography: Mr. Isaac Lage github link

Gottesman, O., Johansson, F., Meier, J., Dent, J., Lee, D., Srinivasan, S., . . . Doshi-Velez, F. (2018). *Evaluating reinforcement learning algorithms in observational health settings.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=7539099852161268532

Lage, I., & Doshi-Velez, F. (2020). *Learning Interpretable Concept-Based Models with Human Feedback.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=644390616868598142

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., & Doshi-Velez, F. (2019, October 28). Human Evaluation of Models Built for Interpretability. *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, 7*(1), 59-67. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=18047780529940000065

Lage, I., Lifschitz, D., Doshi-Velez, F., & Amir, O. (2019). *Exploring computational user models for agent policy summarization.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4185786205509736655

Lage, I., Lifschitz, D., Doshi-Velez, F., & Amir, O. (2019, May 8). Toward robust policy summarization. *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 2081-2083. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15646573855984201979

Lage, I., Ross, A. S., Kim, B., Gershman, S. J., & Doshi-Velez, F. (2018, December 3). Human-in-the-Loop Interpretability Prior. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 10180–10189. Retrieved from https://scholar.google.com/scholar?start=10&hl=en&as_sdt=0,22&cluster=6925363852455924380

Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., & Pan, W. (2021, June). The Promises and Pitfalls of Black-box Concept Learning Models. *ICML Workshop on Theoretic Foundation*, 1-13. Retrieved from https://arxiv.org/pdf/2106.13314.pdf

Ross, A., Lage, I., & Doshi-Velez, F. (2017, December). The neural lasso: Local linear sparsity for interpretable explanations. *Workshop on Transparent and Interpretable Machine Learning in Safety Critical Environments, 31st Conference on Neural Information Processing Systems*, 1-5. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15073022973103931668

Ms. Rathnam, Sarah; PhD 25':

With over a decade of experience in quantitative finance, Ms. Rathnam chose to return to academia. At Harvard, she is co-advised by Finale Doshi-Velez of the Data to Actionable Knowledge (DtAK) Lab and Susan Murphy of the Statistical Reinforcement Learning Lab.

Bibliography: Dr. Weiwei Pan [View My GitHub Profile]

Antorán, J., Yao, J., Pan, W., Hernández-Lobato, J. M., & Doshi-Velez, F. (2020, July 17). Amortised Variational Inference for Hierarchical Mixture Models. *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*, 1-11. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=486508101802156030

Coker, B., Pan, W., & Doshi-Velez, F. (2021). *Wide Mean-Field Variational Bayesian Neural Networks Ignore the Data.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13049947611571259225

Downs, M., Chu, J. L., Yacoby, Y., Doshi-Velez, F., & Pan, W. (2020, July 12). CRUDS: Counterfactual Recourse Using Disentangled Subspaces. *ICML Workshop on Human Interpretability in Machine Learning*, 1-23. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14822408888462715234

Gottesman, O., Pan, W., & Doshi-Velez, F. (2018, March 31). Weighted tensor decomposition for learning latent variables with partial data. *International Conference on Artificial Intelligence and Statistics, 84*, 1664-1672. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11879085412951557959

Gottesman, O., Pan, W., & Doshi-Velez, F. (2019). *A general method for regularizing tensor decomposition methods via pseudo-data.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3178221463209035553

Guénais, T., Vamvourellis, D., Yacoby, Y., Doshi-Velez, F., & Pan, W. (2020). *BaCOUn: Bayesian Classifers with Out-of-Distribution Uncertainty.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=263723785578133163

Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., & Pan, W. (2021, June). The Promises and Pitfalls of Black-box Concept Learning Models. *ICML Workshop on Theoretic Foundation*, 1-13. Retrieved from https://arxiv.org/pdf/2106.13314.pdf

Masood, A., Pan, W., & Doshi-Velez, F. (2016). *An empirical comparison of sampling quality metrics: A case study for bayesian nonnegative matrix factorization.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1863580417329901184

Pan, W., & Doshi-Velez, F. (2016, April 3). *A characterization of the non-uniqueness of nonnegative matrix factorizations.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3099303622001653484

Pradier, M. F., Pan, W., Yao, J., Ghosh, S., & Doshi-Velez, F. (2018, November). *Latent projection bnns: Avoiding weight-space pathologies by learning latent representations of neural network weights.* arXiv. Retrieved from https://deepai.org/publication/latent-projection-bnns-avoiding-weight-space-pathologies-by-learning-latent-representations-of-neural-network-weights

Pradier, M. F., Pan, W., Yao, J., Ghosh, S., & Doshi-Velez, F. (2018, December 7). Projected BNNs: Avoiding weight-space pathologies by projecting neural network weights. *Third workshop on Bayesian Deep Learning*, 1-7.

Pradier, M. F., Pan, W., Yao, J., Ghosh, S., & Doshi-Velez, F. (2019). *Projected BNNs: Avoiding weight-space pathologies by learning latent representations of neural network weights.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=988884354241906835

Pradier, M. F., Pan, W., Yau, M., Singh, R., & Doshi-Velez, F. (2018, December 7). Hierarchical Stick-breaking Feature Paintbox. *3rd Bayesian Nonparametrics Workshop at NeurIPS*, 1-7. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15999838249257899565

Ross, A. S., Pan, W., & Doshi-Velez, F. (2018). *Learning qualitatively diverse and interpretable rules for classification.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17466786373513063009

Ross, A. S., Pan, W., Celi, L. A., & Doshi-Velez, F. (2020, April 3). Ensembles of Locally Independent Prediction Models. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(04), 5527-5536. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14521572664817244396

Schaeffer, R., Bordelon, B., Khona, M., Pan, W., & Fiete, I. (2021, July 27). Efficient Online Inference for Nonparametric Mixture Models. *37th Conference on Uncertainty in Artificial Intelligence*. Retrieved from https://fietelabmit.files.wordpress.com/2021/06/final_camera_ready.pdf

Thakur, S., Lorsung, C., Yacoby, Y., Doshi-Velez, F., & Pan, W. (2020). *Learned Uncertainty-Aware (LUNA) Bases for Bayesian Regression using Multi-Headed Auxiliary Networks.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11848906772340612660

Vaughan, D., Pan, W., Yacoby, Y., Seidler, E. A., Leung, A. Q., Doshi-Velez, F., & Sakkas, D. (2019, September 1). The application of machine learning methods to evaluate predictors of live birth in programmed thaw cycles. *Fertility and Sterility, 112*(3), e273. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16601754021436254072

Yacoby, Y., Pan, W., & Doshi-Velez, F. (2019, June 14). Mitigating Model Non-Identifiability in BNN with Latent Variables. *ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning*. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10193865716492445484

Yacoby, Y., Pan, W., & Doshi-Velez, F. (2020, February 3). Characterizing and Avoiding Problematic Global Optima of Variational Autoencoders. *Symposium on Advances in Approximate Bayesian Inference, 118*, 1-17. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11829077304199688529

Yacoby, Y., Pan, W., & Doshi-Velez, F. (2020). *Failure Modes of Variational Autoencoders and Their Effects on Downstream Tasks.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10882037403625080405

Yacoby, Y., Pan, W., & Doshi-Velez, F. (2021). *Learning deep bayesian latent variable regression models that generalize: When non-identifiability is a problem.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2945272416496962114

Yao, J., Brunskill, E., Pan, W., Murphy, S., & Doshi-Velez, F. (2020). *Power-Constrained Bandits.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9242881172335977666

Yao, J., Pan, W., Ghosh, S., & Doshi-Velez, F. (2019). *Quality of uncertainty quantification for Bayesian neural network inference.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6516073606226527796

Harvard John A. Paulson School of Engineering and Applied Sciences

SEAS, 29 Oxford Street, Cambridge, MA 02138
O: 617 495 3188
Finale@seas.harvard.edu
Finale.seas.harvard.edu

Finale Doshi-Velez
John L. Loeb Associate Professor of Engineering and Applied Sciences

JULY 1, 2021

AI accountability and progress are not odds, as long as mechanisms are appropriately chosen. In the following, we suggest that (1) the current regulatory framework under review could benefit from a **more practical definition of explainability** that focuses on what information needs to be provided to answer the required question, (2) as much or more attention needs to be given to the data that create the models as the models themselves, and (3) the Agencies could use recent research to better define standards for the continuous monitoring of AI by all its stakeholders. We suggest having an **AI Model "Check Engine" light** to set standards to monitor their negative externalities **so that AI models do not "fail silently."**

Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning*

This is a regulatory comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning to (OCC) Mr. Blake J. Paulson, Acting Comptroller of the Currency, Office of the Comptroller of the Currency [Docket ID OCC–2020–0049] (FRB) Ms. Ann Misback, Secretary of the Board, Board of Governors of the Federal Reserve System [Docket No. OP–1743]; (FDIC), Mr. James P. Sheesley, Assistant Executive Secretary, Federal Deposit Insurance Corporation RIN 3064–ZA24, (BCFP) Mr David Uejio, Acting Director, the Bureau of Consumer Financial Protection [Docket No. CFPB–2021–0004]; (NCUA) Ms. Melane Conyers-Ausbrooks, Secretary of the Board, National Credit Union Administration [Docket No. NCUA–2021–0023], henceforth collectively referenced to as "the Agencies," Dated at Washington, DC, on or about February 25, 2021. [FR Doc. 2021–06607 Filed 3–30–21; 8:45 am], billing codes 4810–33–P; 6210–01–P; 4810–AM–P; 6714–01–P

BORGES TORREALBA CARPI, CARLOS
HARVARD'S DATA TO ACTIONABLE KNOWLEDGE LAB
Harvard John A. Paulson School of Engineering and Applied Sciences, 150 Western Ave, Allston, MA 02134

Table of Contents

Preamble:

Header Note

As per the Agencies Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning ("RFI") Dated at Washington, DC, on or about February 25, 2021, published in the Federal Registrar on March 31st, 2021 [FR Doc. 2021–06607 Filed 3–30–21; 8:45 am], by Blake J. Paulson, Acting Comptroller of the Currency; By order of the Board of Governors of the Federal Reserve System, Ann Misback, Secretary of the Board; Federal Deposit Insurance Corporation, James P. Sheesley, Assistant Executive Secretary; David Uejio, Acting Director, Bureau of Consumer Financial Protection. Melane Conyers-Ausbrooks, Secretary of the Board, National Credit Union Administration, as "Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning,"[1] this comment addresses its 7 of its 17 questions.

Note that the ***views expressed here are solely our own,*** and do not necessarily correspond to the official or unofficial views of Harvard University (or its Harvard John A. Paulson School of Engineering and Applied Sciences).

Executive Summary

Artificial Intelligence accountability does not need to stop AI progress. Demanding explanations or other forms of evidence and transparency does not imply disclosing trade secrets no more than asking people to explain themselves implies disclosing how electricity flows through their neurons. Pragmatically, explanations involve sharing the part of a model's decision-making logic that is relevant for adjudicating the question on hand.[2] Below, we summarize our thoughts relating to the seven questions addressed from the RFI's seventeen.

First, as noted above, we suggest that the definition of AI (1) **"Explainability"** might need to be more pragmatic and less conceptual: an explanation is the "information about the AI provided to the user such that they can make the decision they are trying to make." Different contexts will require different explanations.

Second, the **value of explainability depends on how precisely the need can be quantified.** Explainability can be quite valuable for harder-to-quantify issues such as exposing information, preventing or rectifying errors, or dispute resolution; it can help check if **models are "right" for the "right" reasons.** It may not be needed in contexts where there is a well-understood alternative

[1] Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, Vol. 86 No. 60 Fed. Reg. 16837-16842 (March 31st, 2021).
[2] See 'Local Counterfactual Faithfulness,' "as humans we don't expect these explanations to be the same or even consistent what we do expect is that the explanation holds for similar circumstances" See summary presentation here: https://youtu.be/4lIr8rgo5zE?t=488 ; For a more detailed note see Finale Doshi-Velez, Sam Gershman, et al (2017) "Accountability of AI Under the Law: The Role of Explanation"working draft at https://arxiv.org/pdf/1711.01134 - As part of Harvard's Berkman Klein Center Working Group on AI Interpretability, a collaborative effort between legal scholars, computer scientists, and cognitive scientists)

goal metric available. That said, sometimes both are needed: Although we are not experts in fair-lending, AI **fairness** metrics literature discuss how **simple aggregates are not substitute for AI explainability**. [3]

In this way, explainability is one part of a broader accountability toolkit. For example, concerns about model performance under **'dynamic updating**' could be **remedied with internal & external AI model audits** which look at both metrics and explanations. Regular **third-party oversight** is critical so that models do not fail silently -- we need the equivalent of a "check that engine" light to alert users that a model may need further inspection.

More broadly, many concerns come not from the model but from the data used to train the model. For example, without **sufficiently broad data collection**, the models will likely **overfit**; there might be a need to change data-collection incentives to be more sensitive to diversity and inclusivity.[4] More **intensive data usage and processing** concerns can **be mitigated with dataset documentation,** for example the "Data Nutrition Project" at MIT/Harvard Law School produces "nutrition labels" for the datasets being ingested by AI models.

