Uncertainty Estimation by Human Perception versus Neural Models

Pedro Mendes[1,2], Paolo Romano[2], and David Garlan[1]

[1] Software and Societal Systems Department, Carnegie Mellon University
[2] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa

Abstract. Modern neural networks (NNs) often achieve high predictive accuracy but are poorly calibrated, producing overconfident predictions even when wrong. This miscalibration poses serious challenges in applications where reliable uncertainty estimates are critical. In this work, we investigate how human perceptual uncertainty compares to uncertainty estimated by NNs. Using three vision benchmarks annotated with both human disagreement and crowdsourced confidence, we assess the correlation between model-predicted uncertainty and human-perceived uncertainty. Our results show that current methods only weakly align with human intuition, with correlations varying significantly across tasks and uncertainty metrics. Notably, we find that incorporating human-derived soft labels into the training process can improve calibration without compromising accuracy. These findings reveal a persistent gap between model and human uncertainty and highlight the potential of leveraging human insights to guide the development of more trustworthy AI systems.

1 Introduction

Neural networks (NNs) have achieved remarkable success across a variety of tasks, from image classification to medical diagnosis. However, despite their high predictive accuracy, these models often suffer from poor calibration; their confidence scores do not reliably reflect the likelihood of correctness. This issue becomes especially problematic in high-stakes applications, where overconfident yet incorrect predictions can lead to severe consequences. Due to their black-box nature, it is far from trivial to understand and explain the outputs produced by modern large and complex NNs [6, 34].

Uncertainty estimation has emerged as a key component in building more trustworthy AI systems. Several methods [20, 16, 41, 35, 25, 22, 19, 24, 2, 15, 14, 9, 13] have been proposed to improve model calibration, including temperature scaling, Bayesian approximations, and ensemble techniques. While these approaches often improve alignment between predicted probabilities and observed outcomes, it remains unclear whether the resulting uncertainty estimates align with human-perceived uncertainty.

Further, real-world applications, such as medical diagnostics, autonomous driving, and fraud detection, increasingly leverage model uncertainty to enhance trustworthiness by incorporating human oversight. In these systems, models can

flag predictions with high uncertainty as requiring human review. This approach allows systems to balance automation and human judgment, ensuring that critical decisions are not solely reliant on potentially unreliable predictions. By raising alarms for ambiguous cases, these systems create a feedback loop where humans can validate, correct, or refine predictions, thus improving the model's performance and fostering user trust. However, the deployment of uncertainty-based alarm systems creates a trade-off between the frequency of alarms raised and the associated costs. These costs can be reduced by aligning the model uncertainty estimation with human perception.

In this work, we explore the following question: To what extent do uncertainty estimates from modern NNs reflect the uncertainty perceived by humans? If models and humans perceive uncertainty differently, then even well-calibrated models may fail to support downstream decisions in a human-compatible manner.

To investigate this, we study three visual classification tasks that include human annotations of perceptual uncertainty, either through soft labels (reflecting inter-annotator disagreement) or explicit confidence scores. We compare these human-derived signals with uncertainty estimates from multiple NNs using metrics such as predictive entropy (PE) [36, 18]. We further test whether injecting human soft labels into training can improve the model's alignment with human intuition and its calibration.

Our key contributions are as follows:

- We present a systematic comparison between model-estimated and human-perceived uncertainty across three vision benchmarks.
- We quantify the alignment between human and model uncertainty and analyze how it varies across datasets and uncertainty estimation techniques.
- We show that integrating human soft labels into training can improve model calibration without degrading accuracy, indicating the potential of human-informed training.
- Our findings suggest that model confidence alone may not be sufficient for trustworthy human-AI interaction and motivate future work toward hybrid uncertainty estimation approaches.

