# Proposed Guidelines for the Responsible Use of Explainable Machine Learning

**Patrick Hall**
H2O.ai
Washington, DC
phall@h2o.ai

**Navdeep Gill**
H2O.ai
Mountain View, CA
navdeep@h2o.ai

**Nicholas Schmidt**
BLDS, LLC
Philadelphia, PA
nschmidt@bldsllc.com

## 1 Introduction

Explainable machine learning (ML) enables human learning from ML, human appeal of automated model decisions, regulatory compliance, and security audits of ML models.[1,2,3] Explainable ML (i.e. e*x*plainable *a*rtificial *i*ntelligence or XAI) has been implemented in numerous open source and commercial packages and explainable ML is also an important, mandatory, or embedded aspect of commercial predictive modeling in industries like financial services.[4,5,6] However, like many technologies, explainable ML can be misused, particularly as a faulty safeguard for harmful black-boxes, e.g. *fairwashing* or *scaffolding*, and for other malevolent purposes like stealing models and sensitive training data [1], [38], [40], [42], [45]. To promote best-practice discussions for this already in-flight technology, this short text presents internal definitions and a few examples in Section 2 before covering the proposed guidelines in Subsections 3.1 – 3.4. This text concludes in Section 4 with a seemingly natural argument for the use of interpretable models and explanatory, debugging, and disparate impact testing methods in life- or mission-critical ML systems.

## 2 Definitions and Examples

While the explainable ML community has apparently not yet adopted a clear taxonomy of concepts or a precise vocabulary, many authors have grappled with ideas related to interpretability and explanations. Some of these efforts include: "A Survey of Methods for Explaining Black Box Models" by Guidotti et al. [20], "The Mythos of Model Interpretability" by Lipton [29], ***Interpretable Machine Learning*** by Molnar [33], "Interpretable Machine Learning: Definitions, Methods, and Applications" by Murdoch et al. [35], and "Challenges for Transparency" by Weller [48]. To decrease ambiguity herein, this section uses the review and survey corpus and practical examples to address the terms and phrases *interpretable*, *explanation*, *explainable ML*, *interpretable models*, *model debugging techniques*, *unwanted sociological bias*, and *fairness techniques* before proposing guidelines.

### 2.1 Interpretable and Explanation

Doshi-Velez and Kim [10] define interpretability in ML as, "the ability to explain or to present in understandable terms to a human." Professor Sameer Singh of the University of California at

[1] This text and associated software are not, and should not be construed as, legal advice or requirements for regulatory compliance.

[2] In the U.S., interpretable models, explanations, disparate impact testing, and the model documentation they enable may be required under the Civil Rights Acts of 1964 and 1991, the Americans with Disabilities Act, the Genetic Information Nondiscrimination Act, the Health Insurance Portability and Accountability Act, the Equal Credit Opportunity Act (ECOA), the Fair Credit Reporting Act (FCRA), the Fair Housing Act, Federal Reserve SR 11-7, and the European Union (E.U.) Greater Data Privacy Regulation (GDPR) Article 22 [49].

[3] For various security applications, see: "Proposals for Model Vulnerability and Security".

[4] E.g. open source software listed here: "Awesome machine learning interpretability".

[5] E.g., Datarobot, H2O Driverless AI, SAS Visual Data Mining and Machine Learning, Zest AutoML.

[6] E.g., "Deep Insights into Explainability and Interpretability of Machine Learning Algorithms and Applications to Risk Management".

Robust AI in Financial Services Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Irvine (UCI), co-inventor of the seminal local interpretable model-agnostic explanation (LIME) technique, defines *explanation* as a, "collection of visual and/or interactive artifacts that provide a user with sufficient description of a model's behavior to accurately perform tasks like evaluation, trusting, predicting, or improving a model."[7] And Gilpin et al. [16] posit that a *good explanation* occurs when modelers or consumers "can no longer keep asking why" in regards to some ML model behavior. These three thoughtful characterizations link explainability to interpretability, give clarity on explanation, and provide an abstract goal for any explainability task.

### 2.2 Explainable ML and Interpretable Models

Herein *explainable ML* means mostly post-hoc analysis and techniques used to understand trained model mechanisms or predictions. Examples of common explainable ML techniques include:

- Local and global feature importance, e.g., Shapley and derivative-based feature attribution [3] [26], [31], [39], [43].
- Local and global model-agnostic surrogate models, e.g., surrogate decision trees and LIME [7], [8], [9], [23], [37], [47].
- Local and global visualizations of model predictions, e.g., accumulated local effect (ALE) plots, 1- and 2-dimensional partial dependence plots, and individual conditional expectation (ICE) plots [5], [15], [17].