Finally, as experts in medical data, we note that the **MIMIC dataset** with **anonymized individual-level** hospital health data has provided a foundation for AI for health research. There exists **a great opportunity** to ensure the trust of the American people on the fairness of its financial system -- and democratize improvements -- by creating similar datasets from banking institutions.

[3] More concretely, the FDIC could lead the development of Data Donation Frameworks for CDFIs and MDIs under Mission-Driven Bank Fund to expand academic research to operationalize the regulatory monitoring of systemic discrimination.
[4] For example, by reforming the FFIEC to support a MIMIC-type project for finance.

JULY 1, 2021

Response to the Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning

Abstract

AI accountability and progress are not odds, as long as mechanisms are appropriately chosen. In the following, we suggest that (1) the current regulatory framework under review could benefit from a more practical definition of explainability that focuses on what information needs to be provided to answer the required question, (2) as much or more attention needs to be given to the data that create the models as the models themselves, and (3) the Agencies could use recent research to better define standards for the continuous monitoring of AI by all its stakeholders. We suggest having an AI Model "Check Engine" light to set standards to monitor their negative externalities so that regulators can make sure that AI models do not "fail silently."

BORGES TORREALBA CARPI, CARLOS

Harvard's Data to Actionable Knowledge lab

Harvard John A. Paulson School of Engineering and Applied Sciences, 150 Western Ave, Allston, MA 02134

1. Introduction

1.1. Background: DtAK Lab

The Harvard's Data to Actionable Knowledge (DtAK) lab, led by Finale Doshi-Velez, uses probabilistic machine learning methods to address many decision-making scenarios, with a focus on healthcare applications

1.1.1. PI: Finale Doshi-Velez (She/Her/Hers), the Gordon MacKay Full Professor of Engineering and Applied Sciences

Professor Finale Doshi-Velez received her Ph.D. in Computer Science from MIT and an M.Sc. in Engineering from Cambridge University as a Marshall Fellow. Prior to joining SEAS, she was postdoc at Harvard Medical School. Doshi-Velez has received an Alfred P. Sloan Research Fellowship, an NSF CiTRaCS postdoctoral fellowship, an NSF CAREER award, and an AFOSR Young Investigator award. In 2019, she was awarded the Everett Mendelsohn Excellence in Mentoring Award by the Graduate Student Council for her mentorship and support of graduate students.

1.1.2. Major Areas: Modeling, Decision-Making, and Interpretability

Probabilistic modeling and inference:
We focus especially on Bayesian models

- How can we characterize the uncertainty in large, heterogeneous data?
- How can we fit models that will be useful for downstream decision-making?
- How can we build models and inference techniques that will behave in expected and desired ways?

Decision-making under uncertainty:
We focus especially on sequential decision-making

- How can we optimize policies given batches of heterogeneous data?
- How can we provide useful information, even if we can't solve for a policy?
- How can we characterize the limits of our ability to provide decision support?

Interpretability and statistical methods for validation:

- How can we estimate the quality of a policy from batch data?
- How can we expose key elements of a model or policy for expert inspection?

1.1.3. Expertise

These comments were created via discussion in the Data to Actionable Knowledge Lab, with particularly engaged suggestions from Weiwei Pan, Isaac Lage, Andrew Ross, Beau Coker, Sarah Rathnam, and Shalmali Joshi, as well as Eura Shin and Jiayu Yao.

1.2. Disclaimer

1.2.1. Conflict of Interest Statement

Our principal investigator, Professor Finale Doshi-Velez and the lab often focuses on health-care applications, therefore we do not recognize any substantial conflicts of interests here, outside of noting that (1) some of our researchers have substantial experience working on AI in the finance industry, and (2) our work with a few academic partners,[5] "Summarizing Agent Behavior to People" was recognized with the JP Morgan Faculty Award[6] for 2019.[7] Finale Doshi-Velez also consults for Ethena.

2. Comments on RFI's Definitions

2.1. Explainability: Need for pragmatic, less conceptual definition: *information about the AI provided to the user such that they can make the decision they are trying to make.*

Currently, the RFI defines AI explainability as

> *For the purposes of this RFI,* **explainability refers to how an AI approach uses inputs to produce outputs**. *Some AI approaches can exhibit a "lack of explainability" for their overall functioning (sometimes known as global explainability) or how they arrive at an individual outcome in a given situation (sometimes referred to as local explainability). Lack of explainability can pose different challenges in different contexts. Lack of explainability can also inhibit financial institution management's understanding of the conceptual soundness [6] of an AI approach, which can increase uncertainty around the AI approach's reliability, and increase risk when used in new contexts. Lack of explainability can also inhibit independent review and audit and make compliance with laws and regulations, including consumer protection requirements, more challenging. [***emphasis** added]*

At DtAK we consider defining explanation more pragmatically:

> *Explanation is information about the AI provided to the user such that they can make the decision they are trying to make.*

In this sense, explanation is very context dependent: the explanation necessary to determine whether an AI system will be safe in general may be vastly different than an explanation to assist in determining whether a specific decision is safe.

3. Explainability: Trade-offs based on how well defined are model goals

3.1. Question 1: Not answered directly, see Q3

3.2. Question 2: Not answered directly, see Q3

3.3. Question 3: Explainability is needed in cases where metrics are not enough, such as identifying the overall workings of a model, preventing or rectifying errors, and resolving disputes.

For which uses of AI is lack of explainability more of a challenge? Please describe those challenges in detail. How do financial institutions account for and manage the varied challenges and risks posed by different uses?

---

[5] Professor Ofra Amir Technion – Israel Institute of Technology who is part of the Faculty of Industrial Engineering & Management, and Professor David Sarne Bar-Ilan University, Department of Computer Science and Technology

[6] See https://www.jpmorgan.com/insights/technology/artificial-intelligence/awards/faculty-award-recipients

[7] "The J.P. Morgan AI Research Awards 2019 partners with research thinkers across artificial intelligence. The program is structured as a gift that funds a year of study for a graduate student."

At a high level, the lack of explainability is a challenge for tasks that lack a simple, reliable metric. These include exposing information about the overall workings of a model, preventing or rectifying errors, and resolving disputes. Below, we expand on these situations. We have also curated sources in the word document of Human-Computer interactions research as well as AI explainability research that is more detailed and expansive then this summary.

Explanations may **expose information about the AI models to increase transparency.**[8]

- In many applications, it may be possible to build a fully transparent, compact AI model with high accuracy. In such cases, the AI can be completely inspected for possible flaws. Especially in high-stakes settings, such models should always be the starting point.
- However, for more complex models, this may not be possible. In this case, the explanation may provide only a partial view (e.g., how a particular set of inputs affect the output, or which inputs have the most effect on determining a particular output). This partial view must be aligned with the reason for seeking an explanation.

Explanations can be used to **prevent or rectify errors and increase trust.**[9]

- In some cases, it may be possible to define exactly how and when a user needs to be alerted about a situation. For example, the conditions under which a car's engine light turns on are well-understood, can be precisely defined in advance, and the appropriate response to the engine light is also well-understood.
- However, in many other cases, such as fairness, the notion of appropriate behavior may be more subtle and contextual. Explanations that enable an understanding of an AI's behaviors can help ensure that the AI's behavior aligns with what is desired (or rectify errors).
- That said, as noted above, for a sufficiently complex AI system, this explanation will necessarily be partial, and thus some amount of ex-ante decision-making will still be necessary about what parts of the AI to expose to help check for certain kinds of errors (e.g., errors relating to discriminatory behavior, errors relating to risk, etc.). For example, an explanation might reveal what features are important for a particular decision, but not how they interact (unless designed to). Even a partial explanation, however, can provide insights to augment aggregate statistics.

Explanations can also be used to ascertain whether **certain criteria were used appropriately or inappropriately** in case of a dispute. [10]

- Aggregate measures cannot tell you whether there was a wrongdoing in this particular case; explanations that provide information about how factors were used and what would have happened if the factors changed can be used to determine whether a decision was made appropriately.

[8] See Lage et al (2018) "Human-in-the-Loop Interpretability," Lage et al (2019), "Human Evaluation of Models Built for Interpretability;" Ustun et al (2019) "Actionable Recourse in Linear Classification," For concrete problems related to gender classification for example, see Buolamwini et al (2019), "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification;" Keyes (2018) "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition,"

[9] Ribeiro et al, (2016) "Why Should I Trust You? Explaining the Predictions of Any Classifier;" Yang et al (2017) "Evaluating Effects of User Experience and System Transparency on Trust in Automation;" Yin et al (2019) "Understanding the Effect of Accuracy on Trust in Machine Learning Models."

[10] Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., & Pan, W. (June 2021) "Promises and Pitfalls of Black-Box Concept Learning Models." See For concrete examples of unpacking 'Blackbox' models, for example see Koh er al (2017), "Understanding black-box predictions via influence functions"

- That said, one also needs to look at explanations across a dataset (globally) to check for issues. For example, it would be important to know that an AI often makes discriminatory decisions, but not always, without having to adjudicate multiple individual cases first.

Conversely, the **lack of explainability is not a challenge** when

- The system that the AI is modeling is **well-understood.** For example, computer assistance for aircraft collision avoidance follows from well-understood physics. Such as system needs rigorous testing, but not explanation.
- If the **system's context is not going to change**, that is the training data represents the inputs that will be seen when the system is used, and the outputs of that training data are curated to be correct. In this case, we may be less worried about whether the system has the causal factors/correlation may be sufficient.
- There **are other metrics that can be used for the desired goal**. For example, some notions of fairness are simply aggregate statistics of the model's outputs monitored over time. That said, if the situation is sufficiently high stakes, one may not want to wait to collect a large amount of data to see whether a system is unsafe, unfair, etc.

4. Risks from Broader or More Intensive Data Processing and Usage: Dataset Documentation, for example Data Nutrition Project

4.1. Question 4: Data is one of the biggest sources of AI error; transparency about the data sources is critical for accountability. See also Q8 (AI Model audits)

How do financial institutions using AI manage risks related to data quality and data processing? How, if at all, have control processes or automated data quality routines changed to address the data quality needs of AI? How does risk management for alternative data compare to that of traditional data? Are there any barriers or challenges that data quality and data processing pose for developing, adopting, and managing AI? If so, please provide details on those barriers or challenges.

Data is one of the biggest sources of AI error: while many models may work reasonably well for a task, all models will fail if the data quality and processing are poor. There is an emerging consensus in the literature that the data set is absolutely critical with respect to how the model will perform. Many current concerns revolve around general bias embedded at creation into the state-of-the-art AI model that can be attributable to the data used during model training, even when a "universal" dataset (for example, the 'entire' internet) is ingested by the model uncritically – the "universal" dataset still contains the biases of the people who created it.

Regulators might need to consider how to provide guidance to depository institutions about how to document and supervise the dataset collection, so that financial AI models are not replicating biases that are subtle and hard to detect without sufficiently detailed data documentation. For example, lending data might not have gender information; however, this makes it hard to determine whether the dataset is overwhelmingly male -- and thus leading to a model biased against non-males. Although there is not yet a consensus on the best ways to evaluate and document data sets, we point regulators to "Datasheet for

Datasets" https://arxiv.org/abs/1803.09010 as one approach.[11] A more concrete one is underway at the Berkman Klein center with the Data Nutrition Project.[12]

Ideally, a model would be "right for the right reasons," capturing something immutable about the world. However, this is rarely the case, especially when the data come from human processes. Because the input data are likely shifting as trends change, models can stop working as intended. Thus, is important for regulators to consider model audits and related-governance frameworks. Regulator-created scenarios might be used to "stress-test" the AI models in key data-input regimes.

4.2. Question 5: As we gather more alternative data, we must also gather data about sensitive variables to ensure we are not creating proxies for them. See also Q8 (model audits)

Are there specific uses of AI for which alternative data are particularly effective?

In many cases, *it will be necessary to collect data on sensitive variables* to ensure that systems are not building proxies for them based on ever increasingly sophisticated data streams. Therefore, model audits need sensitive data to make sure prohibited categories (i.e., race etc.) are not re-created using other variables.

5. Overfitting: Better incentives towards broader data collection & publication i.e. MIMIC for Finance.

5.1. Question 6: Continuous audits are needed to manage overfitting risks; the biggest risks are overfitting to a specific population used to train the model rather than the model itself. MIMIC-type project to democratize data.

How do financial institutions manage AI risks relating to overfitting? What barriers or challenges, if any, does overfitting pose for developing, adopting, and managing AI? How do financial institutions develop their AI so that it will adapt to new and potentially different populations (outside of the test and training data)?

Artificial Intelligences do an excellent job of interpolating (making predictions within the training data) and a terrible job of extrapolating. **AIs will not extrapolate to new populations in robust and consistent ways;** the fact that oftentimes some amount of transfer from an old population to a new one is possible does not mean that the transfer is guaranteed or even consistent across all members of the new population. Careful checking and monitoring is necessary for applying Artificial Intelligence models to new settings. (The rare exception is if a causal model of the system is learned, e.g., once one has learned the physics of a pendulum, one can extrapolate to pendulums of different lengths.)

