U-Trustworthy Models. Reliability, Competence, and Confidence in Decision-Making

Ritwik Vashistha*, **Arya Farahi**

The Univertsity of Texas at Austin
ritwik.v@utexas.edu, arya.farahi@austin.utexas.edu

Abstract

With growing concerns regarding bias and discrimination in predictive models, the AI community has increasingly focused on assessing AI system trustworthiness. Conventionally, trustworthy AI literature relies on the probabilistic framework and calibration as prerequisites for trustworthiness. In this work, we depart from this viewpoint by proposing a novel trust framework inspired by the philosophy literature on trust. We present a precise mathematical definition of trustworthiness, termed $\mathcal{U}$-trustworthiness, specifically tailored for a subset of tasks aimed at maximizing a utility function. We argue that a model's $\mathcal{U}$-trustworthiness is contingent upon its ability to maximize Bayes utility within this task subset. Our first set of results challenges the probabilistic framework by demonstrating its potential to favor less trustworthy models and introduce the risk of misleading trustworthiness assessments. Within the context of $\mathcal{U}$-trustworthiness, we prove that properly-ranked models are inherently $\mathcal{U}$-trustworthy. Furthermore, we advocate for the adoption of the AUC metric as the preferred measure of trustworthiness. By offering both theoretical guarantees and experimental validation, AUC enables robust evaluation of trustworthiness, thereby enhancing model selection and hyperparameter tuning to yield more trustworthy outcomes.

Introduction

In recent years, the AI community has expressed growing concerns about bias and discrimination embedded within predictive models. These concerns have prompted a shift in focus from performance to fairness and trustworthiness as a pilar of model evaluation (Mehrabi et al. 2021; Eshete 2021). While fairness aims to mitigate disparate impacts on different population groups, trustworthiness encompasses broader notions of model reliability, robustness, competence, generalization, explainability, transparency, reproducibility, privacy, security, and accountability (Serban et al. 2021; von Eschenbach 2021; Li et al. 2023; Broderick et al. 2023). In this paper, by borrowing from the philosophy literature, we develop a theory of trustworthiness from a lens of reliability and competence and investigate its implications for classification models in the context of decision-making.

*Corresponding Author

Our theoretical framework is rooted in competence-based trust theories, which draws from the philosophical literature on the relation between trust, reliability, and competence (Baier 1986; Jones 1996; Ryan 2020; Alvarado 2022). Epistemologically, trust is commonly understood as the act of placing confidence in a source of information or in an agent to perform a task based on its perceived competence to be accurate and reliable. In the context of predictive models and decision-making, this notion translates into our reliance on a model that consistently demonstrates competence in achieving its goal in a decision-making task. In competence-based trust theories, trust is described through a three-part relationship "A trusts B to do X," (Horsburgh 1960; Ryan 2020; von Eschenbach 2021; Alvarado 2022; Afroogh 2023) where, in our case, A represents the end-user, B represents the predictive model, and X specifies the delegated task. What counts as good reasons for trusting is a matter of considerable debate, but at the very least, that B is capable or competent to do X is necessary for A to trust B. Since our interest lies in the trustworthiness of B rather than A's trust, we reframe the investigation as a two-part inquiry, "B is trustworthy to do X." Formalizing this inquiry provides a foundation for trustworthiness evaluation and is the primary aim of this work.

To achieve this goal, this work first constructs a mathematical framework for trustworthy evaluation; to evaluate the claim of whether "B is trustworthy to do X," where B is a predictive model, and X is a subset of decision-making tasks. Establishing trustworthiness requires guaranteeing *reliance*, *competence*, and *confidence*. Reliance reflects the user may rely on the model to achieve its promised goal(s), while competence provides theoretical guarantees that there exists no other model that can achieve a superior result in task X. Confidence is a statistical claim that, given the existing empirical evidence, B is competent.

Traditionally, probabilistic assessment or risk calibration has been considered a crucial aspect of model trustworthiness evaluation, as it ensures that predicted risks align with observed frequencies of outcomes and provides some information regarding the model uncertainty (Crowson, Atkinson, and Therneau 2016; Pleiss et al. 2017; Hébert-Johnson et al. 2018; La Cava, Lett, and Wan 2022; Afroogh 2023). However, its role in determining the trustworthiness of a model, as described above, remains an open question. Drawing upon competence-based trust theories, we challenge this

prevailing belief that risk calibration alone is sufficient or necessary to establish trustworthiness (Barocas, Hardt, and Narayanan 2017; La Cava, Lett, and Wan 2022; Afroogh 2023). To do so, we analyze the limitations of calibration and closely related metrics as a standalone condition for trustworthiness. This work reveals the shortcomings of the probabilistic framework (Afroogh 2023) in capturing the holistic notion of model trustworthiness. Additionally, our results uncover a limitation in the traditional metrics, such as accuracy, that are used for model comparison and hyperparameter tuning. We find that relying solely on accuracy or related measures can lead to misleading conclusions regarding the trustworthiness of a model.

Related Work

Trustworthy AI has garnered widespread attention from practitioners, policy-makers, and AI developers alike, solidifying its position as a frontrunner in addressing pressing societal concerns surrounding AI bias and discrimination (Li et al. 2023). While prediction accuracy is undoubtedly an essential aspect of trustworthiness, factors such as robustness, transparency, reproducibility, replicability, stability, interpretability, and consistency are integral to trustworthiness (Broderick et al. 2023). These aspects elucidate how an AI system performs under varying conditions, ensures unbiased predictions, and maintains consistent outputs (Varshney 2019; Serban et al. 2021; von Eschenbach 2021).

Calibration, which is closely related to the probability theory of trust (Afroogh 2023), is claimed to be as one of the key requirements for trustworthy AI and used by many practitioners (e.g., Safavi, Koutra, and Meij 2020; Tomani and Buettner 2021). By calibrating a classifier, it can be ensured that the predicted probabilities of the classifier more accurately reflect the true likelihood of each outcome, thereby increasing trust in the model's predictions. The literature in trustworthiness has heavily skewed on evaluating calibration of a predictive model (Murphy and Winkler 1977; Naeini, Cooper, and Hauskrecht 2015; Kumar, Sarawagi, and Jain 2018; Widmann, Lindsten, and Zachariah 2019) and developing post-processing calibration methods for risk management (e.g., Murphy and Winkler 1977; Platt et al. 1999; Zadrozny and Elkan 2002; Guo et al. 2017).

