Quantifying Dataset Trustworthiness from Labeling Bias Using Subjective Logic

Koffi Ismael Ouattara[1,2,*], Ioannis Krontiris[1], Theo Dimitrakos[1] and Frank Kargl[2]

[1] *Huawei Technologies Duesseldorf GmbH, Munich, Germany - Trustworthy Technology and Engineering Lab*

[2] *Ulm University, Ulm, Germany - Department of Distributed Systems*

Abstract

Labeling bias, arising from disagreement and inconsistencies among annotators, threatens the reliability of datasets and downstream AI models. This paper introduces a framework based on Subjective Logic (SL) that quantifies dataset trustworthiness by mapping inter-annotator responses into opinions consisting of trust (belief), distrust (disbelief), and uncertainty. The approach provides an interpretable representation of how annotation volume and disagreement influence trust at the instance level. Using CIFAR-10H, we illustrate how trust and uncertainty evolve with different numbers of annotations and parameter settings, and how annotator-level reliability can be estimated. These findings highlight the potential of SL to support responsible data curation, such as identifying controversial items for re-labeling and profiling annotator consistency. While our study focuses on image classification, the framework is general and can be extended to probabilistic and textual annotations. Benchmarking against alternative aggregation methods is left as future work.

Keywords

Trustworthy AI, Labeling bias, Subjective Logic, Dataset quality, Trust quantification, Annotation reliability

1. Introduction

Artificial intelligence (AI) systems increasingly support high-stakes decisions in healthcare, finance, and autonomous mobility. While research has advanced the trustworthiness of AI models through fairness, explainability, and privacy-preserving techniques, less attention has been paid to the trustworthiness of the training datasets that underpin those models. Empirical studies show that dataset-level issues such as sampling bias, label noise, and privacy vulnerabilities can erode fairness, robustness, and interpretability [1]. In safety-critical domains, even minor defects in collection or curation propagate to harmful outcomes. For example, evaluations of pedestrian-detection systems in autonomous driving report systematic miss-rate disparities, with notably higher miss rates for children and for pedestrians with darker skin tones. These gaps remain after controlling for lighting, occlusion, distance, and other scene-difficulty covariates, revealing dataset-driven inequities with direct safety implications. [2].

While such examples highlight distributional risk (coverage and sampling), labeling bias often appears together with class imbalance and fairness concerns. Labeling bias refers to systematic deviations in how data points are annotated, often arising from subjectivity, inconsistent guidelines, or annotator disagreement. Unlike sampling bias, which affects the dataset at the distribution level, labeling bias is inherently a local or entry-level phenomenon, since it manifests in the reliability of individual labels. Inter-Annotator Agreement (IAA) metrics such as Cohen's Kappa provide a statistical measure of consistency among annotators, capturing uncertainty coming from disagreement [3]. However, they fail to capture other dimensions, like annotator confidence and reliability. They neither represent the strength of evidence behind an agreed label (e.g., peaked vs. diffuse per-item label distributions) nor the trustworthiness of particular annotators.

These limitations motivate an uncertainty-aware representation that preserves per-item label distributions, quantifies residual doubt from limited or conflicting annotations, and accounts for annotator-

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

[*] Corresponding author.

koffi.ismael.ouattara@huawei.com (K. I. Ouattara); ioannis.krontiris@huawei.com (I. Krontiris); theo.dimitrakos@huawei.com (T. Dimitrakos); frank.kargl@uni-ulm.de (F. Kargl)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-16

specific reliability when aggregating multiple judgments. Such a representation should map observed annotation evidence into distinct components of support and contradiction for a decided label and behave predictably as annotation volume or disagreement thresholds change.

In this paper, we adopt Subjective Logic (SL) [4] to transform inter-annotator responses into per-instance trust opinions comprising belief, disbelief, and uncertainty. For a chosen label, votes and label distributions are treated as positive and negative evidence, with the remaining mass encoding uncertainty. This yields instance-level measures of dataset trustworthiness and a principled basis for estimating and weighting annotator reliability, providing an auditable bridge from raw annotation signals to downstream dataset assessment.

More specifically, our work is guided by the following research objectives:

- Derive per instance SL opinions from annotation signals and analyze their sensitivity to disagreement thresholds.
- Examine and investigate how opinions evolve when the number of annotators changes, and how this affects the balance of belief, disbelief, and uncertainty.
- Explore annotator level reliability as an additional signal for dataset quality management.

