Fairness in Human-AI

A
Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of
MASTER OF TECHNOLOGY
By

KRISHNA KANT SINGH
Under the Supervision of
Dr. SHASHI SHEKHAR JHA

Department of Computer Science & Engineering
Indian Institute of Technology Ropar
May 29, 2025

Declaration

This is to certify that the thesis entitled "**Fairness in Human-AI**", submitted by me to the *Indian Institute of Technology Ropar*, for the award of the degree of Master of Technology, is a bonafide work carried out by me under the supervision of Dr. Shashi Shekhar Jha. The content of this thesis, in full or in parts, have not been submitted to any other University or Institute for the award of any degree or diploma. I also wish to state that to the best of my knowledge and understanding nothing in this report amounts to plagiarism.

Sign: _____

Krishna Kant Singh
Department of Computer Science & Engineering,
Indian Institute of Technology Ropar,
Rupnagar-140001, Punjab, India.

Date: _____

Certificate

This is to certify that the thesis entitled "**Fairness in Human-AI**", submitted by Krishna Kant Singh (2023AIM1001), a master's student in the *Department of Computer Science & Engineering*, *Indian Institute of Technology Ropar*, for the award of the degree of Master of Technology, is a record of an original research work carried out by the candidate under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard worthy of the award of the degree. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Sign: _____

Supervisor: Dr. Shashi Shekhar Jha
Department of Computer Science & Engineering,
Indian Institute of Technology Ropar,
Rupnagar-140001, Punjab, India.

Date: _____

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Shashi Shekhar Jha, for his invaluable guidance, encouragement, and support throughout the course of this thesis. His insightful feedback, patience, and constant motivation has helped a lot in shaping both the direction and quality of this research.

I am sincerely thankful for the opportunities that Dr. Shashi provided me to learn and grow, and for always being approachable and generous with his time. It has been a privilege to work under his supervision.

I also extend my thanks to other faculty members and fellow researchers who helped refine this work through their stimulating discussions and constructive input.

Sincerely
Krishna Kant Singh

Abstract

Integrating human expertise with machine learning models in human-in-the-loop (HITL) systems enhances predictive accuracy,, but often overlooks fairness across demographic groups. This thesis proposes a novel method to improve fairness by deferring predictions to humans when the expected cost of unfairness, derived from local accuracy disparities measured via k-nearest neighbors (KNN), exceeds the cost of human intervention. Using the HateSpeech and Adult Census Income datasets, the approach trains separate KNN structures for each demographic group to evaluate local statistical parity. It is then combined with fairness cost to compute fairness risk. If the said fairness risk becomes higher than human cost, the system defers to human predictions, leveraging a combiner to integrate human and model outputs.

In addition, this work introduces a mechanism to explicitly control the deferral rate, enabling programmers to balance fairness and human deferral rate. Recommended values for key hyperparameters, such as neighborhood size, are also provided to guide practical adoption. Experimental results demonstrate consistent improvements in both fairness and overall predictive performance over both human and model. This approach contributes to the development of fairer HITL systems in sensitive domains such as healthcare, law systems, and cheating detection.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past few years, the integration of machine learning (ML) models into decision-making processes has become increasingly prevalent in a variety of fields, including healthcare, finance, and content moderation. Although AI models offer efficiency and scalability, they often lack the understanding and ethical considerations that human judgment provides. Human-in-the-Loop (HITL) systems handles this by combining ML predictions with human expertise, enhancing decision accuracy and responsibility. Although HITL systems slightly improve fairness compared to fully automated models, the improvement is often marginal, leaving significant disparities between demographic groups. This may be important based on the ethic standards required demanded by the application.

This thesis introduces a fairness-aware deferral mechanism for HITL systems, tested with respect to the HateSpeech and Adult Census Income datasets. By using synthetic human predictions and group-specific deferral strategies, the approach aims to reduce bias in sensitive applications. Next sections discuss human-AI collaboration, fairness challenges, synthetic human sampling, and baseline methods, describing the contributions of this research.