However, AI trustworthiness literature extends beyond the probabilistic paradigm. In this context, interpretability is proposed as a gateway to establishing trust (Ribeiro, Singh, and Guestrin 2016). Moreover, trustworthiness is further explored through the creation of diverse trust scores (Jiang et al. 2018; Wong, Wang, and Hryniowski 2020), as well as learning-based approaches like enhancing the loss function (Luo et al. 2021). Our work reexamines the foundation of AI trustworthiness and proposes a novel, competence-based trustworthiness paradigm.

Problem Setup

In our study, we aim to investigate the concept of trustworthiness in the context of "B (a predictive model) is trustworthy to do X (a decision-making task)." This work is only concerned with a subset of tasks whose goal is to maximize a class of utility functions, hence $\mathcal{U}$-Trustworthiness. We define a $\mathcal{U}$-trustworthy model as one that possesses a decision boundary (reliability) capable of achieving maximum utility among all possible models (competence) with empirical guarantees (confidence). Next, we formalize this definition.

Let $\mathbf{x} \in \mathcal{X}$, $Y \in \{0,1\}$, $\widehat{Y} \in \{0,1\}$ denote the input features, binary outcome, and binary decision respectively; and $\mathcal{D}$ be the distribution generating $(\mathbf{x}, Y) \in \mathcal{X} \times \{0,1\}$ pairs. $f_\theta : \mathcal{X} \to [0,1]$ is a predictive model that maps inputs from $\mathcal{X}$ to a score used to assign a binary decision. θ specifies the parameters of the model that will be learned through a learning process. This work is not concerned with parameter estimation, so we assume that the model and its parameters are fixed, thus suppressing the notation θ from now on. As we will see later, the output of f does not need to be interpreted probabilistically. However, after calibration, it can take probabilistic meaning. When necessary, we assume a finite test sample, denoted with $\mathcal{S}$. Finally, let $U(\mathbf{x}, Y, \widehat{Y}) : \mathcal{X} \times \{0,1\} \times \{0,1\} \to \mathbb{R}$ be a utility function that quantifies the desirability or usefulness of a decision outcome. Higher utility values indicate more desirable outcomes. See the Supplementary Materials for examples and discussion on the distinction between Y and $\widehat{Y}$.

The ultimate goal of decision-making is to assign $\widehat{Y}$ based on the output of $f(\mathbf{x})$ and observed covariates $\mathbf{x}$ such that it maximizes the expected value of the utility function. The optimal decision rule is determined by solving

$$g^*(\mathbf{x}; U, f) = \arg\max_{\widehat{g} \in \mathcal{G}} \mathbb{E}_{Y,\mathbf{x}\sim\mathcal{D}} \left[U(\mathbf{x}, Y, \widehat{Y}) \mid f \right]. \quad (1)$$

where $\widehat{Y}$ might be dependent on some decision rule $\widehat{g}$. Suppose the solution to 1 has the form of

$$\widehat{Y} = \begin{cases} 1 & \text{if } f(\mathbf{x}) \geq \widehat{g}(\mathbf{x}; U) \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

This implies that there is a decision rule, denoted with $\widehat{g}(\mathbf{x}; U)$, that separates the prediction scores into regions associated with the binary decision assignment. We denote the maximum expected utility with $U^{(m)}$, where $U_f^{(m)} = \max_{\widehat{g}\in\mathcal{G}} \mathbb{E}_{Y,\mathbf{x}\sim\mathcal{D}} \left[U(\mathbf{x}, Y, \widehat{Y}) \mid f \right]$. As we shall see, for this class of problems, we can provide generalizable trustworthiness guarantees. We acknowledge that there is no fundamental reason that the solution to Equation (1) must take the form of Equation (2); on the contrary, we can imagine examples in which the solution to Equation (2) cannot be expressed in the form of Equation (1).

U-Trustworthy

Next, we seek to formalize the proposition "B is trustworthy to do X" for the class of tasks defined in Equation (1) with the solution of the form in Equation (2).

Definition 1 ($\mathcal{U}$-Trustworthy)**.** *A model, denoted with $\widetilde{f}(.)$, is $\mathcal{U}$-trustworthy if $U_f^{(m)} \leq U_{\widetilde{f}}^{(m)}$ $\forall\, U \in \mathcal{U}$ and $\forall\, f \in \mathcal{F}$.*

In simpler terms, for a class of utility functions $\mathcal{U}$, a $\mathcal{U}$-trustworthy model is one that can be relied on to achieve the

highest possible expected utility. For all $U \in \mathcal{U}$, there exists a decision boundary of a $\mathcal{U}$-trustworthy model effectively separates the input space into regions that lead to the most desirable outcome.

Reliance. The reliance condition is met when a solution to Equation (1) exists. This implies that model $\widetilde{f}(.)$ can be relied on to accomplish the intended goal, setting the foundation for competency, which represents the maximum achievable utility across all possible models.

Competency. The competency condition is met when $U_f^{(m)} \leq U_{\widetilde{f}}^{(m)}$ holds for all $U \in \mathcal{U}$ and $f \in \mathcal{F}$. This ensures that no other model can attain a higher expected maximum utility than what model $\widetilde{f}(.)$ achieves.

Confidence. By subjecting the claim $U_f^{(m)} \leq U_{\widetilde{f}}^{(m)}$ to hypothesis testing using the test data in $\mathcal{S}$, users can establish statistical confidence in the trustworthiness of the model.