Contributions. This paper makes three contributions:

1. We formalize a mapping from inter annotator evidence to Subjective Logic opinions that characterize trust, distrust, and uncertainty at the data instance level.
2. We demonstrate the framework on the CIFAR 10H dataset, showing how opinions evolve with annotation.
3. We discuss how our framework can support responsible AI practices, including identifying items for re labeling and profiling annotator reliability.

Our focus in this work is to demonstrate the feasibility and interpretability of applying SL to labeling bias.

Summary and Paper Structure. We show that trust scores derived from SL provide finer-grained insights than majority voting alone. Our approach enables both per-instance trust estimation and annotator-level reliability profiling, offering actionable signals for dataset refinement and quality assurance. The remainder of this paper is organized as follows: Section 2 reviews related work and the foundations of Subjective Logic. Section 3 presents our methodology. Section 4 outlines the experimental setup. Section 5 presents our results and discusses their implications. Section 6 concludes the paper and identifies future directions.

Code Availability. All code and experiments are available at: https://github.com/Ouatt-Isma/-Trustworthiness-of-AI-Training-Dataset

2. Related Work

Ouattara et al. [5] introduced a general SL-based framework for assessing dataset trustworthiness, focusing on properties such as bias and representativeness. Their work formalized the concept of trust propositions and combined logical reasoning with evidence-based quantification. It also introduced two complementary methods to assess class imbalance in centralized and federated settings. This paper extends their methodology by shifting the focus from sampling bias to labeling bias, requiring fine-grained, per-instance analysis of trust.

Regarding labeling bias, recent work by Vasilakes et al. [6] introduced Subjective Logic Encodings (SLEs), which model inter-annotator disagreement using belief-disbelief-uncertainty triplets. Their formulation motivates our approach, which goes beyond traditional agreement metrics to compute per-instance trust opinions grounded in SL. A key difference, however, is that while SLEs treat the subjective opinion itself as the training label for models, we instead assume a standard label aggregation strategies (e.g., majority vote) to determine a consensus label and then assess the degree of trust in that decided label.

SL has also been used to evaluate trustworthiness in machine learning metrics and model behavior. Herd and Burton [7] applied SL to assess the reliability of ML metrics in safety-critical settings. Ouattara et al. [8] applied SL to model calibration. These works highlight the versatility of SL in encoding uncertainty across various AI system components. In contrast, our contribution centers on the label-level granularity and the modeling of annotator reliability within a unified trust framework.

Efforts to systematize dataset evaluation have also emerged through scorecard frameworks such as METRIC [1] and FRIES [9]. These provide expert-curated frameworks for data quality assessment but lack formal mechanisms to capture doubt or conflicting evidence. Other work on fairness and bias mitigation [10, 11] focuses on improving dataset balance or transparency but does not offer quantitative trust measures. Our approach complements these efforts by embedding uncertainty reasoning into the evaluation pipeline, enabling actionable insights into the reliability of labels and annotators alike.

Beyond global agreement metrics, a long line of work in crowd annotation modeling has developed statistical methods to infer latent ground truth. The seminal Dawid and Skene model [12] pioneered this direction by applying the EM (Expectation–Maximization) algorithm to estimate annotator-specific confusion matrices and infer the most likely true labels. This approach explicitly models observer error rates, but it collapses disagreement into a single consensus probability and does not preserve residual uncertainty.

Subsequent methods extended this framework. GLAD (Generative model of Labels, Abilities, and Difficulties) [13] jointly estimates annotator expertise, item difficulty, and latent true labels, showing robustness to noisy or adversarial annotators. By weighting labelers according to inferred ability and accounting for image difficulty, GLAD achieves more accurate consensus than majority voting. However, its outputs remain point estimates of labels and latent parameters, without a principled decomposition of belief and uncertainty. MACE (Multi-Annotator Competence Estimation) [14] instead focuses on identifying spammers and unreliable annotators in an unsupervised fashion. By introducing latent variables for spamming behavior and optimizing with EM or Variational Bayes, it estimates both the ground-truth labels and annotator trustworthiness. While effective in filtering low-quality annotators, MACE also ultimately reduces annotation signals to label probabilities and competence scores.

In contrast, our framework based on Subjective Logic explicitly retains *belief, disbelief, and uncertainty as first-class quantities*. Instead of just merging annotation evidence into a single label, we additionally compute interpretable trust opinions at the instance level, making both consensus and disagreement directly observable. This perspective complements prior work: while Dawid–Skene, GLAD, and MACE focus on inferring hidden ground truth, our approach emphasizes *auditable trust quantification*, enabling transparent assessment of label reliability and supporting downstream tasks such as dataset curation, targeted re-labeling, and annotator profiling within trustworthy AI pipelines.