1.1 Human-AI Collaboration

Human-in-the-Loop (HITL) Artificial Intelligence, also termed as Human-AI in some literature, refers to a class of machine learning and decision-making systems that integrate human judgment and feedback into the process. In high-stakes domains like healthcare, criminal justice [1] [2], and cheating detection [3], combining machine learning models with human expertise can significantly improve overall performance.

Humans and machines excel at different types of reasoning and noticing different complex features. Although models are consistent and scalable, humans bring contextual understanding and moral reasoning [4]. Humans may also have access to more features than model. An example to this would be a doctor investigating into patient's medical records, ordering more tests, and checking family history, in the case where the doctor is deferred for a decision after AI is not confident.

Human-in-the-loop (HITL) systems use the complementarity of human and AI having different approaches by deferring certain predictions to human experts, particularly in low-confidence or ethically sensitive situations [5]. Current works on Human-AI have shown to get higher accuracy than both, model and human individually [6].

1.2 Fairness and its metrics

Fairness in machine learning ensures similar treatment of individuals across demographic groups, usually defined by race, gender, or economic status. In high-stakes applications like content moderation and income prediction, fairness is evaluated using group fairness metrics, which quantify disparities in model performance across protected groups. Key metrics include:

- **Demographic Parity**: This metric requires that the model's predictions (positive outcomes) are independent of group membership [7], ensuring equal selection rates across all groups. For example, in the Adult Census Income dataset, demographic parity would demand similar income category selection rates for different racial

groups.

$$\mathrm{P}(\hat{\mathrm{Y}} = 1 \mid \mathrm{A} = 0) \;=\; \mathrm{P}(\hat{\mathrm{Y}} = 1 \mid \mathrm{A} = 1). \tag{1.1}$$

It is often measured as

$$\Delta_{\mathrm{DP}} \;=\; \big|\,\mathrm{P}(\hat{\mathrm{Y}} = 1 \mid \mathrm{A} = 0) \;-\; \mathrm{P}(\hat{\mathrm{Y}} = 1 \mid \mathrm{A} = 1)\big| \tag{1.2}$$

- **Equalized Odds Difference (EOD)**: EOD measures the difference in true positive and false positive rates between groups, ensuring that a model's errors are balanced across demographics. In the HateSpeech dataset, EOD assesses whether hate speech detection is equally accurate for minority and majority groups.

  For a binary classifier with predicted label $\hat{\mathrm{Y}} \in \{0, 1\}$, true label $\mathrm{Y} \in \{0, 1\}$, and protected group attribute $\mathrm{A} \in \{\mathrm{a}, \mathrm{b}\}$ (e.g., race or gender), EOD measures the maximum disparity in true positive rates (TPR) and false positive rates (FPR) between groups. The TPR and FPR for group $\mathrm{A} = \mathrm{a}$ are defined as:

$$\mathrm{TPR}_{\mathrm{a}} = \mathrm{P}(\hat{\mathrm{Y}} = 1 \mid \mathrm{Y} = 1, \mathrm{A} = \mathrm{a}), \quad \mathrm{FPR}_{\mathrm{a}} = \mathrm{P}(\hat{\mathrm{Y}} = 1 \mid \mathrm{Y} = 0, \mathrm{A} = \mathrm{a})$$

  Similarly, for group $\mathrm{A} = \mathrm{b}$, we compute $\mathrm{TPR}_{\mathrm{b}}$ and $\mathrm{FPR}_{\mathrm{b}}$. The EOD is given by:

$$\mathrm{EOD} = \max\left(|\mathrm{TPR}_{\mathrm{a}} - \mathrm{TPR}_{\mathrm{b}}|, |\mathrm{FPR}_{\mathrm{a}} - \mathrm{FPR}_{\mathrm{b}}|\right)$$

