Towards Generating Consumer Labels for Machine Learning Models

Christin Seifert
University of Twente
Enschede, The Netherlands
c.seifert@utwente.nl

Stefanie Scherzinger
OTH Regensburg
Regensburg, Germany
stefanie.scherzinger@oth-regenburg.de

Lena Wiese
Fraunhofer ITEM
Hannover, Germany
lena.wiese@item.fraunhofer.de

***Abstract*—Machine learning (ML) based decision making is becoming commonplace. For persons affected by ML-based decisions, a certain level of transparency regarding the properties of the underlying ML model can be fundamental. In this vision paper, we propose to issue consumer labels for pre-trained and published ML models. These labels primarily target machine learning lay persons, such as the operators of an ML system, the executors of decisions, and the decision subjects themselves. Provided that consumer labels comprehensively capture the characteristics of the trained ML model, consumers are enabled to recognize when human intelligence should supersede artificial intelligence. In the long run, we envision a service that generates these consumer labels (semi-)automatically. In this paper, we survey the requirements that an ML system should meet, and correspondingly, the properties that an ML consumer label could capture. We further discuss the feasibility of operationalizing and benchmarking these requirements in the automated generation of ML consumer labels.**

***Index Terms*—Artificial intelligence, machine learning, consumer labels, transparency, x-AI**

I. Introduction

In various domains, machine learning (ML) systems support day-to-day decision making. In creating, operating and executing such systems, humans take on different roles, as illustrated in Figure 2, where the authors of [1] distinguish six different roles. As an illustrative example, let us consider a system assisting the negotiation of property loans. There are the *creators* of the ML system providing a trained ML model. The creators, as well as the *external examiners* (e.g., auditors), are highly skilled professionals. They have insight into the data preparation process and know details about the *data subjects* (e.g., past receivers of loans), who constitute the training data.

A bank teller meeting with a potential customer (the *decision subject*), does not have the same background and insight as the creators and external examiners. As the mere *operator* of the system, the bank teller feeds in descriptive attributes about the decision subject (such as age, income, occupation, equity), and receives a recommendation. It is now up to the teller in the role of the *executor* to propose the conditions for a loan, based on this recommendation. Yet being able to challenge a recommendation requires a certain level of understanding of the automated decision making process.

Previous work suggests ideas for documentary material. Datasheets [2] describe the data subjects; Model Cards [3] describe the specific properties of a trained ML model. However, both are targeted at expert users, such as the creators of an ML system.

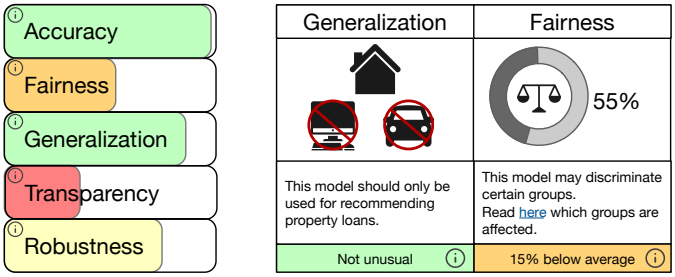

Fig. 1: Sketch of an ML consumer label for a loan prediction application. Left: general overview showing the degree to which certain properties are satisfied (percentages and color-coding), right: details on generalization ability and fairness.

In this vision paper, we build upon these ideas, envisioning that ML systems come complete with *ML consumer labels*, an example of which is sketched in Figure 1. We imagine consumer labels to become as commonplace as nutrition labels. Different from [3], where explainability is tailored towards all roles in a machine learning ecosystem (cf. Figure 2), we specifically focus on the consumers of ML-based recommendations. This concerns the operators, executors, and decision subjects. We further assess to which extent ML consumer labels can be *automatically generated*.

In Figure 1 (right), we sketch parts of an ML consumer label issued to our fictitious bank teller, conveying that

- the model does not generalize well to different data sets. Thus, the teller should not apply it out of context (e.g., for consumer loans instead of property loans). Further,
- since the system is not always fair, the teller will consult an expert for customers from disadvantaged groups.

