NFHA NATIONAL FAIR HOUSING ALLIANCE

Purpose, Process, and Monitoring

A New Framework for Auditing Algorithmic Bias in Housing & Lending

Michael Akinwumi, Lisa Rice, Snigdha Sharma

February 17, 2022

Executive Summary

The National Fair Housing Alliance (NFHA) is proposing a new framework for comprehensively auditing algorithmic systems like credit scoring, insurance scoring, automated underwriting, risk-based pricing, digital advertising, and tenant screening selection tools. The new auditing framework, called Purpose, Process, and Monitoring (PPM), captures the key stages in algorithmic model pre-development, development, and post-development including monitoring. NFHA believes this framework represents the ***gold standard*** for auditing algorithmic systems. The framework provides a roadmap for regulators, businesses, civil rights groups, academics, researchers, and other stakeholders to assess the fairness and efficacy of algorithmic systems to lessen or eradicate harmful effects on society.

Algorithmic systems touch virtually every part of an individual's life from advertisements on their social media feeds to their ability to qualify for a loan and purchase a home. However, many of these systems can also inflict untold harm due to algorithmic bias and discrimination. High-profile examples abound regarding how algorithmic bias and discrimination inflict harm to people in a range of areas including health, housing, law enforcement, credit, employment, and marketing with devastating consequences such as homelessness, job loss, incarceration, health disparities, wealth gaps, and more. Accordingly, the use of algorithms in these areas must be scrutinized to ensure they do not impede equitable access to housing, lending, credit, and other important opportunities. It is essential that algorithmic systems are fair, equitable, explainable, and transparent.

The Purpose stage of the PPM framework examines project goals as well as the expectations, requirements, and objectives of stakeholders collectively referred to as the business problem. During this phase, auditors are guided to seek information to make informed decisions about risks the business problem may pose to consumers, institutions, and society at large. The reality is that there might be some systems that should not be built or put into production because they pose too great a harm. Additionally, during the Purpose stage, auditors are guided to explore how well data is being used to accurately capture and reflect the business problem statement as well as determine if any techniques were used to mitigate risks associated with data paucity or data quality.

The Process stage of the PPM framework evaluates the design, theory, and logic used to develop an algorithmic solution in the context of business use cases and design objectives. This stage includes five elements:

1. The "Staff Profile" element requires that teams working on models are diverse, inclusive, and educated to spot challenges and prevent issues that can lead to unfavorable and deleterious outcomes.
2. The "Data Assessment" element directs auditors to gather information about data sources and data fields considered for the development of the model and conduct other analyses to determine if the data is appropriate, representative, fair, and accurate.
3. The "Model Assessment" element of the Process stage evaluates information about training algorithms, parameters, hyper-parameters, and any fairness constraints used during the development of the model. Additionally, fairness constraints used at the post-modeling stage are evaluated as well as information about searches for Less Discriminatory Alternatives (LDAs) or solutions that present less harmful impacts to consumers.
4. The "Outcome Assessment" element in the Process stage evaluates the performance of the final model in line with the scope and metrics laid out in the Purpose section of the PPM framework to establish whether the final model meets its design objectives, which include minimizing risks to consumers, institutions, and society.
5. The "Model Use and Limitation" element of the Process stage reviews and documents known limitations and assumptions of the model along with identifying circumstances where the model may or may not be used outside of its intended scope of use.

The Monitoring Stage of the PPM Framework is comprised of "Product Model Validation" and "Protection from Confidentiality and Integrity Attacks" and serves to document the validity and robustness of the model as well as consumer fairness, privacy, and related harms, in the production environment. For the "Product Model Validation" element a report on the missingness patterns of the features used to develop the trained model is compared with a similar report on the features used in the production version of the model to inform decisions about whether the production model should be retrained, patched, or retired. The Protection from Confidentiality and Integrity Attack element focuses on the confidentiality and integrity aspects of algorithmic systems to assure consumer risks are minimized through review of defenses built into the system both during the training or development of the model and when the model is in production and making decisions that affect consumers. Under the Monitoring stage of the PPM framework, the model would be reviewed (both during training and in production) 1) to ensure the training data is substantially similar to the data the model see in a production environment; 2) to protect the privacy of records used to either train the model or score the model in production, and 3) to ensure the model incorporates defenses that assure fairness and accountability.