- **Equal Opportunity**: A variant of equalized odds, this metric focuses solely on equal true positive rates across groups [7], emphasizing fairness in correctly identifying positive cases. For instance, in healthcare, equal opportunity ensures equitable diagnosis rates for different demographic groups. After the relaxation of Equalized Odds, the formulation for equal opportunity is:

$$\mathrm{P}(\hat{\mathrm{Y}} = 1 \mid \mathrm{Y} = 1, \mathrm{A} = 0) \;=\; \mathrm{P}(\hat{\mathrm{Y}} = 1 \mid \mathrm{Y} = 1, \mathrm{A} = 1). \tag{1.3}$$

  This ensures qualified individuals receive positives at equal rates.

For current task, we choose Equalized Odds difference since it captures a model being too focused on a demographic, while neglecting other.

1.3 The need of fairness in HITL Systems

Even with improvements in accuracy, HITL systems may often fail to address fairness across demographic groups. Although, combining an unfair model's likelihood vector with human predictions have shown decent improvement over model's fairness, the methods don't aim at fairness as their primary target, and hence may fail to reach certain thresholds of fairness metrics. Fairness considerations require mechanisms that account for group-level disparities in accuracy, treatment, or outcomes, especially in automated decision-making processes.

1.4 Measuring and responding to Local Bias

This thesis introduces a method for local fairness-aware deferral. Instead of deferring based solely on model confidence, the approach identifies local accuracy disparities using group-specific k-nearest neighbors (KNN). The system defers to a human expert when the expected cost of unfairness risk, calculated as the product of a fairness cost and the accuracy disparity between demographic groups, is greater than or equal to the cost of consulting a human. This local view allows detection possibly unfair treatment by model.

1.5 Controlling Deferral Behavior

In practical deployments, over-deferring to humans may be costly or infeasible. To address this, we introduce a mechanism to control the deferral rate by adjusting the fairness cost and human cost parameters. This allows balancing the trade-off between automation and fairness, making the system adaptable to domain-specific constraints. We also provide recommended ranges for key hyperparameters (like KNN size, fairness cost, and human cost), easing real-world adoption.

1.6 Our Contribution

To summarize out contributions, we contribute a fairness-aware deferral method for Human-in-the-Loop systems. Instead of deferring based solely on model confidence, we identify local disparities in accuracy between demographic groups and defer based on that, making the method more practical. Since currently there are no methods focusing on improving fairness, we compare this to previous methods for cost optimization and deferring all data-points. This comparison is done for HateSpeech and Adult Income datasets.

Chapter 2

Related Works

This chapter outlines the existing literature in the field of human-in-the-loop AI with cost-based optimizations and fairness. We also highlight some literature around general fairness in machine learning, and datasets related to the work.

2.1 Human in the loop Artificial Intelligence

Several studies have explored the integration of human expertise into machine learning pipelines to enhance fairness. In recent advancements, Kerrigan et al. [5] propose a method for combining human predictions with model probabilities using confusion matrices and calibration, providing a foundation for integrating human judgment in ML systems. Building on this, Gupta et al. [6] introduce an optimization strategy that minimizes costs for scenarios involving a single human expert, given misclassification and human costs. Their approach provides a way to balance accuracy and resource efficiency, providing reasonable accuracy improvements on partial deferral rate. Further expanding the scope, Singh et al. [8] explore the possibilities of multiple experts, with different costs, and expand the work to coordinate with given such a system. Together,

these studies provide a comprehensive progress in establishing and scaling the human-AI collaboration systems.