Bayes Classifier

Definition 2. *Let $f^\star(\mathbf{x})$ denotes the Bayes classifier, implying $f^\star(\mathbf{x}) = P(Y = 1 \mid \mathbf{x})$; and $Y^\star$ and $g^\star(\mathbf{x}; U)$ be the optimal Bayes decision and decision rule associated with the utility function U.*

Proposition 1. *Let $f^\star(\mathbf{x})$ be the Bayes classifier, then $U_f^{(m)} \leq U_{f^\star}^{(m)} \ \forall\, U \in \mathcal{U}$ and $\forall\, f \in \mathcal{F}$.*

This statement arises from the definition of the Bayes classifier. This proposition implies that the utility of the Bayes classifier equals that of a $\mathcal{U}$-trustworthy classifier. Consequently, we arrive at the following theorem:

Theorem 1. *Model $\widetilde{f}(.)$ is $\mathcal{U}$-trustworthy, if $U_{f^\star}^{(m)} = U_{\widetilde{f}}^{(m)} \ \forall\, U \in \mathcal{U}$ and $\forall\, f \in \mathcal{F}$.*

Although this theorem may seem a straightforward consequence of our definitions, it carries two significant implications. It implies that the maximum utility of a $\mathcal{U}$-trustworthy model aligns with that of the Bayes classifier. Additionally, it streamlines the process of hypothesis testing when the null hypothesis is $U_{f^\star}^{(m)} = U_{\widetilde{f}}^{(m)}$. See Supplementary Materials in the arxiv verison for the proof of all the theorems.

Limitations of Calibration Requirements in U-Trustworthiness Assessment

Calibration, which is closely related to the probability theory of trust (Afroogh 2023), is claimed to be as one of the key requirements for trustworthy AI and used by many practitioners (e.g., Safavi, Koutra, and Meij 2020; Tomani and Buettner 2021). In this section, we revisit this claim.

Definition 3. *Model $f(\mathbf{x})$ is calibrated if $P(Y = 1 \mid f(\mathbf{x}) = \alpha) = \alpha$ for all $\alpha \in [0, 1]$.*

We generated 400 data sets each with a sample size of 15,000 and evaluated the performance of three classifiers: (1) the Bayes classifier (blue), (2) a properly-ranked classifier (green), and (3) a calibrated adversarial classifier (red). See Supplementary Materials for the full description of the simulation.

Results. Properties of these classifiers are illustrated in Figure 1. The top-left present the calibration plot for these classifiers. The Bayes, by definition, and calibrated, by construction, classifiers are calibrated; while the properly-ranked classifier is miscalibrated. Initially, this might lead to the conclusion that the calibrated and Bayes classifiers are more trustworthy for decision-making tasks, which could be reinforced by considering popular metrics like mean calibration error, Brier score (Brier 1950), and accuracy.

Upon closer examination of the utility curve (top row), we observe that both the Bayes and properly-ranked classifiers have similar maximum utilities, while the calibrated classifier consistently exhibits lower maximum utility. This suggests that the properly-ranked classifier may be a $\mathcal{U}$-trustworthy model, while the calibrated classifier falls short. This example motivates us to reevaluate the significance of calibration in trustworthy evaluation and to study the characteristics of $\mathcal{U}$-trustworthy models more closely. We also examine NetTrust score (Wong, Wang, and Hryniowski 2020). NetTrust score ranks the Properly ranked classifier higher than calibrated and Bayes classifier, which can lead to a misleading conclusion that the Bayes classifier is suboptimal.

This example highlights two key findings:

- The notion that a calibrated classifier is inherently trustworthy is challenged, as we demonstrate that a calibrated classifier can actually be incompetent, hence untrustworthy. Conversely, a mis-calibrated classifier can exhibit trustworthy behavior. This counterintuitive result emphasizes the need to reconsider the traditional assumption that calibration is a precursor to trustworthiness.
- The limitations of existing performance measures become evident in the context of trustworthy evaluation. We reveal that these measures fall short in accurately assessing the which classifier is trustworthy. This highlights the importance of reassessing evaluation metrics in capturing the nuanced aspects of trustworthiness.

Characteristics of U-trustworthy Classifiers

Definition 4 (Properly-Ranked Classifier)**.** *Let a classifier $f_{\mathrm{PR}}(.)$ be a properly-ranked classifier if*

$$\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \begin{cases} f^\star(\mathbf{x}_1) > f^\star(\mathbf{x}_2) \Rightarrow f_{\mathrm{PR}}(\mathbf{x}_1) > f_{\mathrm{PR}}(\mathbf{x}_2) \\ f^\star(\mathbf{x}_1) = f^\star(\mathbf{x}_2) \Rightarrow f_{\mathrm{PR}}(\mathbf{x}_1) = f_{\mathrm{PR}}(\mathbf{x}_2) \end{cases}$$

Theorem 2 ($\mathcal{U}$-Competency Theorem)**.** *Suppose for the utility class $\mathcal{U}$ with a solution of the form Equation* (2)*. Any properly-ranked classifier is a $\mathcal{U}$-trustworthy classifier with respect to sample $\mathcal{S}$ with $|\mathcal{S}| < \infty$.*

This theorem has two important implications. First, if the solution to the general problem of Equation (1) that is associated with the utility class $\mathcal{U}$ can be expressed with the form in Equation (2), where there is a decision boundary above which $\widehat{Y} = 1$ and below which $\widehat{Y} = 0$, then any properly-ranked classifier is $\mathcal{U}$-trustworthy. This provides competency criteria. A properly-ranked classifier, even with an incorrect risk estimation, is competent for the decision-making of the class described here. Theorem 2 indicates that $\mathcal{U}$-trustworthiness can be achieved without calibration