The PPM framework enables auditors to conduct a critical analysis of an algorithmic system to identify its assumptions and limitations and produce appropriate recommendations to mitigate consumer fairness and privacy risks that may result from poorly developed models. NFHA hopes it will become the gold standard for auditing algorithmic systems used in the housing and lending sectors as well as in other fields. NFHA's Purpose, Process, and Monitoring framework provides an equity-centered auditing solution at a time when consumers, policy makers, and businesses are calling for fairness, accountability, transparency, explainability, and interpretability.

1 Introduction

Algorithmic systems—like regression models, machine learning models, or Artificial Intelligence (AI)-driven solutions—touch virtually every part of a person's life, whether it is the advertisements on their social media feeds, determining interest rates on their auto loans, or their ability to qualify for a loan to buy a home. However, many of these systems can inflict untold harm. High-profile examples of algorithmic bias and discrimination include but are not limited to applications in health[1], tenant screening[2], criminal justice[3], facial recognition[4], employment screening[5], and marketing.[6] Discriminatory models can lead to incarceration, homelessness, job loss, equity-stripping, debilitating health, and other devastating consequences. Thus, it is crucial that algorithmic systems are fair, equitable, explainable, and transparent to avoid bias and discrimination. The auditing framework presented in this paper provides a roadmap for regulators, businesses, civil rights groups, academics, researchers, and other stakeholders to assess the fairness and efficacy of algorithmic systems to lessen or eradicate harmful effects to society.

The long history of unfair and race-based policies in the U.S. has left a deeply segregated and inequitable landscape which algorithmic systems can perpetuate and exacerbate. For example, one's zip code is often life-impacting. Where people live determine their access to homeownership, the type of credit they use, their ability to attend a well-resourced school, their exposure to toxins and pollutants, and their employment opportunities, all of which are consequential to their economic status and level of wealth. Fair housing research, litigation and other efforts undertaken by the National Fair Housing Alliance (NFHA) over the decades reveal that algorithmic-based systems used in the housing and financial services space, for example in AdTech[7] and credit scoring systems[8], are built using data that is imbued with bias and discrimination, either inherently or due to the proxies created by intricately linked

---

[1] Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science 366*, no. 6464 (2019): 447-453. https://www.science.org/doi/10.1126/science.aax2342

[2] Sisson, Patrick. "Housing Discrimination Goes High Tech." Curbed. Vox Media, December 17, 2019. https://archive.curbed.com/2019/12/17/21026311/mortgage-apartment-housing-algorithm-discrimination.

[3] Richardson, Rashida, Jason M. Schultz, and Kate Crawford. "Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice." *NYUL Rev. Online* 94 (2019): 15. https://www.nyulawreview.org/wp-content/uploads/2019/04/NYULawReview-94-Richardson_etal-FIN.pdf.

[4] Najibi, Alex. "Racial discrimination in face recognition technology." *Harvard Online: Science Policy and Social Jus*tice 24 (2020). https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

[5] Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." Reuters. Thomson Reuters, October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[6] Zang, Jinyan. "Solving the Problem of Racially Discriminatory Advertising on Facebook." Brookings. The Brookings Institution , January 31, 2022. https://www.brookings.edu/research/solving-the-problem-of-racially-discriminatory-advertising-on-facebook/.

[7] NFHA. "Facebook Settlement ." D.C. : National Fair Housing Alliance , March 14, 2019. https://nationalfairhousing.org/facebook-settlement/.

