From: Liz O'Sullivan
Chief Executive Officer
Parity Technologies, Inc.

To: Mark Przybocki
Chief, Information Access Division
U.S. National Institute of Standards and Technology, MS 20899
100 Bureau Drive, Gaithersburg, MD 20899
AIframework@nist.gov

September 15, 2021

To whom it may concern,

Response to AI Risk Management Framework RFI (NIST-2021-0004)

The team at Parity welcome this opportunity to share our years of experience working with and within enterprise AI teams to mitigate AI-driven risk factors across multiple industries, notably in financial services and healthcare.

1. The greatest challenges in improving how AI actors manage AI-related risks—where "manage" means identify, assess, prioritize, respond to, or communicate those risks;

The biggest risks for AI-driven enterprises generally fail to surface not due to some fault of the technology itself, but instead a failure to incorporate sufficient knowledge and insight from a wide enough perspective to constitute real oversight. Some of this challenge stems from the complexity of this newer technology, which is difficult to explain to non-technical stakeholders who are typically responsible for governing risk in large corporations. Even existing so-called "model risk" teams are typically composed of classically trained statisticians who may miss components that AI's enhanced complexity as applied to machine learned models. Consequently, enterprises often fail to move forward at all with AI as a tooling solution, given their fear of unpredicted failure in this poorly understood domain[1].

Moreover, most enterprises are subject to the tools that have been grandfathered into the enterprise over many years or decades. In most of the banks and healthcare insurers we

[1] Terence Tse, Mark Esposito, Takaaki Mizuno, and Danny Goh. The dumb reason your AI project will fail. *Harvard Business Review*, 8 June 2020. URL https://hbr.org/2020/06/the-dumb-reason-your-ai-project-will-fail.

have worked with, teams report a worrying heterogeneity of tools and processes that vary across domains, locations, applications of models, and even industries within the same corporation. This lack of unity across platforms causes disjointed and inconsistent model approval reviews, leaving developers confused and unable to access the information needed to de-risk their tasks. Only recently have certain tools become available to the enterprise to resolve some of this risk, including AI observability tools like Arthur.AI and Fiddler.AI, and model governance framework technologies like Parity (getparity.ai).

One problem in mitigating these risks stems from the fact that large enterprises rarely do de-risking that is not explicitly required, making this new market an exciting one, but one that is also slow to move forward. Part of this slowness is directly correlated to the mentality of some data scientists and executives who believe that certain elements of AI risk, most notably in the form of discriminatory bias, are not worthy of immediate attention. This is a paradigm that has been changing over the last few years, but slowly. This in itself is more evidence that there is a severe lack of consensus among AI practitioners who wish to mitigate risks in their models. Decisions on what to prioritize are often made at the top of the organization with insufficient knowledge of the types of risks to be measured.

Perhaps most importantly of all, as was unanimously agreed upon by our team, there persists a mentality of "what we don't know can't hurt us" in industry as it applies to regulatory oversight of discriminatory biases. This is a line of reasoning we have heard across multiple industries and applications of AI in business, especially in highly regulated industries; there is overwhelming fear that the very act of undertaking some form of model audit will result itself in a regulatory fine, once discrimination is found. Although we know that regulators will jump at the opportunity to investigate purported examples of discrimination in certain industries, as we saw in lending via NY Department of Financial Services' investigation into Apple Card[2], their eventual clearance of wrongdoing provides effective cover for many AI-lending entities. Apple Card is a clear example of where the social issue of discriminatory credit scoring has attracted the attention of the public, and yet where our laws fall far short of their goals in certifying "fair" AI[3]. This connects to the lack of regulatory pressure to measure certain definitions of fairness, notably things like "equality of opportunity" in the form of "equalized odds" for lending.

[2] New York Department of Financial Services, DFS issues findings on the Apple Card and its underwriter Goldman Sachs Bank: no fair lending violations found; broadly, report stresses need for modernizing credit scoring models and updating anti-discrimination laws governing credit access, 23 March 2021, https://www.dfs.ny.gov/reports_and_publications/press_releases/pr202103231

[3] Liz O'Sullivan, How the law got it wrong with Apple Card, *TechCrunch*, 14 August 2021, https://techcrunch.com/2021/08/14/how-the-law-got-it-wrong-with-apple-card

Outside highly regulated industries, we still see pushback against serious engagement with AI risk and discriminatory biases. Even when regulation is sparse, there is overwhelming fear that if unfairness is found, ethically minded actors would push for mitigation. Lack of familiarity with the complexities of mitigation, and the risk that model unfairness could derail projects in their entirety, prevents teams and organizations from looking too closely at potential unfairness in their AI in the first place. The XCheck scandal, where Facebook hid details of a "not publicly defensible" program from its own independent oversight board, is simply this week's example of reluctance to examine bias and a willingness to obstruct those that do[4]. Regulation is key to empower individuals and teams that do this work at the pleasure of reluctant organizational leadership.

