Doubting AI Predictions: Influence-Driven Second Opinion Recommendation

MARIA DE-ARTEAGA, University of Texas at Austin, USA
ALEXANDRA CHOULDECHOVA, Carnegie Mellon University, USA
ARTUR DUBRAWSKI, Carnegie Mellon University, USA

Effective human-AI collaboration requires a system design that provides humans with meaningful ways to make sense of and critically evaluate algorithmic recommendations. In this paper, we propose a way to augment human-AI collaboration by building on a common organizational practice: identifying experts who are likely to provide complementary opinions. When machine learning algorithms are trained to predict human-generated assessments, experts' rich multitude of perspectives is frequently lost in monolithic algorithmic recommendations. The proposed approach aims to leverage productive disagreement by (1) identifying whether some experts are likely to disagree with an algorithmic assessment and, if so, (2) recommend an expert to request a second opinion from.

1 INTRODUCTION

When machine learning is used with the goal of improving experts' decisions, it is often the case that an AI system makes recommendations, but the ultimate decision is made by a human expert. This setup is common in healthcare, human resources, public services, and the criminal justice system. In such contexts of algorithmic decision support, accurate predictions are often not enough. For both the human and the machine to add value to a human-AI team, the human must have effective means at its disposal to make sense of algorithmic recommendations. Without this, algorithm aversion [9], automation bias [21], or other forms of detrimental integration of AI recommendations into decisions may hamper the usefulness of AI predictions. As evidenced by an in-depth field study investigating how AI tools are used by diagnostic radiologists, physicians' professional and legal responsibility for every one of their decisions drive them to look for ways to interrogate AI recommendations, often unsuccessfully [15]. Thus, effective human-AI collaboration remains an elusive goal in many settings. While model interpretability and explainations have been proposed as a way to enable humans to make sense of AI recommendations and to critically integrate them into decisions (e.g. Caruana et al. [5], Ribeiro et al. [20]), it has also been shown that explanations may lead to over-reliance [14], and may fail to improve the quality of decisions [2]. Alternatively, the importance of communicating model uncertainty has been increasingly emphasized as a key feature of algorithmic transparency [2, 3, 18], but this piece of information alone may often not be enough to resolve ambiguity and improve the quality of decisions.

Rather than letting the algorithm speak for (and against) itself, we build on current organizational practices of relying on second opinions provided by other experts. Previous work has explored the use of machine learning to determine *when* to ask for a second opinion [19]. This paper tackles the question of *who* to ask for a second opinion, while also addressing the question of when to do it.

We propose two variants of the methodology, both of which assume access to historical experts' assessments. The first approach builds separate predictive models for each expert and chooses who to ask for a second opinion based on the prediction of these models and their comparison to the recommendation of an AI tool being used for decision support. The second approach leverages influence functions to estimate the influence of individual experts over the predictions of a single model trained to predict experts' decisions. In both variants, the magnitude of the predicted probabilities and the influence, respectively, can also serve as an indication of whether and when to ask. In cases where the AI tool being used for decision support is trained using labels that correspond to human assessments (e.g.

Authors' addresses: Maria De-Arteaga, dearteaga@mccombs.utexas.edu, University of Texas at Austin, Austin, TX, USA, 78712; Alexandra Chouldechova, Carnegie Mellon University, Pittsburgh, USA; Artur Dubrawski, Carnegie Mellon University, Pittsburgh, USA.

radiologists), the latter has the advantage of being directly linked to the AI recommendation, by measuring experts' influence over it. Naturally, this approach is not only useful in identifying who is likely to disagree with an opinion, and can also be used to identify experts who can argue *in favor* of the AI recommendation, by choosing the expert who most positively influences its prediction.

2 RELATED WORK

Human-AI collaboration is a broad space that considers different forms of human-AI teamwork. One type of collaboration considers "division of labor" approaches, in which some instances are routed to a human and some instances are routed to an algorithm [12, 17, 22], with the core idea being that the algorithm can specialize on the instances that are particularly hard for humans to assess. These approaches consider that the algorithm may be the final decision-maker for some instances, in contrast to high-stakes settings where algorithmic recommendations are provided to a human who makes the ultimate decision. In the context of human-in-the-loop frameworks, researchers have studied over- and under-reliance on algorithms [4, 8, 9], the role of explanations [2, 23] and the importance of backward compatibility [1].