Although difficult to quantify, credible research efforts into scientific measures of interpretability are underway [14], [34], and the ability to measure degrees of interpretability implies it is not a binary, on-off quantity. Here, unconstrained, traditional black-box ML models, such as multilayer perceptron (MLP) neural networks and gradient boosting machines (GBMs), are said to be difficult to interpret, potentially unsafe for use in life- or mission-critical applications, but not necessarily completely unexplainable. In this text, *interpretable models* (i.e., white-box models) will include linear models, decision trees, rule-based models, constrained or Bayesian variants of traditional black-box ML models, or novel types of models designed to be directly interpretable. Examples of newer, highly interpretable ML modeling techniques include explainable neural networks (XNNs), explainable boosting machines (EBMs, GA2Ms), monotonically constrained GBMs, scalable Bayesian rule lists, or super-sparse linear integer models (SLIMs), [30], [46], [47], [50].[8,9,10]

### 2.3 Model Debugging Techniques

Herein *model debugging techniques* test ML models to increase trust in mechanisms and predictions. Debugging techniques include model assertions, security audits, variants of sensitivity (i.e., *what-if?*) analysis, variants of residual analysis and residual explanation, and unit tests to verify the accuracy or security of ML models [2], [25].[11] Model debugging should also include remediating any discovered errors or vulnerabilities.

### 2.4 Unwanted Sociological Bias and Fairness Techniques

In this text, *unwanted sociological bias* encompasses several forms of discrimination that may manifest in ML, including overt discrimination, disparate treatment, and disparate impact (DI), i.e., unintentional discrimination. DI may be caused by model misspecification, inaccurate or incomplete data, or data that has differing correlations or dependencies among demographic groups of individuals, driving differences in favorable model outcomes. A model is said to be biased here if, (1) group membership is not independent of the likelihood of a favorable outcome, or (2) under certain circumstances, membership in a *subset* of a group is not independent of the likelihood of a favorable outcome (i.e., *local* bias). Underlying discrimination that causes bias may or may not be illegal, depending on how it arises and applicable discrimination laws. Herein *fairness techniques* are used to diagnose and remediate unwanted sociological bias in ML models. Diagnosis approaches include

[7] From: "Proposed Guidelines for the Responsible Use of Explainable Machine Learning" (presentation only).

[8] EBM, as implemented in the Microsoft `interpret` package.

[9] Monotonic GBM, as implemented in `XGBoost` or `h2o`.

[10] And similar methods, e.g.: `https://users.cs.duke.edu/~cynthia/papers.html`.

[11] And similar methods, e.g.: `https://debug-ml-iclr2019.github.io/`.

DI testing and other tests for bias [12]. Remediation methods tend to involve model selection by minimization of bias, preprocessing training data, e.g., reweighing [24], training unbiased models, e.g., adversarial de-biasing [51], or post-processing model predictions, e.g., by equalized odds [22].[12]

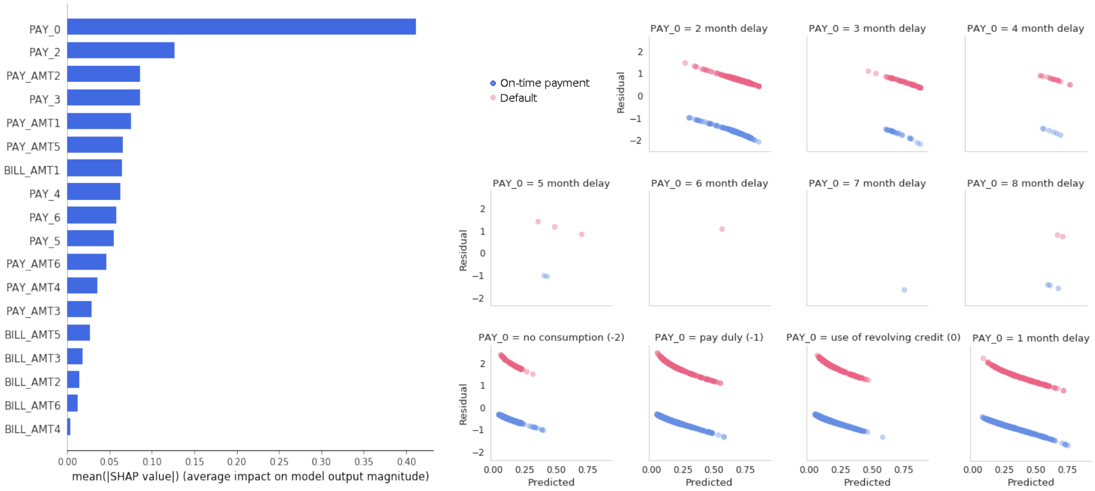
# 3 Proposed Guidelines for the Responsible Use of Explainable ML

Four guidelines are proposed and discussed in Subsections 3.1 – 3.4 to assist practitioners in avoiding unintentional misuse or identifying intentional abuse of explainable ML. The guidelines are:

1. Use explanations to enable understanding.
2. Learn how explainable ML is used for nefarious purposes.
3. Augment surrogate models with direct explanations.
4. Use highly interpretable mechanisms for life- or mission-critical ML.

Important corollaries to the guidelines are also highlighted and simple, reproducible software examples accompany the guidelines to avoid hypothetical reasoning whenever possible.

## 3.1 Guideline: Use Explanations to Enable Understanding



(a) Consistent global Shapley feature importance values for $g_{\text{GBM}}$.

(b) $g_{\text{GBM}}$ deviance residuals and predictions by `PAY_0`.