2 Related Work

A significant body of work has explored methods to enhance calibration of NNs. Early methods leveraged Bayesian Neural Networks (BNNs) [26, 21, 3, 39], which place a prior distribution over the model weights and infer a posterior given the data [33]. While theoretically sound, BNNs are computationally intensive and scale poorly to large models and datasets. To mitigate this, several approximate Bayesian methods have been proposed such as Monte Carlo (MC) Dropout [4] and Variational Inference (VI) [12]. Among these, MC Dropout, due to its higher efficiency, has become the most widely used technique for uncertainty estimation in NNs [22, 1, 4]. This technique estimates uncertainty by applying multiple dropout masks and performing multiple stochastic forward passes, effectively

simulating an ensemble of subnetworks. The resulting predictions are aggregated using statistical measures, such as PE or variance, to quantify the model's uncertainty.

In addition to uncertainty quantification, calibration has become a key property of NNs, as they are often overconfident when their predictions are incorrect [8]. To address this, a range of post-hoc calibration methods have been developed, including Platt Scaling [29], Isotonic Regression [41], Temperature Scaling [8], and Beta Calibration [15], which adjust model confidence scores after training. These techniques are computationally efficient and easy to apply, but do not alter the underlying uncertainty estimates. Consequently, their effectiveness can vary significantly across architectures and datasets [20], and they tend to degrade under distributional shifts, reducing reliability in out-of-distribution (OOD) scenarios [27, 8].

Beyond post-hoc methods, a growing body of work seeks to incorporate uncertainty estimation directly into the training process. These uncertainty-aware training methods aim to jointly optimize for accuracy and calibration. Approaches in this direction include augmenting standard loss functions with uncertainty-regularizing terms, such as Focal Loss [19] and Label Smoothing [24]. Further, Accuracy versus Uncertainty Calibration (AvUC) loss [14] explicitly balances predictive performance with calibration, integrating temperature scaling into the optimization objective. Other advancements propose differentiable calibration metrics, such as Soft Calibration Error [13], which relaxes traditional binning procedures, and binning-free calibration methods [9], which avoid discretization entirely. Additional efforts incorporate conformal prediction frameworks, such as the uncertainty-aware conformal loss [5], to better align model confidence intervals with observed outcomes. Shamsi et al. [35] proposed a composite loss function that accounts for both PE and Expected Calibration Error (ECE). More recently, CALS [20] introduced a class-wise uncertainty weighting scheme that emphasizes harder or more uncertain examples during training. EUAT [22] proposed a dual-loss strategy that differentiates between correct and incorrect predictions, promoting high uncertainty on mispredictions while maintaining low uncertainty for correct ones. At last, CLUE [23] proposes a general framework for model calibration that explicitly aligns predicted uncertainty with observed error during training. While these methods improve statistical calibration, they do not necessarily reflect how humans perceive uncertainty.

Humans intuitively account for ambiguity and express caution under uncertain conditions, suggesting a natural way to navigate complex, unfamiliar, or ambiguous inputs. They evaluate uncertainty in a highly context-dependent manner, based on heuristics, drawing on both perceptual and contextual indications, subjective perceptions, and prior knowledge to estimate confidence levels [38]. Unlike NNs, which only rely on patterns learned from data, humans integrate multiple sources of information, such as experience, visual or sensory ambiguity, and situational perspectives, to assess the certainty of their judgments [11, 30]. This flexible and adaptive uncertainty evaluation allows humans

to balance caution and decisiveness in uncertain situations, adjusting their responses based on the perceived risk and context.

Human-perceived uncertainty has been studied through crowdsourced confidence annotations and disagreement among annotators [28, 37, 40]. Such signals offer a richer view of task difficulty and ambiguity. Recent work has leveraged these signals to generate soft labels for training [31], which can improve robustness and calibration [28]. Further, Peterson et al. [28] have shown that human labels improve generalization, i.e., the quality of the models increases when trained using soft labels obtained by human annotators. However, this work never evaluates the model uncertainty, fundamental for accessing trustworthiness. Additionally, Steyvers et al. [37] developed a Bayesian modeling framework that jointly combines human and models predictions. However, the alignment between these human-derived uncertainty signals and model-predicted uncertainty remains underexplored.

Unlike prior work, we perform a systematic, cross-dataset comparison of model-estimated uncertainty and human-perceived uncertainty. We also investigate whether incorporating soft human labels during training can improve this alignment and enhance model calibration, providing insights into the potential for hybrid human-AI trust pipelines.