3. Method

3.1. Subjective Logic preliminaries

Subjective Logic (SL) [4] is a probabilistic logic framework designed to support reasoning under uncertainty by explicitly modeling beliefs and uncertainty. A binomial opinion in SL is expressed as a tuple $\omega = (b, d, u, a)$, where b is belief, d is disbelief, u is uncertainty (with $b + d + u = 1$), and a is the base rate expressing the prior expectation in the absence of evidence. This structure enables nuanced trust assessments, especially when information is incomplete or conflicting.

To aggregate evidence from multiple sources or points of view, SL provides discounting and several fusion operators [15, 16]. SL offers multiple fusion operators, such as cumulative or averaging fusion, to aggregate trust opinions across sources. The cumulative fusion is used when sources are considered independent, and the averaging operator when they are dependent. Moreover, logical operators such as conjunction (AND), disjunction (OR), and negation are defined to combine opinions over complex propositions. These operators are particularly useful when assessing composite properties of trust, such as fairness, which may depend on multiple sub-properties.

3.2. Opinion construction from annotations

We adopt the four–step trust quantification methodology introduced in previous work [5], adapting it here to the context of *labeling bias*. Whereas sampling bias affects a dataset globally, labeling bias arises locally at the instance level. Each labeled instance is thus treated as a *trust proposition*, and its trust opinion is derived directly from inter–annotator agreement (IAA).

The methodology unfolds in four steps. We first *define the trust proposition*, such as assessing whether the dataset is free from bias. Next, we *decompose this proposition* into multiple atomic propositions, one of which addresses labeling. Focusing on this labeling proposition, we then *identify trust sources and collect evidence*, with one trust source capturing agreement based on annotator responses. Finally, we *quantify subjective opinions* by converting annotation evidence into Subjective Logic opinions, and, when multiple trust sources are present, we *aggregate these opinions* to obtain an overall trust assessment.

To quantify the opinion based on the agreement trust source, we assume the existence of a reference label for each instance (e.g., majority vote or adjudication). For each annotator i, we compute a divergence score d_i measuring the distance between the annotator's label and the reference label. Using a divergence threshold δ, we partition the evidence into *positive evidence* (r_x, the number of annotations with $d_i \leq \delta$), *negative evidence* (s_x, the number of annotations with $d_i > \delta$), and *uncertainty*, modeled either based on the amplitude of the divergence values or based on a prior informative weight W.

From this evidence, two quantification schemes can be applied. The **Constant–Uncertainty Quantification** [5] scheme fixes the uncertainty u and distributes the remaining mass proportionally:

$$\gamma = 1 - u, \quad b = \gamma \cdot r, \quad d = \gamma \cdot,$$

The **Baseline–Prior Quantification** [5] scheme follows the canonical Subjective Logic mapping from evidence counts to opinions:

$$b = \frac{r}{W + r + s}, \quad d = \frac{s}{W + r + s}, \quad u = \frac{W}{W + r + s},$$

where $W > 0$ (usually set to 2) is the prior weight that moderates how quickly uncertainty shrinks as more evidence accumulates.

This formulation enables each instance to be assigned a nuanced trust opinion, which can in turn be aggregated across instances to assess dataset–level trustworthiness. The overall reasoning process can be summarized as follows:

`(1) Trust Proposition` → `(2) Atomic Propositions` → `(3) Trust Sources + Evidence Collection` → `(4) SL Opinion quantification + Aggregated Trust.`

3.3. Examples

Consider an image in the CIFAR–10H dataset annotated by 50 crowdworkers. Suppose 46 annotators agree with the decision label y^* and 4 disagree. With a prior weight $W = 2$, the opinion is

$$b = \frac{46}{52} \approx 0.885, \quad d = \frac{4}{52} \approx 0.077, \quad u = \frac{2}{52} \approx 0.038.$$

This opinion indicates high belief in the correctness of the label, some degree of distrust due to disagreement, and low residual uncertainty. As the number of annotations decreases, uncertainty u increases according to $u = \frac{W}{W+n}$, where $n = r + s$.

3.4. Algorithmic summary

For clarity, Algorithm 1 summarizes the procedure.