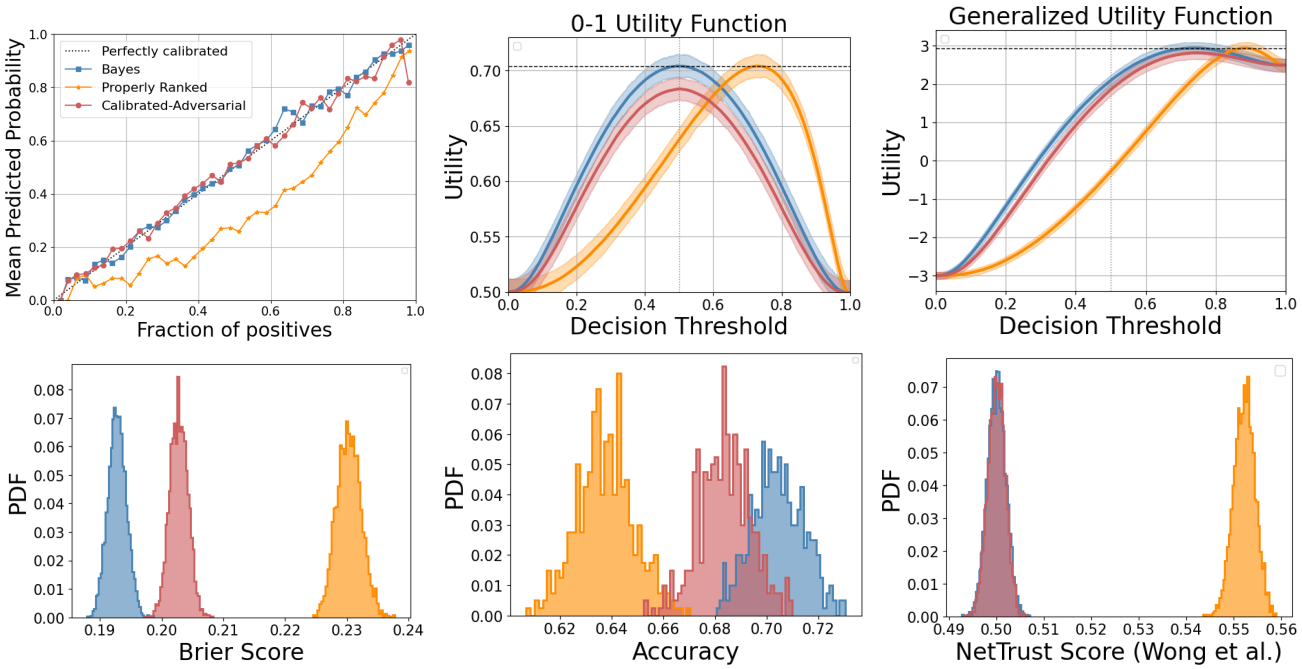

Figure 1: Performance of three models under different criteria. Top row: Calibration plot (left). Utility comparison using 0-1 loss (middle) as a function of decision threshold. Utility comparison using a general utility function as a function of decision threshold (right). The confidence bands represent the 16th and 84th percentiles of 400 data realizations. Bottom row: Distribution of Brier score (left), Accuracy(middle), NetTrustScore (right) under 400 data realizations.

or proper risk estimation. Hence, for this class of problems, calibration alone is neither a necessary nor a sufficient condition for a classifier to be $\mathcal{U}$-trustworthy.

While properly-ranked classifiers possess desirable properties, they do not encompass the most general class of $\mathcal{U}$-trustworthy classifiers. The following proposition serves as an example in this regard.

Proposition 2. *Suppose a group indicator G splits $\mathcal{S}$ into two non-overlapping subsets. Then, if, for each group, the properly-ranked classifier condition is satisfied, then the classifier is $\mathcal{U}$-trustworthy.*

According to the proposition, if the sample $\mathcal{S}$ can be split into two non-overlapping subsets based on a group indicator G, and within each group separately, we get the correct ranking, then the classifier is $\mathcal{U}$-trustworthy. However, it is important to note that the properly-ranked condition may not be satisfied when comparing data points across the two groups. Therefore, while the properly-ranked condition is sufficient, it is not necessary for a classifier to be $\mathcal{U}$-trustworthy. In general, however, we do not consider the classifiers in Proposition 2 as $\mathcal{U}$-trustworthy unless the grouping is known or is learned. If such grouping exists, but it is unknown, additional search and considerations on the user's side are required to guarantee $\mathcal{U}$-trustworthiness.

Cost-sensitive Trustworthy Classifiers

Let $\mathcal{U}$ be a class cost-sensitive utility functions that are given by a weighted combination of the four fundamental population quantities closely related to elements of the "confusion matrix" - true positives, false positives (a.k.a. type-I error), false negatives (a.k.a. type-II error) and true negatives as defined below

$$\mathbf{TP} = \mathbb{I}(Y = 1, \widehat{Y} = 1 \mid \mathbf{x}), \mathbf{FP} = \mathbb{I}(Y = 0, \widehat{Y} = 1 \mid \mathbf{x}),$$
$$\mathbf{FN} = \mathbb{I}(Y = 1, \widehat{Y} = 0 \mid \mathbf{x}), \mathbf{TN} = \mathbb{I}(Y = 0, \widehat{Y} = 0 \mid \mathbf{x}),$$

where $\mathbb{I}$ is the indicator function.

Definition 5. *The cost-sensitive utility family is defined as*

$$\mathcal{U}(\{a_{ij}\}) = a_{11}\mathbf{TP} - a_{01}\mathbf{FP} - a_{10}\mathbf{FN} + a_{00}\mathbf{TN}. \quad (3)$$

where $a_{ij} \geq 0$ for all $i, j \in \{0, 1\}$.

Example 1. *The 0-1 loss function belongs to the family of cost-sensitive utility functions.*

This is straightforward to show by setting $a_{11} = a_{00} = 1$ and $a_{10} = a_{01} = 0$. This is equivalent to 0-1 loss function.

Lemma 1. *The decision rule for the Bayes classifier is*

$$g^\star = \frac{a_{01} + a_{00}}{a_{11} + a_{00} + a_{10} + a_{01}}$$

This lemma is the reliability condition that suggests for cost-sensitive utility function that there exists a solution of form Equation (2). We also note that this class of utility functions is characterized by the coefficients a_{ij}, which in principle can be functions of $\mathbf{x}$, where $\mathbf{x}$ represents additional contextual information. For example, consider the utility associated with the survival of young teens, which might be higher compared to older individuals in certain scenarios. Similarly, the costs associated with the passing away of individuals could also vary based on contextual factors.

