Does Calibration Affect Human Actions?

Extended Abstract

Meir Nizri
Department of Data Science and AI
Ariel University, Israel
meir.nizri@msmail.ariel.ac.il

Amos Azaria
Department of Data Science and AI
Ariel University, Israel
amos.azaria@ariel.ac.il

Chirag Gupta
Carnegie Mellon University
Pittsburgh, Pennsylvania, US
chiragg@andrew.cmu.edu

Noam Hazon
Department of Data Science and AI
Ariel University, Israel
noamh@ariel.ac.il

ABSTRACT

Calibration has been proposed as a way to enhance the reliability and adoption of machine learning classifiers. We study a particular aspect of this proposal: how does calibrating a classification model affect the decisions made by non-expert humans consuming the model's predictions? We performed a Human–Computer Interaction (HCI) experiment to assess the effect of calibration on (i) trust in the model and (ii) the correlation between decisions and predictions. We also propose further corrections to the reported calibrated scores based on Kahneman and Tversky's prospect theory from behavioral economics, and study the effect of these corrections on trust and decision-making. We found that calibration alone was not sufficient: applying the prospect-theory–based correction is crucial for increasing the correlation between human decisions and the model predictions. While this increased correlation suggests higher trust in the model, responses to "Do you trust the model more?" are unaffected by the method used.

For the full paper, see: https://arxiv.org/abs/2508.18317

KEYWORDS

Model Calibration, Prospect Theory, Human-Agent Interaction

ACM Reference Format:
Meir Nizri, Amos Azaria, Chirag Gupta, and Noam Hazon. 2026. Does Calibration Affect Human Actions?: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/YHRI7310

1 INTRODUCTION

Machine learning (ML) systems frequently function as decision-support tools to humans, especially in high-stakes settings such as weather forecasting, medical diagnosis, risk assessment, fraud detection, and financial prediction. In these domains, it is not enough to output a class label; the system should also report a probability that meaningfully reflects predictive certainty. In practice, classification models are probabilistic, and their probabilities are commonly shown to users as *confidence scores*.

A key question is whether the model is *calibrated*: do events predicted with probability p occur about p fraction of the time? Many standard training approaches yield over-confident predictors. Modern neural networks are well known to exhibit such miscalibration [3], and numerous post-hoc methods aim to fix it (see Gupta [4, Chapter 1]).

Despite this progress, we know comparatively little about how calibration affects *human* behavior: trust, reliance, and decision quality. Human-centered evaluation complements standard metrics by capturing contextual judgments [1], usability and expectations [7], and ethical/legal concerns such as transparency, accountability, and fairness [8]. Accordingly, in this paper we study how calibration affects human decision-making and trust.

We also propose an additional calibration layer inspired by prospect theory [5, 6]. Because users act on *perceived* probabilities, we transform calibrated probabilities using the inverse of a prospect-theoretic weighting function, so that reported values better align with human perception (e.g., if 90% is perceived as 80%, then an 80% event is reported as 90%).

We performed a human study to evaluate our approach on two distinct domains: rain forecasting and loan approval. We used a neural network as the uncalibrated model, and the *isotonic regression* method for calibration. We compare our method, which incorporates prospect theory on top of isotonic regression, to four baselines: uncalibrated, isotonic-only, prospect correction on uncalibrated outputs, and a control with matched probabilities but randomized outcomes. Across both studies, reported trust is similar across conditions, but our method yields a significantly higher correlation between participant decisions and model predictions, suggesting practical value in perception-aware calibration.

2 THE PROSPECT THEORY CORRECTION

Incorporating prospect theory into model calibration can improve the accuracy and realism of the model's predictions, making it more effective in capturing human behavior. Since prospect theory accounts for the biases and heuristics that individuals rely on when making decisions, incorporating these biases can better simulate and predict how people respond to different scenarios.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/YHRI7310

Prospect theory uses a non-linear weighting function to describe how people subjectively weigh probabilities. The standard weighting function is

$$w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{\frac{1}{\gamma}}}, \quad (1)$$

with the parameter $\gamma \in (0, 1]$ describing the amount of over and under-weighting. The weighting function has different γ values for gains and losses. We propose to transform each calibrated probability using the inverse of the weighting function. Since there is no simple closed-form formula for the inverse of the weighting function, we compose the following approximation:

$$w^{-1}(p) \approx \frac{p^{\frac{1}{\gamma}}}{(p^{\frac{1}{\gamma}} + (1-p)^{\frac{1}{\gamma}})^{\frac{1}{\gamma}}}. \quad (2)$$

To evaluate how accurately our proposed function approximates the inverse of the weighting function, we calculated the mean absolute error (MAE) between all probabilities $p \in \{0, 0.01, 0.02, \ldots, 1\}$ and $w^{-1}(w(p))$ to be 0.00963, which corresponds to a deviation of less than one percent.