Algorithm 1 Opinion construction from annotations

```
Require: Annotations $\{y_j(x)\}$ or $\{p_j(x)\}$, decision label $y^*$, threshold $\delta$, prior weight $W$
for each instance $x$ do
    $r_x \leftarrow |\{j : \text{divergence}(y_j(x), y^*) \leq \delta\}|$
    $s_x \leftarrow |\{j : \text{divergence}(y_j(x), y^*) > \delta\}|$
    $b_x \leftarrow r_x/(W + r_x + s_x)$
    $d_x \leftarrow s_x/(W + r_x + s_x)$
    $u_x \leftarrow W/(W + r_x + s_x)$
    Output opinion $\omega_x = (b_x, d_x, u_x, a)$
end for
```

4. Experimental Setup

4.1. Dataset

We evaluate the proposed framework on the CIFAR−10H dataset [17], which provides human annotation distributions for the CIFAR−10 test images. Each image is annotated by 50 crowdworkers, yielding a rich source of inter−annotator disagreement.

4.2. Evaluation Protocol

Each image is treated as a trust proposition. We assume the majority label as the decision label y^*, and compute divergence d_i for each annotator. With threshold $\delta = 0$, annotators who agree with y^* contribute to positive evidence r_x, and those who disagree contribute to negative evidence s_x. Uncertainty is controlled by the prior weight W that we set to 2.

To study the effect of annotation volume, we subsample annotations at different rates (5, 10, 20, 50 annotators per instance) while preserving the decision label. This allows us to examine how belief, disbelief, and uncertainty evolve as evidence increases.

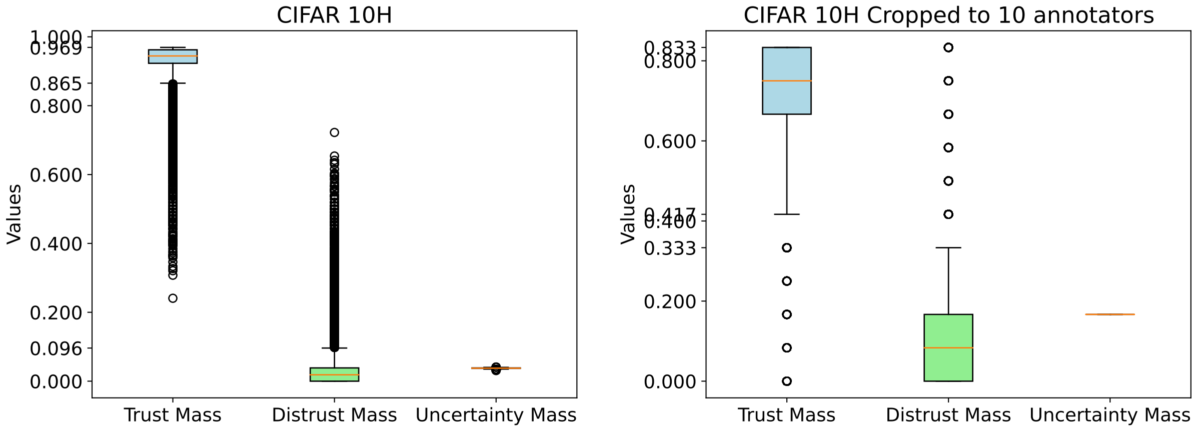
4.3. Evaluation Metrics

We report descriptive statistics of opinion masses across the dataset and analyze correlations between trust t, distrust d and uncertainty u against inter−annotator disagreement (measured as $1-$ majority proportion). This captures whether (t, d, u) reflects the true difficulty of instances. We also visualize the distribution of opinions using boxplots.

5. Results and Discussion

This section presents results from our CIFAR-10H experiments and explores the effects of annotation volume on trust scores. We also discuss how the proposed framework can be extended to richer annotation formats and used for annotator profiling.

Trust Mass Distribution (Figure 1). On the full CIFAR-10H dataset (left), using all available annotations per instance, we observe a high average trust mass, with most instances showing strong annotator agreement. However, this setting may inflate confidence for instances with disproportionately many annotations. After normalizing to 10 annotators per instance via majority-preserving subsampling (right), the relative weight of distrust and uncertainty increases, reducing the dominance of trust mass observed in the full dataset. This shift reflects the trade-off between annotation volume and confidence in label quality, as fewer annotations provide weaker evidence and make disagreement more visible.



(a) Results on the full CIFAR-10H dataset using all available annotators per instance.

(b) Results after cropping each instance to 10 annotators via majority-preserving subsampling.