The visualization in Figure 3 provides further detail. For the total population, approximately ¼th of all predictions are inaccurate (error cases); out of these, false positive and false negative rates (areas with different shades of red) are equally likely. For the specific group of freelancing journalists, the overall error rate is the same, while the percentage of false negatives is higher than that of false positives, indicating unfair treatment of this subgroup. Thus, a freelancing journalist might

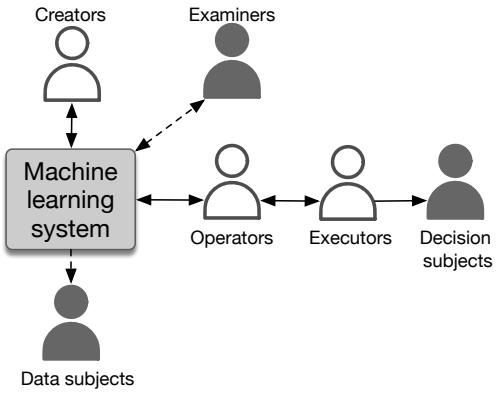

Fig. 2: Machine learning ecosystem as proposed in [1]. Image reproduced with permission by the authors.

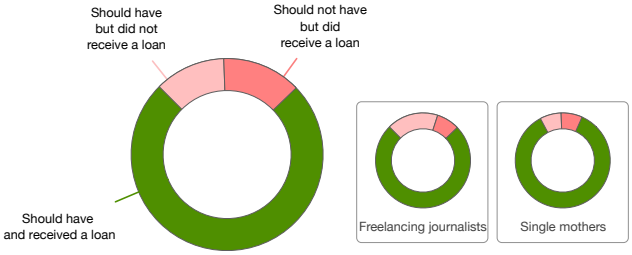

Fig. 3: Visualizing predictive performance for different groups.

see reason to challenge a denied loan. Regarding "single mothers", the algorithm has a lower error rate than for the general population, and the same error distribution.

Contributions: In this vision paper, we propose to establish ML consumer labels and review the requirements on ML models that these labels might capture. We even envision that ML consumer labels might be generated by third-party cloud services, and therefore discuss key properties of ML systems w.r.t. their potential for automated assessment. We compactly present our findings in a table, as a starting point for discussions and new research.

II. Related Work

In this section we review work on advanced ML properties and summarization reports for data and models. Some related work addresses scenarios where traditional metrics are insufficient for characterizing machine learning models. For example, [4], [5] propose to also consider properties like trust, transferability (robustness to adversarial examples), causality (models should find causal relations), informativeness (access to necessary information), usability, and fair and ethical decision making. Some of those subjective desiderata – such as trust, ethics, fairness, and social acceptance – are inherently not (objectively) quantifiable. Our proposed ML consumer labels contribute towards model interpretability for machine-learning lay persons and might act as enablers for such desiderata.

Other related work addresses summarization reports for data and models. The TRIPOD group presented a checklist for reporting multivariate prediction models in the medical domain, covering all stages from study design to data acquisition, to model development, and model update [6].

The work most similar to ours proposes Model Cards, a summarizing characterization of a machine learning model, covering aspects such as model details (e.g., neural network architecture type), intended use, evaluation metrics, ethical considerations and potential caveats [3]. While TRIPOD-conform reporting is targeted towards the scientific community, Model Cards address ML knowledgeable users and should accompany published, trained models. Complementary to TRIPOD and Model Cards, our consumer labels target machine learning lay persons, who are directly or indirectly affected by model decisions. This requires consumer labels to be comprehensive, truthful and understandable by our target audience.

Gebru et al. proposed Datasheets for Datasets [2], a standardized reporting schema for data sets in machine learning, including criteria such as the original motivation for data collection, as well as collection procedure, summary of content, and privacy considerations. Similarly, Bender and Friedman proposed Data Statements [7] specifically tailored toward data sets for natural language processing, adding for instance speaker characteristics. Both Datasheets and Data Statements are manually constructed (whereas we envision generating ML consumer labels semi-automatically).

Extending on Datasheets and Data Statements, the vision of Data Set Nutrition Labels [8] draws on the widely adapted nutrition labels. Data Set Nutrition Labels are automatically constructed from a data set. While our basic intention of communicating internals is the same, we focus on trained machine learning models (rather than data sets), We consider data set descriptions necessary and complementary information.