[8] Rice, Lisa, and Deidre Swesnik. "Discriminatory effects of credit scoring on communities of color." *Suffolk UL Rev*. 46 (2013): 935. https://cpb-us-e1.wpmucdn.com/sites.suffolk.edu/dist/3/1172/files/2014/01/Rice-Swesnik_Lead.pdf, https://nationalfairhousing.org/wp-content/uploads/2021/12/NFHA-credit-scoring-paper-for-Suffolk-NCLC-symposium-submitted-to-Suffolk-Law.pdf

variables. Therefore, the use of algorithms in these areas must be scrutinized such that they do not impede equitable access to housing, lending, credit, and other important opportunities.

There is an imperative to develop an appropriate framework for assessing model fairness[9]. Policymakers at the state and federal levels are already embarking down the path to regulate discriminatory or harmful algorithms more specifically. New York City policymakers recently passed the Automated Employment Decision Act,[10] which calls for regular "bias audits" of automated hiring and employment tools. The District of Columbia Attorney General proposed a bill[11] addressing algorithmic discrimination beyond employment issues. The bill has broad coverage including substantive protections, notice provisions, and auditing requirements. The U.S. Congress has held hearings[12] on and is contemplating legislation to effectively police and regulate various technologies to mitigate bias and harmful impacts. These frameworks and proposals represent an emerging consensus on the need for an acceptable standard for end-to-end algorithmic auditing.

Thus, NFHA is proposing an auditing framework called PPM (**Purpose, Process, and Monitoring**) that captures the key stages in algorithmic model pre-development, development, and post-development including monitoring. This framework is an approach for evaluating internal controls and mitigating risks that may be inherent in algorithmic systems. Traditional fair lending analysis has often included a narrow focus on individual model inputs and statistical outputs. The PPM framework, in contrast, is holistic and system oriented. NFHA hopes it will become a gold standard for auditing algorithmic systems used in the housing and lending sectors as well as in other fields where algorithms are being used to make decisions.

The PPM framework enables auditors to conduct a critical analysis of an algorithmic system to identify its assumptions and limitations and produce appropriate recommendations to mitigate consumer fairness and privacy risks[13] that may result from model risks.[14] Although it has a focus on fairness and

---

[9] With the rise of algorithmic solutions like statistical models, machine learning, or AI, contemporary approaches to evaluating these algorithms still leave much to be desired. A number of products and tools designed to assist corporations or government entities address algorithmic bias or algorithmic discrimination have emerged. Big tech has also taken steps to provide their own frameworks or platforms (like TensorFlow, https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf), including Microsoft (https://fairlearn.org) and IBM (http://aif360.mybluemix.net/). These tools use different fairness metrics, and more research is required to determine their fitness for housing and lending scenarios.

[10] The New York City Council, Committee on Technology, *Automated Employment Decision Tools,* 21AD, New York, New York City: Legislative Research Center, 2021. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=Advanced&Search.

[11] Council of the District of Columbia*, Stop Discrimination by Algorithms Act of 2021*, Phil Mendelson. Sec.2. Washington, D.C.: OAG DC, 2021, https://oag.dc.gov/sites/default/files/2021-12/DC-Bill-SDAA-FINAL-to-file-.pdf.

[12] U.S. Congress. House of Representatives. Committee on Financial Services. *Equitable Algorithms: How Human-Centered AI Can Address Systemic Racism and Racial Justice in Housing and Financial Services, Before the House Financial Services Task Force on Artificial Intelligence*, 117th Cong., virtual hearing, May 7, 2021. (statement of Lisa Rice, President and CEO of the National Fair Housing Alliance). https://financialservices.house.gov/uploadedfiles/hhrg-117-ba00-wstate-ricel-20210507.pdf

[13] Consumer risk is defined as the potential for a consumer to be adversely impacted by an algorithm-based product or service.