The crucial role of uncertainty in humans-in-the-loop settings has been emphasized [3, 18], and it has motivated research on novel statistical methodologies as well as human-computer interaction. Recent work has also explored ways of estimating uncertainty in deep learning [11, 16]. Another line of work studies the impact that communicating model uncertainty may have on human decisions and algorithmic reliance [2, 23].

The use of machine learning to decide *when* to ask for a second opinion has been explored by [19]. In the context studied by Raghu et al. [19] humans are always responsible for making assessments/predictions, and machine learning is used to estimate uncertainty in experts' decisions, in order to determine which cases are most likely to benefit from a second opinion. In this work we focus on the problem of *who* to ask for a second opinion, epsecially considering cases in which the first opinion is provided by an algorithm.

A core element of our methodology relies on influence functions. The local influence method [6] is an approach from robust statistics that estimates the effect of minor perturbations of a model over a functional, such as the loss or the predicted probability. In machine learning, it has been used as a means to explain complex models and as a way to generate adversarial attacks [13]. Most recently, the local influence method has also been used to estimate the influence of individual experts over predictions of a model trained on human assessments, with the goal of bridging the gap between an algorithm's and experts' target objectives [7].

3 METHODOLOGY

3.1 Problem Formulation

We assume a standard supervised learning set up, with features $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$. We also assume there is a set of experts $\{h_1, h_2, ...h_k\}$, for whom historical assessments $d_{h_i}(x) \in \mathcal{D}$ for $x \in \mathcal{X}$ are available. We note that the specific instances $x \in \mathcal{X}$ for which we have human assessments $\mathcal{D}$ and labels $\mathcal{Y}$ need not be the same, nor do we need to have assessments from every expert for each instance. For example, in some cases there may only be one expert assessment available per instance.

Assume that an AI tool used for decision support provides a prediction $\hat{y}(x)$. The task is to identify $h_{ask}(x) \in \{h_1, h_2, ...h_k\}$ such that $d_{h_i}(x) \neq \hat{y}(x)$.

In many domains, the labels $y \in \mathcal{Y}$ used to train the model correspond to human assessments $d \in \mathcal{D}$. Among other domains, this is frequent in some healthcare diagnostic applications, such as radiology. While the proposed approaches

are not constrained to this assumption, they do have additional benefits in such settings, given the tight connection between the experts' assessments and the AI tool. For this reason, and to aid in clarity, we assume for the remainder of the paper that the AI tool is trained to predict historical human assessments, $D \subseteq \mathcal{D}$.

Let $\hat{f}_D$ denote a predictive model of expert decisions, $\hat{f}_D = \hat{P}(D = 1|X)$, which does not differentiate between who provided each label, as is common when training predictive models using human assessments as labels. We assume this to be the AI tool used for decision support, which yields binary predictions $\hat{d}$ based on a threshold τ (in the general case, this predictions correspond to $\hat{y}$).

Given a new case $x \in \mathcal{X}$, the task is to identify an expert to ask a second opinion from, $h_{ask}(x) \in \{h_1, h_2, ...h_k\}$, such that $d_{h_{ask}}(x) \neq \hat{d}(x)$.

3.2 Proposed Approaches

Independent models. The first proposed approach relies on training individual models for each expert, and corresponds to the naive approach of using machine learning to identify sources of second opinion. While simple, this formulation has the advantage of potentially being able to capture heterogeneity across experts' decisions without requiring an increase in the complexity of the algorithms used to model experts, and may be a good fit when there is enough data generated by each expert.

