Evaluation Gaps in Machine Learning Practice

Ben Hutchinson
Google Research
Sydney, Australia
benhutch@google.com

Negar Rostamzadeh
Google Research
Montreal, Canada
nrostamzadeh@google.com

Christina Greer
Google Research
Mountain View, USA
ckuhn@google.com

Katherine Heller
Google Research
Mountain View, USA
heller@google.com

Vinodkumar Prabhakaran
Google Research
San Francisco, USA
vinodkpg@google.com

ABSTRACT

Forming a reliable judgement of a machine learning (ML) model's appropriateness for an application ecosystem is critical for its responsible use, and requires considering a broad range of factors including harms, benefits, and responsibilities. In practice, however, evaluations of ML models frequently focus on only a narrow range of decontextualized predictive behaviours. We examine the evaluation gaps between the idealized breadth of evaluation concerns and the observed narrow focus of actual evaluations. Through an empirical study of papers from recent high-profile conferences in the Computer Vision and Natural Language Processing communities, we demonstrate a general focus on a handful of evaluation methods. By considering the metrics and test data distributions used in these methods, we draw attention to which properties of models are centered in the field, revealing the properties that are frequently neglected or sidelined during evaluation. By studying these properties, we demonstrate the machine learning discipline's implicit assumption of a range of commitments which have normative impacts; these include commitments to consequentialism, abstractability from context, the quantifiability of impacts, the limited role of model inputs in evaluation, and the equivalence of different failure modes. Shedding light on these assumptions enables us to question their appropriateness for ML system contexts, pointing the way towards more contextualized evaluation methodologies for robustly examining the trustworthiness of ML models.

CCS CONCEPTS

• **Computing methodologies → Machine learning approaches**.

KEYWORDS

machine learning, evaluation, applications

ACM Reference Format:
Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation Gaps in Machine Learning Practice. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3531146.3533233

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533233

1 INTRODUCTION

When evaluating a machine learning (ML) model for real-world uses, two fundamental questions arise: *Is this ML model good (enough)?* and *Is this ML model better than some alternative?* Obtaining reliable answers to these questions can be consequential for safety, fairness, and justice concerns in the deployment ecosystems. To address such questions, model evaluations use a variety of methods, and in doing so make technical and normative assumptions that are not always explicit. These implicit assumptions can obscure the presence of epistemic gaps and motivations in the model evaluations, which, if not identified, constitute risky *unknown unknowns*.

Recent scholarship has critiqued the ML community's evaluation practices, focusing on the use of evaluation benchmarks and leaderboards. Although leaderboards support the need of the discipline to iteratively optimize for accuracy, they neglect concerns such as inference latency, robustness, and externalities [51]. The structural incentives of the "competition mindset" encouraged by leaderboards can pose challenges to empirical rigor [154]. For example, over-reliance on a small number of evaluation metrics can lead to gaming the metric (cf. Goodhart's Law "when a measure becomes a target, it ceases to be a good measure") [162]; this can happen unintentionally as researchers pursue models with "state of the art" performance. Benchmarks that encourage narrowly optimizing for test set accuracy can also lead to models relying on spurious signals [31], while neglecting the challenge of measuring the full range of likely harms [22]. Birhane *et al.* find evidence for this in their study of the discourse of ML papers, showing that the field centers accuracy, generalization, and novelty, while marginalizing values such as safety [18]. Given that benchmark evaluations serve as proxies for performance on underlying abstract tasks [152], evaluating against a range of diverse benchmarks for each task might help mitigate biases within each benchmark. However, ML research disciplines seem to be trending towards relying on fewer evaluation benchmark datasets [94], with test set reuse potentially leading to a research community's overfitting with respect to the general task [104, 178]. Furthermore, within each benchmark, items are weighted equally (thus focusing on the head of the data distribution), failing to capture inherent differences in difficulty across items, and hence providing poor measures of progress on task performance [142]. As Raji *et al.* point out, the ML

research discipline's decontextualized and non-systematic use of benchmark data raises serious issues with regards to the validity of benchmarks as measures of progress on general task performance [136].

This paper complements and extends this range of critiques, considering the risks of application developers adopting the ML research community's standard evaluation methodologies. We seek to address challenges in measuring technology readiness (TRAM) [105, 141], while acknowledging this cannot be reduced to a purely technical question [43, 141]. By studying and analyzing the ML research community's evaluation practices, we draw attention to the *evaluation gaps* between ideal theories of evaluation and what is observed in ML research. By considering aspects of *evaluation data* and *evaluation metrics*—as well as considerations of *evaluation practices* such as error analysis and reporting of error bars—we highlight the discrepancies between the model quality signals reported by the research community and what is relevant to real-world model use. Our framework for analyzing the gaps builds upon and complements other streams of work on ML evaluation practices, including addressing distribution shifts between development data and application data [34, 95, 161], and robustness to perturbations in test items [76, 119, 133, 174]. We situate this work alongside studies of the appropriateness of ML evaluation metrics (e.g., [47, 89, 178]), noting that reliable choice of metric is often hampered by unclear goals [44, 98]. In foregrounding the information needs of application developers, we are also aligned with calls for transparent reporting of ML model evaluations [118], prioritizing needs of ML fairness practitioners [78], model auditing practices [137], and robust practices for evaluating ML systems for production readiness [23].

In Section 2, we consider various ideal goals that motivate why ML models are evaluated, discussing how these goals can differ between research contexts and application contexts. We then report in Section 3 on an empirical study into how machine learning research communities report model evaluations. By comparing the ideal goals of evaluation with the observed evaluation trends in our study, we highlight in Section 4 the evaluation gaps that present challenges to evaluations being good proxies for what application developers really care about. We identify six implicit evaluation assumptions that could account for the presence of these gaps. Finally, in Section 5, we discuss various techniques and methodologies that may help to mitigate these gaps.

2 IDEALS OF ML MODEL EVALUATION

> *Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.* — John Tukey [163, pp. 13–14]

Although this paper is ultimately concerned with practical information needs when evaluating ML models for use in applications, it is useful to first step back and consider the ultimate motivations and goals of model evaluation. To evaluate is to form a judgement; however, asking *Is this a good ML model?* is akin to asking such a question of other artefacts—such as *Is this a good glass?*—in that it requires acknowledging the implicit semantic arguments of uses and goals [135]. For example, *Is this a good glass [for my toddler to drink from, given that I want to avoid broken glass]?* is a very different question from *Is this a good glass [in which to serve wine to my boss, given that I want to impress them]?*

In this paper, we will speak of a *model evaluation* as a system of arbitrary structure that takes a model as an input and produces outputs of some form to judge the model. Designing a model evaluation often involves choosing one or more evaluation metrics (such as accuracy) combined with a choice of test data. The evaluation might be motivated by various stakeholder perspectives and interests [92]. The output might, for example, produce a single metric and an associated numeric value, or a table of such metrics and values; it might include confidence intervals and significance tests on metric values; and it might include text. By producing such an output, the evaluation helps to enable transparency by reducing the number of both unknown unknowns and known unknowns.

For the purposes of this paper, it is useful to distinguish between two types of evaluations:

Learner-centric. An ML model evaluation system useful for evaluating the learner (i.e., machine learning algorithm).
Application-centric. An ML model evaluation system useful for evaluating a potential application.

Learner-centric evaluations make conclusions about the quality of the learner or its environment based on the evaluation of the learned model. These including evaluations motivated by novel learning algorithms or model architectures, but also ones that a) aim to shed light on the training data (for example ML model evaluations can shed light on the data-generation practices used by institutions [5]), or b) "Green AI" explorations of how the learner can efficiently use limited amounts of resources [153]. However, when we evaluate a model without a specific application in mind, we lose the opportunity to form judgements specific to a use case. On the other hand, application-centric evaluations are concerned with how the model will operate within an ecosystem consisting of both human agents and technical components (Figure 1), sometimes described as the "ecological validity" [46]. Applications often use scores output by the model to initiate discrete actions or decisions, by applying a specific classification threshold to the scores.[1] In contrast, learner-centric evaluations sometimes care about scores output by models even in the absence of any thresholds.

This distinction between learner-centric and application-centric is related (albeit imperfectly) to the different objectives of model evaluations that concern the engineering and science disciplines [114, 169]. Note that we are not claiming (cf. the debate in [123]) that science lies outside the bounds of statistical/ML methods, but rather that scientific-flavored pursuits have distinct uses of such methods [24]. Debates between AI practitioners about the relationships between AI, science, and statistical methods have a long history, for example Diana Forsythe's studies of 1980s AI labs [56]. Important to this debate regarding the scientific goals of ML is the question of construct validity; that is, whether our measurements actually measure the things that we claim they do [86, 87, 136]. Conversely, consequential validity—which includes the real-world consequences of an evaluation's interpretation and use—is likely

[1] The history of this type of use case extends beyond ML models, e.g., to the use of regression models in university admissions testing [82].

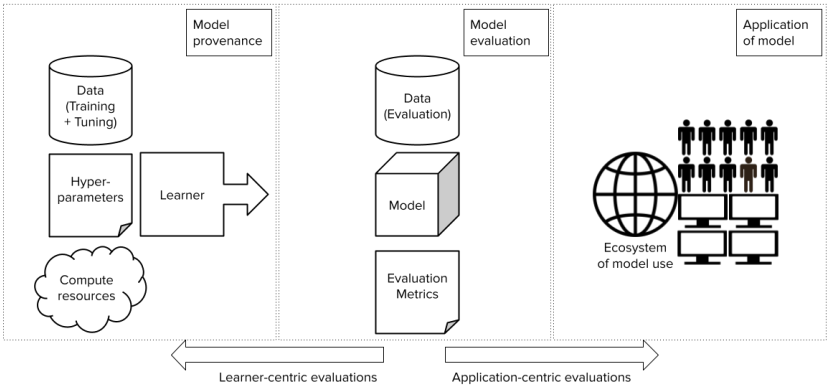

Figure 1: Learner-centric ML model evaluations are concerned with the learner and its environment. Application-centric model evaluations are concerned with how the model will interact with an ecosystem into which it is introduced.

more important to considerations of accountability and governance of ML models in applications [87].