[14] Model risk is defined as the potential for adverse impact of decisions based on incorrect or misused model data, model architecture, model implementation, or model use. This definition is consistent with that in the Supervisory Guidance on Model Risk Management issued by the Office of the Controller of the Currency (OCC) and the Board of Governors of the Federal Reserve System issued on April 4, 2011.

specific consumer risks, the PPM framework builds off existing materials such as the 2011 Supervisory Guidance on Model Risk Management (MRM Guidance) issued by the Board of Governors of the Federal Reserve System and the Office of the Controller of the Currency (OCC),[15] the 2021 National Institute of Standards and Technology's Proposal for Identifying and Managing Bias in Artificial Intelligence (NIST proposal),[16] and the Cross-Industry Standard Process for Data Mining (CRISP-DM).[17] The MRM Guidance articulates to regulated banks supervisory expectations for mitigating model risks, with a focus on model development, model implementation, model use, and model validation. The NIST proposal focuses on avenues where bias may manifest at the pre-design stage; design and development stage; and deployment stage of an AI system. It also contains procedures for extracting auditing evidence that could be used to make recommendations or judgments about the level of consumer risks and model risks in an algorithmic system. CRISP-DM is an open industry standard for the life cycle of an AI or machine learning project, usually used for identifying roles and responsibilities involved in creating an algorithmic, data-driven solution to a business problem. The stages identified as typical stages of a data science project in CRISP-DM are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment/Serving.

The PPM framework evaluates the **Purpose** of an algorithmic system by looking at the expectations, requirements, and objectives related to the business problems being solved, as well as how data can represent those problems.[18] Relatedly, data representativeness issues can pose a major threat to the viability, fairness, and ethical appropriateness of a model given the historical and current inequities in U.S. markets. People of color are often underrepresented in model development data; historical biases driven by racist policies like redlining and segregation are known issues that limit the visibility of marginalized groups in the data that power algorithms in housing and lending sectors of our economy.[19] The **Process** piece of the PPM framework then proceeds to examine the procedure used to develop a solution to the problem that the algorithmic system is designed to address. This stage can include assessments of fairness and discrimination risks. Finally, the **Monitor** piece of the framework examines guardrails put in place to observe patterns in the production environment and compares the patterns with those in the data used to develop and test the model. This component also evaluates fairness metrics on production data and compares the metrics with those reported on development data.

2 Purpose Element of PPM

In the **Purpose** stage of model development, stakeholders decide what business problem to solve, how to formulate the problem as a data science task, and what technology solutions will be employed to solve the problem within the scope of the entity's needs (for example, profit maximization) and

[15] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, *Supervisory Guidance on Model Risk Management,* SR Letter 11-7, Washington, D.C., 2011, https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

[16] U.S. Department of Commerce, National Institute of Standards and Technology, *A Proposal for Identifying and Managing Bias within Artificial Intelligence*, Reva Schwartz, Leann Down, and Elham Tabassi, Gaithersburg, Maryland: NIST Pubs, 2021, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf

[17] "Introduction to CRISP-DM." IBM Docs. IBM https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview.

[18] We use the term "business problem" as a shorthand for any business, policy, or other problem.

[19] NFHA. "Response to Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning." D.C.: National Fair Housing Alliance, July 1, 2021. https://nationalfairhousing.org/wp-content/uploads/2021/07/Federal-Banking-Regulator-RFI-re-AI_Advocate-Letter_FINAL_2021-07-01.pdf.

customers' desired product or service requirements. The **Purpose** element of PPM is an auditing inquiry into the "Business Understanding" and "Data Understanding" structures within CRISP-DM.[20]

2.1 Business Understanding

Project goals, stakeholders' expectations, stakeholders' requirements, and stakeholders' objectives, collectively referred to as the business problem, are set at the Business Understanding stage. Hence, an auditing inquiry at the **Purpose** stage of the PPM framework seeks information that could guide model examiners to make an informed decision about risks that the business problem may pose to consumers, institutions (including financial), and society at large. For example, this inquiry helps to answer questions about whose interests the business problem serves – consumers', shareholders', or business executives' – and how these interests are measured. Seeing as "all models are wrong, but some are useful"[21] according to the Statistician George Box, we also seek to know what constraints or limitations would be imposed on the algorithmic solutions derived for the business problem. The main point is to evaluate the business goals, requirements, and constraints at a high level with clear measurable markers to determine risks that solving the problem may pose to consumers, institutions, and the greater society.