2.2 Accuracy focused approaches in Human-AI

Mozanner et al. [9] show that finding optimal linear classifiers and rejectors is NP-hard, even under ideal conditions, and propose a Mixed-Integer Linear Programming (MILP) solution for small datasets. They also introduce a realizable-consistent surrogate loss function that maintains performance. Verma et al. [10] extend the same for multiple experts, proposing softmax based and one vs all based losses, along with studies on their callibration properties. Okati et al. [11] introduce a method in Human-AI systems, where they change magnitude of loss depending on difference between human and model predictions, giving higher weighted loss to neural network on instances where human is incorrect and lower weights to the instances where human is correct. Wei et al. [12] rethink the standard framework in the system, stating that the system neglects the interdependence between model and expert decisions. They then propose a new new criteria, dependent bayes optimality, which is formed considering the joint behaviour of human and models. After this, they also introduce a dependent cross entropy (DCE) loss which is intended to reach the dependent bayes optimality. Recently, Jie [13] presents a capability-vectors-based architecture for adaptive Human-AI collaboration. The proposed model learns capability vectors representing both AI agents and human experts, which are then used to dynamically assign decision-making responsibility based on context of task. This unified and flexible approach enables robust integration of multiple decision makers with varying expertise levels, and noticeably outperforms previous baselines.

2.3 Fairness approaches in Human-AI

Fairness in Human-AI is still largely unexplored field, however human expertise been utilized in improving fairness in some literature in other forms. Yaghini et al. [4] propose a framework to define a context-aware mathematical notions of fairness, based on

an adapted Equality of Opportunity (EOP) model. Their approach centers on subjective judgment from individuals regarding comparative fairness between different pairs of subjects. These judgment are then used to contextually parameterize a family of fairness measures. Ghai and Mueller [14] introduce a visual tool showing the dependency of various features in the data, and an algorithm to reconstruct data removes some dependencies. This allows humans to determine specific sensitive attributes and select them to remove and reconstruct the dataset, resulting in a fair dataset to train models on. Harris [15] focuses on practical bias mitigation corresponding age in AI job hiring. They use an existing AI fairness toolkit in existing human-in-the loop systems. This approach aims to tackle same challenges as ours and does so in the same framework of Human-AI, but uses proprietary hiring datasets, making it challenging to compare with our results.

2.4 Fairness in Machine Learning

In this section we discuss some recent works in covering general fairness in Machine Learning. Caton and Haas [7] provide a comprehensive survey of fairness in ML, categorizing approaches into pre-processing, in-processing, and post-processing methods across many subcatagories. Their work highlights the complexity of fairness, extending beyond binary classification to regression, recommendation systems, and unsupervised learning, while identifying open challenges such as trade-offs between fairness and utility. This survey serves as a foundational resource for understanding various fairness techniques and metrics. Complementing this, Hardt et al. [16] introduce the concept of equiality of opportunity in supervised learning, proposing equalized odds as a fairness metric that ensures equal true positive and false positive rates across protected groups. This metric is mainly used throughout the project. Bhaskaruni et al. [17] advance fairness through a model ensemble approach, combining multiple classifiers to improve prediction fairness. Their method leverages situation testing, insprired by luong et al. [18], which identifies discrimination by comparing outcomes for similar individuals differing only in protected attributes. Luong's method involves generating test cases to detect bias in decision making systems. Bhaskaruni et al. extend this by training diverse classifiers and aggregating their predictions to minimize bias while maintaining accuracy.

2.5 Datasets available for Fairness

In this section, we discuss the datasets and related techniques used in this paper. Davidson et al. [19] introduce a dataset for automated hate speech detection, comprising social media posts from twitter labeled for hate speech, offensive language, and neutral content. This dataset comes with human labels, allowing us to assess models and HITL systems for performance. However, it does not inherently come with any sensitive labels. Hence, we use Blodgett et al. [20] work to predict demographics over the given tweets, which facilitates us to analyze the dataset from fairness perspective. Ding et al. [21] propose new datasets to replace the Adult dataset, a widely used benchmark for fair ML, addressing its limitations in representativeness and ethical concerns. Adult dataset [22] is a popular dataset for standard comparisons on fairness, but this work expands the dataset and makes it possible to perform regression tasks, or multiclass classification tasks by separating numerical income into bins. This part of their work is based on public-use micro-data by IPUMS population survey dataset [23].