III. Demands on ML algorithms

We next survey classical requirements for ML models, as well as requirements that evolved more recently and that are also debated outside the ML research community. In Table I, we list selected requirements. We discuss these requirements in terms of capturing them in an ML consumer label.

Regarding the source of a requirement, we assess whether the requirement is functional (i.e., required for basic usage), is required by legislation, is a societal requirement, or depends on personal judgment. When operationalizing a requirement, we define a measurement of a phenomenon that may not be directly measurable. In the table, we track whether a requirement can be operationalized, and how it can be measured (quantitatively or qualitatively).

To give a example, the requirement that an ML model must have good predictive performance can be mapped to the property of having high accuracy. As accuracy can be measured quantitatively, this property can then be visualized, e.g., as shown in Figure 3.

		Predictive Performance	Computational Efficiency	Generalization	Online Learning Capability	Privacy Sensitivity	Robustness	Transparency /Interpretability /Explainability	Fairness	Trustworthiness	Accountability	Social acceptability	Morality
Source	**Functional**	✓	✓	✓	✓	(✓)	(✓)						
	Societal	✓				✓	✓	✓	✓			✓	✓
	Legal					✓	✓	✓	✓		✓		✓
	Personal	✓				✓	✓	✓	✓	✓		✓	✓
Definition	**Operational**	✓	✓	✓	✓	○	○	×	○	×	○	×	×
	Measure	qn	qn	ql	qn	qn	qn	ql	qn	ql	ql	qn	ql

TABLE I: Characterizing machine learning models along different requirements. In an operational definition, a mathematical formulation can be directly incorporated in model training (e.g. by adapting the loss function). ○ indicates near-operational definitions (e.g. multiple existing definitions). A cross indicates that we believe that operationalization is not possible, e.g. because of the subjective nature of the requirement. We categorize a measure as quantitative (qn) or qualitative (ql).

A. Classical Demands

a) Predictive Performance: Various metrics quantify the ability of an ML algorithm to learn or re-identify learned patterns. The metrics differ within communities and depend on both the task and the domain. For instance, for a classification problem, accuracy is a prominent metric. It conveys the percentage of correctly classified examples in a hold-out set, but it does not specify the types of errors made. For instance, in the medical domain, some errors can be more severe than others. Accordingly, metrics such as sensitivity and specificity are commonly used. For classification problems with multiple classes and class-imbalance, we may use different averaging strategies – either averaging over classes (macro-averaging) or instance-based (micro-averaging). Also commonly used are precision, recall, and F1 in information retrieval; area under curve (AOC) for medical classification problems; R^2 for regression problems; and error at top 5 in recent computer vision challenges [9]. All of them are in the range $[0, 1]$, and can therefore be represented as percentage values.

b) Computational Efficiency: With massive amounts of training data, computational efficiency in training ML models becomes relevant: Main memory resources must be managed carefully, the overall energy consumption has to be reduced. There is also research on designing special-purpose hardware (e.g. [10], [11]). In terms of capturing computational efficiency on a consumer label, we may turn to complexity theory (the big-O notation), or benchmark runtimes (for example, in milliseconds), main memory usage, and storage requirements (for example, in gigabytes).

c) Generalization: Being able to generalize across different inputs and application scenarios is a basic requirement for trained models [12]. This goal can be addressed during training, preventing overfitting via model selection, regularisation, or data augmentation. Trained models can be tested on various out-of-training data, as well as on cross-domain tasks. This indicates whether the model overfits (e.g., using learning curves), and to which types of input it generalizes to which extent. Such generalization results can be combined with knowledge about the data set (e.g. based on Data Sheets [2]) and integrated in a consumer label.

d) Online Learning Capabilities: In batch learning, models are trained once, remain static, and only adapt after retraining on the complete training data. In contrast, online learning refers to models that adapt instantly to new training samples [12]. Models learned in batch mode can simulate online capabilities, for instance if their training time is short, or if training is done in the background and new models are only deployed after training is complete. Thus, the property of practical interest from a user point of view is the *de facto* online capability of a model.