2.2 Data Understanding

The second focus of the **Purpose** element of the PPM framework is Data Understanding. Any business problem that has an algorithmic solution needs to be formulated as a data problem. Hence, an algorithm auditor should use the **Purpose** element of PPM to further inquire into data formulation or data representation of the business problem to identify how accurately this formulation captures the business problem statement and if any techniques were used to mitigate risks associated with data paucity, data insufficiency, or data quality issues. Data, referred to as features or variables in algorithm parlance, are often used in a predictive or prescriptive algorithmic solution that attempts to use patterns in historical data to estimate formulated business problems. At the **Purpose** stage of the PPM, we seek answers to questions like:

- How representative are features or variables across protected class data like race, sex, and age where applicable?
- Is an intended use case for the model appropriately defined and will the model use be constrained to that use case?
- What is the algorithm optimizing for?[22]
- What metrics would be used to measure model performance and fairness (including potential disparate impact) of the algorithmic solution?
- If there is a cut off score and pass rate at the decision stage of the algorithmic system, what is it and how does it affect model fairness (including disparate impact)?

3 Process Element of PPM

The **Process** stage of the PPM evaluates the design, theory, and logic used to develop an algorithmic solution in the context of business use cases and design objectives. It evaluates the profile of the model

[20] "CRISP-DM Help Overview." SPSS Modeler. IBM. Accessed February 10, 2022. https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview.

[21] Saltelli, Andrea, and Silvio Funtowicz. "When all models are wrong." *Issues in Science and Technology* 30, no. 2 (2014): 79-85. https://issues.org/andrea/

[22] For example, in an underwriting or pricing setting, an algorithm auditor should check if the algorithm is minimizing default risk, maximizing accrued loan interests, or maximizing profits.

team, the data collection and data transformation systems that feed the processing layer of the algorithmic system, the soundness of the methods used to decide features relevant to the model objective and the mathematical architecture of the model used to process the selected features and concludes by evaluating the techniques used to validate the model. This stage is considered under five headings: staff profile; data assessment; model assessment; outcome assessment; and model use and limitations.

3.1 Staff Profile

The design, implementation, and validation of an algorithmic system is done by humans. It is important to apply the PPM framework within the scope of the human profiles involved in decision-making at each stage of model development. Characteristics such as race, ethnicity, sex, and other attributes of the people involved in setting model design objectives and model use cases (i.e., the **Purpose** stage); creating and implementing the process used to actualize the design objectives, (i.e., the **Process** stage); and validating and monitoring the model in the production environment (i.e., the **Monitor** stage), should be acknowledged and accounted for. Entities should ensure that teams working on each stage of model requirement analysis, model development, and model monitoring are diverse and inclusive. Entities should also ensure teams are trained on pertinent topics such as fair housing/lending, privacy, consumer protection, and civil rights. Diverse and well-educated teams can often spot challenges and prevent elements that can lead to unfavorable and deleterious outcomes.

3.2 Data Assessment

The **Process** stage also includes gathering information about the data sources and data fields considered for model development. The auditor should document developmental evidence regarding methodologies used to select features[23] in the final model. Evidence that supports appropriateness of any feature selection method used should be documented and reviewed by the auditor together with information about if (or why) alternative feature selection methods were used.

At the data assessment stage, a data imbalance report that contains the distribution of each model feature across model target or label[24] and available protected class data is generated. For a supervised learning problem with a continuous label, the imbalance report is based on binned categories of the target, and for an unsupervised learning task, the imbalance report is documented only for available protected class data. Missingness patterns in the model label, selected features, and available protected class data should then be evaluated. The data imbalance report and data missingness report would provide evidence that algorithm auditors could evaluate to decide if the selected features are sufficiently representative or inclusive.

In addition, any preprocessing or transformation on model features, values of fairness metrics for each available protected class data, and strategies used to split development data into training and test sets are evaluated at this stage. Feature selection, data imbalance, and data missingness reports are evaluated on training and test splits to ensure that the two splits are similar in patterns and that they reasonably represent the overall data patterns.

[23] In this article, we consider model variable and model feature to be synonymous. A feature is a variable that is used to explain or predict a model outcome.

