Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty

Umang Bhatt[1,2], Yunfeng Zhang[3], Javier Antorán[2], Q. Vera Liao[3], Prasanna Sattigeri[3], Riccardo Fogliato[1,4], Gabrielle Gauthier Melançon[5], Ranganath Krishnan[6], Jason Stanley[5], Omesh Tickoo[6], Lama Nachman[6], Rumi Chunara[7], Adrian Weller[2,8], Alice Xiang[1]

[1]Partnership on AI, [2]University of Cambridge, [3]IBM Research, [4]Carnegie Mellon University, [5]Element AI, [6]Intel Labs, [7]New York University, [8]The Alan Turing Institute

ABSTRACT

Transparency of algorithmic systems entails exposing system properties to various stakeholders for purposes that include understanding, improving, and/or contesting predictions. The machine learning (ML) community has mostly considered explainability as a proxy for transparency. With this work, we seek to encourage researchers to study uncertainty as a form of transparency and practitioners to communicate uncertainty estimates to stakeholders. First, we discuss methods for assessing uncertainty. Then, we describe the utility of uncertainty for mitigating model unfairness, augmenting decision-making, and building trustworthy systems. We also review methods for displaying uncertainty to stakeholders and discuss how to collect information required for incorporating uncertainty into existing ML pipelines. Our contribution is an interdisciplinary review to inform how to measure, communicate, and use uncertainty as a form of transparency.

1 INTRODUCTION

Transparency is one way for an artificial intelligence (AI) system to show stakeholders that the system is trustworthy [140, 183]. Transparency includes a wide variety of efforts to provide stakeholders, such as model developers and end users, with relevant information about how a machine learning (ML) model works [16, 183]. One form of transparency is procedural transparency, which provides information about model development (e.g., code release, model cards, dataset details) [6, 59, 128, 151]. Another form is algorithmic transparency, which exposes information about a model's behavior to various stakeholders [97, 156, 170].

Most of the algorithmic transparency agenda for the Fairness, Accountability, and Transparency (FAccT) community has focused on explainability. Explainability attempts to provide reasons for a model's behavior to stakeholders [16]. However, understanding a model's specific behavior might not be enough for stakeholders to gauge whether the model may be wrong or lack sufficient knowledge to solve the task at hand [15]. **We argue that a complementary form of transparency is to estimate and communicate the uncertainty associated with model predictions.**

There are multiple ways in which uncertainty can be characterized. In regression tasks, uncertainty is often expressed in terms of predictive variance. For example, when predicting the number of crimes in a given city, we could report that the number of predicted crimes is 943 ± 10, where "± 10" represents a 95% *confidence interval* (capturing two deviations on either side of the central, mean estimate). The smaller the interval, the more certain the model. In classification tasks, probability scores are often used to capture how confident a model is in a specific prediction. For example, a classification model may predict that a person is at a high risk for developing diabetes given a prediction of 85% chance of diabetes.

There may be various sources of uncertainty in data-driven decision-making systems [56, 75]. Aleatoric uncertainty is induced by inherent randomness (or noise) in the quantity we want to predict or in our input variables. Epistemic uncertainty can arise due to lack of sufficient data to learn the model precisely. The question of how to quantify different types of uncertainty and communicate them well has long been studied across many domains. For example, in the broader AI community, uncertainty has been leveraged in planning and reasoning tasks [76, 77, 143]. In this paper, we seek to study how uncertainty affects learning models from data.

This work is structured as follows. In Section 2, we motivate uncertainty as a form of transparency by discussing three use cases: designing fairer models, informing decision-making, and calibrating trust in automated systems. In Section 3, we review possible sources of uncertainty and methods for uncertainty quantification. In Section 4, we describe how uncertainty can be leveraged in each use case from Section 2. Next, in Section 5, we discuss how to communicate uncertainty effectively. Finally, we discuss the importance of taking a user-centered approach to collecting requirements for uncertainty quantification and communication in Section 6.

2 WHY DO WE CARE?

Well-calibrated uncertainty helps stakeholders understand when they should trust model predictions and helps developers address fairness issues in models. Uncertainty is crucial in the context of ML-assisted, or automated, decision-making. To that end, we discuss the use of uncertainty for obtaining fairer model outcomes, improved decision-making, and building trust in automation, using the following cancer diagnostics scenario for illustrative purposes.

Suppose we are tasked with diagnosing individuals as having breast cancer or not, as in [38, 46]. Given categorical and continuous characteristics about an individual (medical test results, family medical history, etc.), we estimate the probability of an individual having breast cancer. We can then apply a threshold to classify them into high- or low-risk groups. Specifically, we have been tasked with building ML-powered tools to help three distinct audiences: doctors, who will be assisted in making diagnoses; patients, who will be helped to understand their diagnoses; and review board members, who will be aided in reviewing doctors' decisions across many hospitals.

Throughout the paper, we will refer back to the scenario above to discuss how uncertainty may arise in the design of an ML model

and why, when well-calibrated and well-communicated, it can act as a useful form of transparency for stakeholders.

2.1 Fairness

Developers often aim to assess model fairness to mitigate or prevent unwanted biases. Uncertainty, if not properly quantified and considered in model development, can endanger these efforts. In our example of the breast cancer diagnostic tool, if the model is trained on data where young age groups are underrepresented, the model might be underspecified and would likely be biased towards lower error rates on older patients. Such dataset bias will manifest itself as epistemic uncertainty in the model itself. Yet, in some cases, uncertainty can also be leveraged to diagnose and improve the model. For example, one can use uncertainty to identify portions of the input space where the model is error-prone: regions of high epistemic uncertainty indicate where additional training data could improve model performance. Section 4 details different ways uncertainty interacts with bias in the data collection and modeling stages and how such biases can be mitigated by accounting for uncertainty.

2.2 Decision-making

End users of ML models—e.g., decision-makers using an ML-powered decision-support system—should consider the uncertainty of the model's output if well-calibrated uncertainty measures are available. A user may have to decide whether to accept a model's output in a given interaction or to delegate certain actions to a model. Treating all model predictions the same, independent from their uncertainty, can lead decision-makers to over-rely on the model in cases where it produces spurious outputs or to under-rely on the model in cases where model predictions are likely to be accurate. Conversely, a doctor might observe a model's uncertainty estimates before leveraging the model's output in making a diagnosis. In Section 4, we draw upon the literature of judgment and decision-making (JDM) to discuss the potential implications of showing uncertainty estimates to end users of ML models.

2.3 Trust in Automation

Communicating well-calibrated uncertainty can be seen as a sign of the model's trustworthiness: this communication could in turn improve model adoption and user experience. However, high uncertainty, seemingly arbitrary uncertainty, or incomprehensible uncertainty information could be perceived negatively, spawn confusion, and impair user trust. If our model's recommendations are always accompanied by large error bars (high uncertainty), doctors may choose to always override the model's output, the model's seeming imprecision resulting in an erosion of the doctors' trust. In general, accurately measured and carefully communicated uncertainty estimates should aim to support users in calibrating and forming appropriate trust in an ML model. Appropriately calibrated trust is crucial for avoiding misuse of automated systems [106]. In Section 4, we review prior work on how users form trust in automated systems and discuss the potential impact of uncertainty for user trust in ML models.

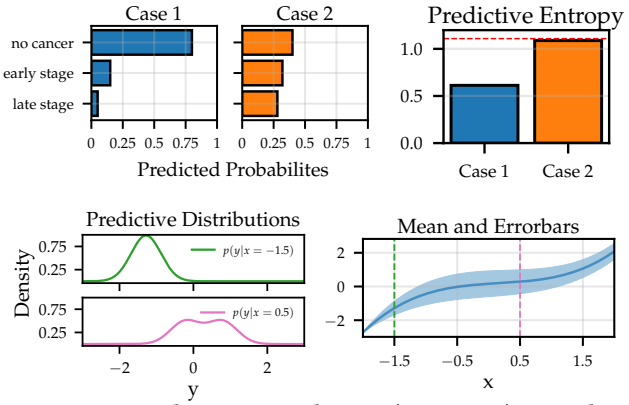


Figure 1: Top: The same prediction (no cancer) is made in two different hypothetical cancer diagnosis scenarios. However, our model is much more confident in the first. This is reflected in the predictive entropy for each case (dashed red line denotes the maximum entropy for 3-way classification). Bottom: In a regression task, a predictive distribution contains rich information about our model's predictions (modes, tails, etc). We summarise predictive distributions with means and error-bars (here standard deviations).

3 MEASURING UNCERTAINTY

In ML, we use the term uncertainty to refer to our lack of knowledge about some outcome of interest. We use the tools of probability to quantify and reason about uncertainty. But what are probabilities? The Bayesian school of thought interprets probabilities as subjective degrees of belief in an outcome of interest occurring [117]. For frequentists, probabilities reflect how often we would observe the outcome if we were to repeat our observation multiple times [18, 144]. Fortunately for end-users, uncertainty from Bayesian and frequentist methods conveys similar information in practice [187], and can be treated interchangeably in downstream tasks.

3.1 Metrics for Uncertainty

The metrics used to communicate uncertainty vary between research communities. As shown in Figure 1, a predictive distribution tells us about our model's degree of belief in every possible prediction. Despite containing a lot of information (prediction modes, tails, etc.), a full predictive distribution may not always be desirable. In our cancer diagnostic scenario, we may want our automated system to abstain from making a prediction and instead request the assistance of a medical professional when its uncertainty is above a certain threshold. When deferring to a human expert, it may not matter to us whether the system believes the patient has cancer or not, just how uncertain it is. For this reason, **summary statistics** of the predictive distribution are often used to convey information about uncertainty. For classification, class probability intuitively communicates our degree of belief in an outcome. On the other hand, predictive entropy decouples our predictions from their uncertainty, only telling us about the latter. For regression, the predictive mean is often given together with error bars (written $\pm\sigma$). These commonly reflect the standard deviation or some percentiles of the predictive distribution. As discussed in Section 6, the best

Table 1: Commonly used metrics for the quantification and communication of uncertainty.

	Full Information	Summary Statistics
Regression	Predictive Distribution	Predictive Variance, Percentile (or quantile) Confidence Intervals
Classification	Predictive Probabilities	Predictive Entropy, Expected Entropy, Mutual Information, Variation Ratio

choice of uncertainty metric will be use case dependent. We show common summary statistics for uncertainty in Table 1 and further discuss them in Appendix A.

3.2 The Different Sources of Uncertainty

While there can be many sources of uncertainty [179], herein we focus on those types that we can quantify in ML models: *aleatoric uncertainty* (also known as indirect uncertainty) and *epistemic uncertainty* (also known as direct uncertainty) [39, 40, 56].

Aleatoric uncertainty stems from noise, or class overlap, in our data. Noise in the data is a consequence of unaccounted-for factors that introduce variability in the inputs or targets. Examples of this could be background noise in a signal detection scenario or the imperfect reliability of a medical test in our cancer diagnosis scenario. Aleatoric uncertainty is also known as irreducible uncertainty: it cannot be decreased by observing more data. If we wish to reduce aleatoric uncertainty, we may need to leverage different sources of data, e.g. switching to a more reliable clinical test. In practice, most ML models account for aleatoric uncertainty through the specification of a noise model or likelihood function. A homoscedastic noise model makes an assumption that all of the input space is equally noisy, $y = f(x) + \epsilon;\ \epsilon \sim p(\epsilon)$. However, this may not always be true. Returning to our medical scenario, consider using results from a clinical test which produces few false positives but many false negatives as an input to our model. A heteroscedastic noise assumption allows us to express aleatoric uncertainty as a function of our inputs $y = f(x) + \epsilon;\ \epsilon \sim p(\epsilon|x)$. Perhaps the most commonly used heteroscedastic noise models in deep learning are those induced by the sigmoid or softmax output layers. These enable almost any ML model to express aleatoric uncertainty (Appendix A.1).

Epistemic uncertainty stems from a lack of knowledge about which function best explains the data we have observed. There are two reasons why epistemic uncertainty may arise. Consider a scenario in which we employ a very complex model relative to the amount of training data available. We will be unable to properly constrain our model's parameters. This means that, out of all the possible functions that our model can represent, we are unsure of which ones to choose. This uncertainty about a model's parameters is known as *model uncertainty*. We might also be uncertain of whether we picked the correct model class in the first place. Perhaps we are using a linear predictor but the phenomenon we are trying to predict is non-linear. This is known as *model specification uncertainty* or *architecture uncertainty*. Epistemic uncertainty can be reduced by collecting more data in input regions where the original training set was sparse. It is less common for ML models to capture epistemic uncertainty. Often, those that do are referred to as probabilistic models.

Given a probabilistic predictive model, aleatoric and epistemic uncertainties can be quantified separately, as described in Appendix A. We depict them both separately in Figure 2. Being aware of which regions of the input space present large aleatoric uncertainty can help ML practitioners identify issues in their data collection process. On the other hand, epistemic uncertainty tells us about which regions of input space we have yet to learn about. Thus, epistemic uncertainty is used to detect dataset shift [141], or adversarial datapoints [190]. It is also used to guide methods that require explorations such as active learning [78], continual learning [134], Bayesian optimisation [70], and reinforcement learning [82].