For each expert $h_i \in \{h_1, h_2, ...h_k\}$, train a (calibrated) model to predict its assessments, $\hat{f}_{h_i}$. The provider of the second opinion, h_{ask} can be selected as:

$$h_{ask}(x) = \begin{cases} \text{argmin}_i(\hat{f}_{h_i}(x)) \text{ if } \hat{d}(x) = 1 \\ \text{argmax}_i(\hat{f}_{h_i}(x)) \text{ if } \hat{d}(x) = 0 \end{cases} \tag{1}$$

The predicted probabilities $\hat{f}_{h_i}(x)$ can also be used as an indication of whether someone is likely to disagree, yielding a formulation that also considers when to ask for a second opinion, thus opening the possibility for $h_{ask}(x)$ to be empty,

$$h_{ask}(x) = \begin{cases} \text{argmin}_i(\{\hat{f}_{h_i}(x) : \hat{f}_{h_i}(x) < \tau\}) \text{ if } \hat{d}(x) = 1 \\ \text{argmax}_i(\{\hat{f}_{h_i}(x) : \hat{f}_{h_i}(x) > \tau\}) \text{ if } \hat{d}(x) = 0 \end{cases} \tag{2}$$

Influence-driven selection. The second proposed approach does not require training of separate models for each expert, which has the advantage of not requiring a large amount of data per expert. Additionally, its choice of second opinion is intricately link to the prediction provided by the AI tool, and to whose decisions would most influence that prediction in a different direction.

Assume the training data X has dimensions $m \times n$, and each instance x has been labeled by an expert h. Which expert labeled each case is stored in a vector $a \in \mathbb{R}^{m \times 1}$, such that each entry of the vector, a^j, denotes the expert who labeled the instance x^j, $a^j \in \{h_1, h_2, ..., h_k\}$. Following the local influence method [6], let $w_{h_i} \in \mathbb{R}^{m \times 1}$ be a perturbation of the training data that marginally up-weights the importance given to decision-maker h_i, defined as follows, where w_h^j denotes the jth entry of the vector w_h:

$$w_{h_i}^j = \begin{cases} 1 + \varepsilon & \text{for} \quad a^j == h_i \\ 1 & \text{for} \quad a^j \neq h_i \end{cases}, \tag{3}$$

The influence of this perturbation over the predicted probability indicates how would the predicted probability change if the training data was perturbed in the direction corresponding to an expert h. This influence can be defined as follows:

$$\begin{aligned} \mathcal{I}_{up,f_D}(w_h, x_{test}) &:= \left.\frac{\partial P(y_{test}|x_{test},\hat{\theta}_{w_h})}{\partial \epsilon}\right|_{\epsilon=0} \\ &= \nabla_\theta P(y_{test}|x_{test},\hat{\theta}_{w_h})^T \left.\frac{\partial \hat{\theta}_{w_h}}{\partial \epsilon}\right|_{\epsilon=0} \end{aligned} \tag{4}$$

Once this influence is estimated, which can be done following the approach in De-Arteaga et al. [7], it can be used to select h_{ask}, based on whose influence would most influence the prediction in a direction opposite to that of the provided recommendation,

$$h_{ask}(x) = \begin{cases} \text{argmin}_i(\mathcal{I}_{up,f_D}(w_{h_i}, x)) & \text{if } \hat{d}(x) = 1 \\ \text{argmax}_i(\mathcal{I}_{up,f_D}(w_{h_i}, x)) & \text{if } \hat{d}(x) = 0 \end{cases} \tag{5}$$

Naturally, the choice to request a second opinion may also be informed by whether any expert has an influence in the direction opposite to $\hat{f}_D(x)$, and even further constrained by choosing a threshold for the minimum magnitude of influence required to request an opinion. Making use of the fact that the influence itself is informative for knowing which cases may be expected to have differing views and conflicting opinions, the choice can be reformulated as,

$$h_{ask}(x) = \begin{cases} \text{argmin}_i(\{\mathcal{I}_{h_i}(x) : \mathcal{I}_{h_i}(x) < 0\}) \text{ if } \hat{d}(x) = 1 \\ \text{argmax}_i(\{\mathcal{I}_{h_i}(x) : \mathcal{I}_{h_i}(x) > 0\}) \text{ if } \hat{d}(x) = 0 \end{cases} \tag{6}$$

where $\mathcal{I}_{h_i}(x)$ denotes $\mathcal{I}_{up,f_D}(w_{h_i}, x)$ for simplification in the notation, and h_{ask} may be empty if nobody is likely to provide a differing opinion.