B. Recent Properties of Interest

a) Privacy-Sensitivity: Protecting sensitive personal data is a major requirement in operating ML systems. In terms of active attacks, the privacy of data subjects can be at risk in multi-party environments [13], [14]. For example, a publicly available trained model can actually leak sensitive information about the training data set to an attacker who is able to run several model executions on appropriately prepared adversarial data. In a different scenario, the trained machine learning model is considered an intellectual property. It should not be accessible to operators and executors who might identify hyperparameters in a model extraction attack. Various notions of privacy-preserving machine learning have been established, but they are usually limited to a specific machine learning task. In general, there is a trade-off between privacy demands and several other requirements (like transparency, fairness or even predictive performance when trained on distorted data). For example, a model may be trained on data containing sensitive attributes and then published to be used on other data sets – releasing the training data set for the sake of transparency is not possible in this case. Yet in certain cases (like k-anonymity or differential privacy), the privacy level can be quantified.

b) Robustness: While generalization is the ability of models to predict out-of training samples taken from the same underlying (unknown) population, robustness refers to model stability during adversarial attacks [14]. Attacks attempt to

misguide the model, for instance by adding noise to the input. In image classification, perturbed noisy images may not be humanly discernible, but nevertheless cause the ML model to misclassify. Examples are automatically generated graffiti-style changes to road signs [15] that lead to misclassification in a self-driving car setting. Especially, in scenarios where failure results in severe consequences, information about model robustness should be conveyed to end-users.

c) Transparency/Interpretability/Explainability: A more recent demand, also recognized by the EU General Data Protection Regulation (GDPR) [16], [17], is model interpretability. Miller defines interpretability as "the degree to which a human can understand the cause of a decision" [18]. Some ML model classes are inherently transparent and thus interpretable [4], such as decision rules [19]. For more complex models, post-hoc explanation strategies have been developed [20]. Interpretability is inherently subjective, as it depends on individual background knowledge. It is thus best evaluated in real-world settings [5]. Since these evaluations are costly, proxies have been proposed, e.g., the fidelity and accuracy of substitute models (how well they approximate the original model); or the model/explanation size, following the assumption that smaller models/explanations are easier to understand. A general advice is that an interpretable model should be used if possible [21]. For more complex models, decisions or general properties can be communicated to end-users by drawing on recent x-AI research results (see e.g., [20]).

d) Fairness: The discussion on algorithmic fairness has reached the public debate at the very latest in 2016 with two popular science books [22], [23]. Fairness in decision making states that decisions should not negatively impact subgroups of people. However, how to establish fairness in decision making is ongoing research. For instance, [24] state that if fairness is ensured for a subgroup defined by a single attribute (e.g. ethnicity="Hispanic"), it is not necessarily ensured for attribute combinations (e.g. elderly Hispanic women). Liu et al. [25] report 21 fairness criteria and show for two criteria that constraining on these fairness criteria may cause harm in the long(er) run. Unwanted algorithmic bias can be quantitatively measured – e.g. by the percentage the true positive/false negative rates deviate for the subgroup – and evaluation software is available [26]. However, fairness depends on the selected fairness criteria and to-be-protected subgroups, and might display a delayed impact that is not immediately quantifiable [25]. Conveying model fairness to end-users requires to select fairness criteria and define potentially discriminated subgroups, for which the fairness criteria are then evaluated.

e) Trustworthiness: The trustworthiness of a system influences user acceptance and behavior. For instance, the "willingness to accept a computer-generated recommendation" has been reported as an observable sign for trust [27]. Various studies suggest a diverse pattern of influence: there is evidence that system transparency might increase trust [28], or on the contrary, might be hindering [29]. Overall, the factors influencing trust are not yet fully explored. Since trust is conveyed to an agent (either automatic or human) by another agent, it is a subjective experience of an individual [30] and can only be measured on a per individual basis. Körber [31] developed a trust metric for automated systems based on a model of human-human trust. It consists of 19 self-report items measuring the trust factors reliability, predictability, the user's propensity to trust, as well as the attitude towards the system's engineers and the user's familiarity with automated systems. We argue that consumer labels are a means to increase trust in applications where automated decision making is applied.