Scientific goal. Evaluating the model can motivate beliefs/explanations about the world (including possibly the learner).
Engineering goal. Evaluating the model can tell us whether the model can be used as a means towards a goal.

This distinction is closely related to one between "scientific testing" and "competitive testing" made by Hooker in 1995, who takes the position that competitive testing a) is unscientific, and b) does not constitute true research but merely development [79]. However, since engineering research has its own goals, distinct from those of science [26], a more defensible position is that evaluations in support of scientific research are distinct from evaluations in support of engineering research.

Table 1 summarizes the above distinctions and the relationships between them. The distinction between learner-centric and application-centric evaluations relates to the question of internal validity and external validity that is more commonly discussed in the social sciences than in ML (see, e.g., [124]) but also sometimes in ML [104]. This is reflected in the ways in which practitioners of the two types of evaluations discuss the topic of robustness. Learner-centric evaluations pay attention to the robustness of the learner to changes in the training data (e.g., distributional shifts, outliers, perturbations, poisoning attacks; and with connections to robust estimation of statistics [102]), while application-centric evaluations pay attention to desired behaviors such as the (in)sensitivity of the model to certain classes of perturbations of the input, or to sensitive input features (e.g., [61]).

Note that nothing in the ideals of evaluation described above has stipulated whether evaluations are *quantitative* or *qualitative*. For example, one could imagine interrogating a chatbot model using qualitative techniques, or adopting methodologies of political critique such as [41]. Similarly, nothing has stipulated what combinations of empirical or deductive methods are used.

3 ML MODEL EVALUATIONS IN PRACTICE

> *Beneath the technical issues lie some differences in* values *concerning not only the meaning but also the relative merit of "science" and "artificial intelligence." —* Diana Forsythe [56]

To shed light on the ML research community's norms and values around model evaluation, we looked at how these communities report their model evaluations. By examining 200 papers from several top conferences in two research disciplines that use ML approaches extensively, we identified patterns regarding choices of metrics, evaluation data, and measurement practices. This empirical study of ML research practices complements several recent studies of ML evaluation practices. These include: a survey 144 research papers studying the properties of models that are tested for [178]; a review of 107 papers from Computer Vision (CV), Natural Language Processing (NLP) and other ML disciplines to diagnose internal and external modes of evaluation failures [104]; an analysis of whether 60 NLP and CV papers pay attention to accuracy or efficiency [153]; and an analysis of the Papers With Code dataset[2] for patterns of benchmark dataset creation and re-use [94].

3.1 Method

3.1.1 Data. We sampled 200 research papers, stratified by discipline, conference and year. 100 papers were selected from each of the NLP and CV disciplines. We selected 20 papers at random from the proceedings of each of the 55th to 59th Annual Meetings of the Association of Computational Linguistics (ACL'2017–ACL'2021), 25 papers from each of the proceedings of the 2019–2021 IEEE Conferences on Computer Vision and Pattern Recognition (CVPR'2019–CVPR'2021), and 25 papers from the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'2021). These conferences represent the pinnacles of their respective research fields.[3]

[2] https://paperswithcode.com
[3] ACL and CVPR are rated A* ("flagship conference"), and MICCAI is rated A ("excellent conference"), by core.edu.au; all three are in the top 30 computer science conferences out of over 900 listed on research.com.

	Learner-centric evaluations	Application-centric evaluations
Typical evaluation goal	Distinguish better learners from poorer ones	Predict ecosystem outcomes
Schematic of goal	$Understand(Learner)$	$Understand(Ecosystem + Model)$
Disciplinary goals	Science or engineering	Primarily engineering

Table 1: Summary of typical goals of the idealized learner-centric and application-centric evaluations.

3.1.2 Analysis. The authors of this paper performed this analysis, dividing the papers among themselves based on disciplinary familiarity. Using an iterative procedure of analysis and discussion, we converged on a set of labels that captured important aspects of evaluations across and within disciplines. Recall from Section 2 that, for our purposes, a single evaluation typically involves choosing one or more metrics and one or more datasets. We coded each of the papers along three dimensions. a) **Metrics**: Which evaluation metrics were reported? After iteration, we converged on the categories of metrics shown in Table 2. b) **Data**: Was test data drawn from the same distribution as the training data, under the Independent and Identically Distributed (I.I.D.) assumption? c) **Analysis**: Was statistical significance of differences reported? Were error bars and/or confidence intervals reported? Was error analysis performed? Were examples of model performance provided to complement measurements with qualitative information?

3.2 Results

Although each of the disciplines and conferences does not define itself solely in terms of ML, the practice of reporting one or more model evaluations in a research paper is ubiquitous. Only five papers did not include evaluations of ML models; of these two were published at ACL (a survey paper, a paper aimed at understanding linguistic features, and one on spanning-tree algorithms), and two at CVPR (a paper with only qualitative results, and one introducing a dataset). Table 3 summarizes the results of the other 195. Counts are non-exclusive, for example papers frequently reported multiple metrics and sometimes reported performance both on I.I.D. test data and on non-I.I.D. test data.

Appendix B contains an overview of the flavors of test data we observed. We found evidence to support the claim that evaluations of NLP models have "historically involved reporting the performance (generally meaning the accuracy) of the model on a specific held-out [i.e., I.I.D.] test set" [20, p. 94].[4] CV evaluations seem to be even more likely to utilize I.I.D. test data, and—consistent with [94]—CV papers typically either introduce a new task (and corresponding benchmark dataset) [88, 103, 143, 175] or present results of a new model on an existing widely-used benchmark [73, 139]. An exception to this trend was CV papers which explored shared representations (e.g., in multi-task learning [53, 100] or domain adaptation [120, 127]).

Evaluations in both disciplines showed a heavy reliance on reporting point estimates of metrics, with variance or error bars typically not reported in our sample. While colloquial uses of phrases like "significantly better" were fairly common, most papers did not report on technical calculations of statistical differences; we considered only those latter instances when coding whether a paper reported significance. Regarding metrics, most of those that were frequently seen in our sample were somewhat insensitive to different types of errors. For example, accuracy does not distinguish between FP and FN; F_1 is symmetric in FP and FN (they can be swapped without affecting F_1); the Overlap metrics are similary invariant to swapping of the predicted bounding box and the reference bounding box; the Distance category of metrics does not distinguish over-estimation from under-estimation on regression tasks.

From our reading of the 200 papers in our sample, one qualitative observation we had was that model evaluations typically do not include concrete examples of model behavior, nor analyses of errors (for a counterexample which includes these practices, see [35]). Also, we noted the scarcity of papers whose sole contribution is a new dataset for an existing task, aligning with previous observations that dataset contributions are not valued highly within the community [148]. We hypothesise that conference reviewers place emphasis on novelty of model, task, and/or metric. We note a general tension between disciplinary values of task novelty and demonstrating state-of-the-art performance by outperforming previous models, and the risk of overfitting from test set re-use discussed by [104].

3.3 Discussion

This small-scale quantitative study of model evaluations provides clues as to the values and goals of the ML research communities. Test data was often old (e.g., the CONLL 2003 English NER dataset [150] used in two papers); optimizing for these static test sets fails to account for societal and linguistic change [14]. Disaggregation of metrics was rare, and fairness analyses were absent despite our sample being from 2017 onward, concurrent with mainstream awareness of ML fairness concerns. Despite being acknowledged by influential thought-leaders in ML to be unrealistic for applications [15], using I.I.D. test data is the norm. These are in alignment with the learner-centric goals of evaluations (Section 2). Similarly, with a few exceptions in our sample, there was general paucity of discussions of tradeoffs such as accuracy vs resource-efficiency that are typical of engineering disciplines [26], suggesting that the ML research disciplines generally aspire to scientific goals concerning understanding and explaining the learner. With this lens, the disciplinary paradigm of measuring accuracy on I.I.D. test data is not surprising: the goal is to assess a model's ability to generalize. This assessment would then give us good guarantees on the application's behavior, if the practical challenges of ascertaining the data distributions in an application ecosystem can be overcome. In practice, however,

[4] Two observed non-I.I.D. evaluation patterns in NLP were: a) testing on a different linguistic "domain" (e.g., training on texts about earthquakes and testing on texts about floods [1]); and b) testing a model's ability to predict properties of a manually compiled lexical resource (e.g., [165]). See also Appendix B.

Metric category	Examples	Description
Accuracy	Accuracy, error rate	Sensitive to the sum TP+TN and to N. Not sensitive to class imbalance.
Precision	Precision, Bleu	Sensitive to TP and FP. Not sensitive to FN or TN.
Recall	Recall, Rouge	Sensitive to TP and FN. Not sensitive to FP or TN.
F-score	F_1, F_β	Sensitive to TP, FP and FN. Not sensitive to TN.
Overlap	Dice, IoU	Sensitive to intersection and overlap of predicted and actual.
Likelihood	Perplexity	Sensitive to the probability that the model assigns to the test data.
Distance	MSE, MAE, RMSE, CD	Sensitive to the distance between the prediction and the actual value.
Correlation	Pearsons r, Spearman's ρ	Sensitive to each of TP, TN, FP and FN, but unlike Accuracy metrics they factor in the degree of agreement that would be expected by chance.
AUC	MAP, AUROC	Does not rely on a specific classification threshold, but instead calculates the area under a curve parameterized by different thresholds.