3 Human and Model Uncertainty Evaluation

Understanding and quantifying uncertainty is crucial to make informed decisions and to develop trustworthy models. By comparing the uncertainty estimation of NNs with human perceptual uncertainty, it is possible to understand how well current methods align with human uncertainty, shedding light on both the strengths and limitations of current uncertainty estimation techniques, leading to the design of improved methods that enhance model trustworthiness.

Therefore, this work conducts a correlational study to compare human perception of uncertainty against uncertainty estimation outputted by NNs. The objectives are twofold: i) to evaluate how well model uncertainty aligns with human perceptual uncertainty, especially in ambiguous or uncertain scenarios, and ii) to explore whether model uncertainty can be enhanced or complemented by human decision-making. By identifying scenarios where human and model uncertainty assessments converge, diverge, or reveal poor calibration, we can tailor the training to develop more trustworthy models.

This study hypothesizes that human perceptual uncertainty strongly correlates with the uncertainty estimates generated by NNs. Specifically, we posit that as human uncertainty increases, model uncertainty should also increase. Furthermore, we hypothesize that the strength of this correlation might be dependent on task complexity, with correlations weakening as task complexity increases. Additionally, we propose incorporating human insights to enhance the quality of model uncertainty evaluations, thereby advancing the development of more trustworthy AI systems.

3.1 Experimental Setup, Benchmarks, and Baselines

Datasets and Models. This work exploits publicly available datasets (namely, Cifar10-H [28], CifarN [40], and ImageNet-16H [37]) in the image recognition domain (normally used to train NNs) that include human annotations from multiple reviewers. The datasets containing human annotations have been labeled by multiple reviewers, with the number of reviews per image varying based on the benchmark used [3]. Further, ImageNet-16H extends the ImageNet [32] dataset by reducing the sample size to 4800 and limiting the number of classes to 16, while introducing noise in the images.

We considered pre-trained models, such as ResNet50 [10] using ImageNet, as well as models trained specifically on the benchmarks used in this study, including ResNet18 on CIFAR10-H and CIFAR-N. For these benchmarks, we randomly split the data into training and testing sets, using an 80/20 ratio, respectively. Additionally, for ResNet50 on ImageNet-16H, we fine-tuned the model using the original dataset inputs and evaluated it on the noisy data containing human annotations.

To train the models, independently of the considered solution, we use stochastic gradient descent to minimize the Cross Entropy (CE) loss function using a learning rate of 0.1, a momentum of 0.9, and a batch size of 64 for all the models. To guarantee reproducibility, ensure a fair comparison, and mitigate the randomness inherent in training NNs, each method was trained using ten different seeds [17]. Throughout this study, when a subset of the dataset is required, stratified sampling techniques are applied to ensure that all classes are adequately represented and that the sample remains representative of the original dataset. Further, the models are evaluated on the same test set used for human annotations to ensure a consistent comparison between human and model uncertainty estimates. All models and training procedures were implemented in Python3 via the Pytorch framework and trained using a single Nvidia RTX A4000 GPU.

Baselines. Furthermore, we resort to several state-of-the-art methods [16, 20, 22, 17, 8] that have been proposed to estimate model uncertainty. Moreover, one important foundation of all these works lies in the computation of uncertainty given a prediction. One of the most widely used techniques is MC Dropout [7], which provides a Bayesian approximation for uncertainty by sampling multiple dropout masks and aggregating the predictions using PE. Like the human uncertainty estimation, MC Dropout combined with PE does not separate epistemic from aleatoric uncertainty, so both human and model uncertainty will reflect the total uncertainty of a prediction. In this study, we compare different baselines to compute the model uncertainty. More in detail, we evaluate both post-processing calibration methods (resorting to Isotonic Regression [41], or DEUP [16]) and uncertainty-aware training algorithms (namely, CE [42], EUAT [22], CALS [20], or deep ensemble [17]).

[3] ImageNet-16H and Cifar-N dataset received approximately four to six judgments per image, while Cifar-10H dataset has fifty annotations per image.

Table 1: Task complexity based on human and model predictions. For each input belonging to class A, the prediction is considered *correct* if classified as A, and *incorrect* if classified as any other class ($\neg A$).