The corollary is that if one expects to apply the AI to a broad population, then the training data must be similarly broad. Examples: Apple facial recognition working poorly for people with darker skin. From a regulatory perspective, it may make sense to have requirements that an AI

[11] See on a much deeper technical level of analysis emerging from language models, notes on normative concerns https://dl.acm.org/doi/10.1145/3442188.3445922
[12] See https://datanutrition.org/

perform similarly on diverse groups, or other measures of fairness, to encourage the collection of appropriately broad datasets.

When it comes to finding effective ways to build robust, anti-discriminatory models, we also point to the fact that democratizing data exploration can be very helpful. In our field, Project MIMIC started in (1992-1999) to "build a collection of multi-parameter recordings of ICU patients."13 Its latest iteration is MIMIC-VI. "It is a large, publicly-available database comprising de-identified health-related data associated with approximately sixty thousand admissions of patients.." (See more details in the Appendix on MIMIC), and now augmented with E-ICU, which contains data across multiple hospitals. Note as well the PhysioNet14 which collects databases under 3 possible access levels (Open, Restricted & Credentialed) [15]in a single place (https://physionet.org/about/database/). The AI and health community has used these data to identify effective algorithms for a large variety of clinical tasks, including how to generalize across hospitals.

Besides general approaches to avoid overfitting, we suggest that such an approach may be valuable in the financial sector. Federal Financial Institutions Examination Council's (the 'Council") is already the "formal interagency body empowered to prescribe uniform principles, standards, and report forms for the federal examination of financial institution,"[16] therefore, its agenda-setting, coordination and convening power give it a responsibility to make sure systematic bias does not go unnoticed by the Agencies. In fact, as early as 2009, it was under the auspices of the FFIEC (74 FR 25240) that determinations about added disclosures from *foreign* banks operating in the US was done.[17] It makes sense that this kind of broad convening power can be harnessed to the cause of making sure the United States financial system does not discriminate against its own people. **A MIMIC-type dataset with anonymized individual-level data has provided a lot to AI researchers in healthcare, and the Agencies have a great opportunity to enhance the trust of the American people in its banking institutions by providing the academic community with similar resources to investigate and measure negative externalities.**

6. Cybersecurity Risk: No comments from our lab.

[13] See https://archive.physionet.org/physiobank/database/mimicdb/ MIMIC-I
[14] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **101**(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13).
[15] Open Access: Accessible by all users, with minimal restrictions on reuse.
Restricted Access: Accessible by registered users who sign a Data Use Agreement. Credentialed Access: Accessible by registered users who complete the credentialing process and sign a Data Use Agreement
[16] See https://www.ffiec.gov/
[17] Namely, it extended the comment period on the "currently approved information collection, the Country Exposure Report for U.S. Branches and Agencies of Foreign Banks (FFIEC 019)." See https://www.ffiec.gov/PDF/FFIEC_forms/FFIEC019_20090812_ffr.pdf

7. Dynamic Updating: Internal & External Audits in AI Model lifecycle with revisions to SR Letter 11-7 A1 on invariances & fallback models

7.1. Question 8: Continuous monitoring is needed to mitigate the risks of dynamic updating. SR Letter 11-7 A1 could benefit from guidance on 'invariances,' not just 'anomalies.' Fall-back models might be important, as well as clarity on penalty mechanisms.

How do financial institutions manage AI risks relating to dynamic updating? Describe any barriers or challenges that may impede the use of AI that involve dynamic updating. How do financial institutions gain an understanding of whether AI approaches producing different outputs over time based on the same inputs are operating as intended?

Dynamic updating poses significant risks. While continuous internal auditing should be part of an AI model's maintenance, AI models will also likely need to externally audit to ensure that outcomes remain as desired. If undesirable outcomes are observed, then an internal group would be required to fix them, which may involve temporarily falling back to another, perhaps simpler, model. In cases where the outcomes are clear, this approach reduces the need for full technical transparency. More specifically:

AI Model auditing and related governance structure. Broadly, AI performance will change over time not only because the AI may be updated but also because the data streams will change (e.g., Google flu trends, Netflix prize). Whether it is an expected change, that one can do rigorous testing in advance, or whether it is a change due to shifts in data properties, continuous monitoring is essential, as suggested in SR-Letter 11-7 from April 2011.[18] The guidance expressly requires that there should be internal mechanisms within an organization to perform regular audits, and the need for regular external audits to ensure rigor, consistency, and keep everyone honest.

However, the SR-Letter 11-7 might benefit from further clarification on invariances. These audits should look for "invariances" e.g. performance that should be met such as a safety level and not just for "anomalies" e.g. any cases that are outside the norm for that model or data, or as the guidance suggests pure "conceptual soundness." **More importantly, SR-Letter 11-7 guidance does not suggest any penalty mechanisms or even what an infraction of these audit principles could entail.** There should be an escalation in penalties where organizations initially have some time to fix an issue—as issues will happen—but issues are not allowed to remain. More concretely, we recommend that regulators look at Professor Watcher's work on using counter-factual explanations that can avoid opening the AI model's black box.[19]

Another piece missing from SR-Letter 11-7 is the need for Fall-back AI Models. In most cases, in case of violation, there will likely be a quick fix that is not ideal – e.g., rolling back to an older version of the AI or replacing the AI with a much simpler algorithm that provides basic functionality – and then the organization will be able to take steps to rectify the problem in a way

[18] As per Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, April 4th 2011, "Supervisory Guidance on Model Risk Management," SR Letter 11-7 and particularly the related appendix attachment A1, sections V and VI.
[19] See https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

that gives the extra functionality (e.g., by collecting more data). Finally, this does not imply that regulators should be overly prescriptive on AI model remedies. It is important that audit processes focus on the outcomes and leave the fixes to the organization.

Use and collect outcome data has proved foundational, for example. in health AI improvements, without an emphasis on specific technologies. Finally, we note that it is important to keep focus on the outcomes that are acceptable and unacceptable, rather than specific models or data collection technologies. The latter change very quickly; a regulation made prior to Fitbits and Apple watches, for example, may not have imagined the kind of personal data that is suddenly easy to collect, but one can and should foresee those certain types of decisions should be made independently of a person's comorbidities (regardless of how those health variables might be detected), independently of their race, etc. **In many cases this will mean, *it will be necessary to collect data on sensitive variables*** to ensure that systems are not building proxies for them based on ever increasingly sophisticated data streams.

8. AI Use by Community Institutions: No answer

9. Oversight of Third Parties: Need for AI "Check Engine" Light

9.1. Question 10: Third Parties need to provide significant information about the training data, metrics, and other audit mechanisms. Current research could be leveraged to create AI model's on-board diagnostics, or an AI mode "Check Engine light," so AI models do not "fail silently."

Please describe any particular challenges or impediments financial institutions face in using AI developed or provided by third parties and a description of how financial institutions manage the associated risks. Please provide detail on any challenges or impediments. How do those challenges or impediments vary by financial institution size and complexity?

Especially in safety-critical domains, such as our work in health, models failing silently is a major danger. Whether bought by a third-party or not, we need (a) the same level of transparency e.g. what data sheets on how the model was trained as one might give to an external auditor, (b) a set of diagnostic suites akin to an version of an AI "Check Engine" light. These would include dashboards for pre-specified outcomes to monitor, the ability to add more items to monitor, and an agreement on how to adjudicate undesired performance. We note that there is ample precedent for federal regulation of "on-board diagnostics," or Malfunction-indicator lamps MIL.[20]

AI "Check Engine" light regulatory framework. We understand that regulators are likely already aware of recent regulatory capture research[21] that suggests that overly complex regulatory frameworks can create perverse incentive. This would make any additional regulatory requirements favor larger institutions over smaller ones who cannot afford the additional compliance. Therefore, we understand the focus of FDIC's applicability of the Fed/OCC SR Letter 11-7, "Supervisory

[20] See EPA 2003, On-Board Diagnostic (OBD) Regulations and Requirements: Questions and Answer see https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100LW9G.txt for short overview, and factsheet EPA 1997 "Environmental Fact Sheet Frequently Asked Questions About On-Board Diagnostics" https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P1009Z15.txt

[21] For example, see from University of Chicago Booth, Luigi Zingales (2014) "A Capitalism for the People: Recapturing the Lost Genius of American Prosperity," on broad anti-trust regulatory reform theoretical proposals for technology companies; See Tim Wu (2018) "The Curse of Bigness: Antitrust in the New Gilded Age," section on "The Rise of the Tech Trust on the side-effect."

Guidance on Model Risk Management" from 2011,[22] that emphasizes a cut-off for depository institutions above USD $1bi AUM.[23]

However, this makes FDIC's regulation of A.i. model's "check engine light" from third parties providing software to smaller depository institutions particularly important. These smaller institutions might be hard pressed to build any alternative internal solutions that would be compliant to these regulations. Therefore, making sure model users understand their models enough to make sense of its disparate or systemic negative impact will fall to how effective are explainability regulation of 3rd party A.i. model providers. By comparison, we do not expect car drivers to have a deep understanding of how their car works, but it would be negligent to not take their car in for repairs if the "check engine light" came on.[24]

10. Fair Lending: Aggregates are not substitute for explainability. The FDIC could lead the development of Data Donation Frameworks for CDFIs and MDIs under Mission-Driven Bank Fund.

10.1. Question 11: Not answered.

What techniques are available to facilitate or evaluate the compliance of AI-based credit determination approaches with fair lending laws or mitigate risks of non-compliance? Please explain these techniques and their objectives, limitations of those techniques, and how those techniques relate to fair lending legal requirements.

10.2. Question 12: Continuous monitoring and regular external audits are essential for identifying bias; internally both quantitative and explanation-based tools will be needed to identify and rectify issues.

What are the risks that AI can be biased and/or result in discrimination on prohibited bases? Are there effective ways to reduce risk of discrimination, whether during development, validation, revision, and/or use? What are some of the barriers to or limitations of those methods?

Continuously monitored metrics are key to check for bias in AI models. We also need explainability at the global level of the model overall and at the local level of the individual decision. Especially when trying to reduce the risk of discrimination during development and revision both are essential. Aggregate statistics can give raise red flags, but they do not point to solutions nor can they adjudicate individual cases. More broadly, research and best practices for building fair models could benefit from FDIC supporting the development of a CDFI and MDI data donation framework and documentation under the Mission-Driven Bank Fund.

Aggregate statistics give a useful summary for a certain concern about, for example, lending patterns to a category of individuals that could be labeled as victims of an AI-driven biased decision-making. That said, the design of these aggregates and alarms is tricky:

[22] As per Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, April 4th 2011, "Supervisory Guidance on Model Risk Management," SR Letter 11-7 and related appendix A1.
[23] Per FDIC's Financial Institution Letter FIL-22-2017 from June 7th, 2017; "Adoption of Supervisory Guidance on Model Risk Management"
[24] We note how similar governance structures are already in place for vehicle emissions, and how gaming such a system could pose significant costs to violating institutions, for example see 2021 Volkswagen usage of 'defeat devices'. https://www.epa.gov/vw/learn-about-volkswagen-violations

- What are the attributes that constitute a threshold for bias overall? In a given lending decision?
- Once the labeling is done, where is the threshold of the aggregate of enough instances to provide causal evidence?
- How sensitive are the alarms to the definitions of the aggregates?

Because AI models and the metrics to evaluate them are so complex, decisions about what statistics to monitor must be broad and with the understanding that one is seeking trends that may cause concern rather than meeting some simple threshold.

More concretely, systematic bias in decision making has already been proven ex-post to disproportionally impact minority communities, as recent research on racial discrimination in auto-lending[25] and access to credit vis-a-vi Minority Depositary Institutions[26] have shown. We urge regulators to not allow decades to pass before many local explanations for bias are aggregated to create a global systematic concern over disparate impact on marginalized communities. Here, inspections of the model globally and locally – in addition to the aggregates – may help identify concerns before the model is even deployed.

Finally, we advocate for ways for the community to build best practices as a whole. **We understand the FDIC's new diversity strategic plan outlining five "C"s – Culture, Career, Communication, Consistency, and Community[27] and for its efforts with the Mission-Driven Bank Fund,[28] and as it builds its operations, it might be important to consider how to provide technical and legal support for how minority depository institutions (MDIs) and Community Development Financial Institutions (CDFIs) can document and donate their data.** Given the FDIC's extensive data tools and API already in place, it puts itself in an ideal position to support this process.[29]

In particular, many entities are willing to undergo research in this essential issue, however these are sensitive data that might need to be anonymized among other various related legal issues given regulatory concerns. [30] In fact, it provides the FDIC with an opportunity to possibly support the expansion of its current data offerings to include diversity-related financial data.[31] This dataset could build a solid foundation for AI fairness research dedicated to remedy these gaps in the current academic understanding of the role MDIs and CDFIs play in combating systemic bias. In fact, allied with data donation documentation frameworks, it could set the financial industry standard for decades to come.