4 EXPERIMENTS

4.1 Results

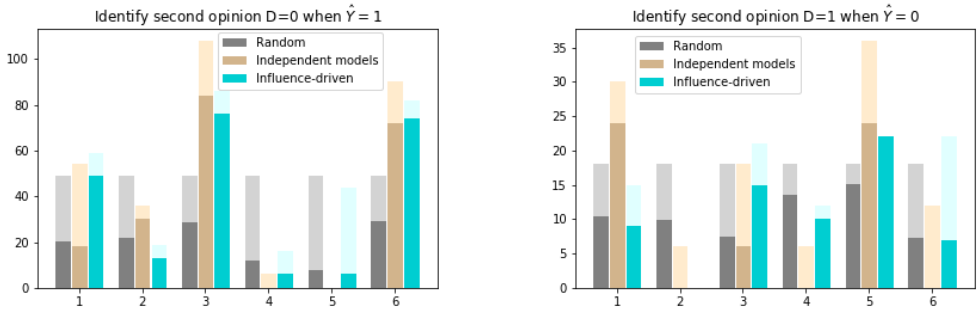

Fig. 1. Performance of proposed approaches to recommend who to ask for a second opinion, assessing whether and to what extent the methods can correctly retrieve disagreement when it exists. Each bar indicates frequency with which a decision maker (1-6) is selected for a second opinion, with shaded portion indicating frequency with which the experts' assessment d_i matches the second opinion sought.

We conduct experiments using a data set containing subjective quality assessment of digital colposcopies [10], which records the independent assessment of six physicians for each case. We use the dataset corresponding to the use of a green filter for feature extraction given its higher degree of disagreement among experts. There is a total of 588

individual assessments, which correspond to 98 unique cases, 66% of which do not have complete agreement across experts.

We first pre-process the data by applying PCA in order to reduce feature colinearity, as this would preclude the reliable estimation of the inverse of the Hessian when estimating influence functions. To estimate $\hat{f}_D$ we train a single logistic regression model that jointly predicts physicians' assessment and estimate each expert's influence as described in Section 3, using 3-fold cross-validation to obtain recommendations for second opnion in the entire data set. Separately, we train six independent logistic regression models (one per physician) and use 3-fold cross-validation to estimate predictions for each $\hat{f}_{h_i}$ in the entire data set.

Figure 1 shows the performance of the two approaches described in Section 3. For each, we assess whether and to what extent they can correctly retrieve a second opinion that disagrees with the AI tool's prediction, whenever disagreement exists. That is, we consider the subset of cases without full agreement, and assess if given a prediction of the AI tool, $\hat{d}(x)$, the methods can correctly identify someone with a differing opinion, h_{ask}, such that $d_{h_{ask}}(x) \neq \hat{d}(x)$. For each expert, Figure 1 shows the frequency with which the expert is chosen and the frequency with which this choice is correct. For comparison, the results of random selection are also shown, where the shown frequency of being correct corresponds to the base rate with which each expert makes each decision. Table 1 shows the aggregate accuracy rate for each of the two proposed models.

Method	Overall	$(\hat{Y} = 1)$	$(\hat{Y} = 0)$
Indep. models	0.64	0.69	0.5
Influence-driven	0.72	0.73	0.69

Table 1. Accuracy of correctly identifying a second opinion that will disagree with the AI tool. Both overall performance and performance disaggregated by models' prediction (and thus by the type of opinion sought) are reported.