3.3 Methods to Quantify Uncertainty

Most ML approaches involve a noise model, thus capturing aleatoric uncertainty. However, few are able to express epistemic uncertainty. When we say that a method is able to quantify uncertainty, we are implicitly referring to those that capture both epistemic and aleatoric uncertainty. These methods can be broadly classified into two categories: Bayesian approaches [5, 20, 32, 51, 57, 63, 69, 92, 118, 184, 194] and Non-Bayesian, or frequentist, approaches [2, 98, 102, 107, 111, 177].

Bayesian methods explicitly define a hypothesis space of plausible models *a priori* (before observing any data) and use deductive logic to update these priors given the observed data. In parametric models, like Bayesian Neural Networks (BNNs) [116, 133], this is most often done by treating model weights as random variables instead of single values, and assigning them a prior distribution $p(\mathrm{w})$. Given some observed data $\mathrm{D} = \{\mathrm{y}, \mathrm{x}\}$, the conditional likelihood $p(\mathrm{y} \mid \mathrm{x}, \mathrm{w})$ tells us how well each weight setting w explains our observations. The likelihood is used to update the prior, yielding the posterior distribution over the weights $p(\mathrm{w}|\mathrm{D})$:

$$p(\mathrm{w}|\mathrm{D}) = \frac{p(\mathrm{y} \mid \mathrm{x}, \mathrm{w})\, p(\mathrm{w})}{\int p(\mathrm{y} \mid \mathrm{x}, \mathrm{w})\, p(\mathrm{w})\, \mathrm{dw}} \tag{1}$$

Predictions for test points x^* are made through marginalization: all possible weight configurations are considered with each configuration's predictions being weighed by that set of weights' posterior density. The disagreement among the predictions from different plausible weight settings induces model (epistemic) uncertainty. Thus, the predictive posterior distribution:

$$p(\mathrm{y}|\mathrm{x}^*) = \int p(\mathrm{y}|\mathrm{x}^*, \mathrm{w}) p(\mathrm{w}|\mathrm{D})\, d\mathrm{w} \tag{2}$$

captures both epistemic and aleatoric uncertainty.

In recent years, the ML community has moved towards favoring NNs as their choice of model due to their flexibility and scalability to large amounts of data. Unfortunately, the more complicated the model, the more difficult it is to compute the exact posterior distribution $p(\mathrm{w}|\mathrm{D})$. For NNs, it is analytically and computationally intractable [69]. However, various approximations have been proposed. Among the most popular are variational inference [20, 57, 71] and stochastic gradient MCMC [32, 184, 194]. Methods that provide more faithful approximations, and thus more calibrated uncertainty estimates, tend to be more computationally intensive and scale worse to larger models. As a result, the best method will vary depending on the use case.

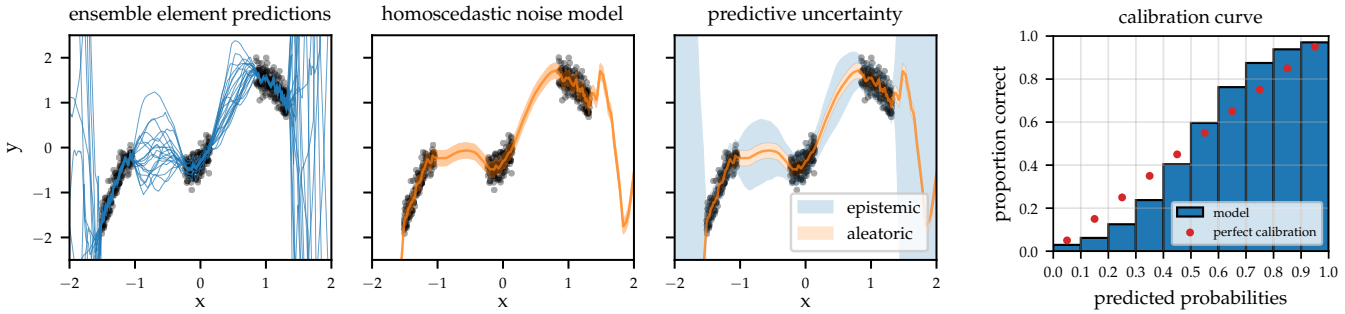


Figure 2: Uncertainty quantification and evaluation. Left three plots: A 15-element deep ensemble provides increased model (epistemic uncertainty) in data sparse regions. A homoscedastic Gaussian noise model provides aleatoric uncertainty matching the noise in the data. Both combine to produce a predictive distribution. Right: calibration plot. Each bar corresponds to a bin in which predictions are grouped. Their height corresponds to their proportion of correct predictions.

Often, the predictive distribution, Equation (2), is also intractable. In practice, for parametric models like NNs, it is approximated by making predictions with multiple plausible models, sampled from $p(\mathrm{w}|\mathrm{D})$. We see this in the leftmost plot from Figure 2: the predictions from different ensemble elements can be interpreted as samples which, when combined, approximate the predictive posterior [186]. Different predictions tend to agree in the data-dense regions but they disagree elsewhere, yielding epistemic uncertainty. Note, because we need to evaluate multiple models, sampling-based approximations of the predictive posterior incur additional computational cost at test time compared to non-probabilistic methods.

Also worth mentioning are Bayesian non-parametrics, such as Gaussian Processes [153]. These are easy to deploy and allow for exact probabilistic reasoning, making their predictions and uncertainty estimates very robust. Unfortunately their computational cost grows cubically with the number of datapoints, making them a very strong choice of model only in the small data regime (≤ 5000 points). Otherwise, approximate algorithms, such as variational inference [122], are required.

Frequentist methods do not specify a prior distribution over hypothesis. They exclusively consider how well the distribution over observations implied by each hypothesis matches the data. Here, uncertainty stems from how we would expect an outcome to change if we were to repeatedly sample different sets of data given our chosen hypothesis. Ensembles [112] train multiple models in different ways to obtain multiple plausible fits. At test time, the disagreement between ensemble elements' predictions yields model uncertainty, as show in Figure 2. Currently, deep ensembles [102] are one of best performing uncertainty quantification approaches for NNs [8], retaining calibration even under dataset shift [141]. Unfortunately, the computational cost involved with running multiple models at both train and test time also make ensembles one of the most expensive methods. Within frequentist methods, post-hoc approaches are especially attractive; they allow us to obtain uncertainty estimates from non-probabilistic models, independently of how these were trained. A principled way to do this is to leverage the curvature of a loss function around its optima. Optima in flatter regions suffer from less variance [73], suggesting our model's parameters are well specified by the data. This is used to draw plausible weight samples by Resampling Uncertainty Estimation (RUE) [160] and local ensembles [119]. Alaa and van der Schaar [3] adapt the Jackknife, a traditional frequentist method, to generating post-hoc confidence intervals in small-scale neural networks. We discuss additional uncertainty quantification approaches in Appendix C.

3.4 Uncertainty Evaluation

Calibration is a form of quality assurance for uncertainty estimates. It is not enough to provide larger error bars when our model is more likely to make a mistake. Our predictive distribution must reflect the true distribution of our targets. Recall our cancer diagnosis scenario, where the system declines to make a prediction when uncertainty is above a threshold, and a doctor is queried instead. Due to the doctor's time being limited, we design our system such that it only declines to make a prediction if it estimates there is a probability greater than 0.05 of the prediction being wrong. If instead of being well-calibrated, our system is underconfident, we would over-query the doctor in situations where the AI's prediction is correct. Overconfidence would result in us taking action on unreliable predictions: delivering unnecessary treatment or abstaining from providing necessary treatment.

Calibration is orthogonal to accuracy. A model with a predictive distribution that matches the marginal distribution of the targets $p(y|x) = p(y)\,\forall\, x$ would be perfectly calibrated but would not provide any useful predictions. Thus, calibration is usually measured in tandem with accuracy through, either a general fidelity metric (most often chosen to be a proper scoring rule [60]) which subsumes both objectives, or two separate metrics. The most common metrics of the former category are *negative log-likelihood* (NLL) and *Brier score* [22]. Of the latter type, *Expected calibration error* (ECE) [131] is popularly used in classification scenarios. ECE segregates a model's predictions into bins depending on their predictive probability. In our example, this would mean grouping all patients who have been assigned a probability of having a cancer $\in [0, k)$ into a first bin, those who have been assigned probability $\in [k, 2k)$ into the second bin, etc. For each bin, calibration error is the difference between the proportion of patients who actually have the disease and the average of the probabilities assigned to that bin. This is illustrated in Figure 2, where we use 10 bins. In this example,

our model presents overconfidence in bins with $p < 0.5$ and underconfidence otherwise. We refer to Appendix B for a discussion of additional calibration metrics, including some for regression.

A transparent ML model requires both well-calibrated uncertainty estimates and an effective way to communicate them to stakeholders. If uncertainty is not well-calibrated, our model cannot be transparent since its uncertainty estimates provide false information. Thus calibration is a precursor to using uncertainty as a form of transparency.

4 USING UNCERTAINTY

In this section, we discuss the use of uncertainty based on the three motivations presented in Section 2. The different uses mentioned are not mutually exclusive; as such, they could all be leveraged simultaneously depending on the use case. In Section 6, we discuss the importance of gathering requirements on how stakeholders, both internal and external, will use uncertainty estimates.

4.1 Uncertainty and Fairness

An unfair ML model is one that exhibits unwanted bias towards or against one or more target classes. Here we discuss possible ways in which bias can be mitigated when uncertainty affects measurement of the features, data collection, and modeling.

Uncertainty in the data

Measurement bias, also known as feature noise, is a case of aleatoric uncertainty (defined in Section 3), and arises when one or more of the features in the data only represent a proxy for the features that were intended to be measured. We briefly describe noise on two types of features. First, noise in the sensitive attribute. In contexts such as our running example, information on race and ethnicity of patients may not be collected [31]. Another example is answers provided by participants in a survey, who may have incentives to misreport their religious or political affiliations. The experimental results of Gupta et al. [65] have shown that enforcing fairness constraints on a noisy sensitive attribute, without assumptions on the structure of the noise, is not guaranteed to lead to any improvement in the fairness properties of the classifier trained on that data. Successive papers have explored which assumptions are needed in order to obtain such guarantees. When the noise only depends on the true unobserved value of the sensitive attribute (i.e., the noise follows the "mutually contaminated learning model" [161]), the measures of demographic parity and equalized odds computed on the observed data are equal to the true metrics up to a scaling factor, which is proportional to the value of the noise [103]; if the noise rates are known, then the true metrics can be directly estimated. When information on the protected class is unavailable, but it can be predicted from an auxiliary dataset, disparity measures are generally unidentifiable [31, 87]. Second, we discuss noise in the outcome. In a case similar to our running example, medical expenses were used as a proxy for illness, and the algorithm severely underestimated the prevalence of the illness for the population of Black patients [138]. Interestingly, this bias has attracted less attention in the fairness community. Notably, Blum and Stangl [19], Jiang and Nachum [84] have shown that fairness constraints are guaranteed to improve the properties of the classifier only under appropriate assumptions on the noise. The work of Fogliato et al. [53] has shown that even a small amount of noise can greatly impact the assessment of the fairness properties of the model. This type of uncertainty can be mitigated by an appropriately specified noise model.

Sampling bias is epistemic in nature and occurs when the observations in the available data are not representative of the true data distribution. Models trained in presence of sampling bias could exhibit unwanted bias towards the under-represented group. For example, the differential performance of gender classification systems across racial groups may be due to under-representation of Black individuals in the sample [26]. Similarly, historical over-policing of certain communities has unavoidable impacts on the data used to train algorithms for predictive policing [49, 114]. In our cancer diagnostics scenario, we would likely prefer to deploy a model that is trained on a population of patients from the same hospital that will treat future patients. This type of bias, often called representation bias [171], can show up due to the difference in the relative distributions of the classes in the training data as compared to the population being represented. In addition, if there is a mismatch between the diversity representation in the data used to build an algorithm and the diversity of the target population, the algorithm can suffer from population bias [139]. This problem also arises when deploying an algorithm trained on one population on a different population. Note that if sampling bias only consists of covariate shift [163] and the model is correctly specified, then it won't affect the quality of the model's predictions. Still, sample size may still represent an issue. Additionally, in many domains, sample size may also not be large enough to assess the existence of biases [50].

Uncertainty in model specification

This type of uncertainty arises when the hypothesis class chosen for modeling the data may not contain the true data-generating process and could result in unwanted bias in model predictions. For example, we might prefer a simple and explainable model for cancer diagnostics that, mistakenly, does not account for the non-linearity present in the data. Similarly, for ethical reasons we might choose to exclude the patient's race as a predictor from the model, when this information could help improve its performance [180]. The hypothesis class or the family of functions used to fit the data is primarily determined using domain knowledge and preferences of the model designer. For these reasons, the resulting model should be seen only as an approximation of the true data-generating process [25]. In addition, using different benchmarks for the measurement of an algorithm's performance can lead to different choices in the final model. As a result, the trained classifier may not achieve high performance, even with potentially unlimited and rich data.