True Class	**Human annotation**	**Model prediction**	**Task complexity**
A	A	A	Easy
A	A	$\neg A$	Medium
A	$\neg A$	A	Difficult
A	$\neg A$	$\neg A$	Difficult

The human annotations contained on the benchmarks used can be exploited to create soft labels, representing the probability of each class given an input. From these soft labels, a distribution can be generated, allowing the computation of uncertainty, for example, using the PE. The datasets primarily consist of images of everyday objects (e.g., cars, airplanes), and thus epistemic uncertainty (stemming from a lack of knowledge) of human predictions can be assumed to be low. In this context, human uncertainty likely arises from noise, ambiguity, or inherent randomness in the data, i.e. the aleatoric uncertainty. However, the proposed method for evaluating uncertainty using human annotations does not differentiate between epistemic and aleatoric uncertainty; it quantifies only the total uncertainty.

Evaluation Metrics. We resort to the PE of the outputted distribution to compute the respective uncertainty. By applying the same uncertainty estimation method for both human and model assessments, a direct comparison of the results becomes possible.

To assess the relationship between human and model uncertainties, Pearson's correlation coefficient is computed for each baseline. Additionally, we compute the correlations for each subgroup composed of inputs with differing complexity. Statistical significance was assessed at a threshold of $p < 0.05$.

3.2 Methods

This study has two independent variables: uncertainty estimation method (human-derived versus model uncertainty estimates) and task complexity (easy, medium, difficult). Each group (humans and models) classified the same set of images to generate the uncertainty estimates.

Moreover, the task complexity (i.e., the complexity/difficulty of classifying an image) is defined based on whether the human and model predictions match the true class. To automate the assessment of the task complexity, we compare the true class of each image with the human reviews (which can be averaged to determine a final class) and model predictions [4]. When discrepancies arise be-

[4] It should be noticed that the uncertainty is not used to evaluate the task complexity but only the predictions.

Table 2: Pearson correlation coefficient between model and human uncertainty using different baselines, models, and benchmarks.

Baseline	**Cifar10-H**	**CifarN**	**ImageNet-16H**
CE	0.22	2.59E-3	0.34
EUAT	0.21	2.60E-3	0.30
Ensemble	0.21	5.07E-3	0.35
CALS	0.22	4.60E-3	0.32
Iso. Reg.	0.24	4.20E-3	0.35
DEUP	0.06	3.90E-3	0.02

tween the true class and the predictions from humans or models, the complexity of the input is determined according to the criteria outlined in Table 1.

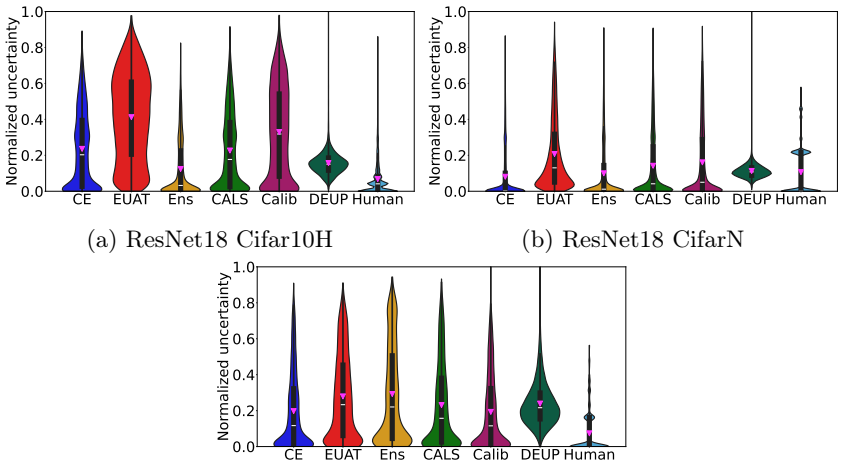
At last, to explore the potential of human insights to improve NNs uncertainty estimates, a subset of human annotations was incorporated into the model's training using soft labels while computing the CE loss. The quality of the updated model's uncertainty estimates was then re-evaluated against the original human uncertainty annotations.