Table 2: Categories of evaluation metrics used in the analysis of the ML research literature. TP=true positives; TN=true negatives; FP=false positives; FN=false negatives; N=total number of data points. See Appendix A for the most common metrics in our data and their categorizations.

Discipline:Venue (# papers with ML evals)	NLP:ACL (97)	CV:CVPR (73)	CV:MICCAI (25)	CV:Combined (98)	NLP+CV:Combined (195)
Most Common Metrics					
Metric category♣ (num. of papers)	Accuracy (47) F-score (45) Precision (43) Recall (25)	AUC (32) Accuracy (25) Overlap (22) Distance (10)	Distance (14) Overlap (9) AUC (6) Accuracy (4)	AUC (38) Overlap (31) Accuracy (29) Distance (24)	Accuracy (76) F-score + Overlap♠ (74) Precision (48) AUC (44)
Data					
I.I.D. test data	78	72	25	97	175
Non-I.I.D. test data	28	21	4	25	53
Analysis					
Reports significance	24	0	7	7	31
Reports error bars◇	10	6	10	16	26

Table 3: Analysis of how Natural Language Processing (NLP) and Computer Vision (CV) research communities perform ML model evaluations. ♣Appendix A provides definitions of commonly observed metrics, and their mappings to categories. ◇Includes any form of error bars/confidence intervals/credible intervals/variation across multiple runs. ♠Reported together here due to the equivalence of the Dice measure (in the Overlap category) and F_1 (in the F-score category) [128].

these challenges can be severe, and the research papers we surveyed do not generally tackle questions of uncertainty regarding data distributions.

4 GAPS AND ASSUMPTIONS IN COMMON EVALUATION PRACTICES

In theory there is no difference between theory and practice, while in practice there is. — Brewster (1881) [25]

We now consider whether the research evaluation practices observed in Section 3 are aligned with the needs of decision-makers who consider whether to use a model in an application. That is, we consider whether the typically learner-centric evaluations, which commonly use metrics such as accuracy or F_1 on test data I.I.D. with the training data, meet the need of application-centric evaluations. In doing so, we expose, in a novel way, the interplay of technical and normative considerations in model evaluation methodologies.

4.1 Assumptions in Model Evaluation

We introduce six assumptions in turn, describing both how they operate individually in evaluations and how they compose and compound. We also call out "evaluation gaps" of concern relevant to each assumption. Appendix C contains a hypothetical example from a specific application domain that illustrates the flavors of the concerns. Our starting point is the observation from Section 2 that the goal of application-centric model evaluations is to understand how a model will interact with its ecosystem, which we denote schematically as:

Understand(Model + Ecosystem)

(Application-centric Evaluation Goal)

Assumption 1: Consequentialism. Consequentialism is the view that whether actions are good or bad depends only on their consequences [158]. The ML research literature often appeals to motivations about model utility to humans (e.g., [19, 27, 51, 59, 77,

84, 108, 122, 125, 180], including papers on fairness in ML such as [29, 36, 38, 39]). In adopting consequentialism as its *de facto* ethical framework, ML prioritizes the greatest good for the greatest number [85] and centers measurable future impacts. Moreover, the consequences that are centered are the *direct* consequences, with little attention given to motives, rules, or public acceptance [158]. This is realised as a focus on the first-order consequences of introducing the model into the ecosystem. Changes to the ecosystem itself—e.g., addressing what social change is perceived as possible and desirable [49, 68, 80]—are assumed to be out of scope, as are concerns for setting of precedents for other ML developers. We denote this assumption schematically as:

$$Understand(Model + Ecosystem) \approx Utility(Model + Ecosystem)$$

(Consequentialism Assumption)

Evaluation Gap 1: Provenance. A focus on future consequences neglects important moral considerations regarding the construction of the model. This excludes both deontological concerns—for example, *Were data consent and sovereignty handled appropriately?* [4, 41, 97] and *Were data workers treated with dignity?* [67]—as well as questions regarding past costs of development—for example, *What were the energy use externalities of model training?* [40, 60, 160] and *Was the labour paid fairly?* [157]. Schwartz *et al.* coin the phrase "Red AI" to describe ML work that disregards the costs of training, noting that such work inhibits discussions of when costs might outweigh benefits [153].

Evaluation Gap 2: Social Responsibilities. Another outcome of focusing primarily on direct consequences is marginalizing the assessment of a model against the social contracts that guide the ecosystem in which the model is used, such as moral values, principles, laws, and social expectations. For instance, *Does the model adhere to the moral duty to treat people in ways that upholds their basic human rights?* [156], *Does it abide by legal mechanisms of accountability?* [115, 138], and *Does it satisfy social expectations of inclusion, such as the "nothing about us without us" principle?* [33].

Assumption 2: Abstractability from Context. The model's ecosystem is reduced to a set of considerations (X, Y), i.e., the inputs to the model and the "ground truth," and in practice X may often fail to model socially important yet sensitive aspects of the environment [5, 10]. The model itself is reduced to a predicted value $\hat{Y}$, ignoring e.g., secondary model outputs such as confidence scores, or predictions on auxiliary model heads.

$$Utility(Model + Ecosystem) \approx Utility(\hat{Y}, X, Y)$$

(Assumption of Abstractability of Context)

Evaluation Gap 3: System Considerations. Equating a model with its prediction overlooks the potential usefulness of model interpretability and explainability. Also, reducing an ecosystem to model inputs and "ground truth" overlooks questions of system dynamics [112, 155], such as feedback loops, "humans-in-the-loop," and other effects "due to actions of various agents changing the world" [15]. Also overlooked are inference-time externalities of energy use [28, 60], cultural aspects of the ecosystem [146], and long term impacts [29].

Evaluation Gap 4: Interpretive Epistemics. By positing a variable $Y = y$ which represents the "ground truth" of a situation—even in situations involving social phenomena—a positivist stance

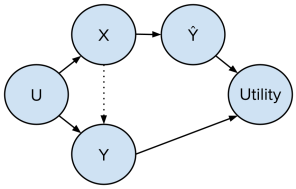

Figure 2: Causal graph illustrating the Input Myopia Assumption.

on knowledge is implicitly adopted. That is, a "true" value $Y = y$ is taken to be objectively singular and knowable. This contrasts with anthropology's understanding of knowledge as socially and culturally dependent [57] and requiring interpretation [63]. In the specific cases of CV and NLP discussed in Section 3, cultural aspects of image and language interpretation are typically marginalized (cf. [11, 16, 90, 101], for example), exemplifying what Aroyo and Welty call AI's myth of "One Truth" [7]. Furthermore, the positivist stance downplays the importance of questions of construct validity and reliability [58, 87].

Assumption 3: Input Myopia. Once the input variable X has been used by the model to calculate the model prediction $\hat{Y}$, X is typically ignored for the remainder of the evaluation. That is, the utility of the model is assumed to depend only on the model's prediction and on the "ground truth." We illustrate this with a causal graph diagram in Figure 2, which shows Utility as independent of X once the effects of $\hat{Y}$ and Y are taken into account.

$$Utility(\hat{Y}, X, Y) \approx Utility(\hat{Y}, Y)$$

(Input Myopia Assumption)

Evaluation Gap 5: Disaggregated Analyses. By reducing the variables of interest to the evaluation to the prediction Y and the ground truth $\hat{Y}$, the downstream evaluation is denied the potential to use X. This exacerbates Evaluation Gap 3 by further abstracting the evaluation statistics from their contexts. For example, X could have been used to disaggregate the evaluation statistics in various dimensions—including for fairness analyses, assuming that sociodemographic data is available and appropriate [6, 9]—or to examine regions of the input space which raise critical safety concerns (e.g., distinguishing a computer vision model's failure to recognise a pedestrian on the sidewalk from failure to recognise one crossing the road) [3]. Similarly, robustness analyses which compare the model predictions for related inputs in the same neighborhood of the input space are also excluded.

Assumption 4: Quantifiability. We have not yet described any modeling assumptions about the mathematical or topological nature of the implied $Utility$ function, which up to now has been conceived as an arbitrary procedure producing an arbitrary output. We observe, however, that when models are evaluated, there is a social desire to produce a small number of scalar scores. This is reinforced by "leaderboardism" [51], and extends to the point of averaging different types of scores such as correlation and accuracy [170]. We identify two assumptions here: first, that impacts on each individual can be reduced to a single numeric value (and thus different dimensions of impacts are commensurable[5]); second,

[5] E.g., one machine learning fairness paper says "c is the cost of detention in units of crime prevented" [39].

that impacts across individuals are similarly commensurable. We define $\hat{y} \in \hat{Y}$ and $y \in Y$ to be a specific model prediction, and a specific "ground truth" value respectively, leading to the Individual Quantifiability Assumption and the Collective Quantifiability Assumption, respectively.

$$Utility(\hat{y}, y) \in \mathbb{R} \qquad \text{(Individual Quantifiability Assumption)}$$

$$Utility(\hat{Y}, Y) \approx E_{(\hat{y},y)}\{Utility(\hat{y}, y)\}$$
(Collective Quantifiability Assumption)

Composing these assumptions with the previous ones leads to the belief that the evaluation can be summarized as a scalar statistic: $Utility(\hat{Y}, Y) \in \mathbb{R}$.