10.3. Question 13: Not answered.

10.4. Question 14: Not answered directly, see Q8 (AI model audits).

10.5. Question 15: Not answered directly, see Q3 (AI explainability in dispute resolution)

[25] See https://bcf.princeton.edu/wp-content/uploads/2020/11/Racial-Discrimination-in-the-Auto-Market-9-10-2020.pdf
[26] For their technological challenges see https://bcf.princeton.edu/wp-content/uploads/2020/11/MDI-9-10-2020.pdf
[27] See https://www.fdic.gov/news/press-releases/2021/pr21016.html.
[28] See https://www.fdic.gov/news/press-releases/2020/pr20125.html
[29] See https://www.fdic.gov/resources/data-tools/
[30] For a brief overview of data donations in healthcare https://blogs.ischool.berkeley.edu/w231/blog/
[31] "MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over 40,000 patients…" See https://physionet.org/content/mimiciii-demo/

11. Additional Considerations: Broader & Better Data Collection, see Q6 and "Overfitting" section 5.

12. Conclusion & Action Agenda

We have made many comments in this document. Most importantly, we suggest that the current regulatory framework under review could benefit from a more practical definition of explainability, while the Agencies could use recent research to better define standards for the continuous monitoring of AI. Leveraging the current research, **a useful regulatory framework to consider would be defining standards for AI model's on-board diagnostics, or an AI model's "Check Engine" light**

Explainability: Trade-offs based on how well defined are model goals

Q3: Explainability is needed in cases where metrics are not enough, such as identifying the overall workings of a model, preventing or rectifying errors, and resolving disputes

Risks from Broader or More Intensive Data Processing and Usage: Dataset Documentation, for example Data Nutrition Project

Q4: Data is one of the biggest sources of AI error; transparency about the data sources is critical for accountability. See also Q8 (AI Model audits)

Q5: As we gather more alternative data, we must also gather data about sensitive variables to ensure we are not creating proxies for them. See also Q8 (model audits)

Overfitting: Better incentives towards broader data collection & publication i.e. MIMIC for Finance.

Q6: Continuous audits are needed to manage overfitting risks; the biggest risks are overfitting to a specific population used to train the model rather than the model itself. MIMIC-type project to democratize data.

Dynamic Updating: Internal & External Audits in AI Model lifecycle with revisions to SR Letter 11-7 A1 on invariances & fallback models

Q8: Continuous monitoring is needed to mitigate the risks of dynamic updating. SR Letter 11-7 A1 could benefit from guidance on 'invariances,' not just 'anomalies.' Fall-back models might be important, as well as clarity on penalty mechanisms.

Oversight of Third Parties: Need for AI "Check Engine" Light, so AI models do not "fail silently."

Q10: Third Parties need to provide significant information about the training data, metrics, and other audit mechanisms. Current research could be leveraged to create AI model's on-board diagnostics, or an AI mode "Check Engine light," so AI models do not "fail silently."

Fair Lending: Aggregates are not substitute for explainability. The FDIC could lead the development of Data Donation Frameworks for CDFIs and MDIs under Mission-Driven Bank Fund.

Q12: Continuous monitoring and regular external audits are essential for identifying bias; internally both quantitative and explanation-based tools will be needed to identify and rectify issues.

Bibliography

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. In S. Bengio, & H. M. Wallach (Eds.), *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 9525–9536). Red Hook, NY, US: Curran Associates Inc. Retrieved from https://papers.nips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf

Alkhatib, A., & Bernstein, M. (2019). Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), Paper 530*, 1-13. Retrieved from https://hci.stanford.edu/publications/2019/streetlevelalgorithms/streetlevelalgorithms-chi2019.pdf

Alvarez-Melis, D., Daumé, H., Vaughan, J. W., & Wallach, H. (2019). *Weight of Evidence as a Basis for Human-Oriented Explanations.* NeurIPS 2019 Workshop on on Human-Centric Machine Learning. Retrieved from https://arxiv.org/pdf/1910.13503.pdf

Amir, O., Doshi-Velez, F., & Sarne, D. (2018, July 9). Agent strategy summarization. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1203-1207. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6255044771133571112

Amir, O., Doshi-Velez, F., & Sarne, D. (2019, September 1). Summarizing agent strategies. *Autonomous Agents and Multi-Agent Systems, 33*(5), 628-644. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17814175098339224659

André, P., Kittur, A., & Dow, S. P. (2014). *Crowd Synthesis: Extracting Categories and Clusters from Complex Data.* ACM. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.9116&rep=rep1&type=pdf

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/3442188.3445922

Buolamwini, J., & Gebru, T. (2019). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency*, 77-91. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Byrne, R. M. (2019). *Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning.* IJCAI. Retrieved from https://www.ijcai.org/proceedings/2019/0876.pdf

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., . . . Terry, M. (2019, April). Human-centered tools for coping with imperfect algorithms during medical decision-making. Retrieved from https://arxiv.org/ftp/arxiv/papers/1902/1902.02960.pdf

Cai, C., Guo, P. J., Glass, J. R., & Miller, R. C. (2015). *Wait-Learning: Leveraging Wait Time for Second Language Education.* ACM. Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/112662/Miller_Wait%20Learning.pdf?sequence=1&isAllowed=y

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., & Olah, C. (2019, March 6). *Exploring Neural Networks with Activation Atlases.* Retrieved from Distill: https://distill.pub/2019/activation-atlas/

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In L. Cao, C. Zhang, T. Joachims, G. Webb, D. D. Margineantu, & G. Williams, *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). New York, NY, US: Association for Computing Machinery. Retrieved from https://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf

Cheng, J., & Bernstein, M. S. (2015). *Flock: Hybrid Crowd-Machine Learning Classifiers.* Stanford University. ACM. Retrieved from https://hci.stanford.edu/publications/2015/Flock/flock_paper.pdf

Cook, C., Bregar, W., & Foote, D. (1984). A Preliminary Investigation of the Use of the Cloze Procedure as a Measure of Program Understanding. *Information Processing and Management, 20*(1), 199-208. Retrieved from https://pdf.sciencedirectassets.com/271647/1-s2.0-S0306457300X01537/1-s2.0-0306457384900505/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEF0aCXVzLWVhc3QtMSJIMEYCIQD0rtc7WtpjaVqknyNnmtGpUDQ%2Fmrv2ZDTvaJyZa9y%2BygIhAJJt6NO7xy2hp6sMbfs7xArmRU0rYUyQVLi4GZClU

Davis, N., Hsiao, C.-P., Popova, Y., & Magerko, B. (2015). Chapter 7: An Enactive Model of Creativity for Computational Collaboration and Co-creation. In N. Zagalo, & P. Branco (Eds.), *Creativity in the Digital Age* (pp. 109-33). Springer. Retrieved from https://link-springer-com.ezp-prod1.hul.harvard.edu/content/pdf/10.1007%2F978-1-4471-6681-8.pdf

Delalande, F. (2007, December). Towards an analysis of compositional strategies. *Circuit Musiques contemporaines, 17*(1), 11-26. Retrieved from https://www.erudit.org/fr/revues/circuit/2007-v17-n1-circuit1896/016771ar.pdf

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2016, May 23). Learning and policy search in stochastic dynamical systems with bayesian neural networks. *arXiv preprint arXiv:1605.07127*, 1-14. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14568373457880481181

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2017, November). *Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4564853263001192019

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., & Udluft, S. (2018, July 3). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. *Proceedings of the 35th International Conference on Machine Learning, 80*, 1184-1193. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13563599882871713230

Doshi, F., & Roy, N. (2008, May 12). The permutable POMDP: fast solutions to POMDPs for preference elicitation. *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, 1*, 493–500. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13547905812581203405

Doshi, F., Brunskill, E., Shkolnik, A., Kollar, T., Rohanimanesh, K., Tedrake, R., & Roy, N. (2007, October). Collision detection in legged locomotion using supervised learning. *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 317-322. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4414783219246469999

Doshi, F., Wingate, D., Tenenbaum, J. B., & Roy, N. (2011, June 28). Infinite dynamic Bayesian networks. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 913–920.

Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6524581654986168933

Doshi-Velez, F. (2009). *The Indian buffet process: Scalable inference and extensions.* University of Cambridge. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9215175749629133787

Doshi-Velez, F. (2009, December 7). The infinite partially observable Markov decision process. *Proceedings of the 22nd International Conference on Neural Information Processing Systems, 22*, 477-485. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10768110427383167189

Doshi-Velez, F., & Ghahramani, Z. (2009, June 14). Accelerated sampling for the Indian buffet process. *Proceedings of the 26th annual international conference on machine learning*, 273-280. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15310891466322889089

Doshi-Velez, F., & Ghahramani, Z. (2009, June 18). Correlated non-parametric latent feature models. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 143–150. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=7094049166815964911

Doshi-Velez, F., & Ghahramani, Z. (2011, July). A comparison of human and agent reinforcement learning in partially observable domains. *Proceedings of the Annual Meeting of the Cognitive Science Society, 33*(33), 2703-2708. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6703778567975104250

Doshi-Velez, F., & Kim, B. (2017, March). *A roadmap for a rigorous science of interpretability.* arXiv. Retrieved from https://arxiv.org/pdf/1702.08608.pdf

Doshi-Velez, F., & Kim, B. (2017, February 28). *Towards A Rigorous Science of Interpretable Machine Learning.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8789025022351052485

Doshi-Velez, F., & Kim, B. (2018). Considerations for evaluation and generalization in interpretable machine learning. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. v. Gerven, *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 3-17). Springer Nature. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=12095494514140397496

Doshi-Velez, F., & Konidaris, G. (2016, July 9). Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 1432–1440. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3270924689080010306

Doshi-Velez, F., & Perlis, R. H. (2019, November 12). Evaluating machine learning articles. *Journal of the American Medical Association, 322*(18), 1777-1779. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15452153815990615586

Doshi-Velez, F., & Roy, N. (2007, March 10). Efficient model learning for dialog management. *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, 65-72. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5759348188893787326

Doshi-Velez, F., & Roy, N. (2008, December 1). Spoken language interaction with model uncertainty: an adaptive human–robot interaction system. *Connection Science, 20*(4), 299-318. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17586990007894260810

Doshi-Velez, F., & Williamson, S. A. (2017, September 1). Restricted Indian buffet processes. *Statistics and Computing, 27*(5), 1205-1223. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=14673419880409111956

Doshi-Velez, F., Avillach, P., Palmer, N., Bousvaros, A., Ge, Y., Fox, K., . . . Kohane, I. (2015, October 1). Prevalence of inflammatory bowel disease among patients with autism spectrum disorders. *Inflammatory bowel diseases, 21*(10), 2281–2288. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9787709197227402824

Doshi-Velez, F., Ge, Y., & Kohane, I. (2014, January). Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics, 133*(1), e54-e63. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8952823131498776530

Doshi-Velez, F., Kortz, M., Budish, R., Chris Bavitz, S. G., O'Brien, D., Scott, K., . . . Wood, A. (2017, November 3). *Accountability of AI under the law: The role of explanation.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13535939933778439444

Doshi-Velez, F., Li, W., Battat, Y., Charrow, B., Curthis, D., Park, J.-g., . . . Teller, S. (2012, July 1). Improving safety and operational efficiency in residential care settings with WiFi-based localization. *Journal of the American Medical Directors Association, 13*(6), 558-563. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16620197873028630726

Doshi-Velez, F., Miller, K., Gael, J. V., & Teh, Y. W. (2009, April 15). Variational inference for the Indian buffet process. *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, 5*, 137-144. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=12982039394924101433

Doshi-Velez, F., Mohamed, S., Ghahramani, Z., & Knowles, D. A. (2009, December 7). Large scale nonparametric Bayesian inference: Data parallelisation in the Indian buffet process. *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 1294-1302. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8022303325169102009

Doshi-Velez, F., Pfau, D., Wood, F., & Roy, N. (2013, October 1). Bayesian nonparametric methods for partially-observable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 37*(2), 394-407. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9789702462404251311

Doshi-Velez, F., Pineau, J., & Roy, N. (2008, July 5). Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs. *Proceedings of the 25th international conference on Machine learning*, 256-263. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4585349008724983456

Doshi-Velez, F., Wallace, B., & Adams, R. (2015, January 25). Graph-sparse lda: a topic model with structured sparsity. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2575–2581. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16621623251555081538

Du, J., Futoma, J., & Doshi-Velez, F. (2020). *Model-based reinforcement learning for semi-markov decision processes with neural odes.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6882030783154485592

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 263-74. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/3301275.3302316

Elibol, H. M., Nguyen, V., Linderman, S., Johnson, M., Hashmi, A., & Doshi-Velez, F. (2016, January 1). Cross-corpora unsupervised learning of trajectories in autism spectrum disorders. *The Journal of Machine Learning Research, 17*(1), 4597–4634. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16607715928057211631

Elmalech, A., Sarne, D., Rosenfeld, A., & Erez, E. S. (2015). When Suboptimal Rules. *In proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*, 1313-19. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9931/9322

Entin, E. B. (1984). Using the cloze procedure to assess program reading comprehension. *Proceedings of the fifteenth SIGCSE technical symposium on Computer science education*. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/952980.808621