[24] A model label or target is the event or value being predicted by the model. In this article, we prefer to use model label.

3.3 Model Assessment

Information about training algorithms, parameters, hyper-parameters, and any fairness constraints used during model development is evaluated at the model assessment stage. If a technique is used to mitigate possible impacts of data imbalance (across model label or across available protected class), then the methodology should be evaluated at this stage. An algorithm auditor should evaluate outcomes of any test used to identify proxies for available protected class data. If multiple models are trained, the model selection criteria should also be reviewed.

Any fairness constraint used at the data stage, model stage, or post-modeling stage should be evaluated at the model assessment stage. The auditor should also request any information showing that the selected model is the least discriminatory alternative (LDA) that meets the design objective. Any method used to evaluate model fitness and used to fix or fit model hyperparameter should be reviewed by the auditor. In addition, any sensitivity analysis of how model selection and hyper-parameter tuning affect the search for LDAs in the context of model performance should be evaluated.

3.4 Outcome Assessment

At the outcome assessment stage, the performance of the final model is evaluated in line with the scope and metrics laid out in the **Purpose** section of the PPM. The goal of the evaluation is to establish whether the final model meets its design objectives, which includes minimizing risks to consumers, institutions, and society. Model performance and fairness metrics on training and test data are compared to assess model robustness and stability. It is important to identify market conditions, the range of model feature values, and other scenarios that may stretch model performance. To this end, an auditor should review any reports on local and global sensitivity analysis as well as conduct other checks for model stability and robustness.

Outcomes from a sound, trustworthy model should be explainable[25], and the structure of the model should enable the explainability of its outcomes[26]. While assessing elements of each decile, the auditor could focus on the top and bottom deciles of model predictions on the test data and review explainability reports on representative samples from the deciles. Special attention should be given to the soundness of the explainability technique(s) used and distribution of protected class data in the selected samples. For example, where race is the protected class of interest, the examined records from the deciles should contain members from all racial groups and explainability reports of the predicted outcomes should be reviewed.

3.5 Model Use and Limitations

Any algorithm-based system assumes that its processing unit, usually a statistical model or a machine learning model, represents the reality of its business objective. Known limitations and assumptions of the model should be documented and reviewed. In addition, circumstances where the model may or may not be used outside the scope of its intended uses should also be documented and reviewed.

---

[25] Explainability is the degree to which a human can understand the cause of a decision (*see,* Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38. https://www.sciencedirect.com/science/article/pii/S0004370218305988) or the degree to which a human can consistently predict the model's result (see Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! criticism for interpretability." *Advances in neural information processing systems* 29 (2016). https://papers.nips.cc/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html).

[26] Molnar, Christoph. "A guide for making black box models explainable." URL: https://christophm.github.io/interpretable-ml-book/index.html

4 Monitor Element of PPM

All models are abstractions of realities, and their usefulness depends on how well they represent reality, not in a way that replicates prejudices that may exist in our society, but rather in a way that removes biases from the underlying data and mitigates injuries to consumers. The **Monitor** stage of the PPM framework serves to document model validity and robustness, as well as consumer fairness, privacy, and related harms, in the production environment (i.e., while using the model to make business decisions). The monitoring of statistical models, machine learning systems, and AI solutions is necessary to ensure total end-to-end development of a utility that is truly fair. Monitoring a solution after its original development allows developers to detect changes between patterns of the data used to develop the deployed model (i.e., training stage), and patterns of the data used to make business decisions post-development (i.e., serving stage). Differences in patterns between the training and serving stages should be monitored and can inform when and how a model should be retrained, redeveloped, or retired.