Figure 1: An unconstrained GBM probability of default model, $g_{\text{GBM}}$, generally over-emphasizes the importance of the input feature `PAY_0`, a customer's most recent repayment status, in the UCI credit card data. $g_{\text{GBM}}$ often produces large positive residuals when `PAY_0` indicates on-time payments (`PAY_0` $\leq 1$) and large negative residuals when `PAY_0` indicates late payments (`PAY_0` $> 1$). Combining explanatory and debugging techniques shows that $g_{\text{GBM}}$ is explainable, but probably not trustworthy.

Explanations are often discussed in the context of trust (e.g., Ribeiro et al. [37]), but explanations alone are not sufficient for trust in ML models. Explanation, as a general concept, is related more directly to understanding and transparency than to trust.[13] Simply put, one can understand and explain a model without trusting it. One can also trust a model and not be able to understand or explain it. Consider the following example scenarios.

- **Explanation and understanding without trust**: In Figure 1, global Shapley explanations and residual analysis identify a pathology in an unconstrained GBM model, $g_{\text{GBM}}$, trained on the UCI credit card dataset [27].[14] $g_{\text{GBM}}$ over emphasizes the input feature `PAY_0`, or a

[12] And similar methods, e.g.: `http://www.fatml.org/resources/relevant-scholarship`.

[13] The Merriam-Webster definition of *explain*, accessed Sept. 8th 2019, does not mention *trust*.

[14] Code to replicate Figure 1: `https://bit.ly/2m58Lxl`.

customer's most recent repayment status. Due to over-emphasis of PAY_0, $g_{\text{GBM}}$ is often unable to predict on-time payment if recent payments are delayed (PAY_0 $> 1$), causing large negative residuals. $g_{\text{GBM}}$ is also often unable to predict default if recent payments are made on-time (PAY_0 $\leq 1$), causing large positive residuals. In this example scenario, $g_{\text{GBM}}$ is explainable, but likely untrustworthy.

- **Trust without explanation and understanding**: Years before reliable explanation techniques were widely acknowledged and available, black-box predictive models, such as autoencoder and MLP neural networks, were used for fraud detection in the financial services industry [18]. When these models performed well, they were trusted.[15,16] However, they were not explainable or well-understood by contemporary standards.

Explanations typically increase trust in models as a side-effect when they are acceptable to human users by various criteria. As illustrated in Figure 4, in an ideal scenario, explanation techniques should be used to directly increase understanding in ML models, while debugging and DI testing methods should be used to directly promote trust.

### 3.2 Guideline: Learn How Explainable ML is Used for Nefarious Purposes

When used disingenuously, explainable ML methods can provide cover for misused or intentionally abusive black-boxes [1], [38], [42]. Explainable ML methods can also enable hacking or stealing of models or data through public prediction APIs or other endpoints [40], [45]. Moreover, explainable ML methods are likely to be used for other nefarious purposes in the future and may be used for unknown destructive purposes now. Responsible practitioners need to understand the malevolent side of this technology to better detect and correct misuse and abuse.

#### 3.2.1 Corollary: Use Explainable ML for Security Audits

Use explainable ML techniques to test ML systems for vulnerabilities to model stealing, inversion, and membership inference attacks.

#### 3.2.2 Corollary: Explainable ML Can be Used to Crack Nefarious Black-boxes

Used as white-hat hacking tools, explainable ML can help draw attention to accuracy or unwanted sociological bias problems in proprietary black-boxes. See Angwin et al. [4] for evidence that cracking proprietary black-box models for oversight purposes is possible.[17]

#### 3.2.3 Corollary: Explainable ML is a Privacy Vulnerability

Recent research shows that providing explanations along with predictions eases attacks that can compromise sensitive training data [41].

### 3.3 Guideline: Augment Surrogate Models with Direct Explanations

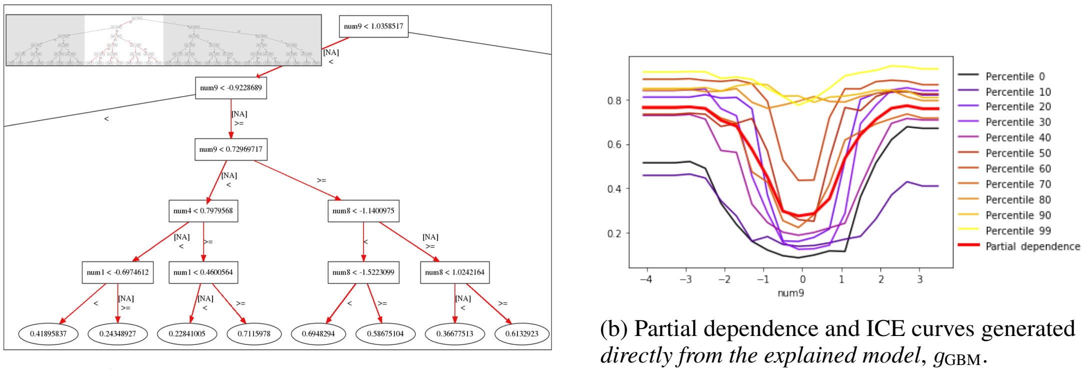
Models of models, or surrogate models, can be helpful explanatory tools, but they are usually approximate, low-fidelity explainers. Aside from (1) a global or local summary of a complex model provided by a surrogate model can be helpful sometimes and (2) much work in explainable ML has been directed toward improving the fidelity and usefulness of surrogate models [7], [8], [9], [23], [47], *many explainable ML techniques have nothing to do with surrogate models*. One of the most exciting breakthroughs for supervised learning problems in explainable ML is the application of a coalitional game theory concept, Shapley values, to compute feature attributions which are consistent globally and accurate locally using the trained model itself [31], [43]. An extension of this idea, called Tree SHAP, has already been implemented for popular tree ensemble methods [32].