f) Accountability: For automated decision making, there is a legal demand that decision making agents can be held accountable for their actions, similar to humans. Doshi-Velez et al. identified three tools to achieve algorithm accountability [32]: First, theoretical guarantees can be used in situations in which both, problem and outcome, can be fully formalized (e.g., as it is the case for encryption). Second, statistical evidence is suitable for situations where outcomes can be formalized and measured, but sufficient prior knowledge of the full problem is not available (e.g. some potential biases might only emerge through statistical aggregation). Third, for incompletely specified problems, explanations can be used as tools for accountability, a view that is also taken by the European Parliament Research Service [33]. However, as of today, there is no universal solution for governance of automated decision systems (for existing proposals see [33]). Conveying the accountability of ML models to end-users would require legislation and its operationalization. If both were available, ML models could be certified so that a consumer label can be equipped with with a certificate.

g) Social Acceptability: Social factors play an important role for technology adoption; the social acceptability of a technology might impact its uptake in the general population. The APA dictionary of psychology defines social acceptability as the "absence of social disapproval" [34] and it has been recognized as part of general system acceptability in usability research [35]. The lack of a clear definition [36] and the (yet) vague notions of influencing factors [37] makes social acceptability hard to quantify. While it has been measured using average scores derived from user questionnaires, the scales are specific to certain applications (e.g., wearable devices [38]). Future research is required to identify factors for socially-acceptable ML-models. We hypothesize that interpretability, trust and accountability play a role. Consumer labels could be a driving factor for socially-acceptable ML models.

h) Morality: Ubiquitous automated decision making faces decisions that are inherently value-laden. For instance, Freeman et al. incorporated moral concepts in their AI for deciding matches for a kidney donor exchange program [39]. Judgment of morality and matching behaviour varies across and within cultures and societies [40], an observation that has also been made based on 40 million decisions in the "moral machine experiment", where users where faced with moral dilemmas in the context of a hypothetical self-driving car [41]. Similarly as with social acceptability and trust, consumer labels could be a means for end-users to judge the morality of ML-based applications.

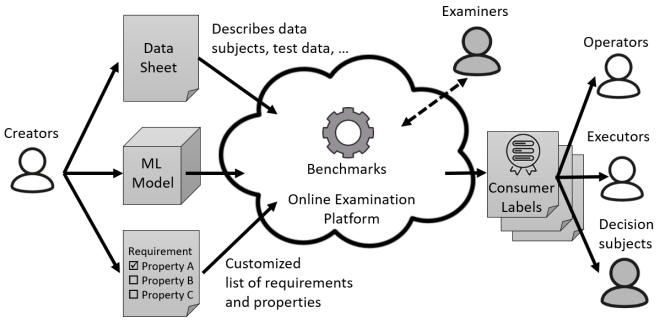

Fig. 4: Generating ML consumer labels as a service.

IV. Vision: Generating Consumer Labels

We propose to generate ML consumer labels automatically. This requires us to discuss the introduced requirements regarding their potential for automation.

a) Consumer labels as-a-service: Ideally, an ML consumer label is issued by a trusted third party, such as an independent examiner (cf. Figure 2). However, we may not be able to rely on humans alone to generate the consumer labels, there are simply too many algorithms, hyperparameters, and other factors influencing a model. Rather, we envision that consumer labels are generated (semi-)automatically. Figure 4 sketches our vision of such a service. The creators of the ML system provide (1) a data sheet that describes the data subjects in a summarized fashion; (2) the trained ML model; (3) a specification which requirements (and specific properties) should be captured by the consumer label.

The service then generates the consumer labels, customized for operators, executors, and decision subjects. While some of the requirements discussed lend themselves nicely to automation, others will require human examiners to intervene (cf. the examiners in Figure 4), as we will elaborate on shortly.