4 Results

This section reports the results obtained in this correlation study. Table 2 shows the correlation between human perceptual uncertainty and model uncertainty evaluated using the different techniques considered (namely, CE with MC dropout, EUAT, CALS, deep ensemble, isotonic regression, and DEUP) training a ResNet18 on Cifar10-H and CifarN, and a ResNet50 on ImageNet-16H. In these experiments, human reviews are not incorporated into the training process.

The results show no significant correlation between human and model uncertainty. The Pearson correlation coefficient between the model and human uncertainty is up to 0.24, $5.07\text{x}10^{-3}$, and 0.35 for ResNet18 trained on Cifar10-H and CifarN, and ResNet50 on ImageNet-16H, respectively. The corresponding p-values for ResNet18 on CIFAR10-H and ResNet50 on ImageNet-16H are below the significance threshold of 0.05, confirming the statistical validity of these results. However, for ResNet18 with CIFAR-N, the p-value exceeds 0.05, indicating no statistically significant correlation. In conclusion, while statistically significant but weak correlations were observed for ResNet18 on CIFAR10-H and ResNet50 on ImageNet-16H, the analysis found no significant correlation between human perceptual uncertainty and model uncertainty estimates for ResNet18 on CIFAR-N. This suggests that there is no correlation between human perception uncertainty and model-derived uncertainty estimates.

Further, Figure 1 plots the distribution of the normalized uncertainty of the models using several different uncertainty estimation techniques and the human uncertainty. Through the analysis of those, we see very different distributions, which visually reinforces the previous results.



(a) ResNet18 Cifar10H

(b) ResNet18 CifarN

(c) ResNet50 ImageNet-16H

Fig. 1: Normalized uncertainty distribution using different baselines, models, and benchmarks. (the average value of each distribution is marked with a pink triangle).

Next, we evaluate the relationship between task complexity and the correlation of uncertainty estimates by reporting, in Table 3, the Pearson coefficient of correlation between humans and multiple uncertainty estimation techniques for the different benchmarks. Once again, the results show no correlation between the task complexity and the uncertainty of humans vs. models.

While both humans and models exhibited higher uncertainty for ambiguous or noisy inputs, their uncertainty levels did not consistently align across the dataset. Furthermore, no clear pattern emerged when evaluating the relationship between task complexity and the correlation of uncertainty estimates. Although both humans and models showed increased uncertainty for more complex inputs, their responses remained uncorrelated, suggesting that task complexity does not significantly impact the alignment between human and model uncertainty. These results highlight a fundamental divergence between how humans and models assess uncertainty, even under varying task conditions.

5 Discussion

In this study, we analyzed the relationship between human perceptual uncertainty and the uncertainty estimates generated by NNs. Contrary to our initial hypothesis, the results reveal no clear correlation between human and model uncertainty. While there are occasional instances of agreement, the overall lack of alignment suggests that models and humans rely on fundamentally different mechanisms to assess uncertainty. This misalignment raises important questions

Table 3: Pearson correlation coefficient between model and human uncertainty using different task complexity, baselines, models, and benchmarks.

Baseline	**Complexity**	**Cifar10-H**	**CifarN**	**ImageNet-16H**
CE	Easy	0.21	0.01	0.22
	Medium	0.03	0.0	0.08
	Difficult	0.24	0.0	0.29
EUAT	Easy	0.21	0.0	0.17
	Medium	0.0	0.0	0.05
	Difficult	0.20	0.0	0.3
Ensemble	Easy	0.19	0.02	0.21
	Medium	0.02	0.04	0.01
	Difficult	0.22	0.0	0.19
CALS	Easy	0.22	0.02	0.19
	Medium	0.01	0.03	0.06
	Difficult	0.26	0.0	0.27
Calibration	Easy	0.22	0.01	0.22
	Medium	0.05	0.02	0.10
	Difficult	0.32	0.0	0.23
DEUP	Easy	0.03	0.02	0.06
	Medium	0.0	0.03	0.01
	Difficult	0.09	0.01	0.04

Table 4: Pearson correlation coefficient using CE with soft labels.