There are many other explainable ML methods that operate on trained models directly such as partial dependence, ALE, and ICE plots [5], [15], [17]. Surrogate models and explanatory techniques that

[15] E.g., "Reduce Losses from Fraudulent Transactions".

[16] E.g., "SAS Secures Technology Patent for Better Fraud Detection Performance".

[17] This text makes no claim on the quality of the analysis in Angwin et al. (2016), which has been criticized [13]. This now infamous analysis is presented only as evidence that motivated activists can crack proprietary black-boxes using surrogate models and other explanatory techniques. Moreover, such analyses would likely improve with established best-practices for explainable ML.



(a) Naïve $h_{\text{tree}}$, *a surrogate model*, forms an approximate overall flowchart for the explained model, $g_{\text{GBM}}$.

(b) Partial dependence and ICE curves generated *directly from the explained model*, $g_{\text{GBM}}$.

Figure 2: $h_{\text{tree}}$ displays known interactions in $f = X_{\text{num1}} * X_{\text{num4}} + |X_{\text{num8}}| * X_{\text{num9}}^2$ for $\sim -1 < X_{\text{num9}} <\sim 1$. Modeling of the known interactions in $f$ by $g_{\text{GBM}}$ is also highlighted by the divergence of partial dependence and ICE curves for $\sim -1 < X_{\text{num9}} <\sim 1$. Explanations from a surrogate model have augmented and confirmed findings from a direct model visualization technique.

operate directly on trained models can also be combined, for instance by using partial dependence, ICE, and surrogate decision trees to investigate and confirm modeled interactions [21]. In Figure 2, an unconstrained GBM, $g_{\text{GBM}}$, models a known signal generating function $f$:

$$f(\mathbf{X}) = \begin{cases} 1 & \text{if } X_{\text{num1}} * X_{\text{num4}} + |X_{\text{num8}}| * X_{\text{num9}}^2 + e \geq 0.42 \\ 0 & \text{if } X_{\text{num1}} * X_{\text{num4}} + |X_{\text{num8}}| * X_{\text{num9}}^2 + e < 0.42 \end{cases} \tag{1}$$

where $e$ signifies the injection of random noise in the form of label switching for roughly 15% of the training and validation observations.[18] $g_{\text{GBM}}$ is then trained such that $g_{\text{GBM}}(\mathbf{X}) \approx f(\mathbf{X})$ in training and validation data. $h_{\text{tree}}$, displayed in Figure 2a, is extracted such that $h_{\text{tree}}(\mathbf{X}) \approx g_{\text{GBM}}(\mathbf{X}) \approx f(\mathbf{X})$ in validation data. Partial dependence and ICE plots are generated directly for $g_{\text{GBM}}$ in the same validation data and overlaid in Figure 2b. The parent-child node relationships displayed in $h_{\text{tree}}$ for $\sim -1 < X_{\text{num9}} <\sim 1$ in 2a and the divergence of ICE and partial dependence curves in 2b for $\sim -1 < X_{\text{num9}} <\sim 1$ help confirm and explain how $g_{\text{GBM}}$ learned the interactions in $f$. As in Figure 1, combining different approaches provided additional, beneficial information about a ML model.

### 3.3.1 Corollary: Augment LIME with Direct Explanations

LIME is important, imperfect (like every other ML technique), and vulnerable to adversarial manipulation [42]. LIME, in its most popular implementation, uses local linear surrogate models fit to perturbed, locally weighted samples to explain regions of machine-learned decision boundaries or response functions [37]. Like other surrogate models, LIME can be combined with model-specific methods for validation and to yield deeper insights. Consider that Tree SHAP can provide locally accurate and consistent point estimates for local feature importance as in 3b below. LIME can then provide approximate information about modeled local linear trends around the same point. Table 1 contains LIME $h_{\text{GLM}}$ coefficients for a local region of a validation set sampled from the UCI credit card data defined by `PAY_0 > 1`, or customers with a fairly high risk of default due to late most recent payments.[19] $h_{\text{GLM}}$ models the predictions of a simple interpretable decision tree model, $g_{\text{tree}}$, displayed in 3a. $h_{\text{GLM}}$ coefficients show linear trends between features in the sampled set $\mathbf{X}_{\text{PAY\_0>1}}$ and $g_{\text{tree}}(\mathbf{X}_{\text{PAY\_0>1}})$. Because $h_{GLM}$ is relatively well-fit (0.73 $R^2$) and has a logical intercept (0.77), it can be used along with Shapley values to reason about the modeled average behavior for risky customers, to differentiate the behavior of any one specific risky customer from their peers under the model, or to validate LIME results. Such additional information can be useful for model debugging, compliance purposes, or for verifying LIME-based feature importance.