Fan, A., Doshi-Velez, F., & Miratrix, L. (2017). *Promoting domain-specific terms in topic models with informative priors.* arXiv. Retrieved from https://www.semanticscholar.org/paper/Promoting-Domain-Specific-Terms-in-Topic-Models-Fan-Doshi-Velez/7f992c8ea80b7ee9640d67be92f377ee11cd01a1

Fan, A., Doshi-Velez, F., & Miratrix, L. (2019, June). Assessing topic model relevance: Evaluation and informative priors. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 12*(3), 210-222. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4628884776712761559

Fang, F., Nguyen, T. H., Pickles, R., Lam, W. Y., Clements, G. R., An, B., . . . Lemieux, A. (2016). Deploying PAWS: Field Optimization of the Protection Assistant for Wildlife Security. *Proceedings of the Twenty-Eighth AAAI Conference on Innovative Applications (IAAI-16)*, 3966-73. Retrieved from https://www.cais.usc.edu/wp-content/uploads/2017/07/Fang-et-al-IAAI16_PAWS-1.pdf

Fast, E., Steffee, D., Wang, L., Brandt, J., & Bernstein, M. S. (2014). *Emergent, Crowd-scale Programming Practice in the IDE.* ACM. Retrieved from https://hci.stanford.edu/publications/2014/Codex/codex-paper.pdf

Fleischhauer, M., Enge, S., Brocke, B., Ullrich, J., Strobel, A., & Strobel, A. (2010, January). Same or Different? Clarifying the Relationship of Need for Cognition to Personality and Intelligence. *Personality and Social Psychology Bulletin, 36*(1), 82-96. Retrieved from https://journals-sagepub-com.ezp-prod1.hul.harvard.edu/doi/pdf/10.1177/0146167209351886

Furnham, A., & Allass, K. (1999). The Influence of Musical Distraction of Varying Complexity on the Cognitive Performance of Extroverts and Introverts. *European Journal of Personality, 13*, 27-38. Retrieved from http://diyhpl.us/~bryan/papers2/neuro/music-distraction/The%20influence%20of%20musical%20distraction%20of%20varying%20complexity%20on%20the%20cognitive%20performance%20of%20extroverts%20and%20introverts%20-%201999.pdf

Futoma, J., Hughes, M. C., & Doshi-Velez, F. (2020, August). Popcorn: Partially observed prediction constrained reinforcement learning. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 108*, 3578-3588. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2544924681479461357

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020, September). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health, 2*(9), e489-e492. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17128416078609966243

Gafford, J., Doshi-Velez, F., Wood, R., & Walsh, C. (2016, September 1). Machine learning approaches to environmental disturbance rejection in multi-axis optoelectronic force sensors. *Sensors and Actuators A: Physical, 248*, 78-87. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1497978661435603804

Galhotra, S., Brun, Y., & Meliou, A. (2017). *Fairness Testing: Testing Software for Discrimination.* ACM. Retrieved from https://people.cs.umass.edu/~brun/pubs/pubs/Galhotra17fse.pdf

Gao, T., Dontcheva, M., Adar, E., Liu, L. Z., & Karahalios, K. (2015). DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. 489-500. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2807442.2807478

Garcia, J., Tsandilas, T., Agon, C., & Mackay, W. E. (2014, June). Structured Observation with Polyphony: a Multifaceted Tool for Studying Music Composition. *DIS: Conference on Designing Interactive Systems*, 199-208. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2598510.2598512

Garrod, S., & Pickering, M. J. (2004, January). Why is conversation so easy? *Trends in Cognitive Sciences, 8*(1), 8-11. Retrieved from https://pdf.sciencedirectassets.com/271877/1-s2.0-S1364661300X00733/1-s2.0-S136466130300295X/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEFwaCXVzLWVhc3QtMSJGMEQCIBwASnopfZp%2BvIcqeFP%2BsDaBk%2FAufCt6XSqKTc8dUzRjAiAUI6Kebuo7Vv%2BOxEidnhqi8VpRw8ZPMq8LO4LY

Geramifard, A., Doshi-Velez, F., Redding, J., Roy, N., & How, J. P. (2011, June 28). Online discovery of feature dependencies. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 881-888. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5757284957182508738

Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., & Szolovits, P. (2014, August 24). Unfolding physiological state: Mortality modelling in intensive care units. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 75-84. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5798843096311556054

Ghassemi, M., Wu, M., Hughes, M. C., Szolovits, P., & Doshi-Velez, F. (2017, July 26). Predicting intervention onset in the ICU with switching state space models. *AMIA Summits on Translational Science Proceedings, 2017*, 82-91. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9563364752254266911

Ghosh, S., & Doshi-Velez, F. (2017). *Model selection in Bayesian neural networks via horseshoe priors.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9439867866709144171

Ghosh, S., Yao, J., & Doshi-Velez, F. (2018, July 10). Structured variational learning of Bayesian neural networks with horseshoe priors. *Proceedings of the 35th International Conference on Machine Learning, 80*, 1744-1753. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9128418694635359827

Ghosh, S., Yao, J., & Doshi-Velez, F. (2019). Model selection in Bayesian neural networks via horseshoe priors. *Journal of Machine Learning Research, 20*(182), 1-46. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13995759821761819294

Glowacka, D., Ruotsalo, T., Konyushkova, K., Athukorala, K., Kaski, S., & Jacucci, G. (2013). *Directing Exploratory Search: Reinforcement Learning from User Interactions with Keywords.* ACM. Retrieved from https://www.cs.helsinki.fi/u/jacucci/directing.pdf

Goldstein, D. G., Mcafee, R. P., & Suri, S. (2014, June). The Wisdom of Smaller, Smarter Crowds. *Proceedings of the fifteenth ACM conference on Economics and computation*, 471-488. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2600057.2602886

Gottesman, O., Futoma, J., Liu, Y., Parbhoo, S., Celi, L., Brunskill, E., & Doshi-Velez, F. (2020, November 21). Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. *Proceedings of the 37th International Conference on Machine Learning, 119*, 3658-3667. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3979059661142155029

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (2019, January 7). Guidelines for reinforcement learning in healthcare. *Nature medicine, 25*, 16-18. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=970534608763260270

Gottesman, O., Liu, Y., Sussex, S., Brunskill, E., & Doshi-Velez, F. (2019, May 24). Proceedings of the 36th International Conference on Machine Learning. *International Conference on Machine Learning, 97*, 2366-2375. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5066391292071299163

Green, S., Heer, J., & Manning, C. D. (2013). *The Efficacy of Human Post-Editing for Language Translation.* ACM. Retrieved from http://vis.stanford.edu/files/2013-PostEditing-CHI.pdf

Grice, P. (1975). Logic and Conversation. In P. Grice, *Syntax and Semantics* (pp. 41-48). Harvard University Press. Retrieved from https://courses.media.mit.edu/2005spring/mas962/Grice.pdf

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010, August 1). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357-364. Retrieved from https://pdf.sciencedirectassets.com/271877/1-s2.0-S1364661310X00079/1-s2.0-S1364661310001129/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEF0aCXVzLWVhc3QtMSJGMEQCIGr7J6eXF2ag11Yh5P6mO2A7%2BCKDFtqo1EbiPSEpRbYZAiAbqD0528pw6BfX8wJIIEmMBwG%2BjP8UOMnfekQGfWuz

Guzzi, A., Bacchelli, A., Riche, Y., & Deursen, A. v. (2015, February). Supporting Developers' Coordination in The IDE. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 518-32. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2675133.2675177

Hara, K., Sun, J., Moore, R., Jacobs, D., & Froehlich, J. E. (2014). *Tohme: Detecting Curb Ramps in Google Street View Using Crowdsourcing, Computer Vision, and Machine Learning.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2642918.2647403

Harrison, B., & Riedl, M. O. (2016). *Learning From Stories: Using Crowdsourced Narratives to Train Virtual Agents.* Burlingame, California: AAAI. Retrieved from https://www.cc.gatech.edu/~riedl/pubs/harrison-aiide16.pdf

Hayes, B. K., Hawkins, G. E., & Newell, B. R. (2015). Why do people fail to consider alternative hypotheses in judgments under uncertainty? (R. P. Cooper, Ed.) *Cognitive Science*, 890-5. Retrieved from https://cogsci.mindmodeling.org/2015/papers/0160/paper0160.pdf

Hilgard, S., Rosenfeld, N., Banaji, M., Cao, J., & Parkes, D. C. (2020). *Learning Representations by Humans, for Humans.* arXiv. Retrieved from https://arxiv.org/pdf/1905.12686.pdf

Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 579-91. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/3290605.3300809

Hoque, E., & Carenini, G. (2015). ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations. *Proceedings of the 20th International Conference on Intelligent User Interfaces., 2015*, 169-180. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2678025.2701370

Horvitz, E. (1999). *Principles of Mixed-Initiative User Interfaces.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/302979.303030

Hottelier, T., Bodik, R., & Ryokai, K. (2014). *Programming by Manipulation for Layout.* Technical Report No. UCB/EECS-2014-161, University of California at Berkeley. Retrieved from https://www2.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-161.pdf

Hudson, S. E., & Mankoff, J. (2014). Concepts, Values, and Methods for Technical. In J. S. Olson, & W. A. Kellogg (Eds.), *Ways of Knowing in HCI* (pp. 69-93). Springer. Retrieved from https://link-springer-com.ezp-prod1.hul.harvard.edu/content/pdf/10.1007%2F978-1-4939-0378-8.pdf

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021, February 4). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry, 11*(1), 1-9. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9602941028157969885

Jain, A., Lupfer, N., Qu, Y., Linder, R., Kerne, A., & Smith, S. M. (2015). *Evaluating TweetBubble with Ideation Metrics of Exploratory Browsing.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2757226.2757239

Jenna Wiens, S. S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., . . . Goldenberg, A. (2019, September). Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine, 25*(9), 1337-1340. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1921488170773999899

Jeuris, S., Houben, S., & Bardram, J. (2014). Laevo: A Temporal Desktop Interface for Integrated Knowledge Work. *Proceedings of the 27th annual ACM symposium on User*, 1-10. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2642918.2647391

Jin, L., Doshi-Velez, F., Miller, T., Schuler, W., & Schwartz, L. (2018, October). Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2721–2731. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=8591212071315725016

Jin, L., Doshi-Velez, F., Miller, T., Schuler, W., & Schwartz, L. (2018, April 1). Unsupervised grammar induction with depth-bounded PCFG. *Transactions of the Association for Computational Linguistics, 6*, 211-

224. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2125035178814160197

Jin, L., Doshi-Velez, F., Miller, T., Schwartz, L., & Schuler, W. (2019, July). Unsupervised learning of PCFGs with normalizing flow. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2442–2452. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=9683284076824164052

Joseph, J., Doshi-Velez, F., & Roy, N. (2012, May 14). A Bayesian nonparametric approach to modeling battery health. *2012 IEEE International Conference on Robotics and Automation*, 1876-1882. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11238423758068122176

Joseph, J., Doshi-Velez, F., Huang, A. S., & Roy, N. (2011, November 1). A Bayesian nonparametric approach to modeling motion patterns. *Autonomous Robots, 31*(4), 383-400. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11402996606679976479

Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., & Doshi-Velez, F. (2019). Explainable reinforcement learning via reward decomposition. *Proceedings at the International Joint Conference on Artificial Intelligence*, 1-7. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=4762608020035398667

Kane, S. K., Bigham, J. P., & Wobbrock., J. O. (2008). Slide rule: making mobile touch screens accessible to blind people using multi-touch interaction techniques. *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, 73-80. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/1414471.1414487

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). *When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2858036.2858558

Kay, M., Nelson, G. L., & Hekler, E. B. (2016). *Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI.* ACM. Retrieved from http://www.mjskay.com/papers/chi_2016_bayes.pdf

Keyes, O. (2018). *The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/3274357

Killian, T., Konidaris, G., & Doshi-Velez, F. (2017, December 4). Robust and efficient transfer learning with hidden parameter markov decision processes. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6251–6262. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6997382761578063659

Kim, B., Rudin, C., & Shah, J. (2014). The Bayesian case model: a generative approach for case-based reasoning and prototype classification. In Z. Ghahramani, M. Welling, & C. Cortes (Eds.), *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems* (Vol. 2, pp. 1952–1960). Cambridge, MA, US: MIT Press. Retrieved from https://proceedings.neurips.cc/paper/2014/hash/390e982518a50e280d8e2b535462ec1f-Abstract.html

Kim, B., Shah, J. A., & Doshi-Velez, F. (2015, December 7). Mind the gap: A generative approach to interpretable feature selection and extraction. *Proceedings of the 28th International Conference on Neural*

Information Processing Systems, 2, 2260–2268. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5774689255374329461

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)*. arXiv. Retrieved from https://arxiv.org/pdf/1711.11279.pdf

Kim, E., & Schneider, O. (2020). *Defining Haptic Experience: Foundations for Understanding, Communicating, and Evaluating HX*. ACM. Retrieved from https://uwspace.uwaterloo.ca/bitstream/handle/10012/15721/DefiningHX_2019_CopyrightUpdate.pdf?sequence=3