Bias arising from model uncertainty can potentially be mitigated by considering enlarging the hypothesis class considered, such as by using deep neural networks, when the datasets are sufficiently large. In general, this kind of bias is hard to disentangle from data uncertainty and therefore it can rarely be detected or analysed.

Uncertainty and bias mitigation

We now present some of the methods to mitigate data bias and the possible implications of using uncertainty. These methods are typically categorized as pre-, in-, or post-processing based on the stage at which the model learning is intervened upon.

Pre-processing techniques modify the distribution of the data the classifier will be trained on, either directly [27] or in a low-dimensional representation space [191]. Implicitly, these techniques reduce the uncertainty emerging from the features, outcome, or sensitive variables. Uncertainty estimation at this stage involves representing and comparing the training data distribution and random population samples with respect to target classes. The uncertainty measurements represented as distribution shifts of targeted classes between the training data and the population samples give a measure of training data bias which can be corrected by data augmentation of under-represented classes or by collecting larger and richer datasets [30]. The equalized odds post-processing method of Hardt et al. [66] is guaranteed to reduce the bias of the classifier under an assumption on the noise in the sensitive attributes, namely the independence of the classifier prediction and the observed attribute, conditional on both the outcome and the true sensitive attribute [9].

In-processing methods modify the learning objective by introducing constraints through which the resulting classifier can achieve the desired fairness properties [1, 43, 192]. Comparisons of the distribution of the features can be used to detect uncertainty in the model during training. When little or no information about the sensitive attribute is available, distributionally robust optimization can be used to enforce fairness constraints [67, 181]. Algorithmic fairness approaches that employ active learning to either acquire features [137] or samples [4] with accuracy and fairness objectives also fall under this category.

Post-processing techniques essentially modify the model's predictions post-training to satisfy a chosen fairness criterion [66]. Frameworks like [88] can use this uncertainty information to de-bias the output of such models. Here, the predictions that are associated with high uncertainty can be skewed to favor a sensitive class as a de-biasing post-processing measure. Uncertainty in predictions can also be used to abstain from making decisions or defer the decisions to experts, which can lead to overall improvement in accuracy and fairness of the predictions [120]. An optimization-based approach is proposed in [182] to transform the scores with suitable trade-off between utility of the predictions and fairness, with the assumption that the scores are well-calibrated.

Often there exist trade-offs between the different notions of fairness [33, 36, 95] (see Appendix E for the definitions of the commonly used fairness metrics). For example, it has been shown that calibration and equalized odds cannot be achieved simultaneously when the base rates of the sensitive groups are different [148]. Interestingly, this impossibility can be overcome, as shown in [28], by deferring uncertain predictions to experts. It is important that the uncertainty measurements produced by the prediction models are meaningful and unbiased [157] for reliable functioning of such bias mitigation methods and for communication to decision-makers.

4.2 Uncertainty and Decision-making

While ML can be used for many purposes, one use is to support or augment human decision-making. Depending on the context, the forms of decision support vary. For example, a model could recommend a product, suggest a risk score, detect potential abnormalities, predict a future event, etc. All of these situations require the end user to make a decision weighing uncertainty, whether the uncertainty is explicitly presented or not, asking themselves: Should I accept or rely on the model's output? Sometimes when users are given output from multiple models, they may ask: Which model should I accept or rely on? These questions correspond to the prototypical tasks of decision-making under uncertainty/risk [1] as studied in the Judgment and Decision-Making (JDM) literature, i.e., action threshold decision and multi-option choices [52]. While recent work has only begun to examine how uncertainty estimates might affect user interactions with ML models and decision task performance [7, 196], we highlight a few conclusions from the bulk of JDM literature that suggest how uncertainty estimates might be used in decision-making.

For classification tasks, decision-makers can use the probability score representing uncertainty as explicit risk information–the chance that the model prediction is wrong. Prospect Theory suggests that risk is not considered independently but together with the expected outcome [86, 174]. That is, a prediction with a small uncertainty but leading to a large loss might be perceived more negatively than a prediction with medium uncertainty for a small loss. How people value choices in terms of their risks and outcomes is also non-linear and asymmetrical. Specifically, when faced with a risky choice leading to gains, people are risk-averse; when the choice leads to losses, people are risk-seeking. Since uncertainty in ML-mediated decisions is more often than not framed as a gain (e.g., a 95% chance or confidence of the model being correct, instead of a 5% chance of the model being wrong), people may have a non-linear risk-averse tendency: As the stake of the decision-outcome increases, people's tolerance for the magnitude of uncertainty, i.e., the acceptable level of confidence, could decrease at a speedier rate [178].

How people actually assess risk, however, also depends on how the uncertainty estimates are communicated and perceived. Both lay people and experts trained in statistics rely on mental shortcuts, or heuristics, to *interpret* uncertainty [175]. This could lead to biased understanding or appraisal of uncertainty even if it is accurately measured. In Section 5, we discuss some of these biases and their implications for communication of uncertainty. There are also individual differences in one's acceptable or preferred level of uncertainty [149], depending on many factors, such as expertise (experts might tolerate less uncertainty [68]), personality (e.g., uncertainty-orientation [164]) and cognitive style [127].

As discussed in Section 3, uncertainty in ML models may come from different sources. To our knowledge, the empirical understanding of how decision-makers make use of aleatory versus epistemic uncertainty is limited. There is some prior work in the JDM literature showing that in the face of epistemic uncertainty about the occurrence of future events, people may postpone their decision [176]. Understanding how people react to different types of uncertainty, and uncertainty expressed as a range or confidence intervals as in regression tasks, could be important gaps to fill by future work on human-AI interaction.

[1] Social scientists often use the term "risk" for chances of negative events known to the decision-maker, and "uncertainty" for the unmeasurable likelihood of events that are uncertain (but can be assessed by the decision-maker) [96, 152]

4.3 Uncertainty and Trust Formation

While trust could be implicit in a decision to rely on a model's suggestion, the communication of a model's uncertainty can also affect people's general trust in an ML system. At a high level, communicating uncertainty is a form of model transparency that can help gain people's trust. However, a look into the underlying construct of trust, and how people form trust, paints a more complex picture of how end users and stakeholders might use uncertainty estimates to form trust in an ML system.

While not limited to ML-powered systems, the HCI and Human Factors communities have a long history of studying trust in automation [74, 99, 106]. These models of trust often build on Mayer et al. [123]'s classic ABI model of inter-personal trust, which postulates that the perceived trustworthiness of the trustee is determined by three attributes: 1) Ability: The level of competencies that enable the trustee to have influence within the targeted domain. 2) Benevolence: The extent to which a trustee is perceived to want to do good to the trustor. 3) Integrity: The extent to which the trustee consistently adheres to a set of principles that the trustor finds acceptable. Taking into account some fundamental differences between interpersonal trust and trust in automation, Lee and See [106] adapted the ABI model to posit that trustworthiness of automated systems is determined by three underlying dimensions: Competence, Intention of Developers, and Predictability/Understandability. We speculate that communicating uncertainty estimates could be relevant to all three of these dimensions. If a model always has high uncertainty, it will harm the model's perceived Competence. If a model shows uncertainty that could not be understood or expected, it will be negatively perceived in Predictability. If uncertainty is not communicated or intentionally mis-communicated, users or stakeholders will hold a negative opinion on the Intention of Developers.

To anticipate how uncertainty estimates and ways to communicate them could impact user trust, it is also useful to consider process models on *how* people develop trust. Rooted in information-processing and decision-making theories [29, 85, 147], process models differentiate between an analytic, or systematic, process of trust formation, and an affective, or heuristic, process of trust formation [106, 125, 168]. Specifically, the former process involves rational evaluation of a trustee's characteristics, while the later process relies on feelings or heuristics to form a quick judgment to (un)trust. When lacking either the ability or motivation to perform an analytic evaluation, people rely more on the affective or heuristic route [147, 169]. While detailed probabilistic uncertainty estimates could facilitate analytic evaluation of model trustworthiness, it is important to note that users, especially lay people, might simply rely on some kind of heuristics or feelings that are invoked by the presentation of uncertainty information. For example, for some users the mere presence of uncertainty information could signal that the ML engineers are being transparent and sincere, which then enhances their trust [79]. For others, uncertainty could invoke negative heuristics as a lack of expertise [179]. Even the style of communication matters. Prior work suggests that politely communicating the existence of uncertainty could promote trust [142]. How uncertainty estimates are processed for trust formation, and what kind of affective impact or heuristics related to trust they could invoke, remain open questions and merit future research.

Lastly, we highlight a non-trivial point that the goal of presenting uncertainty estimates to end users and stakeholders should support forming *appropriate* trust, rather than blindly enhancing trust. A well-measured and well-communicated uncertainty estimate should not only facilitate the *calibration* of overall trust on a system, but also *resolution* of trust [35, 106], referring to how precisely the judgment of trust could differentiate types of model capabilities, for example in what situations the system is more or less trustworthy.

5 COMMUNICATING UNCERTAINTY

Treating uncertainty as a form of transparency also requires accurately communicating it to the stakeholders. However, even well-calibrated uncertainty estimates could be perceived inaccurately by people because (a) people have varying levels of understanding about probability and statistics, and (b) human perception of uncertainty quantities is often biased by their decision-making heuristics. In this section, we will review some of these issues that hinder people's understanding of uncertainty estimates and will discuss how various communication methods may help address these issues. We will first describe how to communicate uncertainty in the form of confidence or prediction probabilities for classification tasks, and then more broadly in the form of ranges, confidence intervals, or full distributions. We will then dive into a case study on the utility of uncertainty communication during the COVID-19 pandemic.

5.1 Issues in Understanding Uncertainty

Many application domains involve communicating uncertainty estimates to the general public to help them make decisions, e.g., weather forecasting, transit information system [89], medical diagnosis and interventions [149]. One key issue in these applications is that a great deal of their audience may not have high numeracy skills and may not understand uncertainty correctly. In a survey [58] conducted in 2010 on statistical numeracy across the US and Germany, it was found that many people do not understand relatively simple statements that involve statistics concepts. For example, 20% of the German and US participants could not say "which of the following numbers represents the biggest risk of getting a disease: 1%, 5%, or 10%," and almost 30% could not answer whether 1 in 10, 1 in 100, or 1 in 1000 represents the largest risk. Another study [197] found that people's numeracy skills significantly affect how well they comprehend risks. Many of the aforementioned decision-making scenarios involve high-stake decisions, thus it is vital to find alternative ways to communicate uncertainty estimates to people with low numeracy skills.

Besides numeracy skills, research (cf. [85, 155, 165]) shows that humans in general suffer from a variety of cognitive biases, some of which hinder our understanding of uncertainties. One of them is called ratio bias, which refers to the phenomenon that people sometimes believe a ratio with a big numerator is larger than an equivalent ratio with a small numerator. For example, people may see 10/100 as a larger odds of having breast cancer than 1/10. This same phenomenon is sometimes manifested as an underweighting of the denominator, e.g. believing 9/11 is smaller than 10/13. This is also called denominator neglect.

In addition to ratio biases, people's perception of probabilities are also distorted in that they tend to underweight high probabilities

while overweighting low probabilities, and this distortion prevents people from making optimal decisions. Zhang and Maloney [193] showed that when people are asked to estimate probabilities or frequencies of events based on memory or visual observations, their estimates are distorted in a way that follows a log-odds transformation of the true probabilities. Research [174, 195] also found that this bias occurs when people are asked to make decisions under risk and that their decisions imply such distortions. Therefore, when communicating probabilities, we need to be aware that people's perception of high risks may be lower than the actual risk, while that of low risks may be higher than actual.

A different kind of cognitive bias that impacts people's perception of uncertainty is framing [85]. Framing has to do with how information is contextualized. Typically, people prefer options with positive framing (e.g., a 80% likelihood of surviving breast cancer) than an equivalent option with negative framing (e.g., a 20% likelihood of dying from breast cancer). This bias has an effect on how people perceive uncertainty information. A remedy to this bias is to always describe the uncertainty of both positive and negative outcomes, rather than relying on the audience to infer what's left out of the description.

5.2 Communication Methods

Choosing the right communication methods can address some of the above issues. van der Bles et al. [179] categorize the different ways of expressing uncertainty into nine groups with increasing precision, from explicitly denying that uncertainty exists to displaying a full probability distribution. While high-precision communication methods help experts understand the full scale of the uncertainty of the ML models, low precision methods require less numeracy skill and can be used for lay people. In this paper, we focus on the pros and cons of the four more precise methods of communicating uncertainty: 1) describe the degree of uncertainty using a predefined categorization, 2) describe a numerical range, 3) show a summary of a distribution, and 4) show the full probability distribution. The first two methods can be communicated verbally, while the last two often require visualizations.

A predefined, ordered categorization of uncertainty and risk levels reduces the cognitive effort needed to comprehend uncertainty estimates, and therefore is particularly likely to help people with low numeracy skills [145]. A great example of how to appropriately use this technique is the GRADE guidelines [11], which introduce a four-category system, from high to very low, to rate the quality of evidence for medical treatments. GRADE has provided definitions for each category and a detailed description of the aspects of studies to evaluate for constructing quality ratings. Uncertainty ratings are also frequently used by financial agencies to communicate the overall risks associated with an investment instrument [42].