Baseline	**Cifar10-H**	**CifarN**	**ImageNet-16H**
CE with soft labels	0.34	0.09	0.47

about the interpretability and reliability of model uncertainty estimates, particularly in tasks where human intuition plays a critical role.

We also analyzed whether task complexity influences the correlation between human and model uncertainty. Once again, the results do not support a consistent relationship. While we observed that both humans and models tend to exhibit higher uncertainty for more complex inputs, their responses remain largely uncorrelated across different levels of complexity. This finding indicates that neural networks struggle to capture the nuanced patterns of uncertainty that humans intuitively recognize, even when task complexity is varied. Such discrepancies highlight limitations in current uncertainty estimation methods, particularly in tasks with ambiguous or noisy data where human expertise is crucial.

Given these results, we further evaluated whether incorporating human insights could improve model uncertainty evaluations. Thus, a subset of human annotations was incorporated into the model using soft labels while computing the CE loss, and the correlation of the updated model's uncertainty estimates was then re-evaluated against the human uncertainty. The results, reported in

Table 4 show an improvement in the correlation of the model uncertainty compared to human uncertainty. More in detail, the correlation improves to 0.34 and 0.47 when training a ResNet18 with Cifar10-H and a ResNet50 with ImageNet-16H, respectively. On ResNet18 with CifarN, while there is some improvement, the correlation remains insignificant.

Further, while human annotations occasionally helped refine model predictions, their impact was inconsistent and depended on the specific nature of the task and the quality of the model's baseline estimates. These findings suggest that simply integrating human insights may not be sufficient to bridge the gap between human and model uncertainty assessments. Instead, more sophisticated approaches for uncertainty estimation or designing models that explicitly account for human-like patterns of uncertainty may be necessary to improve alignment.

The lack of correlation between human and model uncertainty has important implications for the development and application of AI systems. It highlights the need for alternative methods that better emulate human uncertainty reasoning, particularly in high-stakes applications such as healthcare or autonomous driving. Without this alignment, model uncertainty estimates may lack the interpretability needed to build trust and reliability among users. These findings underscore the importance of not only improving model performance but also ensuring that uncertainty estimates reflect human-like reasoning, which is essential for societal acceptance and effective integration of AI systems into decision-making processes.

6 Conclusion

This study explored the relationship between human perceptual uncertainty and the uncertainty estimates generated by NNs. Our findings indicate no significant correlation between the two, highlighting a fundamental divergence in how humans and models assess uncertainty. This behavior persisted across tasks with varying levels of complexity. While incorporating human insights into model uncertainty evaluations showed some improvements, the results emphasize the need for more robust methods that better emulate human uncertainty assessment.

Acknowledgments

This work was supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the Carnegie Mellon Portugal Program under grant SFRH/BD/151470/2021, and by projects UIDB/50021/2020, C645008882-00000055.PRR and C628696807-00454142 (Center for Responsible AI), 101189689 (ACHILLES). This work was developed within the scope of the project no. 62 - "Responsible A", financed by European Funds, namely "Recovery and Resilience Plan - Component 5: Agendas Mobilizadoras para a Inovação Empresarial", included in the NextGenerationEU funding program.

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76**, 243–297 (2021)
2. Antorán, J., Allingham, J.U., Hernández-Lobato, J.M.: Depth uncertainty in neural networks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS'20, Curran Associates Inc., Red Hook, NY, USA (2020)
3. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural networks. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (2015)
4. Brach, K., Sick, B., Dürr, O.: Single shot mc dropout approximation. ArXiv **abs/2007.03293** (2020), https://api.semanticscholar.org/CorpusID:220381176
5. Einbinder, B.S., Romano, Y., Sesia, M., Zhou, Y.: Training uncertainty-aware classifiers with conformalized deep learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35 (2022)
6. Fan, F.L., Xiong, J., Li, M., Wang, G.: On interpretability of artificial neural networks: A survey. IEEE Transactions on Radiation and Plasma Medical Sciences **5** (2021)
7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of The 33rd International Conference on Machine Learning. vol. 48. PMLR (2016)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. vol. 70. PMLR (2017)
9. Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., Hartley, R.: Calibration of neural networks using splines. In: International Conference on Learning Representations (2021)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
11. Hink, R.F., Woods, D.L.: How humans process uncertain knowledge: An introduction for knowledge engineers. AI Magazine **8**(3) (1987)
12. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. Journal of Machine Learning Research **14**(40) (2013)
13. Karandikar, A., Cain, N., Tran, D., Lakshminarayanan, B., Shlens, J., Mozer, M.C., Roelofs, B.: Soft calibration objectives for neural networks. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34 (2021)
14. Krishnan, R., Tickoo, O.: Improving model calibration with accuracy versus uncertainty optimization. In: Advances in Neural Information Processing Systems. vol. 33 (2020)
15. Kull, M., Filho, T.S., Flach, P.: Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. vol. 54 (2017)