Theorem 3. *Let $\mathcal{U}$ be the cost-sensitive utility class. Properly-ranked classifiers are $\mathcal{U}$-trustworthy.*

Equity-aware Trustworthy Classifiers

There has been a surge of interest in formulating utility functions that simultaneously account for both efficiency and equity. In this section, we utilize a class of equity-aware functions proposed by Kleinberg et al. (2018) and demonstrate that properly-ranked classifiers are $\mathcal{U}$-trustworthy. First, we define the compatibility criteria and a class of equity-aware utility functions.

Definition 6 (Compatibility)**.** *The utility function ϕ is compatible with the Bayes classifier if the following natural monotonicity condition holds. If S and S' are two sets of $\mathcal{X}$ of the same size, sorted in descending order of $f^\star(\mathbf{x})$, and $f^\star(\mathbf{x})$ of the i^{th} item in S is at least as large as $f^\star(\mathbf{x})$ of the i^{th} item in S' for all i, then $\phi(S_i) \geq \phi(S'_i)$.*

Definition 7 (Equity-aware Utility Class)**.** *Suppose there is a binary variable G that splits the data into two non-overlapping subsets. The equity-aware utility family is defined as $\mathcal{U}(\phi, \gamma) = \phi(S) + \gamma(S)$, where $S \subseteq \mathcal{X}$, $\phi(S) \in \Phi$ is compatible with Bayes probability, and $\gamma(S) \in \Gamma$ is monotonically increasing in the number of items in S who have $G = 1$.*

The first term characterizes the benefit while the second term characterizes the fairness. An equity-aware decision-maker seeks to maximize $U(S) = \phi(S) + \gamma(S)$. This utility function class and the following Lemma follow that of Kleinberg et al. (2018).

Lemma 2 (Theorem 1, Kleinberg et al. (2018))**.** *For some choice of K_0, from group $G = 0$, and K_1, from group $G = 0$, with $K_0 + K_1 = K$, the solution that maximizes utility in the $G = 0$ and in the $G = 1$ group are the ones with the highest $f^\star(\mathbf{x})$.*

This lemma provides the reliability condition.

Theorem 4. *Let $\mathcal{U}$ be an equity-aware utility class. Properly-ranked classifiers are $\mathcal{U}$-trustworthy.*

AUC and Its Relation to U-Trustworthiness

An ROC curve provides a graphical representation of classifier performance by comparing the true positive rate (TPR) to the false positive rate (FPR) across various decision thresholds. AUC – the area under the ROC curve – is a numerical measure of the classifier's performance. It quantifies the classifier's ability to rank a randomly selected positive example higher than a randomly chosen negative example, given that the positive class is ranked higher than the negative class (Hanley and McNeil 1982).

Pairwise estimator of AUC. In probabilistic terms, the AUC represents the probability of correctly ranking the two examples (Hanley and McNeil 1982). The pairwise estimator for the AUC is also known as the Wilcoxon-Mann-Whitney statistic (Agarwal et al. 2005). It is calculated by comparing all possible pairs of observations, where one observation belongs to class 1 and zero to class 0,

$$\text{AUC}(f) = \mathbb{E}\left[\mathcal{H}(f(\mathbf{x}^+) - f(\mathbf{x}^-))\right]. \tag{4}$$

$(\mathbf{x}^+, \mathbf{x}^-)$ is a pair on i.i.d. draws from class 1 and zero. The expected value is computed over $\{(\mathbf{x}^+, \mathbf{x}^-) \in \mathcal{D}^+ \times \mathcal{D}^-\}$ where $\mathcal{D}^{+/-}$ is the distribution over data with class 1/0. $\mathcal{H}(.)$ is the Heaviside step function which returns 1 if the argument is positive, 1/2 if the argument is zero, and 0 otherwise. An estimator of this expected value is

$$\widehat{\text{AUC}}(f) = \frac{1}{|S^+||S^-|} \sum_{i=1}^{|S^+|} \sum_{j=1}^{|S^-|} \mathcal{H}(f(\mathbf{x}_i) - f(\mathbf{x}_j)) \tag{5}$$

where $S^{+/-} = \{\mathbf{x} \in S : y = 1/0\}$. The properties of this estimator are studied extensively in the literature (Airola et al. 2011; Agarwal et al. 2005; Cortes and Mohri 2004).

Theorem 5 ($\mathcal{U}$-Competency Measure)**.** *Let $\mathcal{U}$ be a utility class with a decision boundary of Equation* (2)*. If and only if f_{PR} is a properly-ranked classifiers then* $\text{AUC}(f^\star) = \text{AUC}(f_{\text{PR}})$.

Theorem 5 provides a theoretical justification for why AUC may be used as a measure of competency, implying that if the AUC of a classifier is equal to the Bayes classifier, then the classifier is competent to achieve the maximume possible expected utility. Unlike the measures constructed by the confusion matrix entries, the AUC is the precision of pairwise rankings (Hanley and McNeil 1982). A properly-ranked classifier produces the same ranking as the Bayes classifier. Hence, both classifiers are expected to exhibit similar AUC performance, as implied from the above theorem, while their error rate, accuracy, and calibration can differ significantly (Cortes and Mohri 2004).

The literature regarding the suitability of AUC as an evaluation metric is subject to varying opinions. For instance, Huang and Ling (2005) have supported the use of AUC by providing evidence that it has greater discriminative ability than accuracy, while Lobo, Jiménez-Valverde, and Real (2008) have cautioned against using AUC, primarily because it does not assess goodness-of-fit or might result in poor calibration. However, our results contend that in applications where utility maximization is a priority, AUC or potentially other ranking quality metrics may be more favorable. These applications include tasks where the relative ordering of actions is crucial for decision-making, such as recommendation systems (Schröder, Thiele, and Lehner 2011), information retrieval (Nguyen et al. 2016), or task delegation (Farzaneh et al. 2023). In the next section, we provide empirical evidence that utilizing AUC and accuracy can lead to different results, and when utility maximization is a priority, one should rely on AUC.