The main drawback of communicating uncertainty via predefined categories is that the audience, especially non-experts, might not be aware of or even misinterpret the threshold criteria of the categories. Many studies have shown that although individuals have internally consistent interpretation of words for describing probabilities (e.g., likely, probably), these interpretations can vary substantially from one person to another (cf. [24, 34, 109]). More recently, Budescu et al. [23] investigated how the general public interpret the uncertainty information in the climate change report published by the Intergovernmental Panel on Climate Change (IPCC). They found that people generally interpreted the IPCC's categorical description of probabilities as less likely than the IPCC intended. For example, people took the word "very likely" as indicating a probability of around 60%, whereas the IPCC's guideline specifies that it indicates a greater than 90% probability. To avoid such misinterpretation, both categorical and numerical forms of uncertainty should be communicated, when possible.

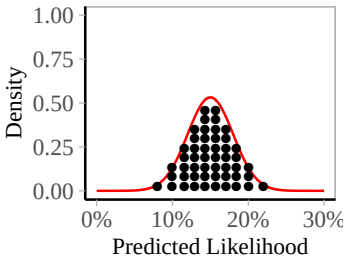
Though numbers and numerical ranges are more precise than categorical scales in communicating uncertainty, as discussed earlier, they are harder to understand for people with low numeracy and can induce ratio biases. However, a few techniques can be used to remediate these problems. First, to overcome the adverse effect of denominator neglect, it is important to present ratios with the same denominator so that they can be compared with just the numerator [166]. Denominators that are powers of 10 are preferred since they are easier to compute. There is so far no conclusive findings on whether frequencies are easier to understand than ratios or percentages, but people do seem to perceive risk probabilities represented in the frequency format as showing higher risk than those represented in the percentage format (c.f. [155]). Therefore, it is helpful to use a consistent format to represent probabilities, and if the audience underestimates risk levels, the frequency format may be preferred.

Uncertainty estimates can also be represented with graphics, which have several advantages over verbal communications, such as attracting and holding the audience's attention, revealing trends or patterns in the data, and evoking mental mathematical operations [110]. Commonly used visualizations include pie charts, bar charts, and more recently, icon arrays (Figure 3a). Pie charts are particularly useful for conveying proportions since all possible outcomes are depicted explicitly. However, it is more difficult to make accurate comparisons with pie charts than with bar charts because pie charts use areas to represent probabilities. Icon arrays vividly depict part-to-whole relationship, and because they shows the denominator explicitly, they can be used to overcome ratio biases.

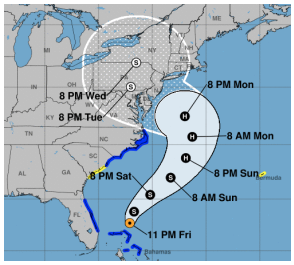
So far, what we have discussed pertains mostly to conveying uncertainty of a binary event, which takes the form of a single number (probability), whereas the uncertainty of a continuous variable or model prediction takes the form of a distribution. This latter type of uncertainty estimate can be communicated either as a series of summary statistics about the distribution, or directly as the full distribution. Commonly reported summary statistics include mean, median, confidence intervals, standard deviation, and quartiles [129]. These statistics are often depicted graphically as error bars and boxplots for univariate data, and two dimensional error bars and bagplots [158] for bivariate data. We describe these summary statistics and plots in detail in Appendix B. Error bars only have a few graphical elements and are hence relatively easy to interpret. However, since they have represented a range of different statistics in the past, they are ambiguous if presented without explicit labeling [185]. Error bars may also overly emphasize the range within the bar [37]. Boxplots and bagplots are less popular in the mass media, and generally require some training to understand.

When presenting uncertainty about a single model prediction, it might be better to show the entire posterior predictive distribution,

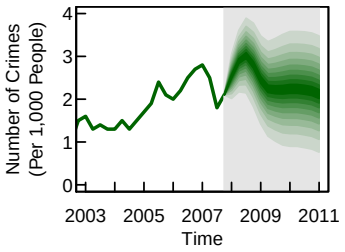
(a) An example icon array chart. This can be used to represent the chance of a patient having breast cancer.



(b) Quantile dot plot, which can be used to show the uncertainty around the predicted likelihood of a patient having breast cancer.



(c) Cone of uncertainty, showing the uncertainty around the predicted path of the center of a hurricane. Taken from [135].



(d) Fanchart, which can be used to show how the predicted crime rate of a city evolves over time.

Figure 3: Examples of uncertainty visualizations.

which can avoid over-emphasis of the within-bar range and allow more granular visual inferences. Popular visualizations of distributions are histograms, density plots, and violin plots ([72] shows multiple density plots side-by-side), but they seem to be hard for an uninitiated audience to grasp. They are often mistaken as bivariate case-value plots in which the lines or bars denote values instead of frequencies [21]. More recently, Kay et al. [89] developed quantile dot plots to convey distributions (see Figure 3b for an example). These plots use stacked dots, where each dot represents a group of cases, to approximate the data frequency at particular values. This method translates the abstract concept of probability distribution into a set of discrete outcomes, which are more familiar concepts to people who have not been trained in statistics. Kay et al. [89]'s study showed that people could more accurately derive probability estimates from quantile dot plots than from density plots.

One very different approach to conveying uncertainty is to individually show random draws from the probability distribution as a series of animation frames called hypothetical outcome plots (HOP) [80]. Similar to the quantile dot plots, HOPs accommodate the *frequency* view of uncertainty very well. In addition, showing events individually does not add any new visual encodings (such as the length of the bar or height of the stacked dots) and thus requires no additional learning from the viewers. Hullman et al. [80] showed that this visualization enabled people to make more accurate comparisons of two or three random variables than error bars and violin plots, presumably because statistical inference based on multiple distribution plots require special strategies while HOP does not. The drawbacks of HOP are: (a) it takes more time to show a representative sample of the distribution, and (b) it may incur high cognitive load since viewers need to mentally count and integrate frames. Nevertheless, because this method is easy to understand for people with low numeracy, similarly animated visualizations are frequently used in the mass media, e.g. [10, 189].

The above methods are designed to communicate uncertainty around a single quantity, so they need to be extended for visualizing uncertainty around a range of predictions, such as those in time-series forecasting. The simplest form of such visualization is a quantile plot, which uses lines to connect predictions at equal quantiles of the uncertainty distribution across the output range.

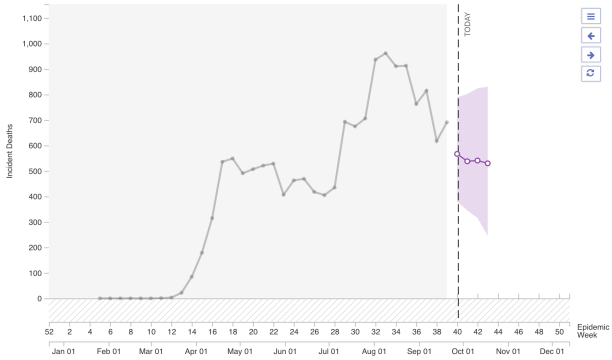


Figure 4: An example uncertainty visualization for projected COVID-19 mortality.

When used in time-series forecasting, such plots are called cone-of-uncertainty plots (see Figure 3c), in which the cone enlarges over time, indicating increasingly uncertain predictions. Gradient plots, or fan charts (see Figure 3d) in the context of time series forecasting, can be used to show more granular changes in uncertainty, but they require extra visual encoding that may not be easily understood by the viewer. In contrast, spaghetti plots simply represent each model's predictions as one line, while uncertainty can be inferred from the closeness of the lines. However, they might put too much emphasis on the lines themselves and de-emphasize the range of the model predictions. Lastly, HOP can also be used to show uncertainty estimates over a range of predictions by showing each model's predictions in an animation frame [185].

5.3 COVID Case Study

Uncertainty communication as a form of transparency is pivotal to garnering public trust during a pandemic. The COVID-19 global pandemic is an exemplar setting: disease forecasts, amongst other tools, have become critical for health communication efforts [83]. In this setting, forecasts are being disseminated to governments, organizations, and individuals for policy, resource allocation, and personal-risk judgments and behavior [48, 108, 146]. The United States Centers for Disease Control and Prevention (CDC) maintains Influenza Forecasting Centers of Excellence, which have recently turned to creating public-facing hubs for COVID-19 forecasts, with

the purpose of integrating infectious disease forecasting into public health decision-making [154]. For example, we may want to forecast the number of deaths due to COVID-19. The CDC repository contains several individual models for this task. The types of models include the classic susceptible-infected-recovered infectious disease model, statistical models fit to case data, regression models with various types of regularization, and others. Each model has its own assumptions and approaches to computing and illustrating underlying predictive uncertainty. Figure 4 shows how uncertainty from one model is visualized as a predictive band with a mean estimate highlighted. Given that such models are disseminated widely in the public via the Internet and other channels, and their result can directly affect personal behavior and disease transmission, this setting exemplifies an opportunity for user-centered design in uncertainty expression. In particular, assessing the forms of uncertainty visualization can be useful (e.g. 95% confidence intervals versus 50% confidence intervals, versus showing multiple different models, etc.). Indeed, the forms of uncertainty in COVID-19 forecasts could be used to inform a study to systematically assess user-specific uncertainty needs (e.g. a municipal public health department may require the most conservative estimate for adequate resource allocation, while an individual may be more interested in trends to plan their own activities).

6 UNCERTAINTY REQUIREMENTS

Familiarity with the findings discussed in Section 5 above will be helpful for teams building uncertainty into ML workflows. Yet, none of these findings should be treated as conclusive when it comes to predicting how usable a given expression of uncertainty will be for different types of users facing different kinds of constraints in real-world settings. Instead, findings from the literature should be treated as fertile ground for generating hypotheses in need of testing with real users, ideally engaged in the concrete tasks of their typical workflow, and ideally doing so in real-world settings.

It is important to recognize just how diverse individual users are, and how different their social contexts can be. In our cancer diagnostic scenario, the needs and constraints of a doctor making a time-pressed and high-stakes medical decision using an ML-powered tool will likely be very different from those of a patient attempting to understand their diagnosis, and different again from those of an ML engineer reviewing model output in search of strategies for model improvement. Furthermore, if we zoom in on any one of these user populations, we still typically observe a tremendous diversity in skills, experience, environmental constraints, and so on. For example, among doctors, there can be big differences in terms of statistical literacy, openness to trusting ML-powered tools, time available to consume and decide on model output, and so on. These variations have important implications for designing effective tools.

To design and build an effective expression of uncertainty, we need to begin with an understanding of who the tool will be used by, what goal that user has, and what needs and constraints the user has. Frequently we also need to understand the organizational and social context in which a user is embedded. For example, to understand how an organization calculates and processes risk, which can influence the design of human-in-the-loop processes, automation, where thresholds are set, and so on. This point is not a new one, and it is by no means unique to the field of ML. User-centered design (UCD), human-computer interaction (HCI), user experience (UX), human factors, and related fields have arisen as responses to this challenge across a wide range of product and tool design contexts [61, 62, 150].

UCD and HCI have a firm footing in many software development contexts, yet they remain relatively neglected in the field of ML. Nevertheless, a growing body of research is beginning to demonstrate the importance of user-centered design for work on ML tools (e.g., [81, 173]). For example, Yang et al. [188] draw on field research with healthcare decision-makers to understand why an ML-powered tool that performed well in laboratory tests was rejected by clinicians in real-world settings. They found that users saw little need for the tool, lacked trust in its output, and faced environmental barriers that made it difficult to use. Narayanan et al. [132] conduct a series of user tests for explainability to uncover which kinds of increases in explanation complexity have the greatest effect on the time it takes for users to achieve certain tasks. Doshi-Velez and Kim [44] propose a framework for evaluation of explainability that incorporates tests with users engaged in concrete and realistic tasks. From a practitioner's perspective, Lovejoy [113] describes the user-centered design lessons learned by the team building Google Clips, an AI-enabled camera designed to capture candid photographs of familiar people and animals. One of their key conclusions is that "[m]achine learning won't figure out what problems to solve. If you aren't aligned with a human need, you're just going to build a very powerful system to address a very small — or perhaps nonexistent — problem."

Research to uncover user goals, needs, and constraints can involve a wide spectrum of methods, including but not limited to in-depth interviews, contextual inquiry, diary studies, card sorting studies, user tests, user journey mapping, and jobs-to-be-done workshops with users [61, 62, 159]. It is helpful to divide user research into two buckets: 1) discovery research, which aims to understand what problem needs to be solved and for which type of user; and 2) evaluative research, which aims to understand how well our attempts to solve the given problem are succeeding with real users. Ideally, discovery research precedes any effort to build a solution, or at least occurs as early in the process as possible. Doing so helps the team focus on the right problem and right user type when considering possible solutions, and can help a team avoid costly investments that create little value for users. Which of the many methods a researcher uses in discovery and evaluative research will depend on many factors, including how easy it is to find relevant participants, how easy it is for the researcher to observe participants in the context of their day-to-day work, how expensive and time-consuming it is for the team to prototype potential solutions for the purposes of user testing, and so on. The key take-away is that teams building uncertainty into ML workflows should do user research to understand what problem needs solving and for what type of user.