Figure 1: Box plots of trust, distrust, and uncertainty masses for CIFAR-10H annotations. The full dataset exhibits high trust mass and low uncertainty, reflecting strong annotator agreement. In contrast, the cropped setting shows reduced trust, increased distrust, and higher uncertainty due to fewer annotations per instance. These results illustrate how annotation volume and redundancy affect perceived trustworthiness in labeling.

Annotations per item	Corr(t)	Avg(t)	Corr(d)	Avg(d)	Corr(u)	Avg(u)
5	-1.00	0.69	1.00	0.02	0.63	0.29
10	-1.00	0.80	1.00	0.03	0.18	0.17
20	-1.00	0.87	1.00	0.04	-0.38	0.09
50	-1.00	0.92	1.00	0.04	-0.00	0.04

Table 1
Correlation between SL trust (t), distrust (d), and uncertainty (u) masses and inter-annotator disagreement on CIFAR–10H. Average opinion masses are also reported across different annotation volumes.

Correlation with Disagreement (Table 1). We analyzed how Subjective Logic trust (t), distrust (d), and uncertainty (u) masses correlate with inter-annotator disagreement (defined as one minus the majority proportion). As shown in Table 1, trust mass consistently shows a strong negative correlation with disagreement, while distrust mass shows a strong positive correlation, reflecting their complementary roles. Uncertainty mass exhibits a weaker and less stable relationship: it is moderately positive with few annotations (5), but decreases as annotation volume grows and eventually approaches zero. At the same time, average trust increases steadily with more annotations, while both average distrust and uncertainty remain low. These results indicate that SL trust and distrust masses are reliable indicators of annotator disagreement across annotation volumes, whereas uncertainty mass is most informative in low-evidence settings.

Extension to Probabilistic and Natural Language Annotations. Our current analysis assumes categorical annotations. However, many modern datasets include probabilistic labels or free-text annotations. For probabilistic labels, divergences can be computed using KL divergence or cross-entropy between annotator distributions and the reference label. For text labels, semantic similarity measures (e.g., cosine similarity between embeddings) can provide divergence values relative to a gold ontology or adjudicated reference. These extensions broaden the applicability of our framework to more complex labeling scenarios.

Calibrating the Divergence Threshold δ. The divergence threshold δ determines whether an annotation is counted as positive or negative evidence. It can be calibrated empirically from high-confidence samples, set statistically (e.g., one standard deviation around the mean divergence), or fixed according to domain-specific policy (e.g., $\delta = 0$ for categorical labels, $\delta = 0.1$ for soft labels). This flexibility allows tailoring the framework to different annotation contexts.

Inferring Annotator reliability. The same approach as the one detailed in Section 3.2 can be extended to estimate annotator reliability. For annotator j, we partition the evidence into *positive evidence* (r_j, the number of data with $d_j \leq \delta$) and *negative evidence* (s_j, the number of data with $d_j > \delta$).

By treating each response as positive or negative evidence, we can construct an opinion about the annotator's reliability. Such opinions enable the identification of low quality annotators whose responses may warrant down–weighting or further review.

Implications for Responsible AI. The framework contributes to responsible dataset management in three ways. First, high-uncertainty instances can be flagged for re-labeling (or more annotators), supporting active dataset curation. Second, annotator opinions provide auditable signals for quality control in crowdsourcing pipelines. Third, per-instance trust scores can be used to re-weight data during model training, reducing the influence of uncertain or controversial labels. Finally, systematic disagreement patterns can help detect unreliable annotators (e.g., spammers or consistently inconsistent contributors), enabling targeted interventions such as retraining, filtering, or exclusion. These mechanisms complement existing dataset documentation practices and provide actionable insights for trustworthy AI development.

6. Conclusion

We presented a Subjective Logic-based framework to quantify labeling bias and derive instance-level trust scores from human annotations. By representing inter-annotator disagreement as belief, disbelief, and uncertainty, our method provides a richer characterization of annotation reliability than majority voting or global agreement metrics.

Our experiments on the CIFAR-10H dataset demonstrated how trust varies with annotation volume, highlighting the value of modeling uncertainty explicitly in dataset diagnostics. In particular, reduced annotations led to increased uncertainty and reduced confidence in label correctness, which could have significant implications for downstream learning tasks.

We also outlined extensions of the framework to handle probabilistic and natural language annotations and proposed how trust signals could inform annotator profiling. These capabilities open avenues for more informed data filtering, re-annotation strategies, and training pipeline integration.