Today, cloud-hosted platforms for machine learning are quite popular, such as Cloud AutoML[1]: Many creators of ML systems are therefore accustomed to storing their data, as well as training their models, in the cloud. Offering cloud services that generate ML consumer labels is a potential next step. However, a major challenge in operating such a service is that when the model and the data sheet become public, we risk adversarial attacks, e.g. [15]. Thus, the owners of an ML system may prefer to download a certified labeling kit that they can run on premise; the challenge is then to design these kits such that they cannot be manipulated.

b) Potential for automation: From our point of view, four properties are straightforward to automate (cf. Table I);

Predictive Performance: Performance metrics can be computed when benchmark data is provided. Advanced settings like k-fold cross-validation can also be tested.

Generalization: How well a model performs on more general datasets can be tested by checking its predictive performance on similar benchmark data, or by automatically pre-processing the input data into a more general representation (for example, oversampling the minority class).

Computational Efficiency: Computational performance can be assessed by monitoring the system resources that the execution of the given model requires on average. This reveals whether the ML system can be run on a commodity PC, or requires dedicated hardware.

Online Learning Capacity: This property is a static property of the training algorithm (or the implementing library) and can therefore be directly assessed.

The evaluation of the following properties could be semi-automated but still requires human involvement – in terms of research, standardization, clarity in legislation, manual input, adaptation or assessment by an independent examiner – in order to enable the formalization of specific notions or definitions on which the ML consumer label should be based.

Privacy Sensitivity: Mechanisms for privacy-preservation in machine learning still require more in-depth analyses; there is no general solution that can be uniformly applied to all machine learning algorithms. However, in a semi-automated setting, we may select a particular privacy definition (such as differential privacy). Then, our cloud service can check whether the ML model complies.

Robustness: While benchmarks and defenses against adversarial attacks have been proposed (e.g., surveyed in [14]), it is difficult to assess for a model whether it is robust against all currently known adversarial attacks – a problem that is similar to attacks in the cyber-security field. However, given a list of attacks, our cloud service could generate appropriate adversarial data sets and automatically compute a robustness score.

Fairness: Fairness requires a precise specification of the underlying subgroups for which a bias-free decision is to be ensured. Moreover, as can be seen from our example in Figure 3, accuracy is not fine-grained enough to assess fairness and a more detailed analysis – e.g. of the false negatives – is necessary.

Accountability: In cases where it is possible to translate legal knowledge into a machine-understandable format, compliance with formally specified regulations can be checked for, in limited cases. However, this will not be possible for all applicable legal requirements.

The other four properties in Table I (transparency, trustworthiness, social acceptability and morality) are inherently subjective. To our best knowledge, we believe they should not be included in a certified consumer label at this point.

V. Conclusion

We call for a united effort on behalf of society, as well as representatives from politics and the judiciary system, to engage in the discussion on how to design and generate consumer labels for ML systems: Evidently, seemingly intuitive notions such as "fairness" and "interpretability" are difficult to capture formally. Yet these fuzzy concepts are the actual enablers for consumers to trust ML systems.

[1]https://cloud.google.com/automl/, accessed Oct. 2019

We hope that sketching our vision inspires academic peers to direct their research towards consumer labels for ML systems. These labels should be comprehensive, intuitive to understand, and – ideally – could even be generated in an automated fashion. Moreover, they will have to be adapted continuously towards changing requirements, e.g., in legislation.

References

[1] R. Tomsett, D. Braines, D. Harborne, A. D. Preece, and S. Chakraborty, "Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems," *CoRR*, vol. abs/1806.07552, 2018.

[2] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. D. III, and K. Crawford, "Datasheets for Datasets," *CoRR*, vol. abs/1803.09010, 2018.

[3] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, 2019, pp. 220–229.

[4] Z. C. Lipton, "The Mythos of Model Interpretability," *Queue*, vol. 16, no. 3, pp. 30:31–30:57, Jun. 2018.

[5] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *ArXiv e-prints*, Feb. 2017.

[6] K. G. Moons, D. G. Altman, J. B. Reitsma, J. P. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins, "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration," *Ann. Intern. Med.*, vol. 162, no. 1, pp. 1–73, Jan 2015.

[7] E. M. Bender and B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018.

[8] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards," *CoRR*, vol. abs/1805.03677, 2018.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, 2015.

[10] V. Sze, Y. Chen, T. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec 2017.

[11] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, no. 1, pp. 239 – 255, 2010.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc., 2008.