Applications and Empirical Evidence

In this section, we illustrate various practical benefits of using AUC as a performance measure in the context of $\mathcal{U}$-trustworthiness. We provide empirical evidence that AUC outperforms popular performance metrics such as accuracy in model comparison and hyperparameter tuning. We recall the definition of $\mathcal{U}$-trustworthiness, a model that achieves the highest maximum expected utility. So a model with the highest maximum expected utility is more trustworthy. When not

Perf.	RF	LR	kNN
AUC	$\mathbf{0.797 \pm 0.003}$	0.788 ± 0.004	0.785 ± 0.003
Acc.	0.731 ± 0.005	$\mathbf{0.736 \pm 0.005}$	0.731 ± 0.004
Brier	0.182 ± 0.001	$\mathbf{0.177 \pm 0.002}$	0.180 ± 0.002
NT	0.580 ± 0.001	0.591 ± 0.002	$\mathbf{0.598 \pm 0.002}$
$U^{(m)}$	$\mathbf{0.744 \pm 0.004}$	0.739 ± 0.004	0.735 ± 0.004

Table 1: Model selection results. NT = NetTrust.

mentioned, we use a 0-1 utility function $U(f) = \mathbf{TP} + \mathbf{TN}$. We note that the decision threshold, $\widehat{g}$, might be different across the models. We first solve for $\widehat{g}$ by varying the decision threshold, and then report the maximum utility. We additionally use a more complex utility function to illustrate the results hold as long as it belongs to the class of utility function discussed above. Our prediction task is to predict which household is a homeowner using the 2019 American Housing Survey data. See Supplementary Materials for the description of data sets and additional experiments.

Model Selection

During model selection, conflicting performance measures can lead to uncertainty about which metric should be given priority. To address this issue, we argued that AUC should be the preferred choice if the ultimate goal is $\mathcal{U}$-trustworthiness. This study compares several popular performance measures across different models using data from the homeownership dataset. Additional experiments and results can be found in the Supplementary Materials. For model comparison, we consider four metrics: AUC, Brier score, accuracy, and NetTrust score, and three models: Logistic Regression, Random Forest, and k-NN. The evaluation results are presented in Table 1. Notably, relying solely on the accuracy or Brier score would select Logistic Regression, and NetTrust would select kNN as the preferred model. This conclusion could be further reinforced by examining the calibration properties of Logistic Regression compared to Random Forest. (Figure 2, left panel). However, when considering AUC, Random Forest emerges as the preferred choice. Given that the ultimate goal is to maximize expected utility ($\mathcal{U}$-trustworthy setting), Random Forest should be the preferred choice, as it aligns with the AUC criteria for $\mathcal{U}$-trustworthiness. This observation highlights the importance of AUC as a reliable measure for model selection when considering the ultimate objective of maximizing expected utility.

The second experiment evaluates the above claim but for a class of utility functions. We now explore the maximum expected utility for a class of utility functions specified by parameter c, defined as

$$\mathcal{U}[c] = \mathbf{TP} + \mathbf{TN} - c \times \mathbf{FP} - 0.5c \times \mathbf{FN}. \quad (6)$$

In the right panel of Figure 2, we present the average maximum utility on the test sample for 20 random test/train realizations. The model selected based on AUC consistently outperforms the model with lower AUC but higher Brier or accuracy score in terms of expected maximum utility. This demonstrates the significance of AUC as a reliable measure for model comparison in the context of $\mathcal{U}$-trustworthiness.

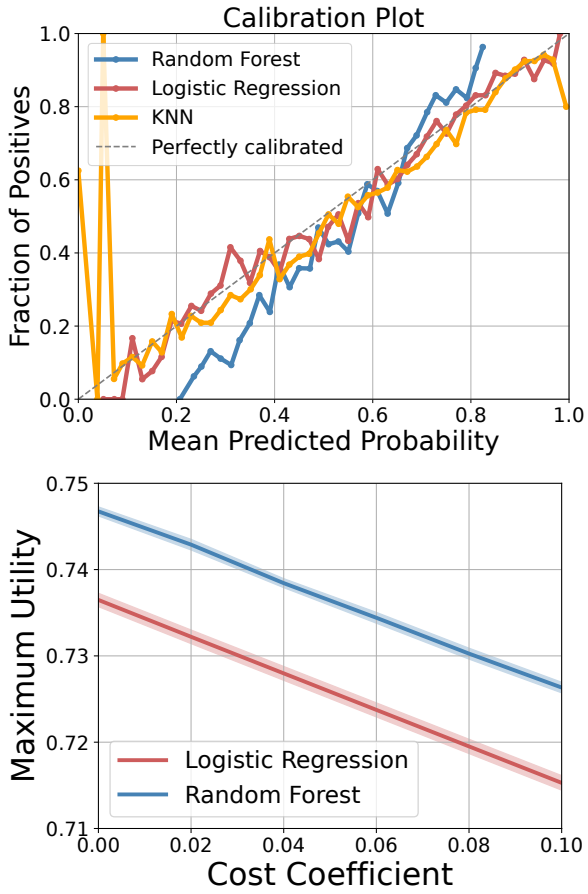

Figure 2: Top: Average calibration curve. Bottom: The performance of Logistic Regression, selected based on accuracy/Brier score, and Random Forest, chosen based on AUC, for a class of utility functions specified by parameter c, cost coefficient. The average maximum utility on the test sample for 20 random test/train realizations and the shaded region is 68% error on the mean.

Hyper-parameter Tuning

Tuning of hyper-parameters is an integral part of the model-building procedure. Next, we focus on a k-nearest neighbor classifier with k as the hyperparameter and investigate the impact of tuning based on different performance measures. Specifically, we compare the use of AUC and accuracy as the performance metrics during the tuning process. To conduct our analysis, we employ a homeownership dataset and perform 20-fold cross-validation to fine-tune k. We evaluated the model's performance using both AUC and accuracy metrics during the tuning process. The left panel of Figure 3 shows the results of tuning using AUC (blue curve) and accuracy (red curve) as performance measures. Next, we maximize the utility by varying the decision threshold. We use a non-trivial utility function that changes with the age of the householder. Let the reward of $\mathbf{TP} = \mathbf{TN}$ be 1, and the costs of the $\mathbf{FP}$ and $\mathbf{FN}$ be $C(\text{FP}) = (1 - \text{Age}/100) \times 3$ and $C(\text{FN}) = (1 - \text{Age}/100) \times 0.5$,