Kimura, K., Hunley, S., & Namy, L. L. (2015). Comparison and Function in Children's Object Categorization. (R. P. Cooper, Ed.) *Cognitive Science*, 1105-10. Retrieved from https://cogsci.mindmodeling.org/2015/papers/0196/paper0196.pdf

Kittur, A., Peters, A. M., Diriye, A., & Bove, M. R. (2014). *Standing on the Schemas of Giants: Socially Augmented Information Foraging*. ACM. Retrieved from https://drive.google.com/file/d/0B9jDvBKRgh6tS3JDN25zcjgtc0U/view?resourcekey=0-PXU3TmsaaREYvLwdoCx_fQ

Kleinberg, J., & Mullainathan, S. (2019). *Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability*. arXiv. Retrieved from https://arxiv.org/pdf/1809.04578.pdf

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In D. Precup, & Y. W. Teh (Eds.), *ICML'17: Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1885-1894). JMLR.org. Retrieved from https://arxiv.org/pdf/1703.04730.pdf

Kruschke, J. K. (2013). Bayesian Estimation Supersedes the t Test. *Journal of Experimental Psychology: General, 142*(2), 573-603. Retrieved from https://jkkweb.sitehost.iu.edu/articles/Kruschke2013JEPG.pdf

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., & Doshi-Velez, F. (2019). Human Evaluation of Models Built for Interpretability. In E. Law, & J. W. Vaughan (Eds.), *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7(1), pp. 59-67). Stevenson, WA, US: AAAI. Retrieved from https://ojs.aaai.org//index.php/HCOMP/article/view/5280

Lage, I., Ross, A. S., Kim, B., Gershman, S. J., & Doshi-Velez, F. (2018, December 3). Human-in-the-Loop Interpretability Prior. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 10180–10189. Retrieved from https://scholar.google.com/scholar?start=10&hl=en&as_sdt=0,22&cluster=6925363852455924380

Lakkaraju, H., & Rudin, C. (2017). Learning Cost-Effective and Interpretable Treatment Regimes. In A. Singh, & J. Zhu (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (Vol. 54, pp. 166-175). Fort Lauderdale, FL, US: Proceedings of Machine Learning Research. Retrieved from http://proceedings.mlr.press/v54/lakkaraju17a/lakkaraju17a.pdf

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable Decision Sets: A Joint Framework for Description and Prediction. In B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, & R. Rastogi, *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1675–1684). New York, NY, US: Association for Computing Machinery. Retrieved from https://www-cs-faculty.stanford.edu/people/jure/pubs/interpretable-kdd16.pdf

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and Customizable Explanations of Black Box Models. In V. Conitzer, G. Hadfield, & S. Vallor, *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 131–138). New York, NY, US: Association for Computing Machinery. Retrieved from https://web.stanford.edu/~himalv/customizable.pdf

Lawrance, J., Bellamy, R., Burnett, M., & Rector, K. (2008). Using Information Scent to Model the Dynamic Foraging Behavior of Programmers in Maintenance Tasks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*, 1-10. Retrieved from https://drive.google.com/file/d/0Bxoj7fgR-gOKVjFVR29RQzE0Z28/view?resourcekey=0-_FIoperANVQmDREWWoX_hA

Lee, D., Srinivasan, S., & Doshi-Velez, F. (2019). *Truly batch apprenticeship learning with deep successor features.* arXiv. Retrieved from https://arxiv.org/abs/1903.10077

Lee, K., Mahmud, J., Chen, J., Zhou, M., & Nichols, J. (2014). *Who Will Retweet This? Automatically Identifying and Engaging Strangers on Twitter to Spread Information.* IUI. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2557500.2557502

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015, September). Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. *Annals of Applied Statistics, 9*(3), 1350-1371. Retrieved from https://arxiv.org/pdf/1511.01644.pdf

Li, O., Liu, H., Chen, C., & Rudin, C. (2017). Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. In *AAAI-18: The Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 3530-3537). Association for the Advancement of Artificial Intelligence. Retrieved from https://arxiv.org/pdf/1710.04806.pdf

Liebman, E., Saar-Tsechansky, M., & Stone, P. (2015). *DJ-MC: A Reinforcement-Learning Agent for Music Playlist Recommendation.* arXiv. Retrieved from https://arxiv.org/pdf/1401.1880.pdf

Lin, C. H., Mausam, & Weld, D. S. (2014). *To Re(label), or Not To Re(label).* AAAI. Retrieved from https://homes.cs.washington.edu/~mausam/papers/hcomp14a.pdf

Lipton, Z. C. (2017). *The Mythos of Model Interpretability.* arXiv. Retrieved from https://arxiv.org/pdf/1606.03490.pdf

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A., Doshi-Velez, F., & Brunskill, E. (2018, December 3). Representation balancing mdps for off-policy policy evaluation. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2649–2658. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=11325221239507449364

Loepp, B., Hussein, T., & Ziegler, J. (2014). *Choice-Based Preference Elicitation for Collaborative Filtering Recommender Systems.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2556288.2557069

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In U. v. Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777). Red Hook, NY, US: Curran Associates Inc. Retrieved from https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Madras, D., Pitassi, T., & Zemel, R. (2018). *Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer.* NIPS. Retrieved from http://www.cs.toronto.edu/~zemel/documents/NIPS_Predict_Responsibly.pdf

Masood, M. A., & Doshi-Velez, F. (2019, May). A Particle-Based Variational Approach to Bayesian Non-negative Matrix Factorization. *Journal of Machine Learning Research, 20*, 1-56. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2027101452273906656

Masood, M. A., & Doshi-Velez, F. (2019). *Diversity-inducing policy gradient: Using maximum mean discrepancy to find a set of diverse policies.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=5089068101117259482

Milkman, K. L., & Berger, J. (2014). The science of sharing and the sharing of science. *Proceedings of the National Academy of Sciences, 111*(Supplement 4), 13642-49. Retrieved from https://www-pnas-org.ezp-prod1.hul.harvard.edu/content/111/Supplement_4/13642.short

Mutlu, B., & Forlizzi, J. (2008). *Robots in organizations: The role of workflow, social, and environmental factors in human-robot interaction.* IEEE. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/1349822.1349860

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018, February 2). *How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=12029304911284070268

Nickerson, D. W. (2008, February). Is Voting Contagious? Evidence from Two Field Experiments. *American Political Science Review, 102*(1), 49-57. Retrieved from https://sites.temple.edu/nickerson/files/2017/07/nickerson.contagion.pdf

Omer Gottesman, e. (2018). *Evaluating reinforcement learning algorithms in observational health settings.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=7539099852161268532

Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., & Freeman, W. T. (2015). *Visually Indicated Sounds.* arXiv. Retrieved from https://arxiv.org/pdf/1512.08512.pdf

Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V., & Doshi-Velez, F. (2017, July 26). Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings, 2017*, 239–248. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=657106867245270507

Parbhoo, S., Gottesman, O., Ross, A. S., Komorowski, M., Faisal, A., Bon, I., . . . Doshi-Velez, F. (2018, November 12). Improving counterfactual reasoning with kernelised dynamic mixing models. *PLOS One, 13*(11), e0205839. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15395330833371325776

Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Lehman, L.-w. H., . . . Doshi-Velez, F. (2018, December 5). Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. *AMIA Annual Symposium Proceedings, 2018*, 887–896. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=3048790530901914801

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn,

& S. Drucker, *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-52). New York, NY, US: Association for Computing Machinery. Retrieved from https://arxiv.org/pdf/1802.07810.pdf

Pradier, M. F., Jr, T. H., Hughes, M., Perlis, R. H., & Doshi-Velez, F. (2020, February 6). Predicting treatment dropout after antidepressant initiation. *Translational psychiatry, 10*(1), 1-8. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2188652166281340500

Pradier, M. F., Pan, W., Yao, J., Ghosh, S., & Doshi-Velez, F. (2019, November). *Latent projection bnns: Avoiding weight-space pathologies by learning latent representations of neural network weights.* arXiv. Retrieved from https://deepai.org/publication/latent-projection-bnns-avoiding-weight-space-pathologies-by-learning-latent-representations-of-neural-network-weights

Raghu, A., Gottesman, O., Liu, Y., Komorowski, M., Faisal, A., Doshi-Velez, F., & Brunskill, E. (2018). *Behaviour policy estimation in off-policy policy evaluation: Calibration matters.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2553179708177407291

Reda, K., Johnson, A. E., Papka, M. E., & Leigh, J. (2015). *Effects of Display Size and Resolution on User Behavior and Insight Acquisition in Visual Exploration.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2702123.2702406

Retelny, D., Robaszkiewicz, S., To, A., Lasecki, W., Patel, J., Rahmati, N., . . . Bernstein, M. S. (2014, October). Expert Crowdsourcing with Flash Teams. *UIST '14: Proceedings of the 27th annual ACM symposium on User interface software and technology*, 75-85. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2642918.2647409

Ribeiro, M., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. In J. DeNero, M. Finlayson, & S. Reddy (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 97–101). San Diego, CA, US: Association for Computational Linguistics. Retrieved from https://arxiv.org/pdf/1602.04938.pdf

Romero, D. M., Huttenlocher, D., & Kleinberg, J. (2015). *Coordination and Efficiency in Decentralized Collaboration.* arXiv. Retrieved from https://arxiv.org/pdf/1503.07431.pdf

Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017, August 19). Right for the right reasons: Training differentiable models by constraining their explanations. *Proceedings of the 26th International Joint Conference on Artificial IntelligenceAugust 2017*, 2662–2670. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1999112949296082528

Ross, A., & Doshi-Velez, F. (2018, April 25). Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1), 1660-1669. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10549843532884126759

Rudin, C. (2019, May). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. Retrieved from https://www.nature.com/articles/s42256-019-0048-x.pdf

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). *I Can Do Better Than Your AI: Expertise and Explanations.* ACM. Retrieved from https://sites.cs.ucsb.edu/~holl/pubs/Schaffer-2019-IUI.pdf

Selbst, A. D., Boyd, D., Friedler, S., Venkatasubramanian, S., & Vertesi, J. (2018). *Fairness and Abstraction in Sociotechnical Systems.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/3287560.3287598

Shahaf, D., Horvitz, E., & Mankoff, R. (2015). Inside Jokes: Identifying Humorous Cartoon Captions. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1065-74. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/2783258.2783388

Shklovski, I., Troshynski, E., & Dourish, P. (2009). *The commodification of location: Dynamics of power in location-based systems.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/1620545.1620548

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., & Findlater, L. (2020). *No Explainability without Accountability: An Empirical.* ACM. Retrieved from https://homes.cs.washington.edu/~wtshuang/static/papers/2020-chi-explain+feedback.pdf

Stock, O., & Strapparava, C. (2015). *Getting Serious about the Development of Computational Humor.* ACM. Retrieved from https://www.ijcai.org/Proceedings/03/Papers/009.pdf

Stock, O., Zancanaro, M., Rocchi, C., Tomasini, D., Koren, C., Eisikovits, Z., . . . Weiss, P. L. (2008). *A Co-Located Interface for Narration to Support Reconciliation in a Conflict: Initial Results from Jewish and Palestinian Youth.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/1357054.1357302

Tam, J., & Greenberg, S. (2006). A Framework for Asynchronous Change Awareness in Collaborative Documents and Workspaces. *International Journal of Human-Computer Studies, 64*(7), 583-98. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/10.1016/j.ijhcs.2006.02.004

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011, March 11). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science, 331*(6022), 1279-85. Retrieved from https://cocosci.princeton.edu/tom/papers/LabPublications/GrowMind.pdf

Toubia, O., & Netzer, O. (2017, January-February). Idea Generation, Creativity, and Prototypicality. *Marketing Science, 36*(1), 1-20. Retrieved from https://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/15027/toubia_netzer_idea_generation.pdf

Tran, D., Ranganath, R., & Blei, D. M. (2015, November 20). *The variational Gaussian process.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13476964332561990649

Ustun, B., & Rudin, C. (2019, June 19). Learning Optimized Risk Scores. *Journal of Machine Learning Research, 20*, 1-10. Retrieved from https://arxiv.org/pdf/1610.00168.pdf

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable Recourse in Linear Classification. In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 10-19). New York, NY, US: Association for Computing Machinery. Retrieved from https://arxiv.org/pdf/1809.06514.pdf

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013, October). Atypical Combinations and Scientific Impact. *Science, 342*(6157), 468-72. Retrieved from https://www.researchgate.net/profile/Satyam-Mukherjee/publication/258044625_Atypical_Combinations_and_Scientific_Impact/links/0deec52b07d0a582b3000000/Atypical-Combinations-and-Scientific-Impact.pdf

Vallee-Tourangeau, F., Steffensen, S. V., Vallee-Tourangeau, G., & Makri, A. (2015). *Insight and Cognitive Ecosystems.* Annual Conference of the Cognitive Science Society. Retrieved from https://cogsci.mindmodeling.org/2015/papers/0422/paper0422.pdf

Wachter, S., Mittelstadt, B., & Russell, C. (Spring 2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841-887. Retrieved from https://arxiv.org/ftp/arxiv/papers/1711/1711.00399.pdf

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). *Designing Theory-Driven User-Centric Explainable AI.* ACM. Retrieved from https://dl-acm-org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1145/3290605.3300831