Evaluation Gap 6: Incommensurables. The Quantifiability Assumptions assume that the impacts on individuals are reducible to numbers, trivializing the frequent difficulty in comparing different benefits and costs [111]. Furthermore, the harms and benefits across individuals are assumed to be comparable in the same scale. These assumptions are likely to disproportionately impact underrepresented groups, for whom model impacts might differ in qualitative ways from the well represented groups [74, 146, 147]. The former groups are less likely to be represented in the ML team [173] and hence less likely to have their standpoints on harms and benefits acknowledged.

Assumption 5: Failures Cases Are Equivalent. For classification tasks, common evaluation metrics such as accuracy or error rate model the utility of $\hat{Y}$ as binary (i.e., either 1 or 0), depending entirely on whether or not it is equal to the "ground truth" Y. That is, for a binary task, $Utility(\hat{Y}{=}0, Y{=}0){=}Utility(\hat{Y}{=}1, Y{=}1){=}1$ and $Utility(\hat{Y}{=}0, Y{=}1){=}Utility(\hat{Y}{=}1, Y{=}0){=}0$. Similarly for regression tasks, common metrics such as MAE and MSE take the magnitude of error into account, yet still treat certain failures as equivalent (specifically, $Utility(\hat{Y}{=}\hat{y}, Y{=}\hat{y}+\delta){=}Utility(\hat{Y}{=}\hat{y}, Y{=}\hat{y}-\delta)$, for all $\delta, \hat{y}$).

$$Utility(\hat{Y} = \hat{y}, Y = y) \approx \mathbb{1}(\hat{y} = y)$$
(Assumption of Equivalent Failures [Classification])

$$Utility(\hat{y}, y) \text{ is a function of } |\hat{y} - y|$$
(Assumption of Equivalent Failures [Regression])

Taken together with the previous assumptions, this yields $Utility(\hat{y}, y) = P(\hat{y} = y)$ for classification tasks.

Evaluation Gap 7: Disparate harms and benefits. Treating all failure cases as equivalent fails to appreciate that different classes of errors often have very different impacts [32, 134]. In multiclass classification, severely offensive predictions (e.g., predicting an animal in an image of a person) are given the same weight as inoffensive ones. In regression tasks, insensitivity to either the direction of the difference $\hat{y} - y$ or the magnitude of y can result in evaluations being possibly poor proxies for downstream impacts. (One common application use case of regression models is to apply a cutoff threshold t to the predicted scalar values, for which both the direction of error and the magnitude of y are relevant.)

Assumption 6: Test Data Validity. Taken collectively, the previous assumptions might lead one to use accuracy as an evaluation metric for a classification task. Further assumptions can then be made in deciding *how* to estimate accuracy. The final assumption we discuss here is that the test data over which accuracy (or other metrics) is calculated provides a good estimate of the accuracy of the model when embedded in the ecosystem.

$$P(\hat{y} = y) \approx P(\hat{y}' = y')$$
(Assumption of Test Data Validity [Classification])

where $Y' = y'$ and $\hat{Y}' = \hat{y}'$ are the ground truth labels and the model predictions on the test data, respectively.

Evaluation Gap 8: Data Drifts. A simple model of the ecosystem's data distributions is particularly risky when system feedback effects would cause the distributions of data in the ecosystem to diverge from those in the evaluation sample [93, 107]. In general, this can lead to overconfidence in the system's reliability, which can be exacerbated for regions in the tail of the input distribution.

4.2 Discussion

We have described six assumptions that simplify the model evaluation task. Taken together, they would cause one to believe—with compounding risks—that a model's accuracy is a good proxy for its fitness for an application. We sketch this composition of assumptions in Figure 4, along with questions that illustrate the gaps raised by each assumption. Our reason for teasing apart these assumptions and their compounding effects is *not* to attack the "strawman" of naive application-centric evaluations which rely solely on estimating model accuracy. Rather, our goal is to point out that most model evaluations, even sophisticated ones, make such assumptions to varying degrees. For example:

- Some robustness evaluations (for surveys, see [54, 171]) explicitly tackle the problem of distribution shifts, rejecting the Assumptions of Test Data Validity without questioning the other assumptions we have identified.
- Some sensitivity evaluations consider the effect on the model predictions of small changes in the input, but use accuracy as an evaluation metric, rejecting the Input Myopia Assumption without questioning the others [140].
- Some fairness evaluations perform disaggregated analyses using the Recall or Precision metrics, sticking by all assumptions other than Input Myopia and Equivalent Failures [37, 71].

It may not be possible to avoid all of the assumptions all of the time; nevertheless unavoidable assumptions should be acknowledged and critically examined. The six assumptions we have identified also provide a lens for assessing the consistency of some evaluation metrics with other assumptions that have been made during the evaluation, for example

- *Is F-score consistent with an utilitarian evaluation framework?* The F-score is mathematically a harmonic mean— which is often appropriate for averaging pairs of rates (e.g., two speeds). When applied to Precision and Recall, however, the F-score constitutes a peculiar averaging of "apples and oranges," since, when conceived as rates, Precision and Recall measure rates of change of different quantities, [131]. F-score is thus difficult to interpret within an evaluation framework that aims to maximize model utility.
- *Do threshold-free evaluations such as the Area Under the Receiver Operating Characteristic (AUROC) abstract too much of*

	Assumption	Considerations that might be Overlooked
$Understand(Model + Ecosystem)$	Application-centric evaluation	Opportunities for scientific insights.
$\approx Utility(Model + Ecosystem)$	Consequentialism	Data sourcing and processing; invisible labour; consultation with impacted communities; motives; public acceptance; human rights.
$\approx Utility(\hat{Y}, X, Y)$	Abstractability from Context	System feedback loops; humans-in-the-loop.
$\approx Utility(\hat{Y}, Y)$	Input Myopia	Disaggregated analyses; sensitivity analyses; safety-critical edge cases.
$\approx E_{(\hat{Y}=\hat{y}, Y=y)}\{Utility(\hat{y}, y)\}$	Quantitative Modeling	Different flavors of impacts on a single person; different flavors of impacts across groups.
$\approx P(\hat{y} = y)$	Equivalent Failures	Severe failure cases; confusion matrices; topology of the prediction space.
$\approx P(\hat{y}' = y')$	Test Data Validity	Data sampling biases; distribution shifts.

Table 4: Sketch of how the six assumptions of Section 4—when taken collectively—compose to simplify the task of evaluating a model for an application ($Understand(Model + Ecosystem)$) to one of calculating accuracy over a data sample. A pseudo-formal notation (akin to pseudo-code) is used to enable rapid glossing of the main connections. $Y = y$ and $\hat{Y} = \hat{y}$ denote the true (unobserved) distributions of ground truth and model predictions, respectively, while the variables $Y' = y'$ and $\hat{Y}' = \hat{y}'$ denote the samples of reference labels and model predictions over which accuracy is calculated in practice. The order of the assumptions reflects an increasing focus on technical aspects of model evaluation, and a corresponding minimizing of non-technical aspects. Appendix C illustrates how each of the sets of considerations might apply in a hypothetical application of a computer vision model.

the deployment context? Since AUROC is calculated by averaging over a range of possible threshold values, it "cannot be interpreted as having any relevance to any particular classifier" [130] (which is not saying AUROC is irrelevant to evaluating *the learner*, cf. Section 2, nor to a learned model's propensity to correctly rank positive instances above negative ones). The same argument can be made for the Mean Average Precision metric used in image classification (see Appendix A). For useful application-centric evaluations, it is more meaningful to report pairs of $(Precision, Recall)$ values (for all classes) for a range of threshold values [129].

In both cases, we ask whether such metrics are of limited utility in application-centric evaluations and whether they are better left to learner-centric ones.

5 CONTEXTUALIZING APPLICATION-CENTRIC MODEL EVALUATIONS

> *the ornithologists were forced to adapt their behavior (for the sake of "science") to the most primitive evaluation method which was the only one considered or known, or else throw their data away.* — Hampel [70]

When applications of ML models have the potential to impact human lives and livelihoods, thorough and reliable evaluations of models are critical. As discussed in Section 3, the different goals and values of academic ML research communities mean that research norms cannot be relied upon as guideposts for evaluating models for applications. In this section, we propose steps towards evaluations that are rigorous in their methods and aim to be humble about their epistemic uncertainties. In doing so, we expand on the call by Raji *et al.* to pay more attention not just to evaluation metric values but also to the quality and reliability of the measurements themselves, including sensitivity to external factors [136].

5.1 Minding the Gaps between Evaluation Goals and Research Practice

Documenting assumptions made during model evaluation is critical for transparency and enables more informed decisions. If an assumption is difficult to avoid in practice, consider augmenting the evaluation with signals that may shed complementary light on questions of concern. For example, even a handful of insightful comments from members of impacted communities can be an invaluable complement to evaluations using quantitative metrics. We now consider specific mitigation strategies for each of the gaps in turn.

Minding Gap 1: Evaluate More than Consequences. To reduce the gap introduced by the Consequentialism Assumption, evaluate the processes that led to the creation of the model, including how datasets were constructed [151]. We echo calls for more reflexivity around social and intentional factors around model development [117], more documentation of the complete lifecycle of model development [83, 168], and greater transparency around ML models and their datasets [13, 62, 118]. It may be appropriate to contemplate whether the model is aligned with the virtues the organization aspires to [166]. Consider the question of whether any ML model could be a morally appropriate solution in this application context, e.g., whether it is appropriate to make decisions about one person on the basis of others' behaviors [49].

Minding Gap 2: Center Obligations. Since reasoning about uncertain future states of the world is fraught with challenges [29], evaluations should consider indirect consequences and assess how the model upholds social obligations within the ecosystem. This may involve processes such as assessments of human rights, social and ethical impact [110, 115], audits of whether the ML system upholds the organization's declared values or principles [137], and/or assessments of the potential for privacy leakage (e.g., [30, 176]).