We hope that updated, industry-specific guidance from NIST will go beyond the "bare minimum" of risk identification mandated by existing laws, and will guide teams on how better to think about identifying and eliminating bias with an eye towards social equality, rather than simply effective compliance to existing or proposed laws.

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

We propose that one of the most important elements of introducing risk to the public through AI is the result not of some technical measurement like robustness or resilience, but instead through incorrectly formed problems and intents when creating AI. There are many examples of this, with online proctoring systems claiming to infer "cheating" through eye movements, using computer vision that is by its very definition ableist against those with visual disabilities (e.g. the blind), and subject to cultural differences across groups in how this intent manifests.

Even in less egregious cases, the intentional use of proxies as stand ins for "ground truth" in a model's proposed quantification of accuracy can result in its own newly introduced risks and biases. If the goal of your model is to predict whether your members might need future preventative healthcare, one might suggest that the enterprise confirm whether their subjects have, in fact, sought care from a hospital for emergency interventions in the timeframe of the model's guess. However, due to commercial, organizational, and

[4] Horwitz, J., Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt. *Wall Street Journal*. 13 September 2021. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353?mod=searchresults_pos1&page=1 See also this thread from the author.

operational constraints or other, this information may not be immediately available to the enterprise. As such, in order to confirm the model's guess, enterprises may construct a definition of accuracy that looks instead at some other model's output, guessing at the overall health of the member/user. This stacking of predictions relying on other predictions can create a wholly self-contained loop, where any issues with one model will skew the outputs of another.

We hope that NIST's risk framework will take great care in their recommendations around testing, since this critical moment between lab conditions and a production environment is poorly understood, and itself a great source of AI failures that correspond to enterprise risk. The creation or manipulation of testing datasets can mean the difference between models that work and those that fail entirely. This can be introduced in many forms, notably: (1) in the use of an external vendor whose marketing claims have been manipulated, (2) in the result of loosely correlated proxies serving as "ground truth", (3) in the result of randomness (e.g. an 80/20 split) failing to accurately characterize the demographic makeup of the public served, and as you mention in the RFI, (4) the abuse of an AI tool in practice that is different from the original modeller's intent.

Finally, we propose that all AI deciding on sensitive criteria should be subject to some form of human oversight corresponding to the associated risks. As such, one missing element of your proposed concept list is that of the level of recommended human involvement. We know that all models are subject to concept drift, and require proactive interrogation to ensure their continued relevance. Yet, methods for human involvement, which range from test-triggered human reviews to human-in-the-loop training and evaluation, are underutilized. Decisions should be made early on to tie levels of human involvement to the highest-risk categories, and to define the "bare minimum" of human oversight to lower-risk kinds of models.

6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;

To our knowledge, the existing model risk frameworks in financial services are the oldest such regulations and form an instructive reference regarding the operational challenges around compliance and enforcement[5], and yet banks face enormous challenges in operationalizing these requirements. One such regulation is the *Supervisory Guidance on Model Risk Management*[6], describing requirements for model risk management required

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

[6] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. Supervisory Guidance on Model Risk Management, 4 April 2011. https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf

for AI/ML models in the consumer finance industry, and is enforced by the Office of the Comptroller of Currency (OCC 2011-12), Federal Reserve System (SR 11-7), and the Federal Deposit Insurance Corporation (FIL-22-1017). This regulation outlines the roles and responsibilities of the "three lines of defense model", defining different adversarial roles of organizational stakeholders of business teams/model development teams (first line), model risk management (second line), and audit (third line).
The resulting needs for model risk management lead to a complex risk management process that can add many months, if not years, to the model development process[7]. Similarly, large investment banks need to comply with Basel standards such as the *Principles for effective risk data aggregation and risk reporting* [8]. This standard (BCBS 239) requires banks to maintain reproducibility in certain critical risk computations, and must confront responsible AI challenges such as managing data lineage and ontology drift[9]. Despite the original goal of compliance by 2016 and years of effort, most banks still lag behind in implementing the necessary business processes[10]. Finally, we note that the Apple Card finding described above demonstrates the ageing utility of existing fair lending laws like the Equal Credit Opportunity Act, which date back to the Civil Rights Act and have not been comprehensively updated for the current age of ubiquitous AI.