4.2 Analysis

Across experts, the influence-driven approach almost always outperforms both the independent model approach and the baseline comparison, as shown in Figure 1. This is also true when considering performance conditioned on the AI tool's predictions, as shown in both Figure 1 and Table 1. Here, we consider separately the set of cases where $\hat{d} = 1$ and thus we seek an opinion $d_{h_{ask}} = 0$, and the cases where $\hat{d} = 0$ and thus we seek an opinion $d_{h_{ask}} = 1$. The gains in performance are particularly large for the set $\hat{d} = 0$, which is the less prevalent assessment.

In addition to raw performance metrics, other properties of the approaches may impact their usability in practice. In particular, it is worth noting that the influence-driven approach distributes the choice of second opinion requests more smoothly across experts. This has important implications for its usability, as it would be likely unfeasable and undesirable to overburden a single expert with requests. Furthermore, this may also translate into benefits for the individuals who are subjected to the algorithm. Different experts will have different areas of expertise, and some may display different performance and biases across subpopulations. Thus, overrelying on a single expert could disadvantage cases that would be better served by other experts.

5 CONCLUSION AND FUTURE WORK

This paper proposes the use of machine learning for second opinion recommendation as a form to augment human-AI collaboration. Grounded on organizational and domain-specific practices of requesting and seeking alternative

perspectives during the decision-making process, the proposed approach seeks to support experts by identifying who is most likely to provide alternative points of view.

Future work will seek to further validate the proposed approaches across datasets, assessing the model's ability to correctly identify experts who can provide a differing second opinion. An important component of these experiments will be to include larger datasets that allow for the use of more complex models.

Additionally, future work will seek to assess the benefits of the proposed framework in dimensions that extend beyond accuracy. In particular, exploring the benefits for fairness considerations is a core direction for future work. Some experts may be better positioned than others to advocate for members of historically underserved communities, who may also be more likely to be incorrectly classified by the AI tool. In such cases, correctly identifying the expert(s) who can more accurately assess these cases may have important implications for the fairness of decisions resulting from human-AI collaboration.

Finally, human subject studies will be important to determine if quality of decisions can be improved with this type of decision support. In such experiments, the use of the proposed approaches to identify both differing opinions and opinions that align with the AI tool may be explored.

REFERENCES

[1] Gagan Bansal, Besmira Nushi, Ece Kamar, Dan Weld, Walter Lasecki, and Eric Horvitz. 2019. A case for backward compatibility for human-ai teams. *ICML Workshop on Human in the Loop Learning* (2019).

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[3] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2020. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *arXiv preprint arXiv:2011.07586* (2020).

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.

[6] R Dennis Cook. 1986. Assessment of local influence. *Journal of the Royal Statistical Society: Series B (Methodological)* 48, 2 (1986), 133–155.

[7] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2021. Leveraging Expert Consistency to Improve Algorithmic Decision Support. *arXiv preprint arXiv:2101.09648* (2021).

[8] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[10] Kelwin Fernandes, Jaime S Cardoso, and Jessica Fernandes. 2017. Transfer learning with partial observability applied to cervical cancer screening. In *Iberian conference on pattern recognition and image analysis*. Springer, 243–250.

[11] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*. 1050–1059.

[12] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-AI Collaboration with Bandit Feedback. *IJCAI* (2021).

[13] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1885–1894.

[14] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 275–284.

[15] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. 2020. To incorporate or not to incorporate AI for critical judgments: The importance of ambiguity in professionals' judgment process. *NYU Stern School of Business* (2020).

[16] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems* 32 (2019), 13153–13164.

[17] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *NeurIPS* 31 (2018), 6147–6157.

[18] Sean McGrath, Parth Mehta, Alexandra Zytek, Isaac Lage, and Himabindu Lakkaraju. 2020. When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making. *arXiv preprint arXiv:2011.06167* (2020).

[19] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2019. Direct uncertainty prediction for medical second opinions. *International Conference on Machine Learning (ICML)* (2019).

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[21] Linda J Skitka, Kathleen Mosier, and Mark D Burdick. 2000. Accountability and automation bias. *International Journal of Human-Computer Studies* 52, 4 (2000), 701–717.

[22] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to Complement Humans. *IJCAI* (2020).

[23] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.