Minding Gap 3: Demarginalize the Context. To address the gap introduced by the Assumption of Abstractability from Context, consider externalities such as energy consumption [75, 153], as well as resource requirements [51]. It is important to think about how the human and technical parts of the system will interact [112, 155]. Note that when substituting one model for another—or for displaced human labor—system stability can itself be a desirable property independent of model accuracy (and perhaps counter to tech industry discourses of "disruption" [64]), and a range of metrics exist for comparing predictions with those of a legacy model [47]. Care should be taken to avoid the "portability trap" of assuming that what is good for one context is good for another [155]. The more attention paid to the specifics of the application context, the better; hence, metrics which assume no particular classification threshold, such as AUC, may provide limited signal for any single context.

Minding Gap 4: Make Subjectivities Transparent. Acknowledge the subjectivities inherent in many tasks [2]. An array of recent scholarship on subjectivity in ML has "embraced disagreement" through practices of explicitly modeling—in both the data model and the ML model—inter-subject variation in interpretations [7, 12, 45, 48, 55]. For the purposes of ML model evaluations, disaggregating labels on test data according to the cultural and socio-demographic standpoints of their annotators enables more nuanced disaggregated evaluation statistics [132].

Minding Gap 5: Respect Differences Between Inputs. A realistic "null hypothesis" is that misclassifications affect people in the application ecosystem disparately. For example, people may differ both in their preferences regarding model predictions $\hat{Y}$ *per se*, as well as their preferences regarding model accuracy $\hat{Y} = Y$ [17].[6] As such—and *independent of fairness considerations*—evaluations should be routinely pay attention to different parts of the input distribution, including disaggregating along social subgroups. Special attention should be paid to the tail of the distribution and outliers during evaluation, as these may require further analysis to diagnose the potential for rare but unsafe impacts. Input sensitivity testing can provide useful information about the sensitivity of the classifier to dimensions of input variation known to be of concern (e.g., gender in text [21, 66, 81, 181]).

Minding Gap 6: Think Beyond Scalar Utility. Resist the temptation to reduce a model's utility to a single scalar value, either for stack ranking [51] or to simplify the cognitive load on decision makers. Instead, include a range of different metrics and evaluation distributions in the evaluation [118]. Acknowledge and report epistemic uncertainty, e.g., the effects of missing data or measurement and sampling error on metrics. Acknowledge qualitative impacts that are not addressed by metrics (e.g., harms to application users caused by supplanting socially meaningful human interactions), and rigorously assess the validity of attempts to measure social or emotional harms. Be conservative in aggregations: consider plotting data rather than reporting summary statistics (cf. Anscombe's quartet); do not aggregate unlike quantities; report multiple estimates of central tendency and variation; and don't assume that all users of an application will have equal benefits (or harms) from system outcomes. Consider applying aggregation and partial ranking techniques from the fair division literature to ML models, including techniques that give greater weight to those with the worst outcomes (e.g., in the extreme case, "Maximin") [50].

Minding Gap 7: Respect Differences Between Failures. If the harms of false positives and false negatives are incommensurable, report them separately. If commensurable, weight each appropriately. For multiclass classifiers, this approach generalizes to a *classification cost matrix* [164], and, more generally, including the *confusion matrix* before costs are assigned; for regression tasks, report metrics such as MSE disaggregated by buckets of Y.

Minding Gap 8: Validate Quality of Test Data. For transparency, do not assume it is obvious to others which datasets are used in training and evaluation; instead, be explicit about the provenance, distribution, and known biases of the datasets in use [6]. Consider Bayesian approaches to dealing with uncertainty about data distributions [91, 99, 116], especially when sample sizes are small or prior work has revealed systematic biases. For example, an evaluation which uses limited data in a novel domain (or in an under-studied language) to investigate gender biases in pronoun resolution should be tentative in drawing strong positive conclusions about "fairness" due to abundant evidence of gender biases in English pronoun resolution models (e.g. [172]).

5.2 Alternate Model Evaluation Methodologies

More radical excursions from the disciplinary paradigm are often worth considering, especially in scenarios with high stakes or high uncertainty.

Evaluation Remits. In 1995, Sparck Jones and Galliers called for a careful approach to NLP evaluation that is broadly applicable to ML model evaluations (see Appendix D) [92]. Their approach involves a top-down examination of the context and goal of the evaluation before the evaluation design even begins, and their call for careful documentation of the evaluation "remit"—i.e., official responsibilities—is in line with more recent work calling for stakeholder transparency for ML [83, 137]. They advocate for establishing whose perspectives are adopted in the evaluation and whose interests prompted it. Appendix D sketches how Sparck Jones and Galliers' framework could be adopted for ML model evaluations.

Active Testing. Active Testing aims to iteratively choose new items that are most informative in addressing the goals of the evaluation [69, 96] (cf. its cousin Active Learning, which selects items that are informative for the learner). Active Testing provides a better estimate of model performance than using the same number of test instances sampled I.I.D. Exploring Active Testing in pursuit

[6] Note that in many real-world applications the "ground truth" variable Y may be a convenient counterfactual fiction, since the system's actions on the basis of the prediction $\hat{Y}$ may inhibit Y from being realised—for example, a finance ML model may predict a potential customer would default on a loan if given one, and hence the system the model is deployed in may prevent the customer getting a loan in the first place.

of fairness testing goals seems a promising direction for future research.

Adversarial Testing. In many cases, there is great uncertainty regarding an application deployment context. One cautious and conservative approach—especially in the face of great uncertainty—is to simulate "adversaries" trying to provoke harmful outcomes from the system. Borrowing adversarial techniques from security testing and privacy testing, adversarial testing of models requires due diligence to trigger the most harmful model predictions, using either manually chosen or algorithmically generated test instances [52, 145, 177, 179].

Multidimensional Comparisons. When comparing candidate models, avoid the "Leaderboardism Trap" of believing that a total ordering of candidates is possible. A multidimensional and nuanced evaluation may provide at best a partial ordering of candidate models, and it may require careful and accountable judgement and qualitative considerations to decide among them. The Fair Division literature on Social Welfare Orderings may be a promising direction for developing evaluation frameworks that prioritize "egalitarian" considerations, in which greater weighting is given to those who are worst impacted by a model [50].

5.3 Evaluation-driven ML Methodologies

In this section, we follow Rostamzadeh *et al.* in drawing inspiration from test-driven practices, such as those of software development [144]. Traditional software testing involves significant time, resources, and effort [72]; even moderate-sized software projects spend hundreds of person-hours writing test cases, implementing them, and meticulously documenting the test results. In fact, software testing is sometimes considered an art [121] requiring its own technical and non-technical skills [113, 149], and entire career paths are built around testing [42]. *Test-driven development*, often associated with agile software engineering frameworks, integrates testing considerations in all parts of the development process [8, 65]. These processes rely on a deep understanding of software requirements and user behavior to anticipate failure modes during deployment and to expand the test suite. (In contrast, ML testing is often relegated to a small portion of the ML development cycle, and predominantly focuses on a static snapshot of data to provide performance guarantees.) These software testing methodologies provide a model for ML testing. First, the model suggests anticipating, planning for, and integrating testing in all stages of the development cycle, research problem ideation, the setting of objectives, and system implementation. Second, build a practice around bringing diverse perspectives into designing the test suite. Additionally, consider participatory approaches (e.g., [112]) to ensure that the test suite accounts for societal contexts and embedded values within which the ML system will be deployed.

An important principle in test-driven software development is visibility into the test data. Typically, engineers working on a system can not only see the details of the test suites but also often develop those test suites themselves. In contrast, the paradigm of ML evaluation methodologies is that the ML practitioner should not inspect the test data, lest their observations result in design decisions that produce an overfitted model. How, then, can these two methodologies be reconciled? We believe that incentives are one important consideration. In the ML research community, the "competition mindset" might indeed lead to "cheating" via deliberate overfitting. In contrast, in real-world applications model developers might benefit from a healthy model ecosystem, for example when they are members of that ecosystem. (However, when developers come from a different society altogether there may be disinterest or disalignment [146].)

Software testing produces artifacts such as execution traces, and test coverage information [72]. Developing practices for routinely sharing testing artifacts with stakeholders provides for more robust scrutiny and diagnosis of harmful error cases [137]. In being flexible enough to adapt to the information needs of stakeholders, software testing artifacts can be considered a form of boundary object [159]. Within an ML context, these considerations point towards adopting ML transparency mechanisms incorporating comprehensive evaluations, such as model cards [118]. The processes that go into building test cases should be documented, so the consumer of the ML system can better understand the system's reliability. Finally, as for any high-stakes system—software, ML or otherwise—evaluation documentation constitutes an important part of the chain of auditable artifacts required for robust accountability and governance practices [137].