7 CONCLUSION

Throughout this paper, we have argued that uncertainty is a form of transparency and is pertinent to the FAccT community. We surveyed the machine learning, visualization/HCI, decision-making

and fairness literature. We reviewed how to quantify uncertainty and leverage it in three use cases: (1) for developers reducing the unfairness of models, (2) for experts making decisions, and (3) for stakeholders placing their trust in ML models. We then described the methods for and pitfalls of communicating uncertainty, concluding with a discussion on how to collect requirements for leveraging uncertainty in practice. In summary, well-calibrated uncertainty estimates improve ML model transparency. In addition to calibration, it is important that these estimates are applied coherently and communicated clearly to various stakeholders considering the use case at hand. Future work could study the interplay between FAccT topics and uncertainty. For example, one could explore how communicating uncertainty to a stakeholder affects their perception of a model's fairness, or one could study how to best measure the calibration of uncertainty in regression settings. We hope this work inspires others to study uncertainty as transparency and to be mindful of uncertainty's effects on models in deployment.

8 ACKNOWLEDGMENTS

The authors would like to thank the following individuals for their advice, contributions, and/or support: James Allingham (University of Cambridge), McKane Andrus (Partnership on AI), Hudson Hongo (Partnership on AI), Terah Lyons (Partnership on AI), Elena Spitzer (Google), Kush Varshney (IBM), and Carroll Wainwright (Partnership on AI).

UB acknowledges support from DeepMind and the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence (CFI), and from the Partnership on AI. JA acknowledges support from Microsoft Research. AW acknowledges support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via CFI.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).

[2] Nilesh A Ahuja, Ibrahima Ndiour, Trushant Kalyanpur, and Omesh Tickoo. 2019. Probabilistic modeling of deep features for out-of-distribution and adversarial detection. *arXiv preprint arXiv:1909.11786* (2019).

[3] Ahmed M. Alaa and Mihaela van der Schaar. 2020. Discriminative Jackknife: Quantifying Uncertainty in Deep Learning via Higher-Order Influence Functions. arXiv:cs.LG/2007.13481

[4] Hadis Anahideh and Abolfazl Asudeh. 2020. Fair Active Learning. *arXiv preprint arXiv:2001.01796* (2020).

[5] Javier Antorán, James Urquhart Allingham, and José Miguel Hernández-Lobato. 2020. Depth Uncertainty in Neural Networks. *arXiv preprint arXiv:2006.08437* (2020).

[6] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, A Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[7] Syed Z Arshad, Jianlong Zhou, Constant Bridon, Fang Chen, and Yang Wang. 2015. Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. 352–360.

[8] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. 2019. Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning. In *International Conference on Learning Representations*.

[9] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. 2020. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*. 1770–1780.

[10] Emily Badger, Claire Cain Miller, Adam Pearce, and Kevin Quealy. The New York Times. Income Mobility Charts for Girls, Asian-Americans and Other Groups. Or Make Your Own. (The New York Times).

[11] Howard Balshem, Mark Helfand, Holger J. Schünemann, Andrew D. Oxman, Regina Kunz, Jan Brozek, Gunn E. Vist, Yngve Falck-Ytter, Joerg Meerpohl, Susan Norris, and Gordon H. Guyatt. [n. d.]. GRADE Guidelines: 3. Rating the Quality of Evidence. 64, 4 ([n. d.]), 401–406. https://doi.org/10/d49b4h

[12] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[13] Peter L Bartlett and Marten H Wegkamp. 2008. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research* 9, Aug (2008), 1823–1840.

[14] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[15] Umang Bhatt, McKane Andrus, Adrian Weller, and Alice Xiang. 2020. Machine Learning Explainability for External Stakeholders. *arXiv preprint arXiv:2007.05408* (2020).

[16] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.

[17] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[18] J Martin Bland and Douglas G Altman. 1998. Bayesians and frequentists. *BMJ* 317, 7166 (1998), 1151–1160. https://doi.org/10.1136/bmj.317.7166.1151 arXiv:https://www.bmj.com/content/317/7166/1151.1.full.pdf

[19] Avrim Blum and Kevin Stangl. 2019. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094* (2019).

[20] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight Uncertainty in Neural Network. In *International Conference on Machine Learning*. 1613–1622.

[21] Lonneke Boels, Arthur Bakker, Wim Van Dooren, and Paul Drijvers. [n. d.]. Conceptual Difficulties When Interpreting Histograms: A Review. 28 ([n. d.]), 100291. https://doi.org/10/ghbw6s

[22] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

[23] David V. Budescu, Han-Hui Por, and Stephen B. Broomell. [n. d.]. Effective Communication of Uncertainty in the IPCC Reports. 113, 2 ([n. d.]), 181–200. https://doi.org/10/c93bb5

[24] David V. Budescu and Thomas S. Wallsten. [n. d.]. Consistency in Interpretation of Probabilistic Phrases. 36, 3 ([n. d.]), 391–405. https://doi.org/10/b9qss9

[25] Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, Linda Zhao, et al. 2019. Models as approximations ii: A model-free theory of parametric regression. *Statist. Sci.* 34, 4 (2019), 545–565.

[26] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[27] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[28] Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. 2019. From soft classifiers to hard decisions: How fair can we be?. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 309–318.

[29] Shelly Chaiken. 1999. The heuristic—systematic. *Dual-process theories social psychology* 73 (1999).

[30] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *Advances in Neural Information Processing Systems*. 3539–3550.

[31] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*. 339–348.

[32] Tianqi Chen, Emily Fox, and Carlos Guestrin. 2014. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*. 1683–1691.

[33] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[34] Dominic A. Clark. [n. d.]. Verbal Uncertainty Expressions: A Critical Review of Two Decades of Research. 9, 3 ([n. d.]), 203–235. https://doi.org/10/ck4kmb

[35] Marvin S Cohen, Raja Parasuraman, and Jared T Freeman. 1998. Trust in decision aids: A model and its training implications. In *in Proc. Command and Control Research and Technology Symp.* Citeseer.

[36] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[37] Michael Correll and Michael Gleicher. [n. d.]. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. 20, 12 ([n. d.]), 2142–2151. https://doi.org/10/23c

[38] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 7403 (2012), 346–352.

[39] Stefan Depeweg. 2019. *Modeling Epistemic and Aleatoric Uncertainty with Bayesian Neural Networks and Latent Variables*. Ph.D. Dissertation. Technical University of Munich.

[40] Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural safety* 31, 2 (2009), 105–112.

[41] Terrance DeVries and Graham W Taylor. 2018. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865* (2018).

[42] Andreia Dionisio, Rui Menezes, and Diana A Mendes. 2007. Entropy and uncertainty analysis in financial markets. *arXiv preprint arXiv:0709.0668* (2007).

[43] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*. 2791–2801.

[44] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[45] Kevin Dowd. 2013. *Backtesting Market Risk Models*. John Wiley & Sons, Ltd, Chapter 15, 321–349. https://doi.org/10.1002/9781118673485.ch15 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118673485.ch15

[46] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[47] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[48] Sibel Eker. 2020. Validity and usefulness of COVID-19 models. *Humanities and Social Sciences Communications* 7, 1 (2020), 1–5.

[49] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*. 160–171.

[50] Kawin Ethayarajh. 2020. Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds. *arXiv preprint arXiv:2004.12332* (2020).

[51] Sebastian Farquhar, Michael Osborne, and Yarin Gal. 2020. Radial Bayesian Neural Networks: Beyond Discrete Support in Large-Scale Bayesian Deep Learning. *Proceedings of the 23rtd International Conference on Artificial Intelligence and Statistics* (2020).

[52] Baruch Fischhoff and Alex L Davis. 2014. Communicating scientific uncertainty. *Proceedings of the National Academy of Sciences* 111, Supplement 4 (2014), 13664–13671.

[53] Riccardo Fogliato, Max G'Sell, and Alexandra Chouldechova. 2020. Fairness Evaluation in Presence of Biased Noisy Labels. *arXiv preprint arXiv:2003.13808* (2020).

[54] Andrew Y. K. Foong, David R. Burt, Yingzhen Li, and Richard E. Turner. 2019. On the Expressiveness of Approximate Inference in Bayesian Neural Networks. *arXiv e-prints*, Article arXiv:1909.00719 (Sept. 2019), arXiv:1909.00719 pages. arXiv:stat.ML/1909.00719

[55] Linton G Freeman. 1965. *Elementary applied statistics*. John Wiley and Sons. 40–43 pages.

[56] Yarin Gal. 2016. Uncertainty in deep learning. *University of Cambridge* 1, 3 (2016).

[57] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. 1050–1059.

[58] Mirta Galesic and Rocio Garcia-Retamero. [n. d.]. Statistical Numeracy for Health: A Cross-Cultural Comparison With Probabilistic National Samples. 170, 5 ([n. d.]), 462–468. https://doi.org/10/fmj7q3

[59] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[60] Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* 102, 477 (2007), 359–378.

[61] Elizabeth Goodman. 2009. Three environmental discourses in human-computer interaction. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. 2535–2544.

[62] Elizabeth Goodman, Mike Kuniavsky, and Andrea Moed. 2012. *Observing the user experience: A practitioner's guide to user research*. Elsevier.

[63] Alex Graves. 2011. Practical variational inference for neural networks. In *Advances in neural information processing systems*. 2348–2356.

[64] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*. 1321–1330.

[65] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness. *arXiv preprint arXiv:1806.11212* (2018).

[66] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[67] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010* (2018).

[68] Chip Heath and Amos Tversky. 1991. Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of risk and uncertainty* 4, 1 (1991), 5–28.

[69] José Miguel Hernández-Lobato and Ryan Adams. 2015. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*. 1861–1869.

[70] José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. 2014. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*. 918–926.

[71] Geoffrey E. Hinton and Drew van Camp. 1993. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory (COLT '93)*. ACM, New York, NY, USA, 5–13. https://doi.org/10.1145/168304.168306

[72] Jerry L. Hintze and Ray D. Nelson. [n. d.]. Violin Plots: A Box Plot-Density Trace Synergism. 52, 2 ([n. d.]), 181–184. https://doi.org/10/gf5hpq

[73] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Flat minima. *Neural Computation* 9, 1 (1997), 1–42.

[74] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[75] Stephen C Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety* 54, 2-3 (1996), 217–223.

[76] Eric Horvitz, David Heckerman, Bharat Nathwani, and Lawrence Fagan. 1986. The Use of a Heuristic Problem-Solving Hierarchy to Facilitate the Explanation of Hypothesis-Directed Reasoning. , 27-31 pages. https://www.microsoft.com/en-us/research/publication/use-heuristic-problem-solving-hierarchy-facilitate-explanation-hypothesis-directed-reasoning/

[77] Eric J Horvitz, David E Heckerman, Keung-Chi Ng, and Bharat N Nathwani. 1989. Heuristic abstraction in the decision-theoretic Pathfinder system. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 178.

[78] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745* (2011).

[79] Carl Iver Hovland, Irving Lester Janis, and Harold H Kelley. 1953. Communication and persuasion. (1953).

[80] Jessica Hullman, Paul Resnick, and Eytan Adar. [n. d.]. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. 10, 11 ([n. d.]), e0142444. https://doi.org/10/f3tvsd

[81] Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric PS Baumer. 2019. Where is the human? Bridging the gap between AI and HCI. In *Extended abstracts of the 2019 chi conference on human factors in computing systems*. 1–9.

[82] David Janz, Jiri Hron, Przemysł aw Mazur, Katja Hofmann, José Miguel Hernández-Lobato, and Sebastian Tschiatschek. 2019. Successor Uncertainties: Exploration and Uncertainty in Temporal Difference Learning. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 4507–4516. http://papers.nips.cc/paper/8700-successor-uncertainties-exploration-and-uncertainty-in-temporal-difference-learning.pdf

[83] Nicholas P Jewell, Joseph A Lewnard, and Britta L Jewell. 2020. Predictive mathematical models of the COVID-19 pandemic: Underlying principles and value of projections. *Jama* 323, 19 (2020), 1893–1894.

[84] Heinrich Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*. 702–712.

[85] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[86] Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 99–127.

[87] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2019. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285* (2019).

[88] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929.

[89] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. [n. d.]. When (Ish) Is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human*

Factors in Computing Systems (2016-05-07) *(CHI '16)*. Association for Computing Machinery, 5092–5103. https://doi.org/10/b2pj

[90] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680* (2015).

[91] Alex Kendall and Roberto Cipolla. 2016. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*. IEEE, 4762–4769.

[92] Durk P Kingma, Tim Salimans, and Max Welling. 2015. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*. 2575–2583.

[93] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*. 7026–7037.

[94] Jon Kleinberg. 2018. Inherent trade-offs in algorithmic fairness. In *ACM SIGMETRICS Performance Evaluation Review*, Vol. 46. ACM, 40–40.

[95] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[96] Frank Hyneman Knight. 1921. *Risk, uncertainty and profit*. Vol. 31. Houghton Mifflin.

[97] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70 (ICML 2017)*. Journal of Machine Learning Research, 1885–1894.