Wang, T., Rudin, C., Doshi, F., Liu, Y., Klampfl, E., & MacNeille, P. (2015). *Bayesian or's of and's for interpretable classification with application to context aware recommender.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=7372300651005179578

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2015). *Or's of and's for interpretable classification, with application to context-aware recommender systems.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=1841876631120950361

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2017, January 1). A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research, 18*(1), 2357-2393. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=17888042361257282144

Weld, D. S., & Bansal, G. (2018). *The Challenge of Crafting Intelligible Intelligence.* arXiv. Retrieved from https://arxiv.org/pdf/1803.04263.pdf

Wu, M., Ghassemi, M., Feng, M., Celi, L. A., Szolovits, P., & Doshi-Velez, F. (2017, May). Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association, 24*(1), 488-495. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=13394270005196874867

Wu, M., Hughes, M., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018, April 25). Beyond sparsity: Tree regularization of deep models for interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1), 1670-1678. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=15277305634094866318

Wu, M., Parbhoo, S., Hughes, M., Kindle, R., Celi, L., Zazzi, M., . . . Doshi-Velez, F. (2019, August 13). *Regional tree regularization for interpretability in black box models.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=10080078237025095815

Yang, W., Lorch, L., Graule, M. A., Srinivasan, S., Suresh, A., Yao, J., . . . Doshi-Velez, F. (2019). *Output-constrained Bayesian neural networks.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=2484614356474647950

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating Effects of User Experience and System Transparency on Trust in Automation. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 1-9. Retrieved from https://interactive.mit.edu/sites/default/files/documents/Yang_HRI_2017.pdf

Yao, J., Pan, W., Ghosh, S., & Doshi-Velez, F. (2019). *Quality of uncertainty quantification for Bayesian neural network inference.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=6516073606226527796

Yessenov, K., Tulsiani, S., Menon, A., Miller, R. C., Gulwani, S., Lampson, B., & Kalai, A. T. (2013, October). A Colorful Approach to Text Processing by Example. *UIST '13 Proceedings of the 26th annual ACM symposium on User interface software and technology*, 1-10. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/2013-uist-colorful-approach-to-pbe.pdf

Yi, K., & Doshi-Velez, F. (2017). *Roll-back hamiltonian monte carlo.* arXiv. Retrieved from https://scholar.google.com/scholar?oi=bibs&hl=en&cluster=16133475510257504120

Yin, M., Vaughan, J. W., & HannaWallach. (2019). *Understanding the Effect of Accuracy on Trust in Machine Learning Models.* ACM. Retrieved from http://www.jennwv.com/papers/accuracy-trust.pdf

Zhao, Q., & Hastie, T. (2021). Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistic, 39*(1), 272-281. Retrieved from https://web.stanford.edu/~hastie/Papers/pdp_zhao.pdf

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). *Visualizing Deep Neural Network Decisions: Prediction Difference Analysis.* arXiv. Retrieved from https://arxiv.org/pdf/1702.04595.pdf

Zou, J. Y., Chaudhuri, K., & Kalai, A. T. (2015). *Crowdsourcing Feature Discovery via Adaptively Chosen Comparisons.* arXiv. Retrieved from https://arxiv.org/pdf/1504.00064.pdf

Bibliography Note

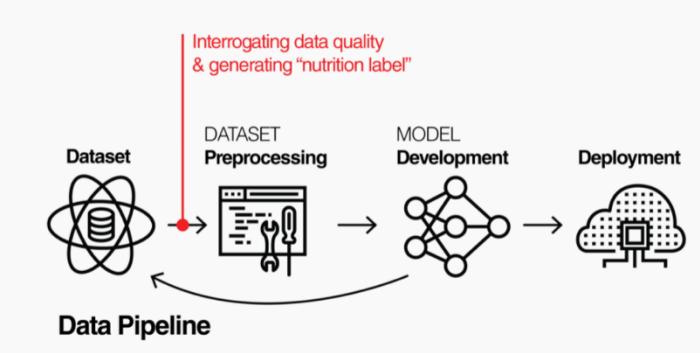
Various sources contributed by the Harvard community and the broader work in human computer interactions, AI explainability and related fields. Their work supports these efforts, but do not necessarily reflect theirs or their institutions official or unofficial views and opinions.

Appendix

Data Nutrition Project

Various efforts look to document dataset and label them, a recent example that might merit greater consideration would be the Data Nutrition Project from the MIT Media lab (https://datanutrition.org/).

Given the need for equity and responsible usage for data, their work emphasizes a "belief that technology should help us move forward without mirroring existing systemic injustice." The work founded in 2018 (https://www.berkmankleinassembly.org/) aims to create "standard labels for interrogating datasets."[32] This would help put in place data governance structures to allow for data-sharing without greater demands for centralization on the part of Treasury.



MIMIC

What is MIMIC[33]

MIMIC-III is a large, publicly-available database comprising de-identified health-related data associated with approximately sixty thousand admissions of patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, nurse and physician notes, imaging reports, and out-of-hospital mortality. MIMIC supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development. It is notable for three factors:

- it is publicly and freely available.
- it encompasses a diverse and very large population of ICU patients.
- it contains high temporal resolution data including lab results, electronic documentation, and bedside monitor trends and waveforms.

Recent Updates

MIMIC-III is an update to MIMIC-II v2.6 and contains the following new classes of data:

- approximately 20,000 additional ICU admissions
- physician progress notes
- medication administration records

[32] See their white paper here: http://securedata.lol/camera_ready/26.pdf See prototype here: https://ahmedhosny.github.io/datanutrition/
[33] See https://archive.physionet.org/physiobank/database/mimic3cdb/

- more complete demographic information
- current procedural terminology (CPT) codes and Diagnosis-Related Group (DRG) codes

The MIMIC-III Clinical Database, although de-identified, still contains detailed information regarding the clinical care of patients, and must be treated with appropriate care and respect. Researchers seeking to use the full Clinical Database must formally request access to the MIMIC-III Database.

More information
For more information about the MIMIC-III Clinical Database, please visit http://mimic.physionet.org/.

Left Blank Intentionally.

DATA NUTRITION PROJECT

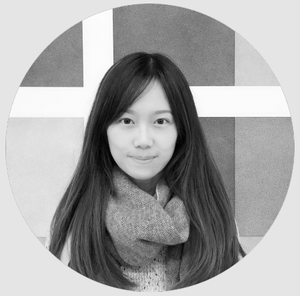
Bias in, Bias out: Nutritional Labels for Datasets

Harvard Kennedy School Responsible Use of Data Workshop
Data Nutrition Project (Launched through Berkman Klein Center (HLS) and MIT Media Lab)
Thursday, May 6th, 2021

DNP’s Mission

We empower data scientists and policymakers with practical tools to **improve AI** outcomes through **products** and **partnerships**, and in an **inclusive and equitable way**

Jess Yurkofsky
Design



Chelsea Qiu
Research

Kasia Chmielinski
Project Lead

Sarah Newman
Research Lead

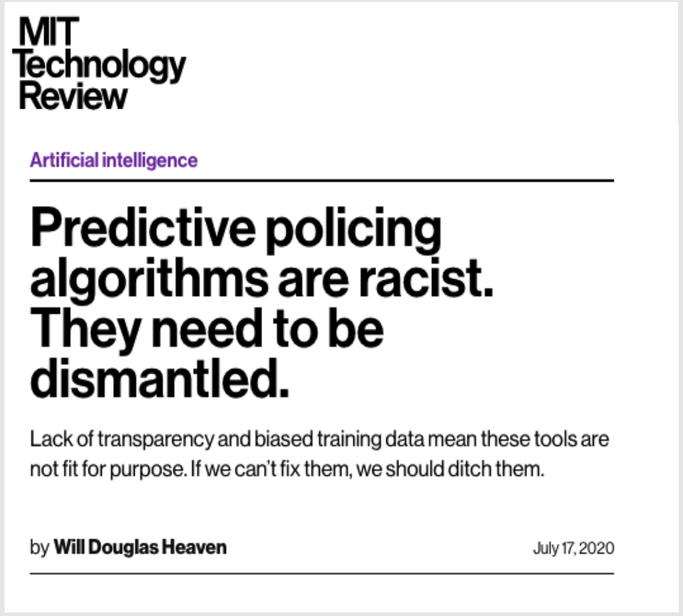
Kemi Thomas
Developer

Josh Joseph
Data Lead

Matt Taylor
Tech Lead

The Problem

Artificial intelligence (AI) systems built on **incomplete or biased data** will often exhibit problematic outcomes.


MIT Technology Review

Artificial intelligence

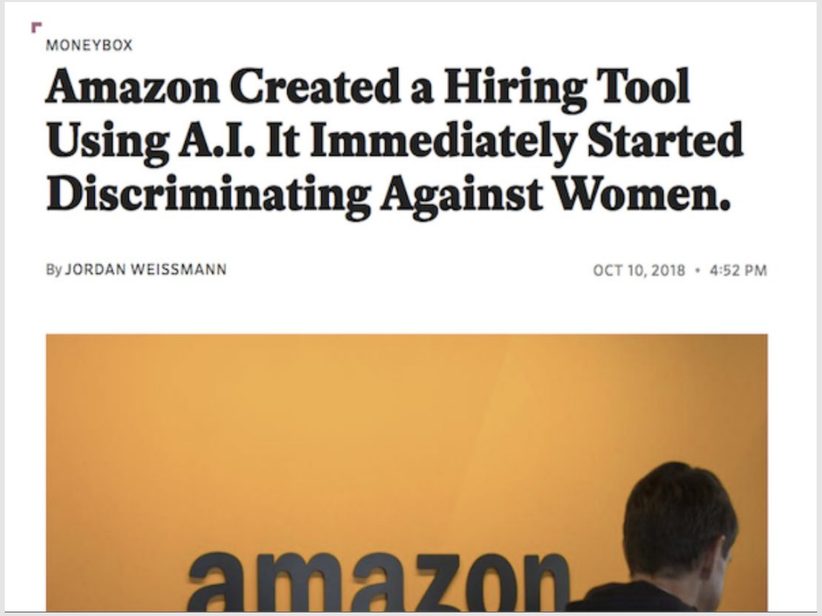
Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

by **Will Douglas Heaven** July 17, 2020


Suicide Risk Prediction Models Could Perpetuate Racial Disparities

Two suicide risk prediction models are less accurate for some minority groups, which could exacerbate ethnic and racial disparities.


MONEYBOX

Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.

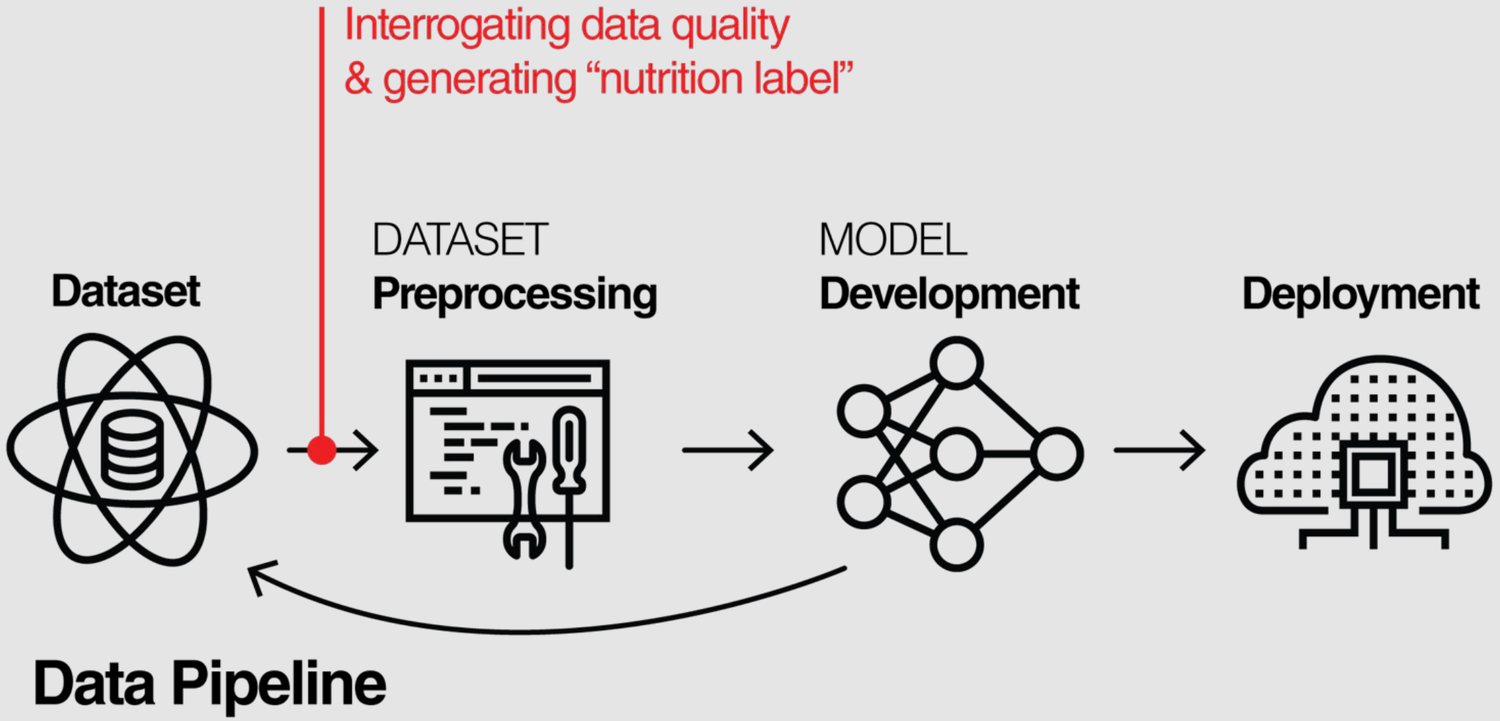
By JORDAN WEISSMANN OCT 10, 2018 • 4:52 PM