6 CONCLUSIONS

In this paper, we compared the evaluation practices in the ML research community to the ideal information needs of those who use models in real-world applications. The observed disconnect between the two is likely due to differences in motivations and goals, and also pressures to demonstrate "state of the art" performance on shared tasks, metrics and leaderboards [51, 94, 162], as well as a focus on the learner as the object upon which the researcher hopes to shed light. One limitation of our methodology is reliance on published papers, and we encourage more human subjects research in the future, in a similar vein to e.g. [78, 109, 148]. We identified a range of evaluation gaps that risk being overlooked if the ML research community's evaluation practices are uncritically adopted when for applications, and identify six assumptions that would have to be valid if these gaps are to be overlooked. The assumptions range from a broad focus on consequentialism to technical concerns regarding distributions of evaluation data. By presenting these assumptions as a coherent framework, we provide not just a set of mitigations for each evaluation gap, but also demonstrate the relationships between these mitigations. We show how in the naive case these assumptions chain together, leading to the grossest assumption that calculating model accuracy on data I.I.D. with the training data can be a reliable signal for real-world applications. We contrast the practices of ML model evaluation with those of the mature engineering practices of software testing to draw out lessons for non-I.I.D. testing under a variety of stress conditions and failure severities. One limitation of our analysis is that we are generally domain-agnostic, and we hope to stimulate investigations of assumptions and gaps for specific application domains. We believe it is fundamental that model developers are explicit about methodological assumptions in their evaluations. We believe that ML model evaluations have great potential to enable interpretation and use by different technical and non-technical communities [159].

By naming each assumption we identify and exploring its technical and sociological consequences, we hope to encourage more robust interdisciplinary debate and, ultimately, to nudge model evaluation practice away from abundant opaque unknowns.

ACKNOWLEDGMENTS

We acknowledge useful feedback from Daniel J. Barrett, Alexander D'Amour, Stephen Pfohl, D. Sculley, and the anonymous reviewers.

REFERENCES

[1] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain Adaptation with Adversarial Training and Graph Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1077–1087.

[2] Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 107–112.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[4] Adam J Andreotta, Nin Kirkham, and Marco Rizzi. 2021. AI, big data, and the future of consent. *AI & Society* (2021), 1–14.

[5] McKane Andrus and Thomas K Gilbert. 2019. Towards a just theory of measurement: A principled social measurement assurance program for machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 445–451.

[6] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 249–260.

[7] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24.

[8] Dave Astels. 2003. *Test driven development: A practical guide*. Prentice Hall Professional Technical Reference.

[9] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. *arXiv preprint arXiv:2103.06076* (2021).

[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NIPS tutorial* 1 (2017), 2017.

[11] Roland Barthes. 1977. *Image-Music-Text*. Macmillan.

[12] Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *arXiv preprint arXiv:2109.04270* (2021).

[13] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[14] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[15] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. 2021. Deep learning for AI. *Commun. ACM* 64, 7 (2021), 58–65.

[16] John Berger. 2008. *Ways of seeing*. Penguin UK.

[17] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. PMLR, 149–159.

[18] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590* (2021).

[19] Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic Inequalities in Language Technology Performance across the World's Languages. *arXiv preprint arXiv:2110.06733* (2021).

[20] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[21] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*. 491–500.

[22] Samuel Bowman and George Dahl. 2021. What Will it Take to Fix Benchmarking in Natural Language Understanding?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4843–4855.

[23] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D Sculley. 2017. The ML test score: A rubric for ML production readiness and technical debt reduction. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1123–1132.

[24] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.

[25] Benjamin Brewster. 1881. *The Yale Literary Magazine* October 1881–June 1882 (1881).

[26] William Bulleit, Jon Schmidt, Irfan Alvi, Erik Nelson, and Tonatiuh Rodriguez-Nikl. 2015. Philosophy of engineering: What it is and why it matters. *Journal of Professional Issues in Engineering Education and Practice* 141, 3 (2015), 02514003.

[27] Razvan Bunescu and Yunfeng Huang. 2010. A utility-driven approach to question ranking in social QA. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING 2010)*. 125–133.

[28] Ermao Cai, Da-Cheng Juan, Dimitrios Stamoulis, and Diana Marculescu. 2017. NeuralPower: Predict and deploy energy-efficient convolutional neural networks. In *Asian Conference on Machine Learning*. PMLR, 622–637.

[29] Dallas Card and Noah A Smith. 2020. On Consequentialism and Fairness. *Frontiers in Artificial Intelligence* 3 (2020), 34.

[30] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.

[31] Brandon Carter, Siddhartha Jain, Jonas W Mueller, and David Gifford. 2021. Overinterpretation reveals image classification model pathologies. *Advances in Neural Information Processing Systems* 34 (2021).

[32] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28, 3 (2019), 231–237.

[33] James I Charlton. 1998. *Nothing about us without us*. University of California Press.

[34] Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. 2021. Mandoline: Model Evaluation under Distribution Shift. In *International Conference on Machine Learning*. PMLR, 1617–1629.

[35] Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. He said "who's gonna take care of your children when you are at ACL?": Reported Sexist Acts are Not Sexist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4055–4066.

[36] Alex Chohlas-Wood, Madison Coots, Emma Brunskill, and Sharad Goel. 2021. Learning to be Fair: A Consequentialist Approach to Equitable Decision-Making. *arXiv preprint arXiv:2109.08792* (2021).

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

[38] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[39] 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.

[40] Kate Crawford and Vladan Joler. 2018. Anatomy of an AI System. (Accessed January, 2022).

[41] Kate Crawford and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY* (2021), 1–12.

[42] Sean Cunningham, Jemil Gambo, Aidan Lawless, Declan Moore, Murat Yilmaz, Paul M Clarke, and Rory V O'Connor. 2019. Software testing: a changing career. In *European Conference on Software Process Improvement*. Springer, 731–742.

[43] Emma Dahlin. 2021. Mind the gap! On the future of AI research. *Humanities and Social Sciences Communications* 8, 1 (2021), 1–4.

[44] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[45] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics* 10 (2022), 92–110.

[46] Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435* (2020).

[47] Leon Derczynski. 2016. Complementarity, F-score, and NLP Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 261–266.

[48] Mark Díaz and Nicholas Diakopoulos. 2019. Whose walkability?: Challenges in algorithmically measuring subjective experience. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–22.

[49] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior* 46, 2 (2019), 185–209.

[50] Ulle Endriss. 2018. Lecture notes on fair division. *arXiv preprint arXiv:1806.04234* (2018).

[51] Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4846–4853.

[52] Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. *arXiv preprint arXiv:1711.01505* (2017).

[53] Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael Curtis Mozer. 2021. Head2Toe: Utilizing Intermediate Representations for Better OOD Generalization. (2021).

[54] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. 2020. A brief review of domain adaptation. *arXiv preprint arXiv:2010.03978* (2020).

[55] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2591–2597.

[56] Diana Forsythe. 2001. *Studying those who study us: An anthropologist in the world of Artificial Intelligence*. Stanford University Press, Chapter Artificial intelligence invents itself: Collective identity and boundary maintenance in an emergent scientific discipline.

[57] Diana Forsythe. 2001. *Studying those who study us: An anthropologist in the world of Artificial Intelligence*. Stanford University Press, Chapter The Construction of Knowledge in Artificial Intelligence.

[58] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[59] Biying Fu, Cong Chen, Olaf Henniger, and Naser Damer. 2022. A deep insight into measuring face image utility with general and face-specific image quality metrics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 905–914.

[60] Eva García-Martín, Crefeda Faviola Rodrigues, Graham Riley, and Håkan Grahn. 2019. Estimation of energy consumption in machine learning. *J. Parallel and Distrib. Comput.* 134 (2019), 75–88.

[61] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 219–226.

[62] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[63] Clifford Geertz. 1973. *The Interpretation of Cultures*. Basic Books.

[64] Susi Geiger. 2020. Silicon Valley, disruption, and the end of uncertainty. *Journal of cultural economy* 13, 2 (2020), 169–184.

[65] Boby George and Laurie Williams. 2004. A structured experiment of test-driven development. *Information and software Technology* 46, 5 (2004), 337–342.

[66] Hila Gonen and Kellie Webster. 2020. Automatically Identifying Gender Issues in Machine Translation using Perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1991–1995.

[67] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

[68] Ben Green. 2020. Data science as political action: grounding data science in a politics of justice. *Available at SSRN 3658431* (2020).

[69] Huong Ha, Sunil Gupta, Santu Rana, and Svetha Venkatesh. 2021. ALT-MAS: A Data-Efficient Framework for Active Testing of Machine Learning Algorithms. *arXiv preprint arXiv:2104.04999* (2021).

[70] Frank Hampel and Eth Zurich. 1998. Is statistics too difficult? *Canadian Journal of Statistics* 26, 3 (1998), 497–513.

[71] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[72] Mary Jean Harrold. 2000. Testing: a roadmap. In *Proceedings of the Conference on the Future of Software Engineering*. 61–72.

[73] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[74] Courtney Heldreth, Michal Lahav, Zion Mengesha, Juliana Sublewski, and Elyse Tuennerman. 2021. "I don't think these devices are very culturally sensitive."—The impact of errors on African Americans in Automated Speech Recognition. *Frontiers in Artificial Intelligence* 26 (2021).

[75] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* 21, 248 (2020), 1–43.

[76] Dan Hendrycks and Thomas Dietterich. 2018. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

[77] Benjamin Hepp, Debadeepta Dey, Sudipta N Sinha, Ashish Kapoor, Neel Joshi, and Otmar Hilliges. 2018. Learn-to-score: Efficient 3D scene exploration by predicting view utility. In *Proceedings of the European conference on computer vision (ECCV)*. 437–452.

[78] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[79] John N Hooker. 1995. Testing heuristics: We have it all wrong. *Journal of heuristics* 1, 1 (1995), 33–42.

[80] Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 591–598.

[81] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 65–83.

[82] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 49–58.

[83] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.

[84] Maximilian Idahl, Lijun Lyu, Ujwal Gadiraju, and Avishek Anand. 2021. Towards Benchmarking the Utility of Explanations for Model Debugging. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*. 68–73.

[85] IEEE. 2019. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. "Classical Ethics in A/IS". In *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition*. 36–67.