[18] Code to replicate Figure 2: `https://bit.ly/2kSuAQD`.

[19] Code to replicate Table 1: `https://bit.ly/2miCPpo`.

Table 1: Coefficients for local linear model, $h_{\text{GLM}}$, with an intercept of 0.77 and an $R^2$ of 0.73. $h_{\text{GLM}}$ is trained on a segment of the UCI credit card dataset containing higher-risk customers with late most recent repayment statuses, $\mathbf{X}_{PAY\_0>1}$, and the predictions of a decision tree, $g_{\text{tree}}(\mathbf{X}_{\text{PAY\_0>1}})$.

| $h_{\text{GLM}}$ Feature | $h_{\text{GLM}}$ Coefficient |
|---|---|
| `PAY_0 == 4` | 0.0009 |
| `PAY_2 == 3` | 0.0065 |
| `PAY_5 == 2` | −0.0006 |
| `PAY_6 == 2` | 0.0036 |
| `BILL_AMT1` | 3.4339e−08 |
| `PAY_AMT1` | 4.8062e−07 |
| `PAY_AMT3` | −5.867e−07 |

### 3.4 Guideline: Use Highly Interpretable Mechanisms for Mission- or Life-Critical ML

Given the known difficulties with explaining black-boxes [38], the existence of unwanted social bias in data and ML models [6], the security vulnerabilities of ML (e.g., Shokri et al. [40], Tramèr et al. [45]), and the potentially surprising behavior of black-boxes (e.g., Nguyen et al. [36], Szegedy et al. [44]), it appears prudent today to use highly transparent ML mechanisms for applications that make life-altering or high-value decisions. Interpretability, as enabled by interpretable models and post-hoc explanations (see Corollary 3.4.1), may be mandated by regulation for some life- or mission-critical applications, but interpretability is also recommended for any ML application in which inevitable wrong decisions should be appealable. This subsection discusses a few details and examples regarding regulated ML applications and appeal, and also advocates for trust-enhancing DI testing (see Corollaries 3.4.2 – 3.4.4) in high-stakes or human-centered applications.

Interpretable ML mechanisms are required under numerous regulatory statutes in the U.S., and explainable ML tools like LIME and other surrogate models, partial dependence plots, and global and local feature importance are already used to document, understand, and validate some predictive models in the financial services industry [23], [47].[2, 6] Moreover, adverse action notices are mandated under the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA) for many credit lending, employment, and insurance decisions in the U.S.[20] If ML is used for such decisions, it must be explained in terms of adverse action notices.[21] Shapley values, and other local feature importance approaches, provide a convenient methodology to rank the direct contribution of input features to final model decisions and potentially generate customer-specific adverse action notices.

Aside from regulatory mandates, interpretable models and explanations enable logical appeal processes for automated decisions made by ML models. Consider being negatively impacted by an erroneous black-box model decision, say for instance being mistakenly denied a loan or parole. How would you argue your case for appeal without knowing how model decisions were made? According to the New York Times, a man named Glenn Rodríguez found himself in this unfortunate position in a penitentiary in Upstate New York in 2016.[22]

#### 3.4.1 Corollary: Use Interpretable Models Along with Explanation Techniques

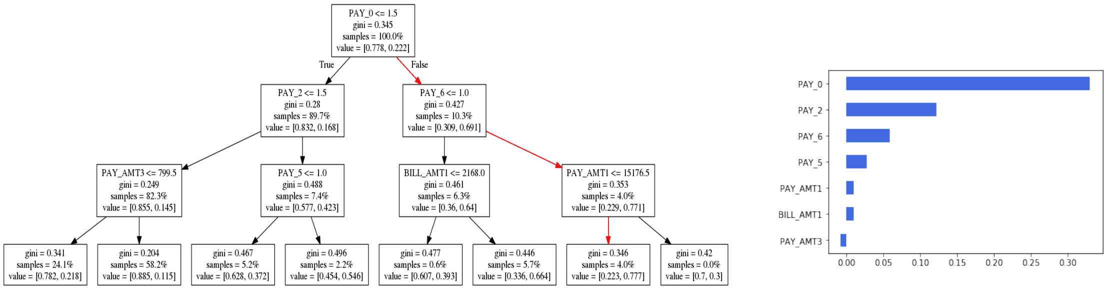
Some well-known publications have focused either on interpretable models (e.g., Ustun and Rudin [46], Yang et al. [50]) or on post-hoc explanations (e.g., Lundberg and Lee [31], Ribeiro et al. [37]), but the two can be used together in the context of a holistic ML workflow, illustrated in Figure 4. Consider the seemingly useful example case of augmenting globally interpretable models with local post-hoc explanations. A practitioner could train a single decision tree, a globally interpretable model, then apply local explanations in the form of Shapley feature importance as in Figure 3.[23] This enables practitioners to see accurate numeric feature contributions for each prediction and the entire directed graph of the decision tree. Even for interpretable models, such as linear models and decision

[20] See: “Adverse Action Notice Requirements Under the ECOA and the FCRA”.

[21] This is apparently already happening: “New Patent-Pending Technology from Equifax Enables Configurable AI Models”.

[22] See: “When a Computer Program Keeps You in Jail”.