[98] Lingkai Kong, Jimeng Sun, and Chao Zhang. 2020. SDE-Net: Equipping Deep Neural Networks with Uncertainty Estimates. In *International Conference on Machine Learning*.

[99] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*. Springer, 13–30.

[100] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*. 12316–12326.

[101] Paul H. Kupiec. 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives* 3, 2 (1995), 73–84. https://doi.org/10.3905/jod.1995.407942 arXiv:https://jod.pm-research.com/content/3/2/73.full.pdf

[102] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*. 6402–6413.

[103] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. 2019. Noise-tolerant fair classification. In *Advances in Neural Information Processing Systems*. 294–306.

[104] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. 2019. Well-calibrated Model Uncertainty with Temperature Scaling for Dropout Variational Inference. *arXiv preprint arXiv:1909.13550* (2019).

[105] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[106] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[107] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*. 7167–7177.

[108] Ruoran Li, Caitlin Rivers, Qi Tan, Megan B Murray, Eric Toner, and Marc Lipsitch. 2020. Estimated Demand for US Hospital Inpatient and Intensive Care Unit Beds for Patients With COVID-19 Based on Comparisons With Wuhan and Guangzhou, China. *JAMA network open* 3, 5 (2020), e208297–e208297.

[109] Sarah Lichtenstein and J. Robert Newman. [n. d.]. Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities. 9, 10 ([n. d.]), 563–564. https://doi.org/10/ghbhk9

[110] I. M. Lipkus and J. G. Hollands. [n. d.]. The Visual Communication of Risk. 25 ([n. d.]), 149–163. https://doi.org/10/gd589v arXiv:10854471

[111] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. 2020. Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness. *arXiv preprint arXiv:2006.10108* (2020).

[112] Daniel Hernández Lobato. 2009. Prediction based on averages over automatically induced learners ensemble methods and Bayesian techniques.

[113] Josh Lovejoy. 2018. The UX of AI: Using Google Clips to understand how a human-centered design process elevates artificial intelligence. In *2018 AAAI Spring Symposium Series*.

[114] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[115] Chao Ma, Yingzhen Li, and Jose Miguel Hernandez-Lobato. 2019. Variational Implicit Processes *(Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 4222–4233. http://proceedings.mlr.press/v97/ma19b.html

[116] David JC MacKay. 1992. A practical Bayesian framework for backpropagation networks. *Neural computation* 4, 3 (1992), 448–472.

[117] David JC MacKay and David JC Mac Kay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

[118] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*. 13153–13164.

[119] David Madras, James Atwood, and Alexander D'Amour. 2020. Detecting Extrapolation with Local Ensembles. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BJl6bANtwH

[120] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*. 6147–6157.

[121] Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*. 7047–7058.

[122] Alexander Graeme de Garis Matthews. 2017. *Scalable Gaussian process inference using variational methods*. Ph.D. Dissertation. University of Cambridge.

[123] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[124] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and A. Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. *ArXiv* abs/1908.09635 (2019).

[125] Miriam J Metzger and Andrew J Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics* 59 (2013), 210–220.

[126] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. 2018. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1–7.

[127] Suzanne M Miller. 1987. Monitoring and blunting: validation of a questionnaire to assess styles of information seeking under threat. *Journal of personality and social psychology* 52, 2 (1987), 345.

[128] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.

[129] Alexander McFarlane Mood, Franklin A. Graybill, and . Boes, Duane C. 1974. *Introduction to the theory of statistics / Alexander M. Mood, Franklin A. Graybill, Duane C. Boes* (3rd ed. ed.). McGraw-Hill New York. xvi, 564 p. : pages.

[130] Jishnu Mukhoti and Yarin Gal. 2018. Evaluating Bayesian Deep Learning Methods for Semantic Segmentation. *arXiv preprint arXiv:1811.12709* (2018).

[131] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[132] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).

[133] Radford M Neal. 1995. Bayesian Learning for Neural Networks. *PhD thesis, University of Toronto* (1995).

[134] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. 2018. Variational Continual Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BkQqq0gRb

[135] NHC. [n. d.]. Definition of the NHC Track Forecast Cone. https://www.nhc.noaa.gov/aboutcone.shtml

[136] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring Calibration in Deep Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 38–41.

[137] Alejandro Noriega-Campero, Michiel A Bakker, Bernardo Garcia-Bulle, and Alex'Sandy' Pentland. 2019. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 77–83.

[138] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[139] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[140] Onora O'Neill. 2018. Linking trust to trustworthiness. *International Journal of Philosophical Studies* 26, 2 (2018), 293–300.

[141] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can You

Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *arXiv preprint arXiv:1906.02530* (2019).

[142] Raja Parasuraman and Christopher A Miller. 2004. Trust and etiquette in high-criticality automated systems. *Commun. ACM* 47, 4 (2004), 51–55.

[143] AD Pearman. 1985. Uncertainty in planning: characterisation, evaluation, and feedback. *Environment and Planning B: Planning and Design* 12, 3 (1985), 313–320.

[144] Jolynn Pek and Trisha Van Zandt. 2020. Frequentist and Bayesian approaches to data analysis: Evaluation and estimation. *Psychology Learning & Teaching* 19, 1 (2020), 21–35. https://doi.org/10.1177/1475725719874542 arXiv:https://doi.org/10.1177/1475725719874542

[145] Ellen Peters, Judith Hibbard, Paul Slovic, and Nathan Dieckmann. [n. d.]. Numeracy Skill And The Communication, Comprehension, And Use Of Risk-Benefit Information. 26, 3 ([n. d.]), 741–748. https://doi.org/10/c77p38

[146] Fotios Petropoulos and Spyros Makridakis. 2020. Forecasting the novel coronavirus COVID-19. *PloS one* 15, 3 (2020), e0231236.

[147] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.

[148] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[149] Mary C. Politi, Paul K. J. Han, and Nananda F. Col. [n. d.]. Communicating the Uncertainty of Harms and Benefits of Medical Interventions. 27, 5 ([n. d.]), 681–695. https://doi.org/10/c9b8g4

[150] Jennifer Preece, Helen Sharp, and Yvonne Rogers. 2015. *Interaction design: beyond human-computer interaction*. John Wiley & Sons.

[151] Inioluwa Deborah Raji and Jingying Yang. 2019. ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles. *arXiv preprint arXiv:1912.06166* (2019).

[152] Tim Rakow. 2010. Risk, uncertainty and prophet: The psychological insights of Frank H. Knight. *Judgment and Decision Making* 5, 6 (2010), 458.

[153] Carl Edward Rasmussen and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[154] Evan L Ray, Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana, Xinyue Xiong, et al. 2020. Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the US. *medRxiv* (2020).

[155] Valerie F. Reyna and Charles J. Brainerd. [n. d.]. Numeracy, Ratio Bias, and Denominator Neglect in Judgments of Risk and Probability. 18, 1 ([n. d.]), 89–107. https://doi.org/10/bdqnh5

[156] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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.

[157] Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J Candès. 2019. With Malice Towards None: Assessing Uncertainty via Equalized Coverage. *arXiv preprint arXiv:1908.05428* (2019).

[158] Peter J. Rousseeuw, Ida Ruts, and John W. Tukey. [n. d.]. The Bagplot: A Bivariate Boxplot. 53, 4 ([n. d.]), 382–387. https://doi.org/10/gg6cp5

[159] Jeff Rubin and Dana Chisnell. 2008. How to plan, design, and conduct effective tests. (2008).

[160] Peter Schulam and Suchi Saria. 2019. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*. 1022–1031.

[161] Clayton Scott, Gilles Blanchard, and Gregory Handy. 2013. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference On Learning Theory*. 489–511.

[162] Claude E Shannon. 1948. A mathematical theory of communication. *Bell system technical journal* 27, 3 (1948), 379–423.

[163] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90, 2 (2000), 227–244.

[164] Richard M Sorrentino, D Ramona Bobocel, Maria Z Gitta, James M Olson, and Erin C Hewitt. 1988. Uncertainty orientation and persuasion: Individual differences in the effects of personal relevance on social judgments. *Journal of Personality and social Psychology* 55, 3 (1988), 357.

[165] David Spiegelhalter. [n. d.]. Risk and Uncertainty Communication. 4, 1 ([n. d.]), 31–60. https://doi.org/10.1146/annurev-statistics-010814-020148

[166] D. Spiegelhalter, M. Pearson, and I. Short. [n. d.]. Visualizing Uncertainty About the Future. 333, 6048 ([n. d.]), 1393–1400. https://doi.org/10/cd4

[167] Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. 2019. FUNCTIONAL VARIATIONAL BAYESIAN NEURAL NETWORKS. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rkxacs0qY7

[168] S Shyam Sundar. 2008. *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative.

[169] S Shyam Sundar and Jinyoung Kim. 2019. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–9.

[170] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70 (ICML 2017)*. Journal of Machine Learning Research, 3319–3328.

[171] Harini Suresh and John V. Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv:cs.LG/1901.10002

[172] Mattias Teye, Hossein Azizpour, and Kevin Smith. 2018. Bayesian Uncertainty Estimation for Batch Normalized Deep Networks. In *International Conference on Machine Learning*. 4907–4916.

[173] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53.

[174] Amos Tversky and Daniel Kahneman. [n. d.]. Advances in Prospect Theory: Cumulative Representation of Uncertainty. 5, 4 ([n. d.]), 297–323. https://doi.org/10/cb57hk

[175] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[176] Amos Tversky and Eldar Shafir. 1992. The disjunction effect in choice under uncertainty. *Psychological science* 3, 5 (1992), 305–310.

[177] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. In *International Conference on Machine Learning*.

[178] Anne Marthe Van Der Bles, Sander van der Linden, Alexandra LJ Freeman, and David J Spiegelhalter. 2020. The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7672–7683.

[179] Anne Marthe van der Bles, Sander van der Linden, Alexandra L. J. Freeman, James Mitchell, Ana B. Galvao, Lisa Zaval, and David J. Spiegelhalter. [n. d.]. Communicating Uncertainty about Facts, Numbers and Science. 6, 5 ([n. d.]), 181870. https://doi.org/10/gf2g9j

[180] Darshali A Vyas, Leo G Eisenstein, and David S Jones. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms.

[181] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael I Jordan. 2020. Robust Optimization for Fairness with Noisy Protected Groups. *arXiv preprint arXiv:2002.09343* (2020).

[182] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio du Pin Calmon. 2019. Optimized score transformation for fair classification. *arXiv preprint arXiv:1906.00066* (2019).

[183] Adrian Weller. 2019. Transparency: motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 23–40.

[184] Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 681–688.

[185] Claus O. Wilke. [n. d.]. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures* (1st edition ed.). O'Reilly Media.

[186] Andrew Gordon Wilson and Pavel Izmailov. 2020. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791* (2020).

[187] Min-ge Xie and Kesar Singh. 2013. Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. *International Statistical Review* 81, 1 (2013), 3–39. https://doi.org/10.1111/insr.12000 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12000

[188] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4477–4488.

[189] Nathan Yau. [n. d.]. Years You Have Left to Live, Probably. https://flowingdata.com/2015/09/23/years-you-have-left-to-live-probably/

[190] Nanyang Ye and Zhanxing Zhu. 2018. Bayesian Adversarial Learning. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 6892–6901. http://papers.nips.cc/paper/7921-bayesian-adversarial-learning.pdf

[191] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[192] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[193] Hang Zhang and Laurence T. Maloney. [n. d.]. Ubiquitous Log Odds: A Common Representation of Probability and Frequency Distortion in Perception, Action, and Cognition. 6 ([n. d.]). https://doi.org/10/fxxssh

[194] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. 2020. Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning. *International Conference on Learning Representations* (2020).
[195] Yunfeng Zhang, Rachel KE Bellamy, and Wendy A. Kellogg. [n. d.]. Designing Information for Remediating Cognitive Biases in Decision-Making. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015). ACM, 2211–2220. https://doi.org/10.1145/2702123.2702239
[196] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.
[197] Brian J. Zikmund-Fisher, Dylan M. Smith, Peter A. Ubel, and Angela Fagerlin. [n. d.]. Validation of the Subjective Numeracy Scale: Effects of Low Numeracy on Comprehension of Risk Communications and Utility Elicitations. 27, 5 ([n. d.]), 663–671. https://doi.org/10/cf2642

A UNCERTAINTY METRICS

In this appendix we detail different metrics with which the uncertainty in a predictive distribution can be summarized. We distinguish between metrics that communicate aleatoric uncertainty, those that communicate epistemic uncertainty, and those that inform us about the combination of both. We also distinguish between the classification and regression setting. Recall that, as discussed in Section 5, predictive probabilities more intuitively communicate a model's predictions and uncertainty to stakeholders than a continuous predictive distribution. Therefore, summary statistics for uncertainty might play a larger role when building transparent regression systems.