We believe that the experience of financial institutions show that even with explicit regulatory requirements, large legacy organizations that are not digital first will face enormous struggles in adopting the Framework. Consequently, the vast majority of enterprises will fail to undertake voluntary frameworks such as this one without sufficient regulatory pressure. As such, we recommend that NIST take great care when developing the framework to be compatible with the requirements set forth by existing antidiscrimination law, which are the public's current best legal defense against discriminatory AI. Even in this scenario, agencies are working to update their rules and requirements for adherence to this set of laws, which may evolve at any time. Our hope is that new guidelines from NIST can serve to help streamline this arduous, time-consuming process that in practice has stymied AI adoption.

[7] Eren Kurshan, Hongda Shen, and Jiahao Chen. Towards self-regulating AI: Challenges and opportunities of AI model governance in financial services. In Proceedings of the 1st International Conference on AI in Finance, 15 October 2020. doi: 10.1145/3383455.3422564. URL https://arxiv.org/abs/2010.04827.
[8] Basel Committee on Banking Supervision, Standard No. 239, Principles for effective risk data aggregation and risk reporting, *Bank for International Settlements*, Basel, Switzerland, January 2013. https://www.bis.org/publ/bcbs239.pdf
[9] Jiahao Chen. Ontology drift is a challenge for explainable data governance, 2021. URL https://arxiv.org/abs/2108.05401.
[10] Basel Committee on Banking Supervision, Progress in adopting the Principles for effective risk data aggregation and risk reporting, *Bank for International Settlements*, Basel, Switzerland, 29 April 2020. https://www.bis.org/bcbs/publ/d501.htm

7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts;

The Parity platform facilitates adherence to any standard of responsibility that emerges from this discussion, including any frameworks that NIST creates. In addition, our partnership opportunities include the Responsible AI Institute (RAI), for whose certification program we act as an expert advisor, and the Algorithmic Justice League.

8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation—and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society.

In our experience, the unfortunate reality is that these large enterprises rarely take such issues into account, due to their homogeneity and common dedication to bettering company KPIs. There are notable exceptions of organizations who have proactively sought our help, with too many to mention that do not. However, even the most socially responsible groups often fail to incorporate a sufficiently inclusive design process into their product development lifecycle.

As described above, financial institutions currently need to comply with AI model risk management regulations, but face operational challenges that the regulations do not address. One major challenge in practice is that fair lending laws are inconsistent in requiring lenders to collect demographic labels (a.k.a. government monitoring information) that are necessary in order to compute bias metrics, which complicates the actual bias measurement[11] and lead to the persistence of discrimination in practice[12]. Under the Fair Housing Act, mortgage lenders must collect GMI either through customer identification or through perceived attributes by a loan officer. The collection of GMI is fraught with difficulties ranging from survey bias in self-identification questionnaires to ethical quandaries around identifying and labeling people, and running the risk of outing people (for example, labeling people as LGBTQ+ when they reside in countries that outlaw homosexuality). On the contrary, the Equal Credit Opportunity Act forbids the collection of GMI for all other consumer loans such as credit cards and auto loans. Without access to GMI, lenders and regulators resort to imputation methods which introduce their own statistical biases. In practice, such GMI labels are imputed from publicly available census data using methods like Bayesian Improved Surname

[11] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2017. URL http://arxiv.org/abs/1707.00075.

[12] Marsha Courchane, David Nebhut, and David Nickerson. Lessons Learned: Statistical Techniques and Fair Lending. *Journal of Housing Research*, 11(2):277–295, 2000.

Geocoding[13], which introduce their own biases and ecological inference errors. These errors are so large that the wrong sign of discriminatory bias can be measured: a model may be measured to be biased in favor of a disadvantaged minority when the ground truth is the exact opposite[14]. Such findings invalidate the standard approaches to measuring and mitigating biases that exist in the academic literature; instead, a careful quantification of the uncertainty in the bias measurement is necessary to avoid erroneous conclusions[15]. In addition, the use of such imputation methods come with their own ethical concerns such as false labeling and risks of compromising individual privacy. Such controversies are not merely academic, but have in fact led to disputes over the legal authority of regulatory agencies[16], with enormous financial consequences over the legality of assessing hundreds of millions of dollars in regulatory penalties[17]. We expect that situations with missing demographic information will be the norm, not the exception, and strongly urge that uncertainty quantification of bias metrics form an integral component of practical and relevant AI risk management frameworks.

A second major challenge in measuring AI fairness and risk in finance is that of reject inference[18]. A falsely approved loan is material and directly measurable, while a falsely denied loan is counterfactual and does not appear on a balance sheet. This data asymmetry between approvals and rejections gives rise to the need to correctly measure false negative error, which is inherently counterfactual and cannot be measured without

[13] (a) Kevin Fiscella and Allen M Fremont. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41(4p1):1482–1500, 2006. (b) Marc N Elliott, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole Lurie. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83, April 2009. doi: 10.1007/s10742-009-0047-1.