Chapter 3

Methodology

This chapter details the methodology developed to enhance fairness in Human-in-the-Loop (HITL) systems, addressing biases across different demographic groups in sensitive applications. The approach trains a neural network K-way classifier and uses a fairness-aware deferral mechanism that uses k-nearest neighbours (KNN) to evaluate local statistical parity. When risk of unfairness (calculated via fairness cost and local disparity) increases human cost, predictions are deferred to human experts, whose outputs are integrated with model probabilities. The final pipeline is represented by the following diagram.

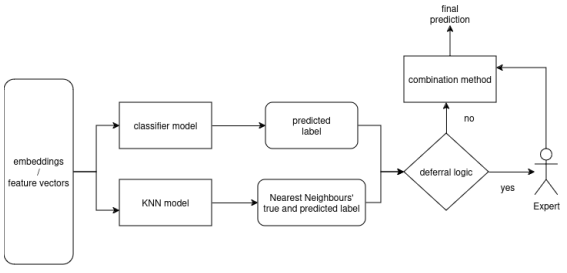

Figure 3.1: Pipeline followed by the methodology

3.1 Assumptions

This method relies on several key assumptions to ensure the validity of fairness-aware deferral mechanism:

- **Human Fairness**: Human experts are assumed to provide fair judgments. This is validated by human evaluations in the hatespeech dataset being much fairer as compared to models. This is also a key to always deferring to human when the local region shows high equalized odds difference.
- **Synthetic Human Predictions**: Synthetic human predictions are generated with varying accuracies but are designed to be completely fair from an equalized odds perspective. This is achieved by ensuring identical accuracy across all of the given demographic groups, resulting in almost no bias in simulated human inputs.
- **Human sampling**: For the HateSpeech dataset, human predictions are given as predictions given by a varying range of humans for each data-point, but parts of dataset is labelled by different humans, making it unreasonable to directly pick a column for single human. To counter this, we have picked an approach common in the field of HITL. The approach is to compute ratio of the number x label appears to total number of labels, and treat it as a probability distribution to sample human labels from. We cam then treat the final human as a possible representative of the human distribution.

3.2 Datasets and Pre-Processing

The experiment is applied to 2 datasets, each requiring specific pre-processing to support the fairness aware deferral mechanism and K-way classification.

3.2.1 HateSpeech Dataset

Description: The HateSpeech dataset [19] is used to classify tweets into 3 labels: Hate-Speech, offensive, and neither. It has around 25,000 different tweets with greater than

or equal to 3 human predictions for each of them. However, the dataset itself doesn't contain any inherent demographic features.

- **Demographic Inference**: To address the lack of demographic data, we use the TwitterAAE repository [20] to infer demographic features, specifically identifying African American English (AAE) tweets. This enables grouping tweets by demographic categories for fairness analysis.
- **Preprocessing**: Tweets are transformed into embeddings using the sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 model [24] . The final vectors are passed to a simple neural network for training purposes.

3.2.2 Adult Reconstruction Dataset

Description: The reconstructed Adult Census Income dataset, provided by folktables [21] and based on microdata provided by [23]. It already includes demographic attributes like race and gender. However, it lacks human labels for now. We have simulated a fair human for different accuracies for this case. This modified dataset contains actual estimate income instead of 2 classes thresholded at 50K dollars.

- **Label Transformation**: Income labels are transformed into three classes: (0-40,000), (40,000-90,000), and (90,000-100,000). we need to transform the prediction label into set classes to comply it with model and methodologies primarily being for K-way classification.
- **Pre-Processing**: The dataset is preprocessed with standard techniques, including normalization, handling missing values, one-hot encoding, etc. Demographic attributes are directly used in the input vector.

3.3 Fairness mechanism

The fairness-aware deferral mechanism is core part of the methodology, ensuring fair performance over all demographic groups by selectively deferring predictions to humans which have high chances of being treated unfairly.