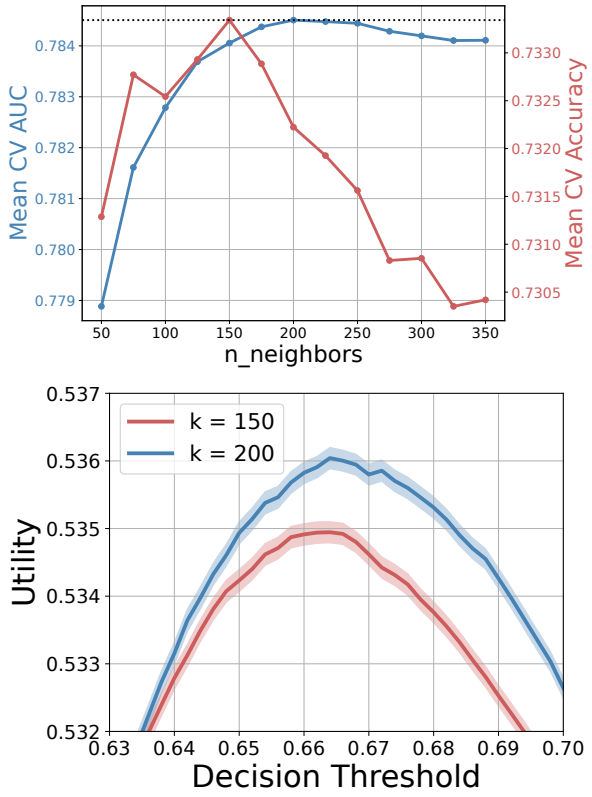

Figure 3: Hyper-parameter Tuning of k-NN. Top: Average cross-validation performance vs. `n_neighbors`. The optimal k for (accuracy) AUC is (150) 200, indicated by the dotted horizontal line. Bottom: Utility as a function of decision threshold for $k = 150$ (red) and $k = 200$ (blue).

and the utility function for the test sample is $U = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left[\mathbb{I}(y_i - \hat{f}_i) \times R(x_i, y_i) - |y_i - \hat{f}_i| \times C(x_i, y_i) \right]$.

Finally, we train two models, one with $k = 150$ and one with $k = 200$, and compute and report the utility of the test sample by varying the decision threshold (The right panel of Figure 3). We find that selecting the hyperparameter based on AUC ($k = 200$) leads to a model with higher utility than one with a hyperparameter based on accuracy ($k = 150$). This observation is evident in both graphs, emphasizing the superiority of AUC as a performance measure when aiming to maximize utility. To account for randomness, we repeated the experiment with 200 random data realizations, and the line shows the mean, and the shaded region is the standard error on the mean. This result supports our claim that AUC is superior to accuracy, where utility maximization is the eventual goal, even for nontrivial utility functions.

Discussion

AUC has faced criticism for its limited consideration of predicted probability values and the model's goodness-of-fit (Lobo, Jiménez-Valverde, and Real 2008). Other findings indicate that AUC surpasses metrics built on the elements of the confusion matrix, such as accuracy or the Matthew correlation coefficient, in terms of discriminative power (Huang and Ling 2005; Halimu, Kasem, and Newaz 2019). Our results provide a clear guideline for selecting the appropriate model performance measure, specifically focusing on AUC, within the class of problems concerned with utility maximization. The findings suggest that for such problems, calibration and goodness-of-fit, although closely related measures are neither necessary nor sufficient conditions for $\mathcal{U}$-trustworthiness. It is important to note that $\mathcal{U}$-trustworthiness definition does not apply to problems requiring an unbiased risk estimate or involving objectives other than maximizing utility. While our proposed trust framework can be extended to a broader class of problems, we acknowledge its limitations in generalizing beyond utility maximization concerns.

Limitations This work comes with certain limitations that should be acknowledged and provides a path for future studies. First and foremost, we recall that our framework is grounded on the competence-based trust theory; but there are other theoretical trust/trustworthiness frameworks proposed and used by the AI community (Toreini et al. 2020; Serban et al. 2021; Li et al. 2023). Our framework establishes a foundation on hypothesis testing the claim "A trusts B to do X." To simplify the problem further, we remove trustor A from the traditional three-part trust relation. This is a first step towards establishing full, three-part trustworthiness. Gaining users' trust requires additional domain-specific efforts in utilizing the predictive model's outcomes (Chatzimparmpas et al. 2020; Boyd et al. 2023; Mittermaier, Raza, and Kvedar 2023). Second, the theorems presented in this work are for binary classifiers and do not generalize to the multi-class case. The $\mathcal{U}$-trustworthy evaluation framework proposed in this work is specifically designed for a subset of tasks that seek to maximize the expected utility with solutions in the form of Equation (2). Consequently, this definition of trustworthiness should not be generalized to other tasks, such as risk mitigation or population inference, potentially rendering a model trustworthy for one set of tasks but not others. Lastly, the use of AUC as an evaluation metric in this work presents its own challenges, as the AUC of a $\mathcal{U}$-trustworthy model remains unknown, making hypothesis testing difficult. Further research is needed to understand better the relationship between the $\mathcal{U}$-trustworthy AUC and performing hypothesis testing, that is related to the last requirement of $\mathcal{U}$-trustworthy which is confidence.

Distinction with Fairness We differentiate between fairness and trustworthiness, as defined here, specifically in the context of the equity-aware utility class. A predictive model that assigns rankings or a risk score to individuals cannot be characterized as fair or unfair. Fairness becomes relevant when this ranking is employed to make decisions about individuals (who will be released on bail, who will be hired). In the absence of decision-making and a decision rule, fairness does not come into play. Conversely, $\mathcal{U}$-trustworthy refers to an inherent characteristic of a predictive model with respect to a utility function class. It asks whether the maximum achievable utility can be realized for every utility function belonging to this class or not, but which utility function

should be used is a concern of fairness.