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

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

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

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

[18] (a) David J. Hand and Niall M. Adams. 2014. Selection bias in credit scorecard evaluation. *Journal of the Operational Research Society* 65, 3 (2014), 408–415. DOI:10.1057/jors.2013.55 (b) Naeem Siddiqi. 2006. *Credit Risk Scorecards*. John Wiley & Sons, Hoboken, NJ. DOI:10.1002/9781119201731

explicit sampling of the assumed negative space by approving customers that the model thinks will be bad. Reject inference also shows up in other fields like epidemiology and the social sciences[19], and is a general challenge of causal inference methods in practice[20].

Finally, we also mention the problem of intra-group variations in fairness and the possible existence of persistently discriminated subgroups. For example, Southeast Asians, Chinese, and Indians are lumped together into the same Asian label at the highest level Census Bureau categorization, even though each subgroup has widely different levels of economic achievement which can average out outcomes in education and employment[21]. A quick look at the history of census demographics will show beyond a shadow of a doubt that these racial category definitions are fluid, change over time, and in some cases even political. There is therefore an application-specific need for a conscious choice to enumerate demographic groups may be on the receiving end of ethical harms.

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Regulations like BCBS 239 acknowledge that systemic risk in the financial industry cannot be solely triaged in algorithms, but must necessarily encompass the data processing pipeline that produces the actual inputs into algorithmic computations. Researchers have

[19] (a) Stephen L Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). Analytical Methods for Social Research, Cambridge University Press, 2015. (b) Maya Sen and Omar Wasow. 2016. Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics. *Annual Reviews of Political Science* 19, (2016), 499–522. DOI:10.1146/annurev-polisci-032015-010015

[20] Isabelle Guyon, Alexander Statnikov, and Berna Bakir Batu. 2019. *Cause Effect Pairs in Machine Learning.* The Springer Series on Challenges in Machine Learning.

[21] (a) Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. *Proceedings of Machine Learning Research* 80, (2018), 2564–2572. URL http://proceedings.mlr.press/v80/kearns18a.html (b) Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax Pareto fairness: a multi-objective perspective. *Proceedings of Machine Learning Research* 119, (2020), 6755–6764. URL http://proceedings.mlr.press/v119/martinez20a.html (c) Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. 2021. Fairness Through Robustness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pp. 466–477. DOI:https://doi.org/10.1145/3442188.3445910 (d) Mark Weber, Mikhail Yurochkin, Sherif Botros, and Vanio Markov. 2020. Black Loans Matter: Distributionally Robust Fairness for Fighting Subgroup Discrimination. *NeurIPS Workshop on Fair AI in Finance*. (2020). URL http://arxiv.org/abs/2012.01193

similarly argued for documentation requirements for model[22] and data[23], so that business stakeholders, legal experts, social scientists, and the algorithms' consumers can understand clearly what limitations and intents are associated with the entire AI system[24].

Closing remarks

In addition to the comments above, we would also like to mention the additional comments that we have provided in response to *Draft NIST Special Publication 1270*[25], which touches upon further themes not explicitly stated in this RFI.

We sincerely hope that when NIST creates the Framework, that it goes further than mere recommendations, and advances strict requirements to include diverse viewpoints not only from employees but experts from the outside world, specifically within the humanities, advocacy groups, civil society, and the organization's intended users. Any AI that is created from the top down without regard for whether AI is wanted or needed by the people it serves will be by its very definition laden with risks, given the likelihood that the problem is poorly formulated. It is far too frequent that current applications of AI only serve to further poorly formulated hypotheses that result in bad outcomes, usually for our society's most marginalized members. If one's goal is to create a model that predicts something about hiring, it must be required that creators speak to those who have studied the realities of hiring in that jurisdiction both from a quantitative and a qualitative perspective, lest we repeat lessons from history that might otherwise have been avoided.

Thank you for your time and attention,

The Parity Team

[22] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 29 January 2019. doi: 10.1145/3287560.3287596. URL http://dx.doi.org/10.1145/3287560.3287596.

[23] (a) Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2020. URL https://arxiv.org/abs/1803.09010. (b) Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions. In *Proceedings of the 23nd International Conference on Extending Database Technology*, November 2019. URL http://arxiv.org/abs/1911.12587. (c) Julia Stoyanovich, Bill Howe, and H. V. Jagadish. Responsible data management. *Proceedings of the VLDB Endowment*, 13(12):3474–3488, 2020. doi: 10.14778/3415478.3415570.

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

[25] Parity AI, Feedback on Draft NIST Special Publication 1270: A Proposal for Identifying and Managing Bias in Artificial Intelligence, 10 September 2021, https://www.nist.gov/system/files/documents/2021/09/13/20210910_S.P.%201270%20Comments%20-%20Parity%20AI.pdf