- **Local Statistical Parity**: For each test instance, the k nearest neighbors are retrieved from the training data for each demographic group using the pre-constructed KNN structures. Local accuracy is computed as the proportion of correct model predictions among these neighbors for each group. The local bias is defined as:

$$\text{Local Bias} = \max_{i}(A_i) - \min_{i}(A_i)$$

  where A_i is the local accuracy for demographic i.

- **Deferral Decision**: The system defers to a human expert when the expected cost of unfairness, calculated as the product of a fairness cost and the local bias, is greater than or equal to the cost of consulting a human. This cost-based approach balances fairness and the deferral rate, making use of human judgments which are assumed to be fair.

- **Combiner**: When deferral occurs, a combiner [5] integrates the human prediction with the model's probabilities. The combiner is fitted during training, using training data to calibrate the combination weights.

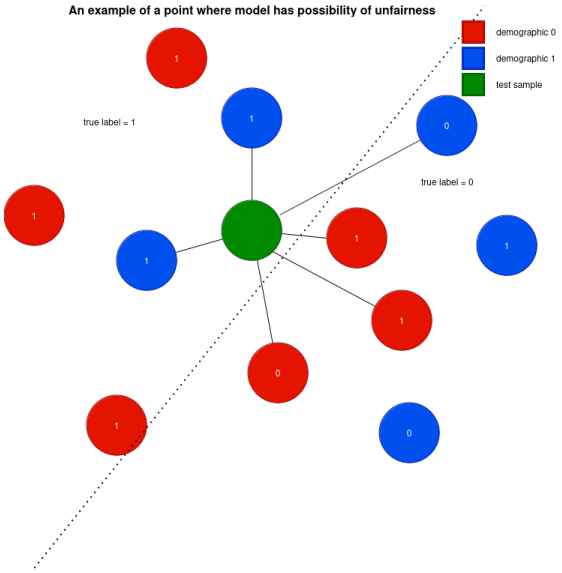

FIGURE 3.2: An example of unfair region

3.4 Algorithm

In this section titled algorithm, we formalize the training and testing phase of our method.

Algorithm 1 Fairness-Aware Deferral Algorithm

Input: Training dataset, test instance x, fairness cost C_f, human cost C_h, number of neighbors k

Output: Final prediction for x

Training Phase:

Train the classifier on the training dataset.

for each demographic group $d \in \{1, \ldots, D\}$ **do**

    Fit a KNN model with k neighbors using training data features, true labels, and model predictions.

    Store the KNN structure for further use.

end for

Testing Phase:

for each test instance x **do**

    for each demographic group d **do**

        Compute model prediction for x.

        Retrieve the k nearest neighbors from the KNN structure.

        Compute local accuracy A_d as the proportion of correct model predictions.

    end for

    Compute local bias: Local Bias = $\max_d(A_d) - \min_d(A_d)$.

    Compute fairness risk: Fairness Risk = $C_f \cdot$ Local Bias.

    if $C_h \leq$ Fairness Risk **then**

        Defer to human expert.

        Use combiner to integrate human prediction with model probabilities.

    else

        Use model's prediction.

    end if

    Output final prediction.

end for

3.5 Hyperparameter Recommendations

To ensure practical implementation, we provide recommendation for key hyperparameters used in the methodology:

- **Number of Neighbors** (k): Set k = log(N), where N is the total number of training datapoints, with a range restricted in [5, 30] to balance computation and accuracy.
- **Fairness Cost** (C_f): Set $C_f = 20$ as a default, with a suggested range of $[5, 50]$ to balance fairness improvements against deferral frequency. Higher values increase deferrals to humans in regions of high disparity.
- **Human Cost** (C_h): Set $C_h = 1$ as a default, with a suggested range of $[0.1, 5]$ to control the cost of human intervention relative to fairness concerns.