3 RESULTS

Overall, we tested five methods for presenting the system's predictions, with each participant assigned to only one method: **Uncalibrated**: The raw probabilistic output from a trained model, with no calibration applied. **Calibrated**: The trained model's output calibrated using isotonic regression. **PT-calibrated**: The calibrated probabilities transformed using our prospect theory method from Section 2. **PT-uncalibrated**: The uncalibrated (raw) probabilities transformed using the prospect theory method. **Random**: A control condition where participants were shown probabilities identical to those in the PT-calibrated method, but the actual outcomes were randomized. This condition was included to verify that any observed effects were due to the model's predictive accuracy in relation to its forecasts, not merely the distribution of the probabilities shown.

Evaluation of the proposed methods relied on two primary metrics: (1) the mean trust ratings attributed to each method, and (2) the mean correlation between system predictions and participant action ratings. To assess whether statistically significant differences existed among the methods, we performed a one-way *Analysis of Variance* (ANOVA) [2, 9]. Results were considered statistically significant at the $\alpha = 0.05$ level.

As anticipated, the random method received the lowest trust rating. However, when comparing the other methods, there is no notable difference in trust. In fact, except for the random model, there is no statistically significant difference between the models. Overall, it cannot be concluded that either calibration or the addition of a prospect theory layer increases people's trust in the model.

We now examine the correlation between the model's predictions and the participants' domain-specific actions. We compute the correlation for each participant and then average all the correlations for each method, in both domains.

Figure 1 illustrates distinct performance disparities among the evaluated models in the rain-forecasting domain. We note that the PT-calibrated model yields a higher correlation than all other methods across both domains. These improvements are statistically

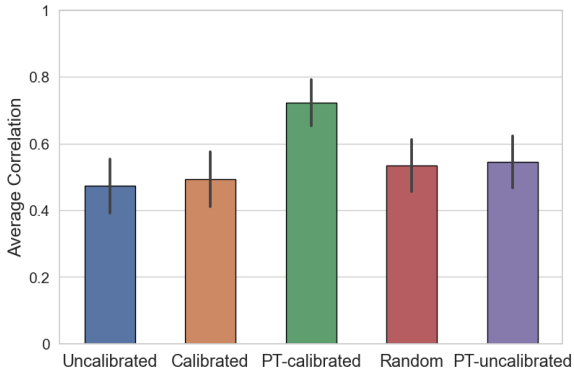

Figure 1: Average correlation in the rain-forecasting domain, between the method predictions and the participants' actions. The PT-calibrated method leads to a statistically significant increase in correlation compared to all other methods, in particular, the calibrated method.

significant for all pairwise comparisons against models that do not incorporate Prospect Theory adjustments. In addition, the calibrated method performed very similarly to the uncalibrated method, showing a marginal improvement in the rain-forecasting domain and a slight degradation in the loan-approval domain. These results indicate that calibrating the model alone does not necessarily influence people to base their decisions on its predictions. However, adding to the calibrated model a layer that adjusts probabilities in line with people's expectations, as per prospect theory, encourages individuals to align their decisions with the model's predictions. Since the results are statistically significant in both domain, it can also be concluded that our approach adds value whether the user is fully dependent on the model or using it as a secondary aid.

Notably, when the final outcomes are random and unrelated to the model's predictions, the correlation is significantly lower (i.e., the random method shows a significantly lower correlation than the PT-calibrated model in both domains). This indicates that the differences between PT-calibrated and other methods depend on the model's true predictive accuracy, not merely on the displayed probability distribution.

4 CONCLUSION

In this paper we study how people react to the probabilities predicted by a machine learning model. We explore whether calibrating the probabilities increases trust in the system and whether it leads to people making decisions that are more aligned with the system's predictions. Furthermore, we introduce a prospect theory-based correction that is applied to the model's calibrated probabilities. We show that while the trust in the system is not significantly affected by the method used, when asked to take action, the resulting correlation between the model's prediction and the human action is significantly higher for the model with calibration and prospect theory correction. Notably, we also observe that calibration alone does not necessarily improve alignment between human decisions and the system's prediction. These results were obtained in two different domains.

ACKNOWLEDGMENTS

This work was supported, in part, by the Ministry of Science and Technology, Israel, and by the Israel Science Foundation under grant 1092/24.

REFERENCES

[1] Nick Bostrom and Eliezer Yudkowsky. 2018. The Ethics of Artificial Intelligence. In *Artificial Intelligence Safety and Security*. Chapman & Hall/CRC, 57–69.
[2] Ronald Aylmer Fisher. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*. Springer, 66–70.
[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*. PMLR, 1321–1330.
[4] Chirag Gupta. 2023. *Post-hoc Calibration Without Distributional Assumptions*. Ph.D. Dissertation. Carnegie Mellon University, Pittsburgh, PA, USA.
[5] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (March 1979), 263–291.
[6] Daniel Kahneman and Amos Tversky. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty* 5, 4 (1992), 297–323.
[7] Andreas M. Kaplan and Michael Haenlein. 2019. Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence. *Business Horizons* 62, 1 (2019), 15–25.
[8] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* 3, 2 (2016).
[9] Douglas C Montgomery. 2017. *Design and analysis of experiments*. John wiley & sons.