A.1 Classification setting

Notation: Let $\mathcal{D} = \{\mathrm{x_i}, \mathrm{y_i}\}_{i=1}^{N}$ represent a dataset consisting of N samples, where for each example i, $\mathrm{x_i} \in \mathcal{X}$ is the input and $\mathrm{y_i} \in \mathcal{Y} = \{c_k\}_{k=1}^{K}$ is the ground-truth class label. Let $p_i\,(\mathrm{y}|\mathrm{x}_i, \mathrm{w})$ be the output from the parametric classifier $f_{\mathrm{w}}\,(\mathrm{y}|\mathrm{x_i})$ with model parameters w. In probabilistic models, the output predictive distribution can be approximated from T stochastic forward passes (Monte Carlo samples), as described in Section 3: $p_i\,(\mathrm{y}|\mathrm{x_i}) = \frac{1}{T}\sum_{t=1}^{T} p_i^t\,(\mathrm{y}|\mathrm{x_i}, \mathrm{w}_t)$, where $\mathrm{w}_t \sim p(\mathrm{w}|\mathrm{D})$.

Predictive entropy: The entropy [162] of the predictive distribution is given by Equation 3. Predictive entropy represents the overall predictive uncertainty of the model, a combination of aleatoric and epistemic uncertainties [130].

$$\mathbb{H}(\mathrm{y}|\mathrm{x}, \mathrm{D}) := -\sum_{k=1}^{K}\left(\frac{1}{T}\sum_{t=1}^{T} p\,(\mathrm{y}{=}c_k|\mathrm{x}, \mathrm{w}_t)\right)\log\left(\frac{1}{T}\sum_{t=1}^{T} p\,(\mathrm{y}{=}c_k|\mathrm{x}, \mathrm{w}_t)\right) \tag{3}$$

In the case of point-estimate deterministic models, predictive entropy is given by Equation 4 and captures only the aleatoric uncertainty.

$$\mathbb{H}(\mathrm{y}|\mathrm{x}, \mathrm{D}) := -\sum_{k=1}^{K} p\,(\mathrm{y} = c_k|\mathrm{x}, \mathrm{w})\log\left(p\,(\mathrm{y} = c_k|\mathrm{x}, \mathrm{w})\right) \tag{4}$$

The predictive entropy always takes positive values between 0 and $\log K$. Its maximum is attained when the probability of all classes is $\frac{1}{K}$. Predictive entropy can be additively decomposed into aleatoric and epistemic components:

Expected entropy: The expectation of entropy obtained from multiple stochastic forward passes captures the aleatoric uncertainty.

$$\mathbb{E}_{p(\mathrm{w}|D)}\left[\mathbb{H}(\mathrm{y}|\mathrm{x}, \mathrm{w})\right] := \frac{1}{T}\sum_{t=1}^{T}\left(-\sum_{k=1}^{K} p\,(\mathrm{y}{=}c_k|\mathrm{x}, \mathrm{w}_t)\log p\,(\mathrm{y}{=}c_k|\mathrm{x}, \mathrm{w}_t)\right) \tag{5}$$

Mutual information: The mutual information [162] between the posterior of model parameters and the targets captures epistemic uncertainty [56, 78]. It is given by Equation 6.

$$MI(\mathrm{y}, \mathrm{w}|\mathrm{x}, D) := \mathbb{H}(\mathrm{y}|\mathrm{x}, \mathrm{D}) - \mathbb{E}_{p(\mathrm{w}|D)}\left[\mathbb{H}(\mathrm{y}|\mathrm{x}, \mathrm{w})\right] \tag{6}$$

The predictive entropy can be recovered as the addition of the expected entropy and mutual information:

$$\mathbb{H}(\mathrm{y}|\mathrm{x}, \mathrm{D}) = \mathbb{E}_{p(\mathrm{w}|D)}\left[\mathbb{H}(\mathrm{y}|\mathrm{x}, \mathrm{w})\right] + MI(\mathrm{y}, \mathrm{w}|\mathrm{x}, D)$$

Variation ratio: Variation ratio [55] captures the disagreement of a model's predictions across multiple stochastic forward passes given by Equation 7.

$$VR := 1 - \frac{f}{T} \tag{7}$$

where, f represents the number of times the output class was predicted from T stochastic forward passes.

A.2 Regression setting

We assume the same notation as in the classification setting with the distinction that our targets $\mathrm{y_i} \in \mathcal{Y} = \mathcal{R}$ are continuous. For generality, we employ heteroscedastic noise models. Recall that this means our aleatoric uncertainty may be different in different regions of input space. We assume a Gaussian noise model with its mean and variance predicted by parametric models: $p(\mathrm{y}|\mathrm{x}, \mathrm{w}) = \mathcal{N}(\mathrm{y}; f_{\mathrm{w}}^{\mu}(\mathrm{x}), f_{\mathrm{w}}^{\sigma^2}(\mathrm{x}))$. Approximately marginalizing over w with T Monte Carlo samples induces a Gaussian mixture over outputs. Its mean is obtained as:

$$\mu \approx \frac{1}{T}\sum_{t=1}^{T} f_{\mathrm{w_t}}^{\mu}(\mathrm{x})$$

Aleatoric and Epistemic Variances: There is no closed-form expression for the entropy of the mixture of Gaussians (GMM). Instead, we use the variance of the GMM as an uncertainty metric. It also decomposes into aleatoric and epistemic components (σ_a^2, σ_e^2):

$$\sigma^2(\mathrm{y}|\mathrm{x}, \mathrm{D}) = \underbrace{\mathbb{E}_{p(\mathrm{w}|\mathrm{D})}\left[f_{\mathrm{w}}^{\sigma^2}(\mathrm{y}|\mathrm{x})\right]}_{\sigma_a^2} + \underbrace{\sigma^2_{p(\mathrm{w}|\mathrm{D})}\left[f_{\mathrm{w}}^{\mu}(\mathrm{y}|\mathrm{x})\right]}_{\sigma_e^2}.$$

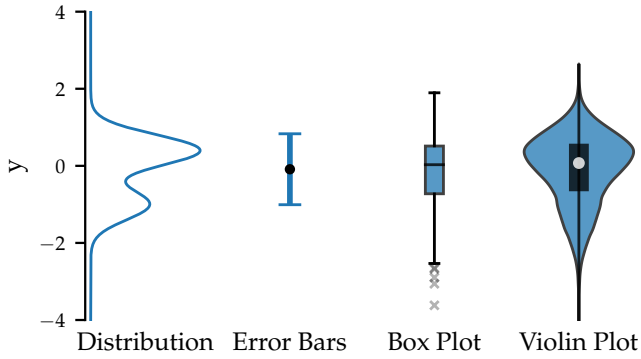


Figure 5: Left: A full predictive distribution over some value of interest y. Center left: A summary of the predictive distribution composed of its mean value and standard deviation error-bars. Center right: A box plot showing quartiles and 1.5× interquartile range whiskers. Right: A violin plot showing quartiles and sampled extrema.

These are also estimated with MC:

$$\sigma^2(\mathrm{y}|\mathrm{x},\mathrm{D}) \approx \underbrace{\frac{1}{T}\sum_{t=1}^{T} f_{\mathrm{w_t}}^{\mu}(\mathrm{y}|\mathrm{x})^2 - (\frac{1}{T}\sum_{t=1}^{T} f_{\mathrm{w_t}}^{\mu}(\mathrm{y}|\mathrm{x}))^2}_{\sigma_e^2} + \underbrace{\frac{1}{T}\sum_{t=1}^{T} f_{\mathrm{w_t}}^{\sigma^2}(\mathrm{y}|\mathrm{x})}_{\sigma_a^2}.$$

Here, σ_e^2 reflects model uncertainty - our lack of knowledge about w - while σ_a^2 tells us about the irreducible uncertainty or noise in our training data. Similarly to entropy, we can express uncertainty in regression as the addition of aleatoric and epistemic components.

We now briefly discuss other common summary statistics to describe continuous predictive distributions [129]. Note that these do not admit simple aleatoric-epistemic decompositions.

Percentiles: Percentiles tell us about the values below which there is a certain probability of our targets falling. For example, if the 20th percentile of our predictive distribution is 5, this means that values ≤ 5 take up 20% of the probability mass of our predictive distribution.

Confidence Intervals: A confidence interval $[a, b]$ communicates that, with a probability p, our quantity of interest will lie within the provided range $[a, b]$. Thus, the commonly used 95% confidence interval tells us that percentile 2.5 of our predictive distribution corresponds to a and percentile 97.5 corresponds to b.

Quantiles: Quantiles divide the predictive distribution into sections of equal probability mass. Quartiles, which divide the predictive distribution into four parts, are the most common use of quantiles. The first quartile corresponds to the 25th percentile, the second to the 50th (or median) and the third to percentile 75.

The summary statistics above are often depicted in error-bar plots, box-plots, and violin plots, as shown in Figure 5. Error bar plots provide information about the spread of the predictive distribution. They can reflect variance (or standard deviation), percentiles, confidence intervals, quantiles, etc. Box plots tell us about our distribution's shape by depicting its quartiles. Additionally, box plots often depict longer error bars, referred to as "whiskers," which tell us about the heaviness of our distribution's tails. Whiskers are most commonly chosen to be of length 1.5 × the interquartile range (Q1 - Q3) or extreme percentile values, e.g. 2-98. Samples that fall outside of the range depicted by whiskers are treated as outliers and plotted individually. Although less popular, violin plots have been gaining some traction for summarizing large groups of samples. Violin plots depict the estimated shape of the distribution of interest (usually by applying kernel density estimation to samples). They combine this with a box plot that provides information about quartiles. However, differently from regular box plots, violin plots' whiskers are often chosen to reflect the maximum and minimum values of the sampled population.

B CALIBRATION METRICS

This section describes existing metrics that reflect the calibration of predictive distributions for classification. There are no widely adopted calibration metrics for regression within the ML community. However, the use of calibration metrics for regression is common in other fields, such as econometrics. We discuss how these can be adapted to provide analogous information to popular ML classification calibration metrics. Calibration metrics should be computed on a validation set sampled independently from the data used to train the model being evaluated. In our cancer diagnosis scenario, this could mean collecting validation data from different hospitals than those used to collect the training data.

Test Log Likelihood (higher is better): This metric tells us about how probable it is that the validation targets were generated using the validation inputs and our model. It is a proper scoring rule [60] that depends on both the accuracy of predictions and their uncertainty. We can employ it in both classification and regression settings. Log-likelihood is also the most commonly used training objective for neural networks. The popular classification cross-entropy and regression mean squared error objectives represent maximum log-likelihood objectives under categorical and unit variance Gaussian noise models, respectively [17].

Brier Score [22] (lower is better): Proper scoring rule that measures the accuracy of predictive probabilities in classification tasks. It is computed as the mean squared distance between predicted class probabilities and one-hot class labels:

$$\mathrm{BS} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}(p(\mathrm{y}=c_k|\mathrm{x},\mathrm{w}) - \mathbb{1}[\mathrm{y}=c_k])^2$$

Unlike log-likelihood, Brier score is bounded from above. Erroneous predictions made with high confidence are penalised less by Brier score than by log-likelihood. This can avoid outliers or misclassified inputs from having a dominant effect on experimental results. On the other hand, it makes Brier score less sensitive.

Expected Calibration Error (ECE) [131] (lower is better): This metric is popularly used to evaluate the calibration of deep classification neural networks. ECE measures the difference between predictive confidence and empirical accuracy in classification. It is computed by dividing the [0,1] range into a set of bins $\{B_s\}_{s=1}^{S}$ and weighing the miscalibration in each bin by the number of points

that fall into it $|B_s|$:

$$\text{ECE} = \sum_{s=1}^{S} \frac{|B_s|}{N} \left|\text{acc}(B_s) - \text{conf}(B_s))\right|$$

Here,

$$\text{acc}(B_s) = \frac{1}{|B_s|} \sum_{\text{x} \in B_s} \mathbb{1}[\text{y} = \arg\max_{c_k} p(\text{y}|\text{x}, \text{w})] \quad \text{and}$$
$$\text{conf}(B_s) = \frac{1}{|B_s|} \sum_{\text{x} \in B_s} \max_{c_k} p(\text{y}|\text{x}, \text{w}).$$

ECE is not a proper scoring rule. A perfect ECE score can be obtained by predicting the marginal distribution of class labels $p(\text{y})$ for every input. A well-calibrated predictor with poor accuracy would obtain low log likelihood values (undesirable result) but also low ECE (desirable result). Although ECE works well for binary classification, the naive adaption to the multi-class setting results in a disproportionate amount of class predictions being assigned to low probability bins, biasing results. [136] and [100] propose alternatives that mitigate this issue.

Expected Uncertainty Calibration Error (UCE) [104] (lower is better): This metric measures the difference in expectation between a model's error and its uncertainty. The key difference from ECE is this metric quantifies model miscalibration with respect to predictive uncertainty (using a single uncertainty summary statistic, Appendix A.1). This differs from ECE, which quantifies the model miscalibration with respect to confidence (probability of predicted class).

$$\text{UCE} = \sum_{s=1}^{S} \frac{|B_s|}{N} \left|\text{err}\,(B_s) - \text{uncert}\,(B_s)\right| \tag{8}$$

Here,

$$\text{err}(B_s) = \frac{1}{|B_s|} \sum_{\text{x} \in B_s} \mathbb{1}[\text{y} \neq \arg\max_{c_k} p(\text{y}|\text{x}, \text{w})] \quad \text{and}$$
$$\text{uncert}\,(B_s) = \frac{1}{|B_s|} \sum_{\text{x} \in B_s} \tilde{\text{u}}$$

where $\tilde{\text{u}} \in [0, 1]$ is a normalized uncertainty summary statistic (defined in Appendix A.1).