Conclusion

Grounded in the philosophy, we integrate competence-based trust theory with the evaluation of predictive models. This work establishes a foundation for modeling and quantifying the trustworthiness of predictive models in decision-making contexts. The outcome of this effort promotes responsible AI development, enhances reliability and competence, and fosters user confidence, contributing to the advancement of ethical and transparent AI systems. Moreover, embracing a non-motives-based perspective aligns with the quest for objectivity and reliability in AI technologies, paving the way for ethically sound and socially beneficial applications of AI systems.

Acknowledgments

This work was supported in part by the NSF AI Institute for Foundations of Machine Learning (IFML).

References

Afroogh, S. 2023. A probabilistic theory of trust concerning artificial intelligence: Can intelligent robots trust humans? *AI and Ethics*, 3(2): 469–484.

Agarwal, S.; Graepel, T.; Herbrich, R.; Har-Peled, S.; Roth, D.; and Jordan, M. I. 2005. Generalization Bounds for the Area Under the ROC Curve. *Journal of Machine Learning Research*, 6(4).

Airola, A.; Pahikkala, T.; Waegeman, W.; De Baets, B.; and Salakoski, T. 2011. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis*, 55(4): 1828–1844.

Alvarado, R. 2022. What kind of trust does AI deserve, if any? *AI and Ethics*, 1–15.

Baier, A. 1986. Trust and Antitrust. *Ethics*, 96: 231.

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *Nips tutorial*, 1: 2017.

Boyd, A.; Tinsley, P.; Bowyer, K.; and Czajka, A. 2023. The value of ai guidance in human examination of synthetically-generated faces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5930–5938.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3.

Broderick, T.; Gelman, A.; Meager, R.; Smith, A. L.; and Zheng, T. 2023. Toward a taxonomy of trust for probabilistic machine learning. *Science Advances*, 9(7): eabn3999.

Chatzimparmpas, A.; Martins, R. M.; Jusufi, I.; Kucher, K.; Rossi, F.; and Kerren, A. 2020. The state of the art in enhancing trust in machine learning models with the use of visualizations. In *Computer Graphics Forum*, volume 39, 713–756. Wiley Online Library.

Cortes, C.; and Mohri, M. 2004. Confidence intervals for the area under the ROC curve. *Advances in neural information processing systems*, 17.

Crowson, C. S.; Atkinson, E. J.; and Therneau, T. M. 2016. Assessing calibration of prognostic risk scores. *Statistical methods in medical research*, 25(4): 1692–1706.

Eshete, B. 2021. Making machine learning trustworthy. *Science*, 373(6556): 743–744.

Farzaneh, N.; Ansari, S.; Lee, E.; Ward, K. R.; and Sjoding, M. W. 2023. Collaborative strategies for deploying artificial intelligence to complement physician diagnoses of acute respiratory distress syndrome. *NPJ Digital Medicine*, 6(1): 62.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Halimu, C.; Kasem, A.; and Newaz, S. S. 2019. Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In *Proceedings of the 3rd international conference on machine learning and soft computing*, 1–6.

Hanley, J. A.; and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1): 29–36.

Hébert-Johnson, U.; Kim, M.; Reingold, O.; and Rothblum, G. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, 1939–1948. PMLR.

Horsburgh, H. J. N. 1960. The ethics of trust. *The Philosophical Quarterly (1950-)*, 10(41): 343–354.

Huang, J.; and Ling, C. X. 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3): 299–310.

Jiang, H.; Kim, B.; Guan, M.; and Gupta, M. 2018. To trust or not to trust a classifier. *Advances in neural information processing systems*, 31.

Jones, K. 1996. Trust as an affective attitude. *Ethics*, 107(1): 4–25.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Rambachan, A. 2018. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, 22–27. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Kumar, A.; Sarawagi, S.; and Jain, U. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, 2805–2814. PMLR.

La Cava, W.; Lett, E.; and Wan, G. 2022. Proportional Multicalibration. *arXiv preprint arXiv:2209.14613*.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9): 1–46.

Lobo, J. M.; Jiménez-Valverde, A.; and Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2): 145–151.

Luo, Y.; Wong, Y.; Kankanhalli, M. S.; and Zhao, Q. 2021. Learning to predict trustworthiness with steep slope loss. *Advances in Neural Information Processing Systems*, 34: 21533–21544.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Mittermaier, M.; Raza, M.; and Kvedar, J. C. 2023. Collaborative strategies for deploying AI-based physician decision support systems: challenges and deployment approaches. *npj Digital Medicine*, 6(1): 137.

Murphy, A. H.; and Winkler, R. L. 1977. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1): 41–47.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Nguyen, A.; Halpern, M.; Wallace, B.; and Lease, M. 2016. Probabilistic modeling for crowdsourcing partially-subjective ratings. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, 149–158.

Platt, J.; et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3): 61–74.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " 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*, 1135–1144.

Ryan, M. 2020. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5): 2749–2767.

Safavi, T.; Koutra, D.; and Meij, E. 2020. Evaluating the Calibration of Knowledge Graph Embeddings for Trustworthy Link Prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Schröder, G.; Thiele, M.; and Lehner, W. 2011. Setting goals and choosing metrics for recommender system evaluations. In *UCERSTI2 workshop at the 5th ACM conference on recommender systems, Chicago, USA*, volume 23, 53.

Serban, A.; van der Blom, K.; Hoos, H.; and Visser, J. 2021. Practices for engineering trustworthy machine learning applications. In *2021 IEEE/ACM 1st Workshop on AI engineering-software engineering for AI (WAIN)*, 97–100. IEEE.

Tomani, C.; and Buettner, F. 2021. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9886–9896.

Toreini, E.; Aitken, M.; Coopamootoo, K.; Elliott, K.; Zelaya, C. G.; and Van Moorsel, A. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 272–283.

Varshney, K. R. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3): 26–29.

von Eschenbach, W. J. 2021. Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, 34(4): 1607–1622.

Widmann, D.; Lindsten, F.; and Zachariah, D. 2019. Calibration tests in multi-class classification: A unifying framework. *Advances in neural information processing systems*, 32.

Wong, A.; Wang, X. Y.; and Hryniowski, A. 2020. How much can we really trust you? towards simple, interpretable trust quantification metrics for deep neural networks. *arXiv preprint arXiv:2009.05835*.

Zadrozny, B.; and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 694–699.