[23] Code to replicate Figure 3: `https://bit.ly/2miCPpo`.



(a) Simple decision tree, $g_{\text{tree}}$, trained on the UCI credit card data to predict default with validation AUC of 0.74. The decision policy for a high-risk individual is highlighted in red.

(b) Locally-accurate Shapley contributions for the highlighted individual's probability of default.

Figure 3: The decision-policy for a high-risk customer is highlighted in 3a and the locally-accurate Shapley contributions for this same individual's predicted probability are displayed in 3b. Due to their consistency properties, Shapley values highlight the local importance of features not on the decision path in this particular encoding, i.e., $g_{\text{tree}}$, of the unknown signal-generating function, which is likely helpful for the generation of consistent adverse action notices across similar data and models.

Table 2: Basic group disparity metrics for females under monotonically constrained GBM model, $g_{\text{mono}}$, trained on the UCI credit card dataset.

| | Adverse Impact Ratio | False Omissions Rate Disparity |
|---|---|---|
| Female | 0.8869 | 0.7866 |

trees, Shapley values present accuracy and consistency advantages over standard feature attribution methods [28], [31], [32]. Shapley values also enable the consistent ranking of input features for each model decision, which is likely helpful for FCRA and ECOA compliance. Another twist on the idea of combining explainable ML methods and interpretable models is described by Gosiewska et al. [19] in "Surrogate Assisted Feature Extraction for Machine Learning (SAFE ML)". In the SAFE ML approach, features learned by more complex models are extracted and used in an explainable fashion to increase the accuracy of more interpretable models. Aren't either of these augmented processes more desirable than either an interpretable model or post-hoc explanations by themselves?

### 3.4.2 Corollary: Use Explanations Along with DI Testing

Like interpretable models, fairness methods are often presented in different articles than post-hoc explanatory methods. However, in banks for example, using post-hoc explanatory tools along with DI testing is often necessary to comply with model documentation guidance and with fair lending regulations.[24,25] To clarify, explanatory techniques should *not* replace DI testing for bias detection purposes, but in general, explanations increase transparency and understanding of model mechanisms and predictions, while DI auditing and remediation increases trust that model predictions are as fair as possible. As in previous sections, trust and understanding are different but complimentary goals achieved by combining multiple approaches.

Table 2 displays basic group disparity metrics for a monotonic GBM model, $g_{\text{mono}}$, trained on the UCI credit card data.[26] In this example, $g_{\text{mono}}$ displays group parity for adverse impact with `male` as the reference level according to the four-fifths rule, but also presents unwanted false omissions rate bias against females, indicating that males may be receiving too much credit they cannot repay, potentially preventing females from receiving that credit.[27] This disparity can be remedied by gently increasing the decision cutoff for $g_{\text{mono}}$, and Shapley values can also explain each $g_{\text{mono}}$ prediction.[26]

[24] See: "Interagency Fair Lending Examination Procedures".

[25] See: "CFPB Consumer Laws and Regulations: ECOA".

[26] Code to replicate Table 2: `https://bit.ly/2lZUlyN`.

[27] The four-fifths rule was delineated by the Equal Employment Opportunity Commission (EEOC) as a measure of DI that would be of concern to regulators. This threshold has been associated with a specific measure of DI, the adverse impact ratio (AIR). While it may be applied to other measures of fairness, as in Table 2, this is often irrelevant in real-world compliance and litigation settings in the U.S.

Beyond explaining predictions, Explainable ML can assist in the difficult problem of determining input features within wide training sets that drive DI. For example, weighted average Shapley values can be analyzed by demographic segment, highlighting the features that have the largest deleterious impact on the protected segment. Explainable ML techniques, especially when paired with clustering, can also be useful for isolating instances of local bias.

### 3.4.3 Corollary: Explanation is Not a Frontline Fairness Tool

Demographic attributes cannot currently be used in predictive models for high-stakes and commercially viable uses of explainable ML in credit lending, insurance, and employment in the U.S. that fall under FCRA, ECOA, or other applicable regulations. Thus their contribution to models cannot be assessed using accurate, direct explainable ML techniques. Even when demographic attributes can be used in models, it has been shown that explanations may not detect unwanted sociological bias [1]. Given these drawbacks, it is recommended that fairness techniques are used to test for and remediate bias, and explanations are used to understand bias when appropriate (see Corollary 3.4.2).

### 3.4.4 Corollary: Use DI Testing Along with Constrained Models

Unconstrained ML models can treat similar individuals differently due to small differences in input data values, causing local bias that is not detectable with standard DI testing methods that measure group fairness [11]. To mitigate local bias when using ML, and to ensure standard bias or DI testing methods are most effective, pair such testing with constrained models.

## 4 Conclusion: a Holistic Approach for Life- or Mission-Critical ML

ML systems are used today to make life-altering decisions about employment, bail, parole, and lending,[28] and the scope of decisions delegated to ML systems seems likely to expand in the future. Many researchers and practitioners are tackling DI, inaccuracy, privacy violations, and security vulnerabilities with a number of brilliant, but sometimes siloed, approaches. By proposing some straightforward explainable ML guidelines, this short text also gives examples of combining innovations from several sub-disciplines of ML research to train understandable and trustworthy predictive modeling systems. As illustrated in Figure 4, these innovations can be used together, and this combination may be better-suited than conventional ML methods for use in business- and life-critical applications.