Chapter 4

Results

This chapter outlines the experimental results of our proposed fairness-aware deferral mechanism for Human-in-the-Loop (HITL) systems. The evaluation focuses on two datasets: the HateSpeech dataset and the Adult Census Income dataset. Our primary objective is to assess improvements in fairness, measured via Equalized Odds Difference (EOD), while maintaining predictive accuracy. We compare our approach against cost-based deferral methods, which serve as baselines, analyzing EOD for each label treated as the positive class.

4.1 Experimental Setup

4.1.1 Datasets

We evaluate our method on the following datasets:

- **HateSpeech Dataset** [19]: Comprising tweets labeled as hate speech, offensive language, or neutral, with demographic features inferred using the TwitterAAE project [20].

- **Adult Census Income Dataset** [21]: A reconstructed dataset with demographic attributes, where income is categorized into three classes for multiclass classification.

4.1.2 Baselines

We compare our method with:

- **Cost-based Deferral** [6]: Defers predictions based on a cost threshold balancing misclassification cost and human cost.
- **All-Deferral**: A baseline where all predictions are deferred to humans using the combination method given by Kerrigan et al. [5].

4.1.3 Evaluation Metrics

Key metrics include:

- **Equalized Odds Difference (EOD)**: The maximum disparity in true positive rates (TPR) and false positive rates (FPR) across demographic groups, computed for each label as the positive class.
- **Accuracy**: Overall predictive accuracy.
- **Deferral Rate**: Proportion of instances deferred to humans.

4.2 Results on HateSpeech Dataset

4.2.1 Fairness comparison

Our method reduces EOD greatly (hate speech, offensive, neutral) compared to baselines, while still achieving competitive accuracy, better than human and model. The comparisons are done at Fairness cost of 20 and Human cost of 1. For consistency in other methods, Cost based deferral baseline uses a misclassification cost of 20 and human cost of 1 as well.

Method	EOD1 %	EOD2 %	EOD3 %	Deferral Rate %	Accuracy %
Human Only	1.40	7.59	7.44	-	90.59
Model Only	19.59	35.36	20.79	-	87.69
Defer All	18.18	22.06	7.85	100	92.77
Cost-Based Deferral	20.76	23.00	7.44	45.27	92.79
Current Work	3.68	12.43	3.79	69.45	90.61

TABLE 4.2: Comparison of metrics across different strategies for Hatespeech Dataset.

It's noticeable that Cost based Deferral is able to provide 0.02% better accuracy than all deferrals. It can be possible due to the method not choosing some deferrals where combining with human possibly results in misclassification.

Our current method, even though providing slightly lower accuracy than other methods, provides much better improvement on Fairness metrics and still keeps the accuracy higher than both human and model.

4.3 Results on Adult Census Income Dataset

For results on Adult Dataset, we use synthetically generated fair humans (from an EOD perspective) with correct chances of 80% on all labels. Humans being fair is also supported by the humans in HateSpeech Dataset.

4.3.1 Fairness comparison

Out method greatly reduces EOD for all classes in the Adult Income Dataset as well. This costs are considered 20 for misclassification/fairnesss and 1 for human, as chosen in previous method as well.

Method	**EOD1 %**	**EOD2 %**	**EOD3 %**	**Deferral Rate %**	**Accuracy %**
Human Only	1.37	1.65	3.07	-	65.18
Model Only	35.45	29.98	10.69	-	65.00
Defer All	18.68	12.61	19.44	100	74.26
Cost-Based Deferral	18.77	12.77	18.65	87.33	74.21
Current Work	2.62	2.91	2.96	82.25	68.15

TABLE 4.4: Comparison of metrics across different strategies for Adult Dataset.

4.4 Discussion

Our fairness-aware deferral mechanism consistently reduces EOD across most labels compared to cost-based methods, leveraging local bias detection via KNN. This methodology enhances fairness with acceptable amount of tradeoff in accuracy. It can help sensitive applications reach their required fairness requirements. The results underscore the potential of our approach for fair HITL systems in sensitive domains.