Conditional Probabilities for Uncertainty Evaluation [130] (higher is better): Conditional probabilities *p(accurate | certain)* and *p(uncertain | inaccurate)* have been proposed in [130] to evaluate the quality of uncertainty estimates obtained from different probabilistic methods on semantic segmentation tasks, but can be used for any classification task.

p(accurate | certain) measures the probability that the model is accurate on its output given that it is confident on the same. *p(uncertain | inaccurate)* measures the probability that the model is uncertain about its output given that it has made inaccurate prediction. Based on these two conditional probabilities, *patch accuracy versus patch uncertainty* (PAvPU) metric is defined as below.

$$p(\text{accurate}|\text{certain}) = \frac{\text{n}_{AC}}{\text{n}_{AC} + \text{n}_{IC}}$$
$$p(\text{uncertain}|\text{inaccurate}) = \frac{\text{n}_{IU}}{\text{n}_{IC} + \text{n}_{IU}}$$
$$PAvPU = \frac{\text{n}_{AC} + \text{n}_{IU}}{\text{n}_{AC} + \text{n}_{IC} + \text{n}_{AU} + \text{n}_{IU}}$$

Here, n_{AC}, n_{AU}, n_{IC}, n_{IU} are the number of predictions that are accurate and certain (AC), accurate and uncertain (AU), inaccurate and certain (IC), inacurate and uncertain (IU) respectively.

Regression Calibration Metrics: We can extend ECE to regression settings, while avoiding the pathologies described by [136]. We seek to assess how well our model's predictive distribution describes the residuals obtained on the test set. It is not straightforward to define bins, like in standard ECE, because our predictive distribution might not have finite support. We apply the cumulative density function (CDF) of our predictive distribution to our test targets. If the predictive distribution describes the targets well, the transformed distribution should resemble a uniform with support $[0, 1]$. This procedure is common for backtesting market risk models [45].

Regression Calibration Error (RCE) (lower is better): To assess the global similarity between our targets' distribution and our predictive distribution, we separate the $[0, 1]$ interval into S equal-sized bins $\{B_s\}_{s=1}^{S}$. We compute calibration error in each bin as the difference between the proportion of points that have fallen within that bin and $\frac{1}{S}$:

$$\text{RCE} = \sum_{s=1}^{S} \frac{|B_s|}{N} \cdot \left|\frac{1}{S} - \frac{|B_s|}{N}\right|; \quad |B_s| = \sum_{n=1}^{N} \mathbb{1}[CDF_{p(y|\text{x}^{(n)})}(y^{(n)}) \in B_s]$$

Tail Calibration Error (TCE) (lower is better): In cases of model misspecification, e.g. our noise model is Gaussian but our residuals are multimodal, RCE might become large due to this mismatch, even though the moments of our predictive distribution might be generally correct. We can exclusively assess how well our model predicts extreme values with a "frequency of tail losses" approach [101]. Only considering calibration at the tails of the predictive distribution allows us to ignore shape mismatch between the predictive distribution and the true distribution over targets. Instead, we focus on our model's capacity to predict on which inputs it is likely to make large mistakes. We specify two bins $\{B_0, B_1\}$, one at each tail end of our predictive distribution, and compute:

$$\text{TCE} = \sum_{s=0}^{1} \frac{|B_s|}{|B_0| + |B_1|} \cdot \left|\frac{1}{\tau} - \frac{|B_s|}{N}\right|;$$
$$|B_0| = \sum_{n=1}^{N} \mathbb{1}[CDF_{p(y|\text{x}^{(n)})}(y^{(n)}) < \tau];$$
$$|B_1| = \sum_{n=1}^{N} \mathbb{1}[CDF_{p(y|\text{x}^{(n)})}(y^{(n)}) \geq (1 - \tau)]$$

We specify the tail range of our distribution by selecting τ. Note that this is slightly different from Kupiec [101], who uses a binomial test to assess whether a model's predictive distribution agrees with the distribution over targets in the tails. RCE and TCE are not

proper scoring rules. Additionally, they are only applicable to one-dimensional continuous target variables.

C UNCERTAINTY QUANTIFICATION METHODS

In this appendix, we describe additional approaches to uncertainty quantification and calibration which were omitted from Section 3.

C.1 Bayesian Methods

Various approximate inference methods have been proposed for Bayesian uncertainty quantification in parametric models, such as deep neural networks. We refer to the resulting models as Bayesian Neural Networks (BNN), Figure 6. Variational inference [20, 51, 63, 71] approximates a complex probability distribution $p(\mathrm{w}|\mathrm{D})$ with a simpler distribution $q_\theta(\mathrm{w})$, parameterized by variational parameters θ, by minimizing the Kullback-Leibler (KL) divergence between the two $KL(q_\theta(\mathrm{w}) \,||\, p(\mathrm{w}|\mathrm{D}))$. In practice, this is done by maximising the evidence lower bound (ELBO), as given by Equation 9.

$$\mathcal{L}_{\mathrm{ELBO}} := -\mathbb{E}_{q_\theta(\mathrm{w})}\left[\log\, p(\mathrm{y}|\mathrm{x},\mathrm{w})\right] + KL[q_\theta(\mathrm{w})||p(\mathrm{w})] \quad (9)$$

In Bayesian neural networks, this objective can be optimised with stochastic gradient descent optimization [92]. After optimization $q_\theta(\mathrm{w})$ represents a distribution over plausible models which explain the data well. In mean-field variational inference, the approximate weight posterior $q_\theta(\mathrm{w})$ is represented by a fully factorized distribution, most often chosen to be Gaussian. Some stochastic regularization techniques, originally designed to prevent overfitting, can also be interpreted as instances of variational inference. The popular Monte Carlo dropout [57] method is a form of variational inference which approximates the Bayesian posterior with a multiplicative Bernoulli distribution over sets of weights. The stochasticity introduced through minibatch sampling in batch-norm can also be seen in this light [172]. Stochastic weight averaging Gaussian (SWAG) [118] computes a Gaussian approximation to the posterior from checkpoints of a deterministic neural network's stochastic gradient descent trajectory. These approaches are simple to implement but represent crude approximations to the Bayesian posterior. As such, the uncertainty estimates obtained by using these methods may suffer from some limitations [54].

Stochastic gradient MCMC [32, 184, 194] methods allow us to draw biased samples from the posterior distribution over NN parameters in a mini-batch friendly manner.

Recent work has started to explore performing Bayesian inference over the function space of NNs directly. Sun et al. [167] use a stochastic NN as a variational approximation to the posterior over functions. They define and approximately optimize a functional ELBO. Ma et al. [115] use a variant of the wake sleep algorithm to approximate the predictive posterior of a neural network with a Gaussian Process as a surrogate model. The Spectral-normalized Neural Gaussian Process (SNGP) [111] enables us to compute predictive uncertainty through input distance awareness, avoiding Monte Carlo sampling. Neural stochastic differential equation models (SDE-Net) [98] provide ways to quantify uncertainties from the stochastic dynamical system perspective using Brownian motion. More recently, Antorán et al. [5] introduce Depth Uncertainty, an approach that captures model specification uncertainty instead

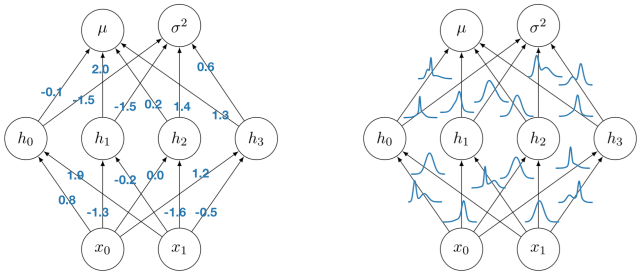


Figure 6: Left: In a regular NN, weights take single values which are found by fitting the data with some optimisation procedure. In a Bayesian Neural Network (BNN), probability distributions are placed over weights. These are obtained by leveraging (or approximating) the Bayesian update Eq(1).

of model uncertainty. By marginalizing over network depth their method is able to generate uncertainty estimates from a single forward pass.

C.2 Non-Bayesian Uncertainty quantification and calibration methods

Deterministic uncertainty quantification (DUQ) [177] builds upon ideas of radial basis function networks [105], allowing one to obtain uncertainty, computed as the distance to a centroid in latent space, with a single forward pass.

Guo et al. [64] show how using a validation set to learn a multiplicative scaling factor (known as a temperature) to output logits is a cheap post-hoc way to improve the calibration of NN models.

D OTHER ALGORITHMIC USE CASES FOR UNCERTAINTY

Epistemic uncertainty can be used to guide model-based search in applications such as active learning [78, 93]. In an active learning scenario, we are provided with a large amount of inputs but few or none of them are labelled. We assume labelling additional inputs has a large cost, e.g. querying a human medical professional. In order to train our model to make the best possible predictions, we would like to identify for which inputs it would be most useful to acquire labels. Here, each input's epistemic uncertainty tells us about how much our model would learn from seeing its labels and therefore directly answers our question.

Uncertainty can also be baked directly into the model learning procedure. In rejection-option classification, models can explicitly abstain from making predictions for points on which they expect to underperform [13]. For example, our credit limit model may have high epistemic uncertainty when predicting on customers with specific attributes (e.g., extremely low incomes) that are underrepresented in the training data; as such, uncertainty can be leveraged to decide whether the model should defer to a human based on the observed income level. Section 4 reviews various ways to use uncertainty to improve ML models.

Distributional shift, which occurs when the distribution of the test set is different from the training set distribution, is sometimes considered as a special case of epistemic uncertainty, although Malinin and Gales [121] explicitly consider it as a third type of

uncertainty: distributional uncertainty. Ovadia et al. [141] compare different uncertainty methods in the specific case of dataset shifts. These techniques all have in common that they use epistemic uncertainty estimates to improve the model by identifying regions of the input space where the model performs badly due to a lack of data. Some work also shows how using uncertainty techniques can directly improve the performance of a model and can out-perform using the soft-max probabilities [57, 90, 91]. Miller et al. [126] discuss how dropout sampling helped in an open-set object detection task where new unknown objects can appear in the frame. Some other papers propose their own uncertainty techniques, and presents how they out-perform different baselines, often focusing on the out-of-distribution detection task [41, 102, 121].

It is also important to note that sometimes a model's output could be used for downstream decision-making tasks by another model, whether an ML model, or an operational research model which will optimize a plan based on a predicted class or quantity. Uncertainty estimates are essential to transparently inform the downstream models of the validity of the input. For instance, stochastic optimization can take distributional input to reduce the cost of the solution in the presence of uncertainty. More generally, in systems where models, humans, and/or heuristics are chained, it is crucial to understand how the uncertainty of each step interacts with each other, and how it impacts the overall uncertainty of the system.

E FAIRNESS DEFINITIONS

Algorithmic fairness is complementary to algorithmic transparency. The ML community has attempted to define various notions of fairness statistically: see Barocas et al. [12] for an overview of the fairness literature. While there is no single definition of fairness for all contexts of deployment, many define ML fairness as absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics [124]. Unfairness can be the result of biases in (a) data used to build the algorithms or (b) the algorithms chosen to be implemented. We discuss some standard definitions of fairness here from the machine learning literature [14, 66]. Note that Kleinberg [94] finds that typically it is not possible to satisfy many fairness notions simultaneously. Let us assume we have a classifier f, which outputs a predicted outcome $\hat{Y} = f(X) \in \{0, 1\}$ for some input X. Let Y be the actual outcome for X. Let A be a binary sensitive attribute that is contained explicitly in or encoded implicitly in X. When we refer to groups, we mean the two sets that result from partitioning a dataset $\mathcal{D}$ based on A (i.e., if Group 1 was $\{X \in \mathcal{D}|A = 0\}$, Group 2 would be $\{X \in \mathcal{D}|A = 1\}$). Let $A = 0$ (Group 1) be considered unprivileged. Below are three common fairness metrics.

(1) Demographic Parity (DP): A classifier f is considered to be fair with regard to DP (also known as statistical parity) if the following quantity is close to 0:

$$P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)$$

The predicted positive rates for both groups should be the same [47].

(2) Equal Opportunity (EQ): A classifier f is considered to be fair with regard to EQ if the following quantity is close to 0:

$$P(\hat{Y} = 1|A = 0, Y = 1) - P(\hat{Y} = 1|A = 1, Y = 1)$$

That is, the true positive rates for both groups should the same [66].

(3) Equalized Odds (EO): A classifier f is considered to be fair with regard to EO if the following quantity is close to 0:

$$\sum_{y \in \{0,1\}} |P(\hat{Y} = 1|A = 0, Y = y) - P(\hat{Y} = 1|A = 1, Y = y)|$$

That is, we want to equalize the true positive and false positive rates across groups [66]. EO is satisfied if $\hat{Y}$ and A are independent conditional on Y.