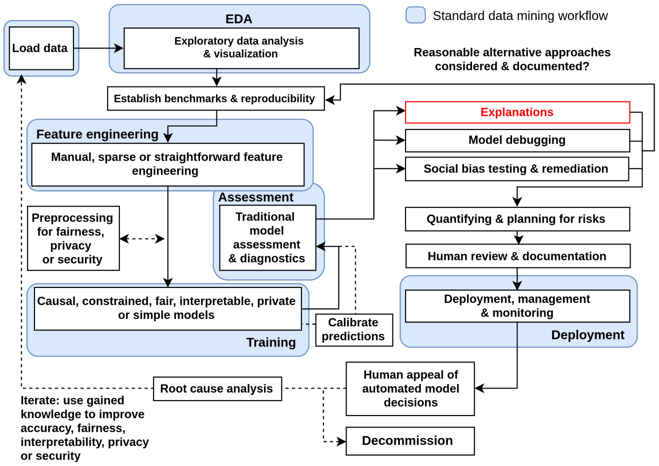

Standard data mining workflow
Load data
EDA
Exploratory data analysis & visualization
Reasonable alternative approaches considered & documented?
Establish benchmarks & reproducibility
Explanations
Model debugging
Social bias testing & remediation
Feature engineering
Manual, sparse or straightforward feature engineering
Assessment
Traditional model assessment & diagnostics
Preprocessing for fairness, privacy or security
Quantifying & planning for risks
Human review & documentation
Causal, constrained, fair, interpretable, private or simple models
Training
Calibrate predictions
Deployment, management & monitoring
Deployment
Root cause analysis
Human appeal of automated model decisions
Iterate: use gained knowledge to improve accuracy, fairness, interpretability, privacy or security
Decommission

Figure 4: A diagram of a proposed holistic ML workflow in which explanations (highlighted in red) are used along with interpretable models, DI analysis and remediation techniques, and other review and appeal mechanisms to create an understandable and trustworthy ML system.

[28] See: "Debugging Machine Learning Models".

## Acknowledgements

The authors thank Przemyslaw Biecek, Pramit Choudhary, Benjamin Cox, Lingyao Meng, Christoph Molnar, Sameer Singh, and Bryce Stephens for their helpful input and insights.

## References

[1] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the Risk of Rationalization. *arXiv preprint arXiv:1901.09749*, 2019. URL: `https://arxiv.org/pdf/1901.09749.pdf`.

[2] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. Modeltracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 337–346. ACM, 2015. URL: `https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/amershi.CHI2015.ModelTracker.pdf`.

[3] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards Better Understanding of Gradient-based Attribution Methods for Deep Neural Networks. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018. URL: `https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/249929/Flow_ICLR_2018.pdf`.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *ProPublica*, 2016. URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[5] Daniel W. Apley. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv preprint arXiv:1612.08468*, 2016. URL: `https://arxiv.org/pdf/1612.08468.pdf`.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL: `http://www.fairmlbook.org`.

[7] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting Blackbox Models via Model Extraction. *arXiv preprint arXiv:1705.08504*, 2017. URL: `https://arxiv.org/pdf/1705.08504.pdf`.

[8] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable Reinforcement Learning Via Policy Extraction. In *Advances in Neural Information Processing Systems*, pages 2494–2504, 2018. URL: `http://papers.nips.cc/paper/7516-verifiable-reinforcement-learning-via-policy-extraction.pdf`.

[9] Mark W. Craven and Jude W. Shavlik. Extracting Tree-Structured Representations of Trained Networks. *Advances in Neural Information Processing Systems*, 1996. URL: `http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf`.

[10] Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017. URL: `https://arxiv.org/pdf/1702.08608.pdf`.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012. URL: `https://arxiv.org/pdf/1104.3913.pdf`.

[12] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the $21^{st}$ ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015. URL: `https://arxiv.org/pdf/1412.3756.pdf`.

[13] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation*, 80:38, 2016. URL: `https://bit.ly/2Gesf9Y`.

[14] Sorelle A. Friedler, Chitradeep Dutta Roy, Carlos Scheidegger, and Dylan Slack. Assessing the Local Interpretability of Machine Learning Models. *arXiv preprint arXiv:1902.03501*, 2019. URL: `https://arxiv.org/pdf/1902.03501.pdf`.

[15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. ***The Elements of Statistical Learning***. Springer, New York, 2001. URL: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf.

[16] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *arXiv preprint arXiv:1806.00069*, 2018. URL: https://arxiv.org/pdf/1806.00069.pdf.

[17] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1), 2015. URL: https://arxiv.org/pdf/1309.6392.pdf.

[18] Krishna M. Gopinathan, Louis S. Biafore, William M. Ferguson, Michael A. Lazarus, Anu K. Pathria, and Allen Jost. Fraud Detection using Predictive Modeling, October 6 1998. US Patent 5,819,226. URL: https://patents.google.com/patent/US5819226A.