Bibliography

[1] I. G. Cohen, B. Babic, S. Gerke, Q. Xia, T. Evgeniou, and K. Wertenbroch, "How ai can learn from the law: putting humans in the loop only on appeal," *npj Digital Medicine*, vol. 6, no. 1, p. 160, 2023.

[2] N. Lettieri, A. Guarino, R. Zaccagnino, and D. Malandrino, "Keeping judges in the loop: a human–machine collaboration strategy against the blind spots of ai in criminal justice," *Soft Computing*, vol. 27, no. 16, pp. 11 275–11 293, 2023.

[3] Y.-S. Shih, M. Liao, R. Liu, and M. B. Baig, "Human-in-the-loop ai for cheating ring detection," *arXiv preprint arXiv:2403.14711*, 2024.

[4] M. Yaghini, A. Krause, and H. Heidari, "A human-in-the-loop framework to construct context-aware mathematical notions of outcome fairness," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1023–1033. [Online]. Available: https://doi.org/10.1145/3461702.3462583

[5] G. Kerrigan, P. Smyth, and M. Steyvers, "Combining human predictions with model probabilities via confusion matrices and calibration," 2021. [Online]. Available: https://arxiv.org/abs/2109.14591

[6] S. Gupta, S. Jain, S. Jha, P.-A. Hsiung, and M.-H. Wang, "Take expert advice judiciously: Combining groupwise calibrated model probabilities with expert predictions," 09 2023.

[7] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Comput. Surv.*, vol. 56, no. 7, Apr. 2024. [Online]. Available: https://doi.org/10.1145/3616865

[8] S. Singh, S. Jain, and S. S. Jha, "On subset selection of multiple humans to improve human-ai team accuracy," in *Proceedings of the 2023 international conference on autonomous agents and multiagent systems*, 2023, pp. 317–325.

[9] H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. Sontag, "Who should predict? exact algorithms for learning to defer to humans," 2023. [Online]. Available: https://arxiv.org/abs/2301.06197

[10] R. Verma, D. Barrejón, and E. Nalisnick, "Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles," 2023. [Online]. Available: https://arxiv.org/abs/2210.16955

[11] N. Okati, A. De, and M. Gomez-Rodriguez, "Differentiable learning under triage," 2021. [Online]. Available: https://arxiv.org/abs/2103.08902

[12] Z. Wei, Y. Cao, and L. Feng, "Exploiting human-ai dependence for learning to defer," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[13] R. Jie, "Learning to collaborate: A capability vectors-based architecture for adaptive human-ai decision making," 2025. [Online]. Available: https://arxiv.org/abs/2502.15196

[14] B. Ghai and K. Mueller, " D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias ," *IEEE Transactions on Visualization & Computer Graphics*, vol. 29, no. 01, pp. 473–482, Jan. 2023. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TVCG.2022.3209484

[15] C. G. Harris, "Combining human-in-the-loop systems and ai fairness toolkits to reduce age bias in ai job hiring algorithms," in *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2024, pp. 60–66.

[16] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[17] D. Bhaskaruni, H. Hu, and C. Lan, "Improving prediction fairness via model ensemble," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019, pp. 1810–1814.

[18] B. Thanh, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," 08 2011, pp. 502–510.

[19] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ser. ICWSM '17, 2017, pp. 512–515.

[20] S. L. Blodgett, L. Green, and B. O'Connor, "Demographic Dialectal Variation in Social Media: A Case Study of African-American English," in *Proceedings of EMNLP*, 2016.

[21] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[22] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[23] S. Flood, M. King, R. Rodgers, S. Ruggles, J. Warren, and M. Westberry, "Integrated public use microdata series, current population survey: Version 8.0 [dataset]. ipums, minneapolis," 2020.

[24] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084