Future work includes refining trust threshold calibration, validating annotator trust metrics on larger and more diverse datasets, and exploring how trust scores can be used to dynamically weight training data. Overall, this framework contributes toward principled, transparent tools for assessing dataset reliability within trustworthy AI development pipelines.

Acknowledgements

This research has received funding from the European Union's Horizon 2020 EU Research & Innovation program under Grant Agreement No 101069688 (H2020-EU CONNECT). The authors thank the project partners for their collaboration and feedback.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 in order to: Grammar and spelling check.

References

[1] D. Schwabe, K. Becker, M. Seyferth, A. Klaß, T. Schaeffter, The metric-framework for assessing data quality for trustworthy ai in medicine: a systematic review, NPJ Digital Medicine 7 (2024) 203.

[2] X. Li, Z. Chen, J. M. Zhang, F. Sarro, Y. Zhang, X. Liu, Bias Behind the Wheel: Fairness Testing of Autonomous Driving Systems, 2024. URL: https://arxiv.org/abs/2308.02935. arXiv:2308.02935.

[3] F. Yang, G. Zamzmi, S. Angara, S. Rajaraman, A. Aquilina, Z. Xue, S. Jaeger, E. Papagiannakis, S. K. Antani, Assessing Inter-Annotator Agreement for Medical Image Segmentation, IEEE Access 11 (2023) 21300–21312. doi:10.1109/ACCESS.2023.3249759.

[4] A. Jøsang, Subjective Logic: A Formalism for Reasoning under Uncertainty, 1 ed., Springer, 2016.

[5] K. I. Ouattara, I. Krontiris, T. Dimitrakos, F. Kargl, Assessing Trustworthiness of AI Training Dataset using Subjective Logic: A Use Case on Bias, in: Proceedings of the Bias and Fairness in AI Workshop at ECML/PKDD 2025, Lecture Notes in Computer Science, Springer, 2025. URL: https://arxiv.org/abs/2508.13813, also available as arXiv preprint arXiv:2508.13813.

[6] J. Vasilakes, C. Zerva, S. Ananiadou, Subjective Logic Encodings, arXiv preprint arXiv:2502.12225 (2025). URL: https://arxiv.org/abs/2502.12225.

[7] B. Herd, S. Burton, Can You Trust Your ML Metrics? Using Subjective Logic to Determine the True Contribution of ML Metrics for Safety, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, 2024, pp. 1579–1586.

[8] K. I. Ouattara, I. Krontiris, T. Dimitrakos, F. Kargl, Quantifying Calibration Error in Neural Networks Through Evidence-Based Theory, in: Proceedings of the 28th International Conference on Information Fusion (FUSION), 2025.

[9] J. Rutinowski, S. Klüttermann, J. Endendyk, C. Reining, E. Müller, Benchmarking Trust: A Metric for Trustworthy Machine Learning, in: World Conference on Explainable Artificial Intelligence, Springer, 2024, pp. 287–307.

[10] A. Chouldechova, A. Roth, A snapshot of the frontiers of fairness in machine learning, Communications of the ACM 63 (2020) 82–89. URL: https://doi.org/10.1145/3376898. doi:10.1145/3376898.

[11] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, K. Crawford, Datasheets for datasets, Communications of the ACM 64 (2021) 62–71. URL: https://doi.org/10.1145/3458723. doi:10.1145/3458723.

[12] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the EM algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1979) 20–28.

[13] J. Whitehill, T. Wu, J. Bergsma, J. Movellan, P. Ruvolo, Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, in: Advances in Neural Information Processing Systems, volume 22, Curran Associates, Inc., 2009.

[14] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, E. Hovy, Learning whom to trust with MACE, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2013, pp. 1120–1130.

[15] A. Jøsang, T. Bhuiyan, Optimal Trust Network Analysis with Subjective Logic, in: 2008 International Conference on Emerging Security Information, Systems and Technologies, 2008, pp. 179–184.

[16] K. I. Ouattara, A. Petrovska, A. Hermann, N. Trkulja, T. Dimitrakos, F. Kargl, On Subjective Logic Trust Discount for Referral Paths, in: Proceedings of the 27th International Conference on Information Fusion (FUSION), 2024, pp. 1–8.

[17] J. C. Peterson, R. M. Battleday, T. L. Griffiths, O. Russakovsky, Human Uncertainty Makes Classification More Robust, in: Advances in Neural Information Processing Systems (NeurIPS), 2019. URL: https://arxiv.org/abs/1908.07086.