[86] Abigail Z Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. The meaning and measurement of bias: lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 706–706.

[87] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 375–385.

[88] Yasamin Jafarian and Hyun Soo Park. 2021. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12753–12762.

[89] Nathalie Japkowicz. 2006. Why question machine learning evaluation methods. In *AAAI workshop on evaluation methods for machine learning*. 6–11.

[90] Tony Jappy. 2013. *Introduction to Peircean visual semiotics*. A&C Black.

[91] Disi Ji, Padhraic Smyth, and Mark Steyvers. 2020. Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. *Advances in Neural Information Processing Systems* 33 (2020), 18600–18612.

[92] Karen Sparck Jones and Julia R Galliers. 1995. *Evaluating natural language processing systems: An analysis and review*. Vol. 1083. Springer Science & Business Media.

[93] Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 240–248.

[94] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *NeurIPS Dataset & Benchmark track* (2021).

[95] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *CoRR* abs/2012.07421 (2020). https://arxiv.org/abs/2012.07421

[96] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. 2021. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*. PMLR, 5753–5763.

[97] Tahu Kukutai and John Taylor. 2016. *Indigenous data sovereignty: Toward an agenda*. ANU press.

[98] Hiroshi Kuwajima, Hirotoshi Yasuoka, and Toshihiro Nakae. 2020. Engineering problems in machine learning systems. *Machine Learning* 109, 5 (2020), 1103–1126.

[99] Alexandre Lacoste, Thomas Boquet, Negar Rostamzadeh, Boris Oreshkin, Wonchang Chung, and David Krueger. 2017. Deep prior. *arXiv preprint arXiv:1712.05016* (2017).

[100] Alexandre Lacoste, Boris Oreshkin, Wonchang Chung, Thomas Boquet, Negar Rostamzadeh, and David Krueger. 2018. Uncertainty in multitask transfer learning. *arXiv preprint arXiv:1806.07528* (2018).

[101] George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

[102] Guillaume Lecué and Matthieu Lerasle. 2020. Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics* 48, 2 (2020), 906–931.

[103] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4641–4650.

[104] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[105] Chien-Hsin Lin, Hsin-Yu Shih, and Peter J Sher. 2007. Integrating technology readiness into technology acceptance: The TRAM model. *Psychology & Marketing* 24, 7 (2007), 641–657.

[106] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[107] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158.

[108] Chi-kiu Lo and Dekai Wu. 2010. Evaluating Machine Translation Utility via Semantic Role Labels.. In *LREC*. Citeseer.

[109] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[110] Alessandro Mantelero. 2018. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review* 34, 4 (2018), 754–772.

[111] Marrkula Center. 2019. Approaches to Ethical Decision-making. https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/

[112] Donald Martin, Jr., Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. 2020. Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. arXiv:2006.09663 [cs.CY]

[113] Gerardo Matturro. 2013. Soft skills in software engineering: A study of its demand by software companies in Uruguay. In *2013 6th international workshop on cooperative and human aspects of software engineering (CHASE)*. IEEE, 133–136.

[114] Fulvio Mazzocchi. 2015. Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO reports* 16, 10 (2015), 1250–1255.

[115] Lorna McGregor, Daragh Murray, and Vivian Ng. 2019. International human rights law as a framework for algorithmic accountability. *International & Comparative Law Quarterly* 68, 2 (2019), 309–343.

[116] Douglas S McNair. 2018. Preventing disparities: Bayesian and frequentist methods for assessing fairness in machine learning decision-support models. *New Insights into Bayesian Inference* (2018), 71.

[117] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 161–172.

[118] 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*. 220–229.

[119] Milad Moradi and Matthias Samwald. 2021. Evaluating the Robustness of Neural Language Models to Input Perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1558–1570.

[120] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. 2018. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4500–4509.

[121] Glenford J Myers, Corey Sandler, and Tom Badgett. 2011. *The art of software testing*. John Wiley & Sons.

[122] Michael Neumann, Oliver Roessler, David Suendermann-Oeft, and Vikram Ramanarayanan. 2020. On the utility of audiovisual dialog technologies and signal analytics for real-time remote monitoring of depression biomarkers. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*. 47–52.

[123] Peter Norvig. 2017. On Chomsky and the two cultures of statistical learning. In *Berechenbarkeit der Welt?* Springer, 61–83.

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

[125] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. 2018. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8466–8475.

[126] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[127] Pedro O Pinheiro, Negar Rostamzadeh, and Sungjin Ahn. 2019. Domain-adaptive single-view 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7638–7647.

[128] David Martin Ward Powers. 2011. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* 2, 1 (2011), 37–63.

[129] David Martin Ward Powers. 2012. The problem of area under the curve. In *2012 IEEE International conference on information science and technology*. IEEE, 567–573.

[130] David Martin Ward Powers. 2012. The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 345–355.

[131] David Martin Ward Powers. 2014. What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. *Technical report, Beijing University of Technology, China & Flinders University, Australia, Tech. Rep.* (2014).

[132] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*. 133–138.

[133] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5740–5745.

[134] Foster Provost and Tom Fawcett. 1997. Analysis and visualization of classifier performance with nonuniform class and cost distributions. In *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*. 57–63.

[135] James Pustejovsky. 1998. *The generative lexicon*. MIT press.

[136] Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[137] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[138] Filippo A Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. 2018. Artificial intelligence & human rights: Opportunities & risks. *Berkman Klein Center Research Publication* 2018-6 (2018).

[139] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[140] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).

[141] Shalaleh Rismani and Ajung Moon. 2021. How do AI systems fail socially?: an engineering risk analysis approach. In *2021 IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)*. 1–8. https://doi.org/10.1109/ETHICS53270.2021.9632769

[142] Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation Examples are not Equally Informative: How should that change NLP Leaderboards?. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4486–4503. https://doi.org/10.18653/v1/2021.acl-long.346

[143] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. 2018. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317* (2018).

[144] Negar Rostamzadeh, Ben Hutchinson, Christina Greer, and Vinodkumar Prabhakaran. 2021. Thinking Beyond Distributions in Testing Machine Learned Models. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

[145] Nataniel Ruiz, Adam Kortylewski, Weichao Qiu, Cihang Xie, Sarah Adel Bargal, Alan Yuille, and Stan Sclaroff. 2022. Simulated Adversarial Testing of Face Recognition Models. *CVPR* (2022).

[146] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-Imagining Algorithmic Fairness in India and Beyond.

In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 315–328. https://doi.org/10.1145/3442188.3445896

[147] Nithya Sambasivan and Jess Holbrook. 2018. Toward responsible AI for the next billion users. *Interactions* 26, 1 (2018), 68–71.

[148] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[149] Mary Sánchez-Gordón, Laxmi Rijal, and Ricardo Colomo-Palacios. 2020. Beyond Technical Skills in Software Testing: Automated versus Manual Testing. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*. 161–164.

[150] Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147.

[151] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[152] David Schlangen. 2021. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 670–674.

[153] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM* 63, 12 (2020), 54–63.

[154] David Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's curse? On pace, progress, and empirical rigor. In *Proceedings of ICLR 2018*.

[155] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[156] Henry Shue. 2020. *Basic rights: Subsistence, affluence, and US foreign policy*. Princeton University Press.

[157] M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM* 61, 3 (2018), 39–41.

[158] Walter Sinnott-Armstrong. 2021. Consequentialism. *The Stanford Encyclopedia of Philosophy* Winter 2021 Edition (2021). https://plato.stanford.edu/archives/win2021/entries/consequentialism/

[159] Susan Leigh Star and James R Griesemer. 1989. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3 (1989), 387–420.

[160] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3645–3650.

[161] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. 2007. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. *Advances in Neural Information Processing Systems* 20 (2007).

[162] RL Thomas and D Uminsky. 2020. Reliance on metrics is a fundamental challenge for AI. In *Proceedings of the Ethics of Data Science Conference*.

[163] John W Tukey. 1962. The future of data analysis. *The annals of mathematical statistics* 33, 1 (1962), 1–67.

[164] Peter D Turney. 1994. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of artificial intelligence research* 2 (1994), 369–409.

[165] Dmitry Ustalov, Alexander Panchenko, and Chris Biemann. 2017. Watset: Automatic induction of synsets from a graph of synonyms. In *55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*. Association for Computational Linguistics, 1579–1590.

[166] Shannon Vallor. 2016. *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

[167] Cornelis Joost Van Rijsbergen. 1974. Foundation of evaluation. *Journal of documentation* (1974).

[168] Andreas Vogelsang and Markus Borg. 2019. Requirements engineering for machine learning: Perspectives from data scientists. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 245–251.

[169] Hanna Wallach. 2018. Computational social science≠ computer science+ social data. *Commun. ACM* 61, 3 (2018), 42–44.

[170] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 353–355.

[171] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. 2021. Generalizing to Unseen Domains: A Survey on Domain Generalization. In *Proceedings of IJCAI 2021*.

[172] Kellie Webster, Marta R Costa-jussà, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (GAP) shared task at the Gender Bias in NLP Workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 1–7.

[173] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems. *AI Now* (2019).

[174] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. 2020. Contrastive Training for Improved Out-of-Distribution Detection. *arXiv e-prints* (2020), arXiv–2007.

[175] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11307–11317.

[176] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 268–282.

[177] Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. OpenAttack: An Open-source Textual Adversarial Attack Toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. 363–371.

[178] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* (2020).

[179] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 3 (2020), 1–41.

[180] Benjamin Zi Hao Zhao, Mohamed Ali Kaafar, and Nicolas Kourtellis. 2020. Not one but many tradeoffs: Privacy vs. utility in differentially private machine learning. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*. 15–26.