DATA
NUTRITION
PROJECT

Introducing the Data Nutrition Project

Model Development

There is an opportunity to **interrogate data quality for bias** before building the model



It’s a total free for all. *When there isn't a best practice that translates well, it takes some time to discover you might need one.*

— Survey Respondent

The Importance of Transparency & Choice

People and practitioners can make informed decisions when they know what's inside

Nutrition Facts
Serving Size 1 Cup (240mL)
Serving Per Container 8

Amount Per Serving	
Calories 60	Calories from Fat 15
	% Daily Value*
Total Fat 2g	**3%**
Saturated Fat 0g	**0%**
Trans Fat 0g	
Polyunsaturated Fat 1g	
Monounsaturated Fat 0.5g	
Cholesterol 0mg	**0%**
Sodium 115mg	
Potassium 340mg	
Total Carbohydrate 5g	
Dietary Fiber 1g	
Sugars 3g	
Protein 6g	**12**
Vitamin A 10% • Vitamin C 0%	
Calcium 45% • Iron 6%	
Vitamin D 30% • Riboflavin 30%	
Folate 10% • Vitamin B12 50%	

Should I eat this?

"*From reviewing 60 intervention studies, food labeling reduces consumer dietary intake of selected nutrients and influences industry practices to reduce product contents of sodium and artificial trans fat.*"

- *American Journal of Preventive Medicine*

A Nutritional Label for Datasets (2018)

Dataset Fact Sheet

Metadata

Title COMPAS Recidivism Risk Score Data

Author Broward County Clerk's Office, Broward County Sherrif's Office, Florida

Email browardcounty@florida.usa

Description Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

DOI 10.5281/zenodo.1164791

Time Feb 2013 - Dec 2014

Keywords risk assessment, parole, jail, recidivism, law

Records	7214
Variables	25

priors_count: *Ut enim ad minim veniam, quis nostrud exercitation* **numerical**

two_year_recid: *Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.* **nominal**

. . .

Missing Units	15452 (8%)

⚠ This dataset contains variables named "age," "race," and "sex."

Probabilistic Modeling

Analysis

12

Dependency Probability **Pearson R**

Missing Units

Clustering Variable race

Missing Variable r_days_from_arrest

missing_earthquake
missing_drought
missing_coal_pm1
missing_broadband
missing_jail_received
missing_air_accident
missing_immune_hepatitis
missing_professional_birth
missing_mortality_kid
missing_gdp_per1k
missing_food
missing_population
missing_continent
missing_tsunami
missing_sanitation
missing_water_source
missing_completion
missing_sugar
missing_food
missing_invest_domestic_per1k
missing_mortality_maternal
missing_invest_foreign_per1k
missing_spending_health_per1

The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards

Sarah Holland[1]*, Ahmed Hosny[2]*, Sarah Newman[3], Joshua Joseph[4], and Kasia Chmielinski[1*†]

[1]b and Berkman Klein Center at Harvard University, [2]Dana-Farber Cancer Institute, Harvard Medical
[3]rvard, Berkman Klein Center for Internet & Society, Harvard University, [4]33x.ai
**authors contributed equally*
[†]nutrition@media.mit.edu

ABSTRACT

telligence (AI) systems built on incomplete or biased data will often exhibit
s. Current methods of data analysis, particularly before model development, are
ardized. The Dataset Nutrition Label[1] (the Label) is a diagnostic framework that
standardized data analysis by providing a distilled yet comprehensive overview of
before AI model development. Building a Label that can be applied across domains
ires that the framework itself be flexible and adaptable; as such, the Label is
e qualitative and quantitative modules generated through multiple statistical and
ng backends, but displayed in a standardized format. To demonstrate and advance
erated and published an open source prototype[2] with seven sample modules on the
or Docs dataset. The benefits of the Label are manyfold. For data specialists, the
e robust data analysis practices, provide an efficient way to select the best dataset for
ncrease the overall quality of AI models as a result of more robust training datasets
eck for issues at the time of model development. For those building and publishing
reates an expectation of explanation, which will drive better data collection practices.
limitations of the Label, including the challenges of generalizing across diverse
of using "ground truth" data as a comparison dataset. We discuss ways to move
mitations identified. Lastly, we lay out future directions for the Dataset Nutrition
ding research and public policy agendas to further advance consideration of the
concept.

DATA NUTRITION PROJECT

Nutritional Label for Datasets (2020)

https://datanutrition.org/labels/

Dataset Nutrition Label

2020 SIIM-ISIC Melanoma Classification Challenge Dataset

About

The 2020 SIIM-ISIC Melanoma Classification challenge dataset was created for the purpose of conducting a machine learning competition to identify melanoma in lesion images. As the leading healthcare organization for informatics in medical imaging, the Society for Imaging Informatics in Medicine (SIIM)'s mission is to advance medical imaging informatics through education, research, and innovation in a multi-disciplinary community. SIIM is joined by the International Skin Imaging Collaboration (ISIC), an international effort to improve melanoma diagnosis. The ISIC Archive contains the largest publicly available collection of quality-controlled dermoscopic images of skin lesions.

Data Creation Range: 1998 - 2019
Created By: International Skin Imaging Collaboration (ISIC)
Content: The 2020 SIIM-ISIC Melanoma Classification challenge dataset was created for the purpose of conducting a machine learning competition to identify melanoma in lesion images. As the leading healthcare organization for informatics in medical imaging, the Society for Imaging Informatics in Medicine (SIIM)'s mission is to advance medical imaging informatics through education, research, and innovation in a multi-disciplinary community. SIIM is joined by the International Skin Imaging Collaboration (ISIC), an international effort to improve melanoma diagnosis. The ISIC Archive contains the largest publicly available collection of quality-controlled dermoscopic images of skin lesions.
Source: https://challenge2020.isic-archive.com/

Alert Count	5*
Completeness	4
Racial Bias	2
Socioeconomic Bias	1
Gender Bias	1
Provenance	0
Collection	0
Description	0
Composition	1
Racial Bias	1

* Please refer to the Objectives and Alerts section for more details

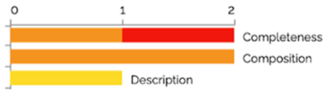
Use Cases

Potential real-world applications of the dataset

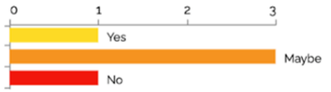
1 Identify melanoma in lesion images
2 Predict incidence of melanoma in a population

Badges

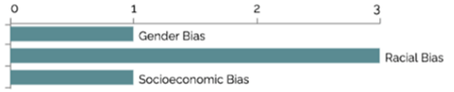
Alert Count by Category



Alert Count by Mitigation Potential



Alert Count by Potential Harm

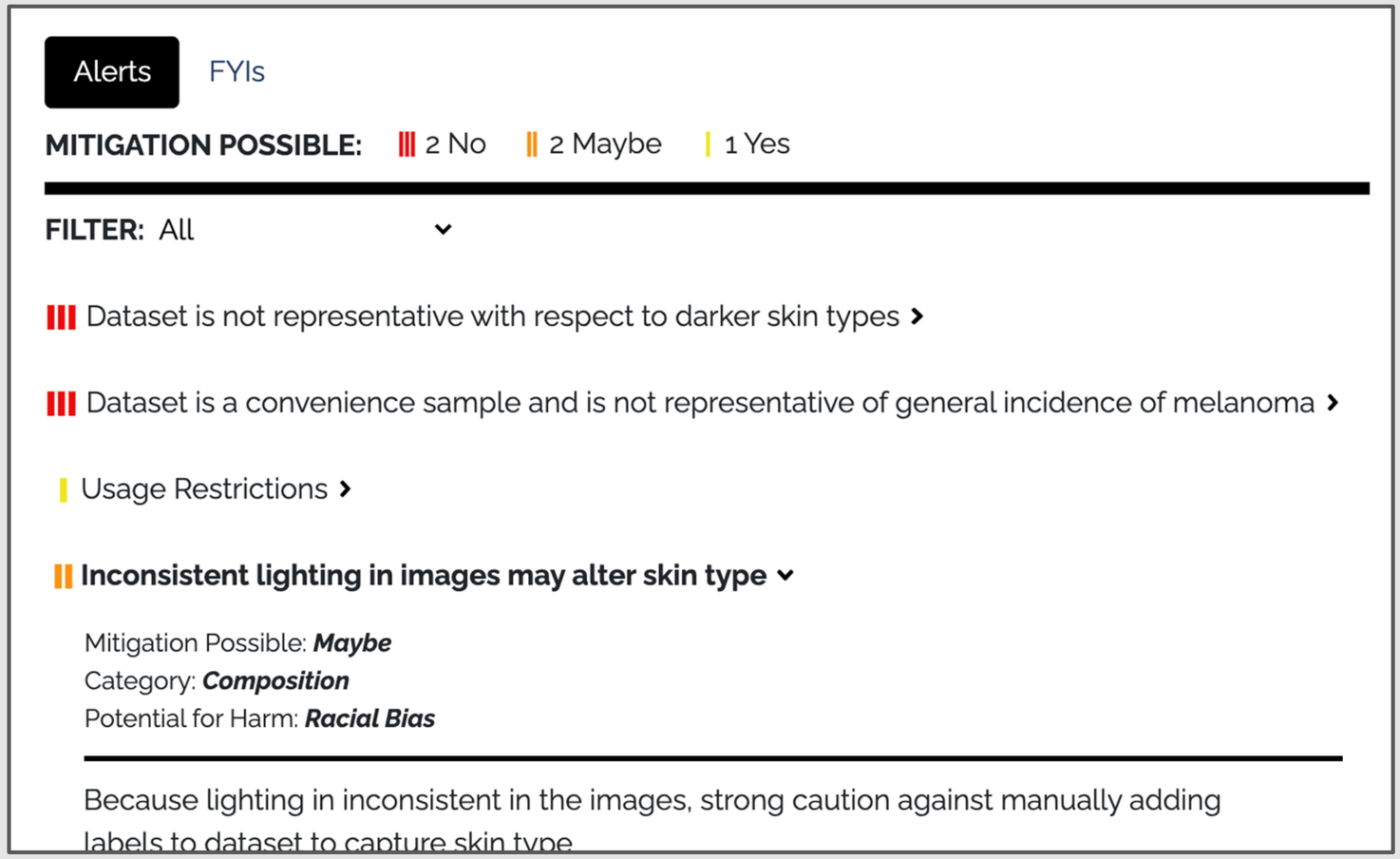


DATA NUTRITION PROJECT

Nutritional Label for Datasets (2020)

https://datanutrition.org/labels/

The tool is dynamic and built for data practitioners and those who are selecting datasets for advanced stats / AI purposes



... While I know that the primary mission of the DNP is to improve the understanding, searching, and consumption of datasets by users of datasets, ***it has also been key to improving my dataset design moving forward.***

— Dataset Partner

Impact of the approach, methodology, and standard

RAI Certification Beta

The world's first independent, accredited certification program of its kind. Developed under the Global AI Action Alliance for the World Economic Forum (WEF), along with a diverse community of leading experts, RAI certification is based on objective assessments of fairness, bias, explainability, and other concrete metrics of responsibly built AI systems. The Schwartz Reisman Institute for Technology and Society (SRI) at University of Toronto is serving as a business partner on the development phase of the initiative. =

NeurIPS | 2021

Thirty-fifth Conference on Neural Information Processing Systems

- Submission introducing new datasets must include the following in the supplementary materials:
  - Dataset documentation and intended uses. Recommended documentation frameworks include datasheets for datasets, dataset nutrition labels, data statements for NLP, and accountability frameworks.
  - URL to website/platform where the dataset/benchmark can be viewed and downloaded by the reviewers.
  - Author statement that they bear all responsibility in case of violation of rights, etc., and confirmation of the data license.

The Vision

We believe that Nutritional Labels on Datasets will:

1. Drive **robust data analysis practices** by making it easier and faster for data scientists to interrogate and select datasets.

2. Increase **overall quality of models** by driving the use of better and more appropriate datasets for those models

3. Enable the **creation and publishing of responsible datasets** by those who collect, clean and publish data

DATA
NUTRITION
PROJECT

Thank You!

Contact: info@datanutrition.org
Twitter: @makedatahealthy
Website: datanutrition.org