16. Lahlou, S., Jain, M., Nekoei, H., Butoi, V.I., Bertin, P., Rector-Brooks, J., Korablyov, M., Bengio, Y.: DEUP: Direct epistemic uncertainty prediction. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=eGLdVRvvfQ
17. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (2017)
18. Learned-Miller, E.G.: Entropy and mutual information. Department of Computer Science University of Massachusetts (2013)
19. Liu, B., Ayed, I.B., Galdran, A., Dolz, J.: The devil is in the margin: Margin-based label smoothing for network calibration. In: Computer Vision and Pattern Recognition Conference (2022)
20. Liu, B., Rony, J., Galdran, A., Dolz, J., Ben Ayed, I.: Class adaptive network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16070–16079 (2023)
21. MacKay, D.J.C.: A Practical Bayesian Framework for Backpropagation Networks. Neural Computation **4**(3), 448–472 (1992)
22. Mendes, P., Romano, P., Garlan, D.: Error-driven uncertainty aware training. In: Proceedings of the 27th European Conference on Artificial Intelligence (2024)
23. Mendes, P., Romano, P., Garlan, D.: Clue: Neural networks calibration via learning uncertainty-error alignment. arXiv preprint arXiv:2505.22803 (2025)
24. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. In: Advances in Neural Information Processing Systems. vol. 33 (2020)
25. Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
26. Neal, R.M.: Bayesian Learning for Neural Networks. Springer-Verlag (1996)
27. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
28. Peterson, J.C., Battleday, R.M., Griffiths, T.L., Russakovsky, O.: Human uncertainty makes classification more robust. 2019 IEEE/CVF International Conference on Computer Vision (2019)
29. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers (1999)
30. Pouget, A., Drugowitsch, J., Kepecs, A.: Confidence and certainty: distinct probabilistic quantities for different goals. Nature Neuroscience **19**(3) (2016)
31. Reidsma, D., op den Akker, R.: Exploiting 'subjective' annotations. In: Artstein, R., Boleda, G., Keller, F., Schulte im Walde, S. (eds.) Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics (Aug 2008)
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
33. Segalman, D.J., Brake, M.R., Bergman, L.A., Vakakis, A.F., Willner, K.: Epistemic and aleatoric uncertainty in modeling. In: International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. vol. Volume 8: 22nd Reliability, Stress Analysis, and Failure Prevention Conference; 25th Conference on Mechanical Vibration and Noise (2013)

34. Seuss, D.: Bridging the gap between explainable ai and uncertainty quantification to enhance trustability. ArXiv **abs/2105.11828** (2021)
35. Shamsi, A., Asgharnezhad, H., Tajally, A., Nahavandi, S., Leung, H.: An uncertainty-aware loss function for training neural networks with calibrated predictions. ArXiv **abs/2110.03260** (2023)
36. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27** (1948)
37. Steyvers, M., Tejeda, H., Kerrigan, G., Smyth, P.: Bayesian modeling of human–ai complementarity. Proceedings of the National Academy of Sciences **119**(11) (2022)
38. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science **185**(4157) (1974)
39. Wang, H., Yeung, D.Y.: A survey on bayesian deep learning. ACM Comput. Surv. **53**(5) (2020)
40. Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., Liu, Y.: Learning with noisy labels revisited: A study using real-world human annotations. In: International Conference on Learning Representations (2022)
41. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01 (2001)
42. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31 (2018)