[181] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2.

APPENDIX A: METRICS IN ML MODEL EVALUATIONS

Here we give definitions and categorizations of some of the metrics reported in the study in Section 3. In practice, there was a long tail since many metrics were used in only a single paper. Here we include only the metrics which were most frequently observed in our study.

Metric	Example Task(s)	Metric category	Definition
Accuracy	Classification	Accuracy	A metric that penalizes system predictions that do not agree with the reference data ($\frac{TP+TN}{TP+TN+FP+FN}$).
AUC	Classification	AUC	The area under the curve parameterized by classification threshold t, typically with y-axis representing recall and x-axis representing false positive rate ($\frac{FP}{FP+TN}$).
Bleu	Machine translation	Precision	A form of "n-gram precision," originally designed for machine translation but also sometimes used for other text generation tasks, which measures whether sequences of words in the system output are also present in the reference texts [126].
Dice	Image segmentation	Overlap	Equivalent to F_1 ($\frac{2TP}{2TP+FP+FN}$). More commonly used for medical image segmentation.
Error rate	Classification	Accuracy	The inverse of accuracy ($1 - accuracy = \frac{FP+FN}{TP+TN+FP+FN}$).
F (or F_1)	Text classification	Overlap	The harmonic mean of recall and precision ($\frac{2PR}{P+R}$), originally developed for information retrieval [167] but now widely used in NLP.
$F_{0.5}$	Text classification	Overlap	A weighted harmonic mean of recall and precision, with greater weight given to recall ($(1+\beta^2)\frac{PR}{\beta^2 P+R}$ with $\beta = 0.5$).
Hausdorff distance	Medical Image Segmentation	Distance	A measure of distance between two sets in a metric space. Two sets have a low Hausdorff distance if every point in each set is close to a point in the other set.
IoU	Image segmentation	Overlap	$\frac{TP}{TP+FP+FN}$. Equivalent to Jaccard.
Matthew's Correlation Coefficient		Correlation	Has been argued to address shortcomings in F_1's asymmetry with respect to classes ($\frac{TP.TN-FP.FN}{\sqrt{(TP+TN)(TP+FN)(TN+FN)(TN+FP)}}$).
Mean absolute error	Regression	Distance	$\frac{1}{N}\sum_{i=1}^{N}\lvert\hat{y}_i - y_i\rvert$
Mean Average Precision (MAP)	Information retrieval (NLP)	AUC	In information retrieval, the average over information needs of the average precision of the documents retrieved for that need.
Mean average precision (mAP)	Object detection (CV)	AUC	The area under the Precision-Recall tradeoff curve, averaged over multiple IoU (intersection over union) threshold values, then averaged across all categories (https://cocodataset.org/#detection-eval).
Mean reciprocal rank	Information retrieval	Other	A measure for evaluating processes that produces an ordered list of possible responses. The average of the inverse rank of the first relevant item retrieved.
MSE	Image Decomposition	Distance	Mean squared error (MSE) measures the average of the squared difference between estimated and actual values.
Normalized Discounted Cumulative Gain (NDCG)	Recommendation or ranking tasks	Other	A measure of ranking quality which takes into account the usefulness of items based on their ranking in the result list.

Metric	Example Task(s)	Metric category	Definition
Pearson's r	Quality Estimation	Correlation	A measure of linear correlation between two sets of data.
Perplexity	Language modeling	Perplexity	Information-theoretic metric (measured in bits-per-unit, e.g., bits-per-character or bits-per-sentence) often used for language models, inversely related to the probability assigned to the test data by the model. Closely related to the cross-entropy between the model and the test data. Can be thought of as how efficiently does the language model encode the test data.
Precision	Classification	Precision	A metric that penalizes the system for predicting a class (if class is unspecified, by default the "positive" class) when the reference data did not belong to this class ($\frac{TP}{TP+FP}$).
PSNR	Super Resolution	Distance	Peak Signal-to-Noise ratio (PSNR) is the ratio between the maximum possible value of a signal and the power of distorting noise (Mean Squared Error) that impacts the quality of its representation.
Recall	Classification	Recall	Also known as "sensitivity", this metric that penalizes the system for failing to predict a class (if class is unspecified, by default the "positive" class) when the reference data did belong to this class ($\frac{TP}{TP+FN}$); a.k.a. true positive rate.
RMSE	Depth Estimation	Distance	Root Mean Square Error (RMSE) is the square root of the MSE.
Rouge	Text summarization	Recall	A form of "n-gram recall," originally designed for text summarization but also sometimes used for other text generation tasks, which measures whether sequences of words in the reference texts are also present in the system output[106].
Spearman's ρ	Graph Edit Distance	Correlation	A measure of monotonic association between two variables–less restrictive than linear correlations.
Specificity	Classification	Other	Like Precision, this metric that penalizes the system for failing to predict a class (if class is unspecified, by default the "positive" class) when the reference data did belong to this class; unlike Precision it rewards true negatives rather than true positives ($\frac{TN}{TN+FN}$).
SSIM	Super Resolution	Distance	The Structural Similarity Method (SSIM) is a perception-based method for measuring the similarity between two images. The formula is based on comparison measurements of luminance, contrast, and structure.
Top-n accuracy	Face recognition	Accuracy	A metric for systems that return ranked lists, which calculates accuracy over the top n entries in each list.
Word error rate	Speech recognition	Accuracy	The inverse of word accuracy: $1 - \textit{word accuracy}$ (which is not technically always in $[0, 1]$ due to the way *word accuracy* is defined but which is categorized as "Accuracy" here because both insertions and deletions are penalized).

Table 5: Definitions and categorizations of metrics reported in Section 3. TP, TN, FP and FN indicate the number of true positives, true negatives, false positives and false negatives, respectively. y and $\hat{y}$ represent actual values and values predicted by the system, respectively.

APPENDIX B: TYPES OF EVALUATION DATA USED IN ML MODEL EVALUATIONS

Type of Test Data	Example Task(s)	I.I.D. with training data?	Definition
Test split	Classification	yes	Typically, labeled data is partitioned into training and test splits (and often a dev split too), drawn randomly from the same dataset.
Manual resource	Lexical acquisition	no	A manually compiled resource (in NLP, often a word-based resource such as a lexicon or thesaurus), against which knowledge acquired from a dataset is compared.
References	Machine translation	no	Reference outputs (typically obtained prior to building the system) which a generative system is trying to reproduce, typically obtained from humans (e.g., manual translations of input sentences in the case of evaluations using BLEU for machine translation tasks).
Training data	Keyword extraction	yes	Training data that contains labels is used to evaluate an unsupervised algorithm that did not have access to the labels during learning.
Novel distribution	Domain transfer	no	Test data that has the same form as the training data but is drawn from a different distribution (e.g., in the case of NLP training on labeled newspaper data and testing on labeled Wikipedia data).

Table 6: Types of datasets used in ML model evaluations.

APPENDIX C: EXAMPLE OF ASSUMPTIONS AND GAPS FOR A HYPOTHETICAL APPLICATION

Suppose we are evaluating a hypothetical image classification model for use in an application for assisting blind people in identifying groceries in their pantries. Then some application-specific questions related to the assumptions in Section 4 might be:

Consequentialism. Was data ethically sourced and labeled? Were blind people involved in the design process? Does this use of this model encourage high-risk uses of other similar models, such as identifying pharmaceutical products?

Abstractability from Context. Does the application have a human-in-the-loop feature available when the model is uncertain? Will the system nudge purchasing behaviors towards products on which the model performs well?

Input Myopia. Are uncommon grocery products misclassified more often? Does this disproportionately impact home cooks who don't stick to the dominant cuisines, or who have food requirements due to medical conditions?

Quantitative Modeling. Does measuring predictive accuracy fail to take into account dignitary consequences associated with increased independence? Should each user be weighted equally in the evaluation (cf. each image)?

Equivalent Failures. Are there severe risks in confusing certain pairs of products, e.g., food products associated with dangerous allergies? Are some errors only minimally inconvenient, such as confusing different shapes of pasta?

Test Data Validity. Is the evaluation data representative of what the application's users have in their pantries? Are the image qualities (lighting, focus, framing, etc.) representative of images taken by blind photographers?

APPENDIX D: MODEL EVALUATION REMITS AND DESIGN

Model Evaluation Remit
To establish:
motivation — why evaluate the model?
what is the perspective being adopted — task/financial/administrative/scientific/...
whose interests prompted the evaluation — developer/funder/...
who are the consumers of the model evaluation results — manager/user/researcher/...
goal — what do we want/need to discover?
orientation — intrinsic/extrinsic
kind — investigation/experiment
type — black box/glass box
form (of yardstick) — ideal/attainable/exemplar/given/judged
style — suggestive/indicative/exhaustive
mode — quantitative/qualitative/hybrid
Model Evaluation Design
To identify:
ends — what is the model for? what is its objective or function?
context — what is the ecosystem the model is in? what are the animate and inanimate factors?
constitution — what is the structure of the model? what was the training data?
To determine:
factors that will be tested
environment variables
'system' parameters
evaluation criteria
metrics/measures
methods
Evaluation data — what type, status and nature?
Evaluation procedure

Table 7: A sketch of how Karen Sparck Jones and Julia Galliers' 1995 NLP evaluation framework questionnaire [92] can be adapted for the evaluation of ML models. The output of the remit and the design is a strategy for conducting the model evaluation. For a related but simpler framework based on model requirements analysis, see also the "7-step Recipe" for NLP system evaluation (https://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html) developed by the EAGLES Evaluation Working Group in 1999, which considers whether different parties have a shared understanding of the evaluation's purpose.