[19] Alicja Gosiewska, Aleksandra Gacek, Piotr Lubon, and Przemyslaw Biecek. SAFE ML: Surrogate Assisted Feature Extraction for Model Learning. *arXiv preprint arXiv:1902.11035*, 2019. URL: https://arxiv.org/pdf/1902.11035v1.pdf.

[20] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, 51(5):93, 2018. URL: https://arxiv.org/pdf/1802.01933.pdf.

[21] Patrick Hall. On the Art and Science of Machine Learning Explanations. In *KDD '19 XAI Workshop Proceedings*, 2019. URL: https://arxiv.org/pdf/1810.02909.pdf.

[22] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016. URL: http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

[23] Linwei Hu, Jie Chen, Vijayan N. Nair, and Agus Sudjianto. Locally Interpretable Models and Effects Based on Supervised Partitioning (LIME-SUP). *arXiv preprint arXiv:1806.00663*, 2018. URL: https://arxiv.org/ftp/arxiv/papers/1806/1806.00663.pdf.

[24] Faisal Kamiran and Toon Calders. Data Preprocessing Techniques for Classification Without Discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. URL: https://bit.ly/2lH95lQ.

[25] Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. Debugging Machine Learning Models via Model Assertions, 2019. URL: https://www-cs.stanford.edu/~matei/papers/2018/mlsys_model_assertions.pdf.

[26] Alon Keinan, Ben Sandbank, Claus C. Hilgetag, Isaac Meilijson, and Eytan Ruppin. Fair Attribution of Functional Contribution in Artificial and Biological Networks. *Neural Computation*, 16(9):1887–1915, 2004. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.436.6801&rep=rep1&type=pdf.

[27] M. Lichman. UCI Machine Learning Repository, 2013. URL: http://archive.ics.uci.edu/ml.

[28] Stan Lipovetsky and Michael Conklin. Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

[29] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490*, 2016. URL: https://arxiv.org/pdf/1606.03490.pdf.

[30] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate Intelligible Models with Pairwise Interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 623–631. ACM, 2013. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.352.7682&rep=rep1&type=pdf.

[31] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[32] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles. In Been Kim, Dmitry M. Malioutov, Kush R. Varshney, and Adrian Weller, editors, *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*, pages 15–21. ICML WHI 2017, 2017. URL: https://openreview.net/pdf?id=ByTKSo-m-.

[33] Christoph Molnar. ***Interpretable Machine Learning***. christophm.github.io, 2018. URL: `https://christophm.github.io/interpretable-ml-book/`.

[34] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Quantifying Interpretability of Arbitrary Machine Learning Models Through Functional Decomposition. *arXiv preprint arXiv:1904.03867*, 2019. URL: `https://arxiv.org/pdf/1904.03867.pdf`.

[35] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable Machine Learning: Definitions, Methods, and Applications. *arXiv preprint arXiv:1901.04592*, 2019. URL: `https://arxiv.org/pdf/1901.04592.pdf`.

[36] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015. URL: `https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Nguyen_Deep_Neural_Networks_2015_CVPR_paper.pdf`.

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. URL: `http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf`.

[38] Cynthia Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv preprint arXiv:1811.10154*, 2018. URL: `https://arxiv.org/pdf/1811.10154.pdf`.

[39] Lloyd S. Shapley, Alvin E. Roth, et al. *The Shapley value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, 1988. URL: `http://www.library.fa.ru/files/Roth2.pdf`.

[40] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017. URL: `https://arxiv.org/pdf/1610.05820.pdf`.

[41] Reza Shokri, Martin Strobel, and Yair Zick. Privacy Risks of Explaining Machine Learning Models. *arXiv preprint arXiv:1907.00164*, 2019. URL: `https://arxiv.org/pdf/1907.00164.pdf`.

[42] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. How Can We Fool LIME and SHAP? Adversarial Attacks on Post-hoc Explanation Methods. *arXiv preprint arXiv:1911.02508*, 2019. URL: `https://arxiv.org/pdf/1911.02508.pdf`.

[43] Erik Strumbelj and Igor Kononenko. An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, 11(Jan):1–18, 2010. URL: `http://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf`.

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. *arXiv preprint arXiv:1312.6199*, 2013. URL: `https://arxiv.org/pdf/1312.6199.pdf`.

[45] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618, 2016. URL: `https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf`.

[46] Berk Ustun and Cynthia Rudin. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning*, 102(3):349–391, 2016. URL: `https://users.cs.duke.edu/~cynthia/docs/UstunTrRuAAAI13.pdf`.

[47] Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N. Nair. Explainable Neural Networks Based on Additive Index Models. *arXiv preprint arXiv:1806.01933*, 2018. URL: `https://arxiv.org/pdf/1806.01933.pdf`.

[48] Adrian Weller. Challenges for Transparency. *arXiv preprint arXiv:1708.01870*, 2017. URL: `https://arxiv.org/pdf/1708.01870.pdf`.

[49] Mike Williams et al. *Interpretability*. Fast Forward Labs, 2017. URL: `https://www.cloudera.com/products/fast-forward-labs-research.html`.

[50] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian Rule Lists. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. URL: `https://arxiv.org/pdf/1602.08610.pdf`.

[51] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018. URL: `https://arxiv.org/pdf/1801.07593.pdf`.