4.1 Production Model Validation

In the **Monitor** stage of the PPM, a report on the missingness patterns of the features used to develop the trained model is compared with a similar report on the features used in the production version of the model. A statistically significant difference between these missingness patterns can cause injuries to consumers, especially consumers of color, and it can cause a degradation of, lag, and drift in model performance. A check is also conducted at this stage to assure that the same missing data imputation method being used at the training stage is being applied in the serving stage[27]. In addition, the distribution of each feature used in the model and the distribution of model predictions in training and serving stages are compared[28]. Distribution distance metrics[29] like feature stability index (FSI) and population stability index (PSI) are derived from these distributions and then used to make judgments about the level of deterioration in stability and reliability of the deployed models. Model evaluation and fairness metrics in environments like training, testing, and production are compared. Where applicable, fairness metrics across protected class categories before and after the model is deployed in production are compared using appropriate distance metrics. The goal of these comparisons is to inform decisions about whether the production model should be retrained, patched, or retired.

4.2 Protection from Confidentiality and Integrity Attacks (CIAs)

Algorithmic systems must be secured and protected from attacks on their structures and underlying data so that bad actors cannot induce outputs or behaviors that may be adversarial to consumers[30]. It is also important to monitor algorithmic systems for protection from denial of service (DoS) attacks so that consumers in dire need of critical services have access to them. While a business has an incentive to make sure its algorithm-based services are always available to consumers, the **Monitor** section of the PPM focuses on the confidentiality and integrity aspects of algorithmic systems to assure that consumer risks are minimized. We suggest the review of defenses built into the system at training time and

[27] Van Buuren, Stef. *Flexible Imputation of Missing Data*. CRC press, 2018. https://stefvanbuuren.name/publication/2018-01-01_vanbuuuren2018/

[28] Dhinakaran, Aparna. “What Is ML Observability?” Towards Data Science. Medium, August 16, 2021. https://towardsdatascience.com/what-is-ml-observability-29e85e701688

[29] Dhinakaran, Aparna. “Using Statistical Distance Metrics for Machine Learning Observability.” Towards Data Science. Medium, August 16, 2021. https://towardsdatascience.com/using-statistical-distance-metrics-for-machine-learning-observability-4c874cded78

[30] Hu, Hongsheng, Zoran Salcic, Gillian Dobbie, and Xuyun Zhang. "Membership Inference Attacks on Machine Learning: A Survey." *arXiv preprint arXiv:2103.07853* (2021). https://arxiv.org/abs/2103.07853

runtime under three lines[31]: 1) robustness of distribution drift defense; 2) use of privacy-aware techniques like differential privacy; and 3) defenses that assure fairness and accountability. The first line of defense guarantees that model performance is retained in the likely event that input distributions in training and production environments differ. The second line of defense assures that in the event of an adversarial attack like a membership inference attack (MIA)[32], consumers' privacy is guaranteed, and data leakages may not be exploited to their detriment[33]. The third and last line of defense assures that fairness metrics and explainability codes at training time relatively hold during the lifetime of an algorithmic model.

5 Conclusion

End-to-end development of an algorithm-based solution involves the stages covered in NFHA's PPM framework, which creates a standardized and comprehensive approach for auditing an algorithmic system. It builds on existing frameworks, such as CRISP-DM (Cross-Industry Standard Process for Data Mining), MRM (Model Risk Management) supervisory guidance, and NIST's proposal for identifying and managing bias in Artificial Intelligence, in addition to standards NFHA has developed over years of monitoring and addressing algorithmic bias. NFHA's **Purpose Process and Monitoring** framework provides an equity-centered auditing solution at a time when policy makers and civil rights organizations are calling for fairness, accountability, transparency, explainability, and interpretability.

[31] Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman. "Sok: Security and privacy in machine learning." In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 399-414. IEEE, 2018. https://oaklandsok.github.io/papers/papernot2018.pdf

[32] A membership inference attack is an attack that checks if a person's record was part of a model training data. For example, an MIA test may conclude that a person's record was used to train an underwriting model with a likelihood of, say, 95%. This means the bad actor has 95% belief that the person belongs to, say, the "deny" group. If this information is leaked, then it may be easier to identify the credit score category of the person.

[33] Goldsteen, Abigail, Gilad Ezov, and Ariel Farkash. "Reducing Risk of Model Inversion Using Privacy-Guided Training." *arXiv preprint arXiv:2006.15877* (2020). http://export.arxiv.org/pdf/2006.15877