[13] M. Al-Rubaie and J. M. Chang, "Privacy-Preserving Machine Learning: Threats and Solutions," *IEEE Security & Privacy*, vol. 17, no. 2, pp. 49–58, 2019.

[14] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "Sok: Security and privacy in machine learning," in *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2018, pp. 399–414.

[15] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust Physical-World Attacks on Deep Learning Visual Classification," in *Conference on Computer Vision and Pattern Recognition*, June 2018.

[16] B. Goodman and S. R. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[17] S. Wachter, B. Mittelstad, and L. Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, Jun. 2017.

[18] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[19] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An Empirical Evaluation of the Comprehensibility of Decision Table, Tree and Rule Based Predictive Models," *Decis. Support Syst.*, vol. 51, no. 1, pp. 141–154, Apr. 2011.

[20] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 93:1–93:42, Aug. 2018.

[21] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, May 2019.

[22] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY, USA: Crown Publishing Group, 2016.

[23] S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, 2018.

[24] M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, 2018, pp. 2569–2577.

[25] L. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed Impact of Fair Machine Learning," in *Proceedings International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 3156–3164.

[26] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," *CoRR*, vol. abs/1810.01943, 2018.

[27] E. S. Vorm, "Assessing Demand for Transparency in Intelligent Systems Using Machine Learning," in *Innovations in Intelligent Systems and Applications (INISTA)*, 2018, pp. 1–7.

[28] A. Glass, D. L. McGuinness, and M. Wolverton, "Toward establishing trust in adaptive agents," in *Proceedings International Conference on Intelligent User Interfaces*. ACM, 2008, pp. 227–236.

[29] A. Papenmeier, G. Englebienne, and C. Seifert, "How model accuracy and explanation fidelity influence user trust in AI," in *IJCAI Workshop on Explainable Artificial Intelligence (X-AI)*, 2019.

[30] N. G. Mohammadi, S. Paulus, M. Bishr, A. Metzger, H. Könnecke, S. Hartenstein, T. Weyer, and K. Pohl, "Trustworthiness attributes and metrics for engineering trusted Internet-based software systems," in *International Conference on Cloud Computing and Services Science*, 2013, pp. 19–35.

[31] M. Körber, "Theoretical considerations and development of a questionnaire to measure trust in automation," in *Congress of the International Ergonomics Association*, 2018, pp. 13–30.

[32] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. J. Gershman, D. O'Brien, S. Shieber, J. Waldo, D. Weinberger, and A. Wood, "Accountability of AI Under the Law: The Role of Explanation," Berkman Center, Tech. Rep. 18-07, 2017.

[33] A. Koene, C. Clifton, Y. Hatada, H. Webb, and R. Richardson, "A governance framework for algorithmic accountability and transparency," European Parliamentary Research Service, Tech. Rep. PE 624.262, Apr. 2019.

[34] APA Dictionary of Psychology, "Social acceptance," 2019. [Online]. Available: https://dictionary.apa.org

[35] J. Nielsen, *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[36] C. S. Montero, J. Alexander, M. T. Marshall, and S. Subramanian, "Would You Do That?: Understanding Social Acceptance of Gestural Interfaces," in *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, ser. MobileHCI '10, 2010, pp. 275–278.

[37] M. Koelle, T. Olsson, R. Mitchell, J. Williamson, and S. Boll, "What is (Un)Acceptable?: Thoughts on Social Acceptability in HCI Research," *Interactions*, vol. 26, no. 3, pp. 36–40, Apr. 2019.

[38] N. Kelly and S. Gilbert, "The WEAR Scale: Developing a Measure of the Social Acceptability of a Wearable Device," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '16, 2016, pp. 2864–2871.

[39] R. Freedman, J. Schaich Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer, "Adapting a Kidney Exchange Algorithm to Align with Human Values," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18, 2018, pp. 115–115.

[40] J. Graham, P. Meindl, E. Beall, K. M. Johnson, and L. Zhang, "Cultural differences in moral judgment and behavior, across and within societies," *Current Opinion in Psychology*, vol. 8, pp. 125 – 130, 2016.

[41] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, p. 59, 2018.